



Münchener Beiträge zur Politikwissenschaft

herausgegeben vom
Geschwister-Scholl-Institut
für Politikwissenschaft

2012

Lukas Rudolph

Evaluation of Development Programs

Theory of Quantitative Impact
Estimation and Its Application to an
Asset-Based Approach to Poverty
Alleviation

Magisterarbeit betreut von
Prof. Dr. Paul W. Thurner
2012

To Nici and Timo: Thanks for your patience and support.
To Tine, Hanna and especially Steffen: Thanks for giving me time for discussions and feedback. And to Karina, Kathrina, Karin, Ulrich, Charlotte, Marlene and Yui:
Without you, this thesis wouldn't have been possible.

Table of Contents

List of Figures	vi
List of Tables.....	vii
Abbreviations	viii
Acknowledgements	x
1 Introduction	1
2 Methods for Impact Evaluation: Overview on the State of Quantitative Impact Analysis	4
2.1 Understanding What Works and Why it Does	5
2.1.1 The Distinction between Qualitative and Quantitative Approaches	6
2.1.2 The Role of Theory	7
2.1.3 Causation in Quantitative Project Evaluation	9
2.2 Bypassing the Counterfactual-Problem.....	12
2.2.1 Self-Selection as Core Obstacle in Practice	14
2.2.2 The Comparison of Potential Outcomes	16
2.2.3 Approaches to the Counterfactual	18
2.3 The Randomization Approach to Impact Evaluation.....	19
2.3.1 Methodology	19
2.3.2 Notes of Caution for Randomization	20
2.3.2.1 Hard Constraints to the Method	21
2.3.2.2 Ethical Concerns	22
2.3.2.3 Cost efficiency.....	24
2.3.3 Natural experiments	24
2.4 Non-experimental methods	25
2.4.1 Matching.....	25
2.4.2 Regression-discontinuity designs	27
2.4.3 Higher-order difference methods	29
2.4.3.1 Pooling Cross-Sectional Data.....	30
2.4.3.2 Panel Data Analysis	32
2.4.4 Instrumental variables	34
2.4.5 Bounding the Endogenous Placement.....	37
2.4.6 Sensitivity analysis	40
2.4.7 Control function methods.....	40
2.4.8 Combination of approaches.....	40
2.5 General Drawbacks of the Outlined Approaches.....	41

2.5.1	Spill-Overs as Confounding Factor.....	41
2.5.2	Issues of Heterogeneity	43
2.5.2.1	Heterogeneity in Impact Pathways.....	43
2.5.2.2	Heterogeneity in Treatment Effects	44
2.5.3	Data Mining Risks.....	46
2.6	A Hierarchy of Methods?.....	48
2.6.1	Clear-Cut Research Designs.....	48
2.6.2	RCTs Solving Second Tier Problems.....	49
2.6.3	Weakest Link Principle	50
2.6.4	A Hierarchy beyond the Quantitative Approach?	51
3	Evaluating a Pro-Ultra Poor Intervention: The Impact of IFSUP.....	52
3.1	Theoretical Classification and Description of the IFSUP Approach: Promotion, Prevention and Transformation.....	54
3.2	Estimation Strategy of the Project Impact.....	59
3.2.1	The Selection Process.....	59
3.2.2	Data on the Project Population and Its Comparability over Groups and Time	61
3.2.2.1	Descriptives of the Project Population at Baseline	62
3.2.2.2	Differences between Control and Beneficiary Group at Baseline	66
3.2.2.2.1	Differences in Means on Overall and Upazila Level	66
3.2.2.2.2	Differences in Distributions on Overall and Upazila Level.....	67
3.2.2.3	Single Difference as Conservative Impact Measurement	71
3.2.3	Estimation Strategies for the Project Impact.....	72
3.2.3.1	Manski-Bounds as Solution to Non-Random Attrition.....	72
3.2.3.2	Difference-in-Differences as Solution to Time-Invariant Bias	76
3.2.3.3	Fixed Effects as Solution to Skill-Differences	76
3.2.3.4	Specification Including Additional Controls (Appendix)	80
3.2.4	Spill-Overs as Confounders	80
3.2.5	Standard Error Concerns	81
3.3	Impact Estimation	82
3.3.1	Overall project impact.....	83
3.3.2	Project Impact on Physical Asset Holdings	90
3.3.3	Project Impact on Vulnerability of Households.....	95
3.3.4	Findings for Correlates and Summary Statistics	102
3.4	Rounding up the Case for Internal Validity	103

4	External Validity	105
4.1	Theoretical Perspective	105
4.2	Generalizations beyond IFSUP?	107
4.2.1	Transfer of the Approach to Other Implementing Organizations	107
4.2.2	Difficulties to Scaling Up.....	108
4.2.3	Specificity of the Sample and Transfer to Other Contexts.....	109
4.2.4	Comparison with Other Programs.....	110
4.2.5	Relation to Theory.....	111
4.3	Policy Lessons.....	112
5	Conclusion.....	113
	Bibliography.....	115
	Appendix 1: The Point of Departure in IFSUP Analysis.....	128
	Appendix 2: The IFSUP Project Intervention.....	130
	Appendix 3: Bangladesh and Its Northern Regions	132
	Appendix 4: IFSUP Working Area in Bangladesh	134
	Appendix 5: PCA Analysis	135
	Appendix 6: Further Analysis of Internal Validity	140
	Appendix 7: Variable Definitions and Construction Procedure.....	146
	Eigenständigkeitserklärung / Declaration of Academic Honesty	150
	Curriculum Vitae.....	151

List of Figures

Figure 1: Exemplary distribution of a factor X within three populations 11

Figure 2: Development of Bangladeshi retail rice prices per kilogram in taka (national average)..... 13

Figure 3: A differentiation of quantitative estimation strategies..... 18

Figure 4: The principle set-up of difference-in-differences analysis 30

Figure 5: The logical framework of the instrumental variable method..... 34

Figure 6: Conception of strategy and envisioned intervention pathway by the IFSUP project 56

Figure 7: Density of yearly expenditure per capita 67

Figure 8: Density of average yearly expenditure per capita by *upazila*..... 68

Figure 9: Density of socio-economic status by *upazila* 69

Figure 10: Distribution of the size of homestead by *upazila* 70

Figure 11: Distribution of the diversification of income sources..... 70

Figure 12: Attrition in control and target groups by *upazila*..... 74

Figure 13: Difference in fixed household characteristics between control and target group households by *upazila* and in-between 2007 and 2009 74

Figure 14: Attrition in control and target groups by *upazila* for the panel data-set..... 78

Figure 15: Difference in fixed household characteristics between control and target group households by *upazila* and in-between the 2007 cross-section households and 2009 panel households after attrition. 79

Figure 16: Manski bounds and confidence intervals for average meals per day 84

Figure 17: Censoring and low variability in the indicator meals per day thus far used for analysis 129

Figure 18: Map of Bangladesh with Target Districts (Grey) and Upazilas of the IFSUP working area 134

Figure 19: Distribution of scores for the SES index 138

List of Tables

Table 1: Description of beneficiaries’ baseline characteristics and control group differences..... 65

Table 2: Difference estimation with Manski bounds for the overall project impact..... 87

Table 3: Difference-in-Differences estimation for the overall project impact..... 88

Table 4: Within estimation for the overall project impact 89

Table 5: Difference estimation with Manski bounds for physical asset-related variables..... 92

Table 6: Difference-in-Differences estimation for physical asset-related variables 93

Table 7: Within estimation for physical asset-related variables 94

Table 8: Difference estimation with Manski bounds for vulnerability related variables..... 99

Table 9: Difference-in-Differences estimation with for vulnerability related variables 100

Table 10: Within estimation for vulnerability related variables 101

Table 11: Factor loadings and descriptive statistics for variables involved in PCA index construction 139

Table 12: Difference estimation for the overall project impact conditional on GOB/NGO support of BG by year 142

Table 13: Difference-in-Differences estimation for the overall project impact conditional on GOB/NGO support of BG by year 143

Table 14: Within estimation for the overall project impact conditional on GOB/NGO support of BG by year 144

Table 15: Within estimation with additional control in the panel data set for overall outcome related variables..... 145

Abbreviations

ATE	average treatment effect
BBS	Bangladesh Bureau of Statistics
BDT	Bangladeshi Taka
BG	beneficiary group
BRAC	Bangladesh Rural Advancement Committee
comp.	compare
CCT	conditional cash transfer
CG	control group
CI	confidence interval
DC	development cooperation
DCI	direct calorie intake
DDS	dietary diversity score
DIS	diversification of income-sources score
e.i.o.	emphasis in original
FD	first differencing
FE	fixed effects
FN	footnote
GO	government organization
GOB	Government of Bangladesh
IFSUP	Income and Food Security for Ultra-Poor
IV	instrumental variable
JCF	Jagorani Chakra Foundation
LATE	local average treatment effect
MDG	Millennium Development Goal
MFI	micro-financial institution
MPD	meals per day
MTR	monotone treatment response
NETZ	NETZ Partnership for Development and Justice
NGDO	non-governmental development organization
NGO	non-governmental organization
OECD	Organization for Economic Co-operation and Development
OECD-DAC	OECD Development Assistance Committee

OLS	Ordinary Least Squares
pc	per capita
PCA	principle component analysis
pd	per day
PNGO	non-governmental partner organization
PPP	purchasing power parity
RCT	randomized control trial
RD	regression discontinuity
RDD	regression discontinuity design
SD	single difference
SES	socio-economic status
SOS	secondary outcomes status
SUS	Sabalamby Unnayan Samity
SUTVA	Stable Unit Treatment Value Assumption
TLU	tropical livestock units
TOT	average treatment effect on the treated
TOU	average treatment effect on the untreated
TUP	Targeting the Ultra-Poor [development program of the NGDO BRAC]
UN	United Nations
UP	<i>Union Parishad</i>
WFP	World Food Program

Acknowledgements

I would very much like to thank ‘NETZ Partnership for Development and Justice’ for the provision of the dataset on their project ‘Income and Food Security for the Ultra-Poor’. Special thanks go to Peter Dietzel (NETZ Germany), Md. Nuruzzaman Khan and Md. Abdullah-Al-Maamun (NETZ Bangladesh), for the time they took to give me background information on the IFSUP project.

I hope this thesis can make a small contribution to their committed fight against poverty in Bangladesh. Nevertheless, opinions voiced and findings presented in this report are those of the author and neither represent an institution or institutional ideology nor should they be affiliated with NETZ.

1 Introduction

“Moving beyond storytelling” (Watkins & Hicks 2009).

Recent advancements in the project evaluation literature bring with them the great promise to revolutionize development cooperation (DC) and enhance its effectiveness. Aim of the field is to establish credible knowledge on what works. As Khandker et al. (2010b: 3) observe:

“Programs might appear potentially promising before implementation yet fail to generate expected impacts or benefits. The obvious need for impact evaluation is to help policy makers decide whether programs are generating intended effects [and] to promote accountability in the allocation of resources across public programs.”

Main driving force of this development was the pressure project implementers came under by popular works claiming development cooperation doing “so much ill and so little good” (Easterly 2006) or even delivering “Dead Aid” (Moyo 2009).

At the heart of project evaluation is the problem of the counterfactual: “[O]bject of interest is a comparison of the two outcomes for the same unit, when exposed and when not exposed to the treatment” (Imbens & Wooldridge 2009: 6). Therefore, a statement on the effect of a project with certainty could only be taken by a change in the time dimension – through a comparison of the exact same individuals’ performance with and without the program given exactly the same context. Additionally, the very same observation should be conducted on the surrounding vicinity if one were to obtain the impact on the non-participants with certainty as well. As this is obviously impossible, the project evaluation literature aims at finding methods for approaching this counterfactual as far as possible. The last decades saw a quickly growing literature on methods and best-practices, summarized in various works: experimental (e.g. Duflo et al. 2008) or non-experimental (e.g. Blundell & Costa Dias 2000), from a practical side (e.g. Khandker et al. 2010b) or with a detailed reference to the technical literature (e.g. Imbens & Wooldridge 2009), additional to the references in various textbooks (e.g. Angrist & Pischke 2009, Stock & Watson 2007, Wooldridge 2009). Leading developmental institutions are pushing for a widespread adoption of structured project evaluation in DC following clear cut norms and standards – be it OECD recommendations for the macro level (OECD-DAC 2010) or World Bank documents with concrete techniques for practical evaluations (e.g. Baker 2000). The drive for effective, results oriented evaluation has even reached the discussion in parliaments (e.g. Perrin 2011).

Effective DC is back on the agenda. The crucial issue is “to fill gaps in understanding what works, what does not, and how measured changes in well-being are attributable to a par-

tical project or policy intervention” (Khandker et al. 2010b: 3). This change bears the promise of maximizing the effectiveness of scarce funds for the benefit of the poor and the poorest. For this, evaluators have a broad range of qualitative and quantitative techniques at their disposal as well as their fruitful combination by so called ‘Q-squared approaches’ (Kanbur & Shaffer 2007). But in approaching the counterfactual, quantitative approaches allow for a generation of evidence on a large scale. This thesis will therefore concentrate in Chapter 2 on a review of the most promising methods in the field of econometric program evaluation, which are aiming at quantitatively measuring the causal impact of a (development) intervention. The core question of this part subsequently is: How can causality be proven through quantitative methods? Attached to this question is the central problem of selection bias or endogeneity in program placement, a problem that routinely affects any attempts to evaluation. Thus, it will be crucial for an answer to the above mentioned question to identify methods that allow controlling for or circumventing this selection bias.

Chapter 3 will apply some of these econometric methods to data on a development program in Bangladesh.¹ As will be argued, sound project evaluation is essential for finding viable pathways in fulfilling the Millennium Development Goals (MDGs) of the United Nations (UN), especially MDG 1, the eradication of extreme poverty and hunger (UN 2000). The scene of non-governmental development organizations (NGDOs) conducting many small-scale projects could provide an important field of experimentation for effective DC on the ground as well as for methodologically sound program evaluation. One such project is the *Income and Food Security for Ultra-Poor* (IFSUP) project initiated by the German-based NGDO *NETZ Partnership for Development and Justice* (NETZ)². It aims at sustaining the livelihoods of the very poorest segment of society in three districts of northern Bangladesh, intervening with 4,800 households. The project for this uses a three-pillar approach to enhance financial and human capital, reduce vulnerability to shocks and a transform the local social context in a pro-poor direction.³ The project is in line with a general move towards more effective social protection policies in developing countries (Barrientos 2011, Ellis et al.

¹ In this, Chapter 3 is building on past analysis of the same project in Rudolph (2010, 2011). Appendix 1 will describe this point of departure.

² The project was jointly implemented by the NGDOs ‘NETZ Partnership for Development and Justice’ together with its Bangladeshi partner organizations (PNGOs) ASHRAI, ‘Jagorani Chakra Foundation’ (JCF) and ‘Sabalambay Unnayan Samity’ (SUS).

³ The first part of this thesis is already referring to IFSUP for the illustrating of the applicability of different approaches to the counterfactual. Appendix 2 therefore provides a general overview on the project and its population in case a more thorough understanding on the approach of IFSUP is necessary. Even further, Appendix 3 anchors the project in the context of Bangladesh and its northern districts.

2009) and a focus on asset accumulation for providing pathways out of poverty (Moser 2007, Moser & Dani 2008). It thereby is one of several pioneering projects centering on asset transfer for the poorest segment of society.⁴

It will be the goal of Chapter 3 to quantitatively evaluate the impact of IFSUP on the targeted population in various dimensions. An answer is thought to the core question of the second part, whether data on IFSUP can link the project to changes in outcome variables. For establishing causality, data on a control group of 1200 households will be used as counterfactual. But the estimation of the counterfactual is not clear cut: Endogeneity in program placement and attrition problems will be addressed through methods such as Manski bounds (Manski 1990, 2007) and difference-in-differences (Imbens & Wooldridge 2009: Chapter 6.5) as described, amongst other methods, in the first part of this thesis.

As shall be demonstrated, IFSUP has generated a positive impact on outcome variables indicating food security, expenditure, socioeconomic status, physical asset holdings (both productive as well durable), and savings. Outcomes on health indicators are not conclusive. It will be argued that the threefold estimation strategy is leading to internally valid results. Nonetheless, it is too early to draw conclusion on the sustainability of the intervention.

For a relevance of evaluations to the social sciences and knowledge in general, the generalizability of results is crucial. Chapter 4 will therefore address the concept of external validity: How can a transferability of results to other contexts be plausibly established in theory – and is this the case for the project at hand? In comparison with related programs, IFSUP adds to the picture of positive evidence for an asset-based approach to ultra-poverty reduction. This as well supports calls for a more prominent role of promotional elements in social protection schemes.

Chapter 5 concludes this thesis. The arguments laid out support the overall case of a rigorous evaluation of development programs with a credible counterfactual. Only credible evidence can underpin the search on what works in poverty alleviation.

⁴ See for a first overview of different projects in the social protection field Barrientos et al. (2010), for asset transfer projects especially pp. 54 & 72.

2 Methods for Impact Evaluation: Overview on the State of Quantitative Impact Analysis

“Suppose you have 10 premises, 9 of them almost certain, one dicey. Your conclusion is highly insecure, not 90% probable” (Cartwright 2007: 14).

The first part of this thesis, as outlined, focuses on the question of ‘How can causality be established by quantitative methods?’. In the following (Chapter 2.1), differences in the application of quantitative methods will be addressed and delimited against qualitative approaches. Shortly, foundations of the method in probabilistic theory will be outlined. Afterwards, in Chapter 2.2, the focus will rest on the question of how to establish the judgment that a project works: The self-selection or endogeneity problem will be described as major confounder in establishing a counterfactual, a measure how the targeted population would have fared without the program. It will be argued that the credibility of impact estimation strategies depends on their case for successfully establishing this counterfactual. The Rubin Causal Model (Rubin 1974) will be used for analyzing the problem. Subsequently, different methods are outlined that account for or circumvent possible selection bias. Chapter 2.3 describes the randomization approach that provides for a clear-cut solution under ideal circumstances – issues in implementation and possible problems such as ethical considerations are subsequently addressed. Chapter 2.4 describes non-experimental (or observational) methods that allow for controls (matching methods, regression discontinuity designs, higher-order difference methods) or circumvention of the bias (instrumental variable methods, bounds for the selection bias, sensitivity analysis, control function methods). In the discussion of these methods, the IFSUP project will be used for illustration of evaluation problems where feasible. A special emphasize will be put on higher-order differences and Manski bounds as these will be applied for the evaluation of IFSUP. Chapter 2.5 will address common problems to the application of these approaches (confounders such as spill-overs or questions of heterogeneous impact) and discuss them in comparison. Chapter 2.6 concludes with the debate on a hierarchy of methods.

2.1 *Understanding What Works and Why it Does*

The question of impact evaluation is always twofold: What works and why does it so? This question alone separates impact evaluation from mere operational evaluation and monitoring systems. While monitoring systems provide the data⁵, the monitoring is no evaluation as such. Monitoring systems in this are prerequisites for later impact evaluation, especially if they include the monitoring of the performance of a control group.⁶ Operational evaluation on the basis of monitoring systems deducts whether the achievement of project targets was fulfilled – from the number of beneficiaries reached to the amount of inputs disbursed and output levels achieved. But impact evaluation is more: It tries to establish the counterfactual and compare outcomes of participants under the project to what they would have achieved otherwise. Thus, as Khandker et al. (2010b: 18) note, “[o]perational and impact evaluation are complementary rather than substitutes.” In their definition

“*[o]perational evaluation* relates to ensuring effective implementation of a program in accordance with the program’s initial objectives. *Impact evaluation* is an effort to understand whether the changes in well-being are indeed due to project or program intervention. Specifically, impact evaluation tries to determine whether it is possible to identify the program effect and to what extent the measured effect can be attributed to the program and not to some other causes” (Khandker et al. 2010b: 18, e.i.o.).

Impact evaluation in this view provides answers to the question on ‘what works’. Added to this must be the question of ‘why’: The core conceptual debate in impact evaluation is then centered on the question where to place the focus. White (2009: 282) in this light contrasts “black-box” approaches centering around the ‘what’ with “[t]heory based impact evaluation” interested in the ‘why’: The ‘what’-focus implies an evaluation of average effects for the average population, the ‘why’-focus an emphasis on the processes transmitting these effects, the surrounding theory and heterogeneity in treatment responses.

Related to the call for a focus on processes, the debate surrounds the perspective the evaluator should take. Beyond the interest in the mere effect of the single intervention lie questions of significance for the understanding of human coexistence: Out of this perspective, it is then crucial to “ensure that studies which may serve as program evaluations for partners provide more general lessons for social scientists. Scholars will benefit from linking empirical studies to more general theoretical work, using that work to guide hypothesis formation” (Humphreys & Weinstein 2009: 376).

Also underlying this debate is a methodical question concerning the normative basis the researcher adheres to in his view of poverty, leading to the choice of different methods.

⁵ Through primary assessments of the project area, the targets to be achieved, indicators to monitor these targets as well as the means to keep track of them. Monitoring by this as well fulfills the role of providing a guideline for the concrete project implementation and adjusting the policy in case of need (Khandker et al. 2010b: 11f.).

⁶ See Perrin (2011) for an overview on a results-based approach to monitoring that recognizes these mechanisms.

2.1.1 The Distinction between Qualitative and Quantitative Approaches

Following Kanbur & Shaffer (2007), qualitative analysis often implies an approach to poverty denying universal, external benchmarks. ‘Truth’ on the poor has to be founded in arguments or discourse that at the best follows equal participation, taking care of e.g. power relations. This still allows for rankings and comparisons of poor and non-poor households, yet based on perceptions of the respective community members in a given area with methods such as participatory poverty assessment (comp. for an example Kadigi et al. 2007).⁷ Central is the try to, as best as possible, refrain from imposing one’s own conceptual categories on others. In this view, as Chambers (2009: 243f.) puts it,

“[t]he starting point would be to ask about the political economy of the evaluation: who would gain? Who might lose? And how? And, especially, how was it intended and anticipated that the findings would make a difference. This might well require a brainstorming workshop with staff from the funding agency. “

On the other hand, a quantitative and empiricist method is taken by the consumption poverty approach, the “‘gold standard’ in applied poverty analysis in the developing world” (Kanbur & Shaffer 2007: 185).⁸ In this conception a definition of poverty is constructed following a certain level of (non-)fulfillment of basic needs beyond the individual. Methodologically central is the interpersonal comparison of utility. Modern utility theory argues that it is possible to empirically derive a valid poverty line (as well as subsequently trends and causes for poverty) insofar as it is based on an intersubjectively observable assessment of relevant consumption data (Kanbur & Shaffer 2007: 190). If these normative assumptions⁹ are accepted, quantitative approaches have the merit of providing an approximation to the counterfactual with enormous potential – at least under ideal conditions. But the quantitative researcher has to remain cautious in generalizing his results: The context dependency of an intervention is not only important on the individual level¹⁰ but as well in transferring results to different social and cultural contexts – a question that will turn up again when discussing the problem of external validity of evaluations (comp. Chapter 4).

⁷ This approach derives from a discursive normative framework with reference to the Frankfurt School around Habermas and Apel, inspired by Kantian thoughts (Shaffer 2002: 59), and tries to found its results in “dialogue [as] a necessary means for arriving at normative conclusions” (Kanbur & Shaffer 2007: 191).

⁸ As Shaffer (2002: 59f.) outlines, this approach is based on naturalist normative theory, which derived historically from the moral theory of David Hume. See Shaffer (2002: Section 4 & 5) for a short overview.

⁹ Revealed preference theory allows deriving levels of well-being and poverty, money metric utility allows for a numerical representation as preferences are represented by consumption expenditure. Subsequently, this allows for „the aggregation of persons or households below the poverty line and comparisons of well-being across persons or households“ (Kanbur & Shaffer 2007: 191).

¹⁰ This is clearly complicating the comparability of answers to questions such as ‘Is your husband opposing the violation of women’s rights’ between individuals in the IFSUP data-set, as the respective answer will depend on the level of women’s rights existing as perceived by the women in question.

The weighting of the roles that qualitative and the quantitative methods shall play in project evaluation is thus contested. Bamberger et al. (2010) emphasize that both qualitative impact assessments as well as mixed methods have to be kept in mind as important alternatives to a quantitative approach, depending on the evaluation context. Khandker et al. (2010b: 19, e.i.o.) on the other hand argue in line with the widespread conviction that “qualitative assessment on its own cannot assess outcomes against relevant alternatives or *counterfactual* outcomes.” A popular combination of approaches in this latter understanding would give the qualitative approach the role of identifying pathways of the causal effect while the quantitative is seen as way of choice in establishing credible evidence of the counterfactual.

While this combination in a ‘Q-squared approach’ can create the normative tensions pointed to above, it as well generates promising insights: Not only, but importantly an in-depth understanding of motives and situations by qualitative research as well as quantitative information to empirically test general hypotheses (Kanbur & Shaffer 2007: 183). As the practical part of this thesis is relying on quantitative analysis with a ‘what works’-focus, the question of an in-depth understanding of pathways is troublesome – partly, this can be solved by resorting to theory, as outlined below; partly it has to be kept for further research.

2.1.2 The Role of Theory

In line with proponents of a participatory approach to poverty, Deaton (2010b: 3, e.i.o.) calls for a greater focus on processes: He argues that “[f]inding out how people [...] escape from poverty is unlikely to come from the empirical evaluation of actual programs [...] unless such analysis tries to discover *why* projects work rather than *whether* they work.” Related is a call for a broader focus for impact evaluation: As Humphreys & Weinstein (2009: 376) emphasize, the relevance of program evaluation rests on a look beyond the mere analysis to questions of broader interest to the understanding of human (co-)existence. Quantitative methods, especially social experiments as the emerging quantitative method of choice, are then a tool with very high potential. Deaton (2010b: 13, e.i.o.) argues for a linkage between the single evaluation, the processes surrounded and a general understanding of social mechanisms through the use of theory: “[F]ormulate hypotheses and derive predictions, what I have referred to here as ‘acid tests,’ that are hard to explain if the theory is not true, so that we seem to learn at least *something* when the predictions are confirmed”. If empirical work is uninformed by theoretical considerations the case is less clear-cut whether evaluation results are causal relationships or only statistical artifacts. He therefore advocates a hypothetico-

deductive approach:¹¹ According to this understanding, theory and data are connected through a mutually reinforcing process. Theory is leading to (natural) experiments and their econometric analysis, which is then (partially) falsified and in this process theory reformulated and strengthened.

But it has to be kept in mind that the empirical evidence generated in this process is relying on the “*weakest link principle*” (Cartwright 2007: 14, e.i.o): Its great promise is the proof of a causal relationship between factors and outcome and a confirmation or refutation of the theory to be tested. But for this to hold, the researcher is usually required to make assumptions for the validity of empirical results and their applicability in a broader context. The degree of the weakest assumption then determines the overall weakness of the argument. “[G]iven the background assumptions, the hypothesis follows deductively from the results [...]. But if you want credit for this benefit of a clinching method, you must be able to show that the *conjunction* of your premises has high probability *in the case at hand*” (Cartwright 2007: 15, e.i.o.).¹²

But the role and value of theory is not undoubted: Khandker (2010b: 20f.) sees a clear distinction between an approach to evaluation through structural models predicting the behavior of participants and outcome of the intervention – what he calls “*ex ante evaluation*” – and typical *ex post* analysis. Although these approaches can be combined, there certainly is no need for it. Even further, Banerjee (2006b: 19) (concerning the generalization of empirical results) observes a complementary relation in the use of theory against the use of empirical evidence: “[W]e could reduce our dependence on theories by running more experiments.”¹³ This view is taken up by DiNardo (2008), who argues that experiments, in a broader sense evaluation as such, serve foremost for clarification if faced with theoretical ignorance:

“Whatever the validity of the view that one cannot experiment in advance of ‘theory’ in the natural sciences, in the social sciences, it is clear that no theory has the same standing as, say, general relativity in physics. [...O]ne does not need a ‘correct’ theory at hand, nor an understanding as rich as that found in some of the natural sciences to find an experiment useful. Experiments are just ways to use things we (think we) understand to learn about something we do not.”

Latest methodological approaches aim at a combination of general theory and empirical evidence already in their basic approach to estimation: Heckman & Vytlacil (2005) follow the idea of an underlying pre-determined structural model that the researcher aims to estimate as comprehensively as possible depending on the data at hand. Thus, the concrete contribution

¹¹ A detailed overview can be found in Cartwright (2007).

¹² This weakest link principle is the most important underlying limitation of quantitative research and especially important for abstractions from the single experiments as will be discussed in Chapter 4. Additionally, it influences the debate on an underlying hierarchy of methods as will be argued in Chapter 2.6.

¹³ One has to note that Banerjee is in other contexts recognizing the distinct role of theory, even in a normative sense for policy advice (comp. e.g. Banerjee 2002).

the empirical data can make to the theoretical framework is made evident; at the same time, clear limits are set by the theoretical framework for the interpretation of the data beyond their direct applicability.

Duflo (2008: 3902f.) notes an important value added by such an approach involving structural modeling: Evaluation usually contributes only to the understanding of total derivatives of an intervention – the total effect after all agents affected have “re-optimized”. Of additional interest to the policy maker may be the total welfare effect, though, to judge an intervention against possible alternatives – this is better understood by looking at partial derivatives, excluding the re-optimization process from analysis through *ceteris paribus* assumptions. For the IFSUP case, this would be of direct interest: A massive capital influx such as the assets transferred through the project is certainly affecting household choices – if households were to offer less work on the wage labor market due to the IFSUP intervention, they would substitute away from other income sources. The total derivative then measures impacts of the assets transferred in relation to earlier external income opportunities. If one assumes, however, a situation of labor shortages in which these external income opportunities are open for other community members, a total derivative is underestimating the welfare change that actually occurred. This is just one example of how the overall welfare effect of the intervention as measured by partial derivatives can differ from total derivatives.

2.1.3 Causation in Quantitative Project Evaluation

As implied in the discussion so far, measuring effects by quantitative methods is relying on probabilistic prediction. Interest lies in the distribution of outcomes of a certain variable the intervention is, deliberately or not, supposed to be influencing: Even though projects claim to cause increased well-being in a given population covered, this is not literally correct. They usually increase the probability¹⁴ of a certain target group to be (in whatever respect) better off than without the project. Thus, the causality between intervention and envisaged outcome is generally understood as probabilistic effect. As Suppes (1970: Chapter 2) outlines, the attribution of event A to have caused event B is not a straightforward task. *Prima facie*, only few conditions are necessary (Suppes 1970: Chapter 2): A causing event must precede its outcome in time, $t^1 < t^2$. The occurrence of the causing event needs to be probable, $P(A_{t^1}) > 0$. And most important, the occurrence of the event B conditional on A must be more probable than the occurrence of B otherwise, $P(B_{t^2} | A_{t^1}) > P(B_{t^2})$. But this *prima facie* attribution

¹⁴ See Cartwright (2007) for a summarized or Pearl (2009) for a detailed overview on the basis of quantitative program evaluation in probabilistic theory as formalized by, amongst others, Suppes (1970). Comp. as well Manski's (2007: Chapter 1) introduction to conditional predictions.

establishes only a correlation between both events – event A might have caused event B , but it must not. It could especially be the case that the correlation is only due to a third factor C_{t_0} (a “*spurious cause*” in the terminology of Suppes (1970: 21, e.i.o.)).¹⁵ In the IFSUP case, the intervention might have caused higher incomes for the project participants, as noted e.g. by NETZ (2009a: 2): „The average family income per capita per day, adjusted for inflation, increased to 30,1 Taka”. But only looking at the intervention group and their income might be misleading, as other factors are potentially confounding the *prima facie* conclusion: This separation of causes is of high practical concern. Two highly influential studies (Khandker 2005, Pitt & Khandker 1998) indicated a very positive impact of microcredit and were much cited in favour of the method: As e.g. Li et al. (2011: 404) state, through exactly these studies “microcredit’s potential in reducing poverty has been thoroughly examined [for Bangladesh]”. Much less cited are though studies that have shown how these conclusions could be a prime example of confounding *spurious causes* with evidence: Reanalysing the data, they estimate that the positive effect of the very same microcredit intervention may just as much rely on the fact that microcredit villages are different from non-microcredit villages (Roodman & Morduch 2009: 24, 31, 36). The intervention of the microcredit institution A is then correlated with the probability of finding higher incomes (event B) but only because some attributes C are causing both the introduction of microcredit programs as well as implying positive economic dynamics.¹⁶ Roodman & Morduch (2009: 4) therefore emphasize that

“when studying causality in social systems with strong endogeneity, claims of non-experimental identification need to be held to demanding standards. The experience also casts doubt on the power of sophisticated parametric techniques to compensate for the lack of such.”

Further complicated is the attribution of cause to an event by the general existence of “*supplementary causes*”¹⁷ (Suppes 1970: 33, e.i.o.) in social systems that accompany the causer but are not causal in themselves.

This being the case, the researcher has to isolate the intervention effect on the subjects next to other factors accompanying or even interacting with the variables of interest. Only then an intervention can be attributed to the observed higher probability of a differing level of the outcome variable of interest. The higher this probability in comparison is, the larger the project impact. The fact that a project is supposed to cause a certain effect, its impact, must

¹⁵ As Suppes (1970: 21) outlines, this would imply $P(B_{t_2} | A_{t_1}C_{t_0}) = P(B_{t_2} | A_{t_1})$. In reality, it will usually be the case that both event A as well as C will have an influence on B , while influencing each other. The task of the researcher would then be to disentangle the causal relationships and isolate the effects.

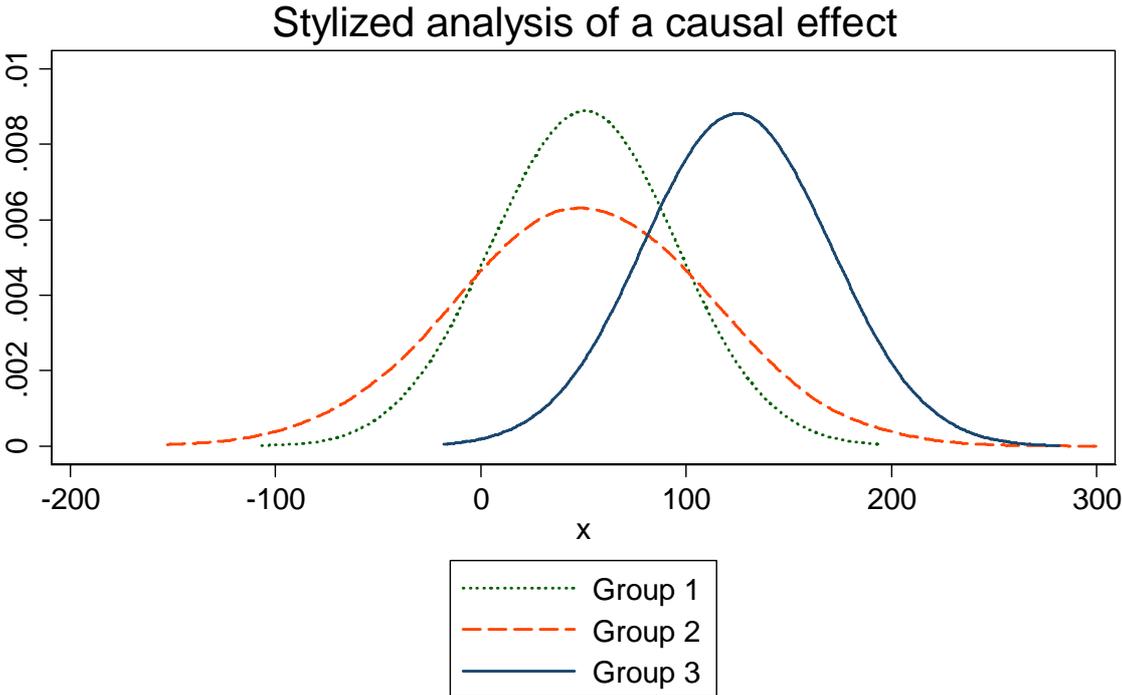
¹⁶ This could for example be better infrastructure, implying better accessibility for the microcredit institution as well as better marketing opportunities for the villagers.

¹⁷ As Suppes (1970: 33) outlines, this would imply an event A and an event C to raise the probability of event B in joint occurrence, thus $P(B_{t_2} | A_{t_1}C_{t_1}) > \max(P(B_{t_2} | A_{t_1}); P(B_{t_2} | C_{t_1}))$.

therefore be stated in terms of probability or, as usually conducted, in an overall aggregation such as by analysis of mean developments as later in this thesis.

Figure 1 displays the stylized distribution of a variable over three different populations. The distribution of group 3 exemplifies a group where an intervention aimed at increasing this variable took place. The distributions of group 1 and group 2 are possible control groups. Central part of program evaluation is an estimation of the average treatment effect (ATE) (e.g. Khandker et al. 2010b: 26): The ATE reflects a shift in the mean allocation as evident in Figure 1. The distribution representing the intervention group and its mean is very distinct from the distributions of control group 1 and 2. But for a convincing assertion on the ATE, a crucial assumption must hold: Only the project occurrence may separate the groups, all other social, political and individual factors being at least on average equally distributed within these groups – this assumption is the crux of impact evaluation.

Figure 1: Exemplary distribution of a factor X within three populations



Own graph
The graph shows the distributions of the density of an arbitrary variable X for three different populations interpreted as control and target groups.

If this assumption can be accepted, the probability aspect shows itself when choosing a random member of either of these groups: The probability of a higher value of factor X is higher for drawings from the intervention group. But the large overlap for the distribution invalidates a comparison of singular observations. Therefore, the mean over all individuals in all three

populations can be compared as indicator of the typical project effect – “[e]ven though other definitions of typical are interesting” (Rubin 1974: 690). Subsequently, Figure 1 as well indicates how a comparison of effects through ATEs can lose important information: While a comparison of group 3 both with group 1 and group 2 leads to the same conclusions concerning the ATE, the apparent compression of distribution 3 in comparison with 1 does not appear in an ATE analysis.

Last but not least, these measurements are typically relying on probabilistic conclusions. As cost efficiency usually prevents full population surveys, the drawing of samples allows then with a certain probability¹⁸ an inference on the probable distributions of variables within the populations – and thereby on the probability that the event in question caused the outcome it was intended to cause as well as its magnitude.

2.2 Bypassing the Counterfactual-Problem

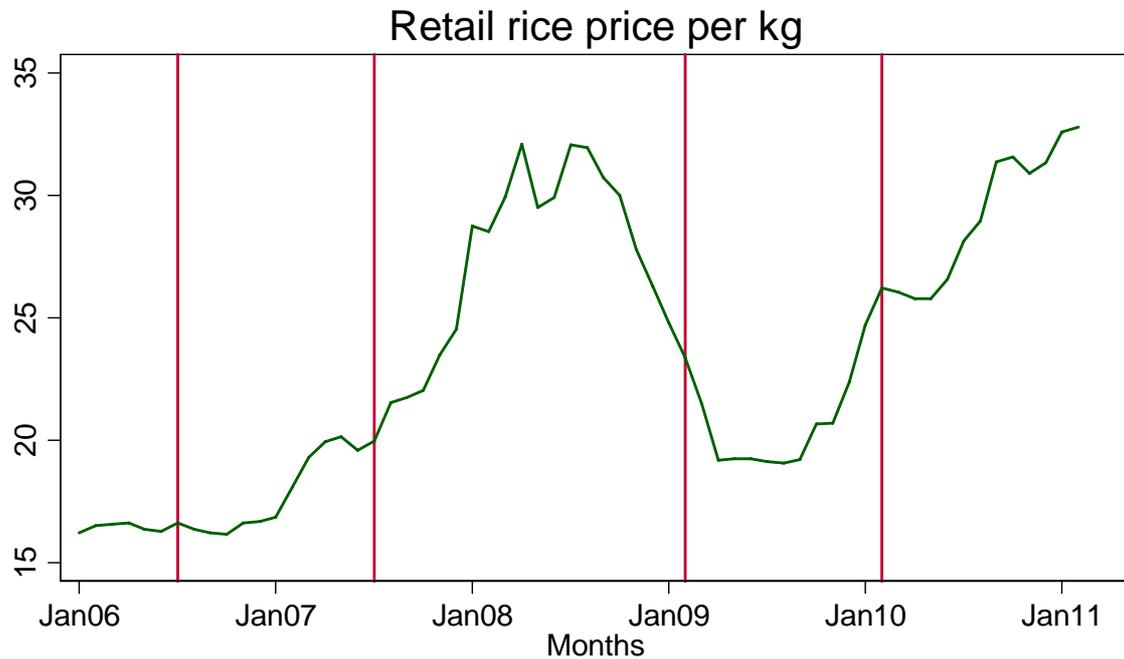
Obvious problem for non-refutable ATEs is the physical impossibility to observe the same individuals in the same context and at the same time but with different levels of treatment. Two obvious possibilities come to mind to bypass this counterfactual-problem: First, a comparison of the same units at different points of time – which requires a good case for structural conditions not changing over time; second, a comparison holding time constant but observing different units – which requires a good case for the similarity, at least on average, of the units under observation.

Both approaches are difficult in practice: In the case of IFSUP, data for both approaches would be present, but unobservable factors complicate their clear-cut application. A major confounder for the unit constant, time varying approach is for example a severe hike of food prices during the project period, as illustrated in Figure 2: The graph outlines the development of national retail rice prices from 2006 to 2011, bars indicating the recall periods for the IFSUP baseline and impact data concerning questions on food security. Food security and hunger are directly related to market prices, as market prices influence the “trade based entitlements” and the “exchange entitlement decline” (Devereux 2001: 246) of the poor. A simple before-after comparison of outcomes in IFSUP can therefore be expected to show only a confounded estimate of the project impact, as the hike in rice prices would diminish the project’s effects and not even show off in the data analysed.¹⁹

¹⁸ Significance levels are by this derived that assess the probability of estimates not having occurred by chance.

¹⁹ This entitlement approach by Sen (1981) offered a new way of thinking about the relationship between hunger and its relationship to the individual’s position on the market through ‘entitlements’. Although his approach is much criticized for laying too much focus on the economic aspects of food distribution, “it shifts the analytical

Figure 2: Development of Bangladeshi retail rice prices per kilogram in taka (national average).



Own graph
Retail rice price (national average) of coarse rice (per kg) in Bangladesh in between January 2006 and January 2011. Bars frame food security related recall periods for the baseline and impact survey of IFSUP.
Data source: Management Information System and Monitoring Unit, Directorate General of Food, GoB, online at <http://www.fao.org/giews/pricetool/>, accessed 08.03.2011.

Similarly, a major confounder of a time constant approach such as earlier analysis of the IFSUP data (comp. Appendix 1: The Point of Departure in IFSUP Analysis) are structural dissimilarity between the group of participants and the group of controls that might not even be present at the point of departure: The IFSUP project is expected to have changed the migration dynamics within these groups, which leads to differences in the observed populations against the actual population of interest, again confounding results. As analyzed in Rudolph (2010: 36),

“[i]n the IFSUP case it could e.g. well be that the drop outs of the control group were due to those who, not being held by the promises of the project’s asset transfer, took their chances in migration [...] In this case, the over- or underestimation of the project impact would depend on the [poverty] strata of migrants: Were it the most able, taking their chances, the project impact is actually overestimated – were it the very endangered, being forced to migrate out of desperateness, the project impact would be underestimated. It is open for further research to apply controls for these possible effects.”

These examples indicate how “[f]inding an appropriate counterfactual constitutes the main challenge of an impact evaluation” (Khandker et al. 2010b: 22). And this challenge is substantial in practice: Tedeschi & Karlan (2010) report for example how leading development or-

focus away from a fixation on food supplies [...] and on to the inability of groups of people to acquire food” (Devereux 2001: 246). Additionally, the exchange entitlement approach sheds light on how market processes can create hunger without the food supply changing through the relative value of the poor’s income (Devereux 2001: 246). In this light, the conclusion of the rice price hike negatively influencing the project participants would of course be premature, if the relative wage rate is not accounted for – but it cannot be ruled out either.

ganizations promote flawed approaches with doubtful internal validity, where groups are compared that are not (reasonably) identical in expectation.²⁰

For a solution to these problems, most methods available draw on the time constant approach comparing a control and beneficiary group – additionally, the internal validity of estimates can be increased, if methods incorporate additional unit-constant elements by looking at differences for the observations over time.

2.2.1 Self-Selection as Core Obstacle in Practice

The starting point for much of the econometric literature are the mentioned problems of endogeneity in program placement or self-selection: Oftentimes, it is not clear how the selection into the participant and control groups has taken place and whether they are really comparable in all aspects relevant for the intervention pathway (Imbens & Wooldridge 2009: 6). And even if it is clear how individuals got into the groups it may be unclear who remains in the sample.

The classical self-selection problem occurs, if individuals by themselves opt to take part in the project. Armendáriz & Morduch (2010: Chapter 9) report several examples complicating the evaluation of development projects: It is e.g. found that microfinance borrowers are wealthier than their neighbors in the vicinity. This must not, but is very likely to cause bias in a comparison of the outcomes of borrowers to non-borrowers. Bias occurs if the fact of higher wealth is related to the outcome measured, as then an estimation of the ATE could be related to either the intervention itself, the pre-intervention wealth differentials or, most likely, both.²¹ These biases can be very substantial, as McKernan (2002: 109) finds: In his estimation of profits of Grameen Bank customers, not controlling for this selection „may overestimate the effect of participation on profits by more than 200 percentage points.” Controlling for this selection might not be possible in all circumstances. But even in such a case, sensible conclusions can be derived if the direction of the bias is known: The inference that a project works is valid if one finds a significant positive effect despite a negative bias in the data and even the magnitude of the effect is then a conservative approach to the true effect. A proper understanding of the impact pathway is therefore important – which provides an important link between the quantitative impact evaluation and qualitative as well as theoretic research. In the case of microfinance, the question would be whether the cause of the wealth discrepancy is to

²⁰ In their case, a leading developmental organization suggested microfinance organizations to measure impacts by comparing veteran to new borrowers. But this approach is flawed as it is not taking into account, just as in the case of IFSUP, that drop-outs are structurally different in both groups which induces bias in estimates.

²¹ In mathematical terms, the error term u is then related to the regressor X_i indicating treatment, and therefore $Cov(X_i, u) \neq 0$ which is violating the crucial assumption for unbiased results (Angrist & Pischke 2009: 35).

be found in microfinance institutions selecting clients with more collateral or in wealthier villagers choosing to borrow from the MFIs due to e.g. better management skills or power relations (self-selection): In the second case, it can be expected that non-MFI borrowers would have fared worse than borrowers even if given the opportunity of taking credit. In the first case, an upward bias is not certain, it might even be the opposite sign given the theoretical expectation of increasing marginal returns to capital (Rudolph 2010: Chapter 1).

The analysis of IFSUP is similarly a good example for self selection problems in two respects: on the one hand side because of deliberate project placement by the project officials, on the other hand because of self-selected differences in the participation of target and control group.

Concerning the first, the fact that IFSUP targets the poorest of the poor²² makes the intervention without further assumptions only comparable to other households in the same circumstances. A comparison with average villagers would e.g. always fail to the concern that unaccounted differences between them influence their response to asset transfer or training, not only but especially with reference to the respective power status and subsequent access to social resources they have (e.g. Bastiaensen et al. 2005: 980f.), and invalidate the estimates. A comparison with other ultra-poor is therefore both necessary to obtain unbiased results as well as highly instructive given the rarity of studies of the response of the most marginalized section of society to poverty alleviation programs.

Concerning the second, the aspect of self-selection, the target and control groups can be supposed to differ by various reasons: A potentially differing probability for migration between the project data gathering periods was already addressed. Additionally, the response rates of households chosen for target and control groups are likely to be much higher in the former given the closer contact to the project officials. It cannot be ruled out that the willingness to respond in the control group is in turn related to personal outcomes and thus that self-selection of reported outcomes is taking place. This bias could be indicated by a first analysis revealing a higher probability for non-minority households to dropping out of the data gathering process (Rudolph 2010: 36). As it can be expected that the local ethnic minorities are discriminated against by the majority of society (Braun 2010, Lehmann 2006) and thus face more difficulties on the local markets as well as via government officials, the control group altogether could be expected to have fared better than the data suggests. This aspect would imply an upward bias for a simple mean difference measurement.

²² As discussed in Appendix 2, IFSUP targets the work-able ‘ultra-poor’, a subgroup of the poor defined as those below an average 1800 kcal/day intake but able to physically work.

At the heart of obtaining valid estimates is thus to ensure what is called the assumption of ‘(conditional) exogeneity of program placement’ (Haughton & Haughton 2011: 243). Exogeneity per se would be implied if one could be certain of structural similarity between both groups. Conditional exogeneity is “the assumption that adjusting treatment and control groups for differences in observed covariates, or pre-treatment variables, remove all biases in comparisons between treated and control units” (Imbens & Wooldridge 2009: 7). This framing of the selection bias problem for impact evaluation is as well called ‘conditional exogeneity of placement’ as well as “‘selection on observables,’ ‘unconfounded assignment’ or ‘ignorable assignment,’ although the latter two terms usually refer to the stronger assumption that Y^T and Y^C [counterfactual outcomes] are independent of T [group status] given X [observables]” (Ravallion 2008: 3792).

2.2.2 The Comparison of Potential Outcomes

To separate issues of endogenous placement on the measured project impact, the evaluation problem is classically formulated following a comparison of potential outcomes – a model developed among others by Rubin (1974). He thereby popularized a common “mathematical and conceptual language” (Sekhon 2009: 489) for the most popular techniques used in the social sciences for program evaluation:²³

Following the standard illustration (drawing primarily on Duflo et al. 2008: Chapter 2, Imbens & Wooldridge 2009: Chapter 2, Khandker et al. 2010b: 25-27, Sekhon 2009: 490-494) of the *Rubin Causal Model* Y_i , the potential outcome Y of individual²⁴ i , $i = 1, \dots, N$, is compared if it received treatment, Y_i^T , as well as if it remained untreated, Y_i^C . Translated to the case of IFSUP, the causal effect of the intervention on the variable of interest, e.g. meals per day, for each individual i would be the difference in meal intakes for the same person, under treatment and if left to itself, thus the difference in its potential outcomes $Y_i^T - Y_i^C$. The overall causal effect of the project as measured by changes in the population mean²⁵, the

²³ Sekhon (2009: 489) stresses that the different techniques of causal inference do not necessarily need to be tied to the, as he calls it, *Neyman-Rubin-Model*. They can be used in the sense of a nonparametric estimator, although then “what exactly has been estimated is unclear”. He similarly makes the important point that alternatives to this specification exist and should not be forgotten, such as the approach of Pearl (2009) to causality – “an alternative whose prominence has been growing in recent years” (Sekhon 2009: 489).

²⁴ Throughout the text, in generally discussing inference frameworks unit i , individual i or household i are used synonymously to refer to the object of treatment of the intervention. Where a distinction would be relevant, this is indicated.

²⁵ Todd (2008: 3851f.) summarizes other parameters of interest: Amongst others direct vs. indirect effects of the program, quantiles of the distribution of impacts, if relevant conditional on specific characteristics, the effect of treatment on the treated as well as on the untreated that are important for project evaluation besides the ATE.

ATE²⁶, could then be displayed by the expected value over all individuals, $ATE = E(Y_i^T - Y_i^C)$. This approach is obviously counterfactual. In practice, only actual meal intakes are observable, although of course counterfactual outcomes are “a quantity that [...] is logically well defined” (Duflo et al. 2008: 3900).

Still, an empirical comparison would only be possible for two distinct groups, one receiving treatment ($G = 1$), one serving as control group ($G = 0$). This approach leads to an observed project effect D , referred to as single difference (SD), of

$$D = E(Y_i^T | G = 1) - E(Y_i^C | G = 0).$$

To clarify the difference between observed project effect and actual project effect, the counterfactual outcome Y_i^C of an IFSUP member, thus $G = 1$, had it not received this treatment is added and subtracted to the equation. The last two terms cancel out and then just simplify to the equation above:

$$D = E(Y_i^T | G = 1) - E(Y_i^C | G = 0) + E(Y_i^C | G = 1) - E(Y_i^C | G = 1).$$

Reformulation leads to²⁷

$$D = E(Y_i^T - Y_i^C | G = 1) + [E(Y_i^C | G = 1) - E(Y_i^C | G = 0)].$$

The first argument, $E(Y_i^T - Y_i^C | G = 1)$, gives the average causal effect of treatment for the group of treated (“*average treatment effect on the treated [TOT]*” (Ravallion 2008: 3790, e.i.o.)). If the project has no effect (e.g. no spill-overs) on the untreated, TOT is equal to ATE (Todd 2008: 3852f.). The second argument of the above equation indicates the bias of D , $[E(Y_i^C | G = 1) - E(Y_i^C | G = 0)]$, if it were used as measurement for TOT. The bias in TOT or ATE is the difference in mean outcomes of the two groups $G = 1$ and $G = 0$ in the case none of them were treated: Thus, for IFSUP, it would be the expected difference in IFSUP members and control group villagers had IFSUP never been implemented.

The crucial question for using D as an estimate of the causal effect is whether this difference can be supposed to be close to zero – or in other words: If one can be certain that the members chosen for the project were not structurally stronger (as e.g. working with them would be more easy) or structurally weaker (as e.g. only the poorest were chosen for project participation while the rest serves as control group) compared to the counterparts in the control group. This, of course, is nothing but a mere reformulation of the initially mentioned condition, namely that the assumption of (conditional) exogeneity of program placement holds.

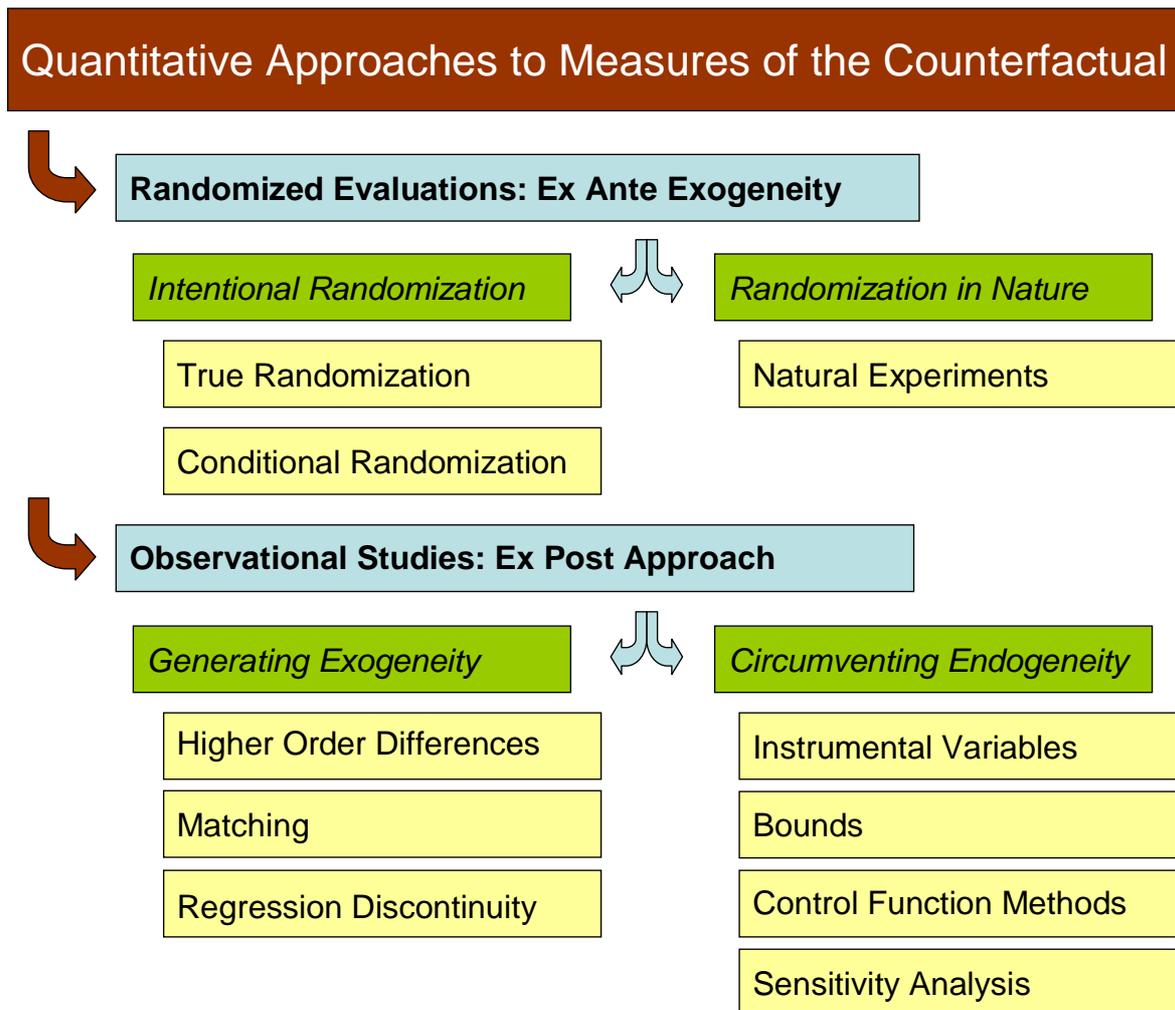
²⁶ Different Names are used for the average treatment effect: Imbens & Wooldridge (2009: 15) e.g. refer to the “Population Average Treatment Effect (PATE)” to give credit to the fact that the observations N are usually one of many subpopulations within an existing superpopulation (subsequently they refer to the later to be introduced treatment effect on the treated (TOT) as “Population Average Treatment effect on the Treated (PATT)”.

²⁷ That, as Deaton (2010a: 439) puts it, the difference of averages is the average of differences is due to the fact that the expectation is a linear operator.

2.2.3 Approaches to the Counterfactual

In the literature, several designs have been proposed that allow for a good case to uphold the case for similarity between beneficiaries and control group members in expectation. These designs are separated by their central design features, as depicted in Figure 3 below. The classification is based on Imbens & Lemieux (2009: 7f.) and Ravallion (2008: 3792).²⁸

Figure 3: A differentiation of quantitative estimation strategies



Own Graph

One set of methods is based on a randomized allocation of units to control and target group status. These are ex ante and by their design ensuring exogeneity in expectation, be it through intentional design, potentially conditional on further observable factors, or as found in nature. The randomized approach can create a clear-cut inference on a causal ATE with a minimum of assumptions (discussed in Chapter 2.3).

²⁸ Note that Ravallion (2008) and Imbens (2009) classify RDD and higher order differences differently. I follow Ravallion (2008) in these cases. Additionally, Todd (2008: 3855f.) mentions “before-after comparisons” as separate strategy. This kind of evaluation would just assume an ignorability of time for the counterfactual which is an assumption difficult to uphold in social systems and therefore not considered separately.

Another set of methods are observational methods, where endogeneity in placement is ex ante present in the data (discussed in Chapter 2.4). Solutions of observational (or non-experimental) methods are on the one hand provided through the introduction of further restrictions on the sample, by this generating exogeneity – as is the case with higher-order-differences, matching and regression discontinuity designs. On the other hand, methods have been developed that allow for a discarding of the bias by circumventing it altogether – as with instrumental variables, bounds, sensitivity analysis or control function methods.²⁹

2.3 The Randomization Approach to Impact Evaluation

2.3.1 Methodology

The most comprehensive solution to the endogeneity problem is the case of a randomization of the intervention, intentionally or by natural hazard. For a truly random assignment, the whole population, or a sub-sample of it, would be randomly, e.g. via a lottery³⁰, separated into treatment and control groups. The case of true randomization makes, in expectation, the average response to treatment or non-treatment by definition equivalent in both groups given a large enough N (Stock & Watson 2007: Chapter 2): Thus, with $E(Y_i^C | G = 1) \equiv E(Y_i^C | G = 0)$, the selection bias cancels out, consequently the observed effect, $D = E(Y_i^T | G = 1) - E(Y_i^C | G = 0)$, constitutes the actual causal effect of treatment: $D = ATE$. As Deaton (2010a: 439) observes:

“The difference in means between the treatments and controls is an estimate of the average treatment effect among the treated, which, since the treatment and controls differ only by randomization, is an estimate of the average treatment effect for all. This standard but remarkable result depends both on randomization and on the linearity of expectations.”

Still, it is important to keep in mind that this result hinges on the assumption of the non-occurrence of spill-overs. Otherwise biases could be present in both directions (Duflo et al. 2008: 3941f.). This is a potential problem for all methods, picked up in Chapter 2.5.1.

Oftentimes, clear-cut randomization is difficult to implement. Pure randomization may not be feasible for the objectives of the implementing organization or even a social scientists’ re-

²⁹ Overall, this discussion focuses on the most important non-experimental methods in both categories. Given the quickly expanding literature it can by no means claim to be exhaustive. Treatment is e.g. assumed to be binary in almost all cases, only the bounding approach touching continuous treatment in passing. See Imbens & Wooldridge (2009: Chapter 7) for further methods for the case of continuous treatment and generally for further interest in program evaluation.

³⁰ Duflo et al. (2008: 3915-3918) discuss various additional methods of randomizing interventions, as lotteries might not be feasible in all cases: A randomized phase-in would for example compare results between those, who were randomly chosen to receive the treatment later than the first treatment group. Popular are as well “encouragement designs”, where the evaluator does not randomize the treatment itself but incentives to take up the treatment that are then used as proxy (comp. as well Khandker et al. 2010b: 38).

search question. In case of IFSUP, it is the explicit aim of targeting certain strata of the ultra-poor, amongst others 40% ultra-poor of aboriginal ethnicity. Randomization would thus be conducted conditional on these observable factors. This approach does not lead to less consistent estimates as these observables can be segregated to arrive at an overall causal effect. In case of randomization conditional on a set of observables X , the expected outcome conditional on the same set X is equal in both groups: $E(Y_i^C | X, G = 1) \equiv E(Y_i^C | X, G = 0)$. This allows for comparing the effect of the project within the respective strata, e.g. the effect of the project within the subgroup of aboriginal ethnicity against the effect within the subgroup of non-aboriginal ethnicity. The overall effect is deducted by weighting the outcomes with the share of the treated against non-treated in the strata it was conditioned on (Duflo et al. 2008: 3934f.)³¹, ensuring conditional exogeneity of placement.

The approach of (conditional) randomization has the great advantage of, in theory, being a fool-proof strategy for internal validity: Internal validity refers to questions, “whether we can conclude that the measured impact is indeed caused by the intervention in the sample” (Duflo et al. 2008: 3950). Thus, evaluators can make clear statements on the mean project effect.

2.3.2 Notes of Caution for Randomization

This case for clear inference has to be mitigated in practice, however – not only for the case of natural or quasi-experiments, where randomization oftentimes is only warranted by assumption as discussed below, but also for randomized experiments, which just like other evaluation methods suffer from confounding with spill-over effects, as addressed, or attrition. Attrition refers to the problem that information cannot be gathered on some of the units receiving treatment or control status (comp. for one of the first popular discussions of the problem Hausman & Wise 1979: 445f.). If these individuals were structurally different from the rest of the population in their outcomes bias is introduced.³²

Even if these problems are circumvented, enabling clear policy advice requires external validity of experiments (comp. Chapter 4): This is the question of a generalizability of results. Deaton (2010b: 445) refers to the issue that usually participants in an experiment are,

³¹ As Duflo et al. (2008: 3935) put it, from $E(Y_i^C | X, G = 1) \equiv E(Y_i^C | X, G = 0)$ and if X takes discrete values, $D = ATE$ follows from

$$E(Y_i^T - Y_i^C | G = 1) = E_x \{ E[Y_i^T | X, G = 1] - E[Y_i^C | X, G = 0] \} = \int \{ E[Y_i^T | x, G = 1] - E[Y_i^C | x, G = 0] \} P(X = x | G = 1) dx.$$

³² Compare Foster & Bickman (1996) for a practical guide on detecting attrition problems.

deliberately or not, not randomly drawn from the parent population. Results in this case may not be a reliable guide to policy, since

“even if the experiment itself is perfectly executed, [...] the selection or omitted variable bias that is a potential problem in nonexperimental studies comes back in a different form and, without an analysis of the two biases, it is impossible to conclude which estimate is better – a biased nonexperimental analysis might do better than a randomized controlled trial if enrollment into the trial is nonrepresentative.”

But drawing on its high internal validity, the method of randomization is called for widespread adoption by leading scholars in the field. Yet the view of these “[r]andomistas”³³ (Ravallion 2009) is not unchallenged, as the method has different limits in its application. As discussed in the subsequent section, these concern especially hard constraints to the questions addressable, ethical concerns and cost efficiency arguments.³⁴ Additionally, natural experiments can be a borderline case to observational studies.

2.3.2.1 Hard Constraints to the Method

First of all, randomized control trials (RCTs) are not a useful design for all types of projects: Humphreys & Weinstein (2009: 373f.) summarize these cases as “hard constraints” to the method. This is the case if real time experiments take too long, such as studies examining the influences of cultural change or intergenerational questions. Other instances include the influence of historical events. Important are additionally cases in which the variable of interest is inherently linked to the subject of interest. Humphreys and Weinstein (2009: 373f.) cite the case of gender or religion where the history of belonging to this category assumably matters. Additionally, treatment assignment itself might change behavior.³⁵ Questions which involve serious political interests might be added to this list, especially issues such as large-scale infrastructure interventions or broad political reform – be it the height of tariffs and tax rates, the design of political institutions such as reforms on government accountancy. It will not only be political will that deters randomization in these cases but as well the mere scale of the intervention: This is a problem for both the introduction of control groups as well as the num-

³³ It has to be noted that these ‘randomistas’ are aware of these concerns: Duflo & Kremer (2005: 206) therefore emphasize that “[i]t is important to note that we are not proposing that all projects be subject to randomized evaluations. We are arguing that there is currently a tremendous imbalance in evaluation methodology, and that increasing the share of projects subject to randomized evaluation from near-zero to even a small fraction could have a tremendous impact.”

³⁴ The question of a comparison between different evaluation designs will be addressed in detail in Chapter 2.6.

³⁵ This is an inherent problem to all treatment evaluations. Known as Hawthorn (for beneficiaries) or John Henry (for control group members) effect (Duflo et al. 2008: 3951). The mere fact of being surveyed changes incentives of the units under observation (comp. for an interesting experiment on the matter, where filling out questionnaires raised awareness to e.g. hygiene and subsequent health outcomes, Zwane et al. 2011). These effects can be especially problematic for the method of randomization if changes in the attitude of subjects towards the treatment differ with randomization compared to other evaluation methods. This effect was revealed in other circumstances: Own laboratory experiments with dictator and ultimatum games give evidence on a significant effect of the institutional design on later game behaviour – in this case the allocation of decision/proposal power via randomization, merit or bargain (Behnke et al. 2010: 174f.).

ber of cases necessary for the statistical power of the experiment. White (2006: 16) therefore summarizes:

“When [...] can randomized impact evaluations be used? Banerjee [2006b] compares aid programs to drugs; the analogy is a good one. Randomized approaches can be used to evaluate discrete, homogenous interventions, much like a pill in a drug trial. But most of the projects of large official agencies – which constitute the bulk of aid – do not resemble the conditions of medical testing.”

This argument has to be softened to some degree: Field experiments can find very ingenious ways of introducing variation in complicated cases. Humphreys & Weinstein (2009: 371f.) highlight various works concerning e.g. participation on political responsiveness, designs of anti-corruption institutions, or the role of media on interethnic relations.

2.3.2.2 Ethical Concerns

Ethical concerns for the application of randomized designs are twofold: First, a common concern is whether it is feasible to withhold benefits from potentially needy individuals, especially as they are carefully included in the data gathering process. The second objection concerns the acceptance of a randomized design as viable allocation method on the ground.

Concerning the first, a priori reasoning usually leads to first hypotheses about the benefits of a policy or program that then should not be withheld from the control group. On the one hand, Duflo et al. (2008: 3915) argue comprehensibly that randomization in these cases is only unethical as such if full coverage were physically possible: In most cases, social interventions are only designed to cover a section of the population, due to financial constraints or questions of organizational capacity. In these cases, randomization, via e.g. a public and transparent lottery, could be a very just mechanism of allocation. On the other hand, even in cases where only a part of society is addressed, a hierarchy of needs even amongst the poorest is usually discernable. Many organizations are dedicated to serving the beneficiaries they perceive as having the greatest need or, after their subjective assessment, the highest chance of success (Armendáriz & Morduch 2010: 307).³⁶ This approach to DC would deter fully randomized designs and other methods for evaluation have to be found.

Concerning the second, a randomized design might be neither acceptable from the perspective of the local population nor from the perspective of a just project design on the ground.

³⁶ This being Armendáriz & Morduch (2010: 307) caution that a “selection bias in the choice of field partner[s]” could be introduced if organizations willing to perform randomized evaluations are structurally different from other organizations.

Problems with non-acceptance are e.g. already evident in IFSUP: Program managers report difficulties for gathering data on control group subjects. As they put it, these subjects saw how the program transformed the live of their neighbors while they still live in despair. They even reported the fear that elites without interest in project success could capture the opinion of control-group ultra-poor and instigate them against the project and its beneficiaries. This indicates how a target versus control group approach even without randomization can raise substantial tensions. NETZ out of these reasons has for example abandoned the practice of gathering control group data for impact measurement in the same geographical areas, although this hampers the evaluation of their projects to a great deal.

But non-acceptance of a RCT might as well be the consequence of a needs-based project development process. Ownership of the processes by the affected is a central aspect of DC, acknowledged by all major donors (OECD 2005). Out of this, Chambers (1995) makes a strong point for equitable and emancipated DC, arguing that it requires

“[a]nalysis and action by local people, and putting first the priorities of the poor: central to the paradigm is the basic human right of poor people to conduct their own analysis. People centered development starts not with analysis by the powerful and dominant outsiders - the ‘North’, uppers and professionals, but with enabling local people, especially the poor, to conduct theirs.

But such a ‘people centered development’ may run counter the intention of a RCT, as it would not be chosen by the involved. The positive value of increased knowledge is of no direct concern for the affected population. When confronted with the proposal of a RCT granting a major asset transfers to some, whereas others would only ‘receive’ questionnaires, program managers hold it more likely that the potential target group would rather suggest splitting the assets in half for everybody – what seems a reasonable solution for ultra-poor households.³⁷

This is not to say that rigorous program evaluation should not take place. But in some cases flipping coins might not be feasible. Deviations from the ‘perfect’ randomizations or observational methods are then necessary. Randomized phase-ins of programs could be one possible solution, where different subsections receive the treatment at different times and thereby providing information on the counterfactual while waiting for the program to start (Khandker et al. 2010b: 38), although one has to weigh the higher costs involved. Additionally, with every deviation from an optimal RCT, causal inference is complicated by definition and the clear-cut case for internal validity diminished – which reduces the benefits of non-optimal RCTs against observational methods.

³⁷ Segers et al. (2010: 533) provide additional examples on the ingenuity of the poor in this respect: For the context of Ethiopia, they describe how households appropriate microcredit programs to fit their need if allowed to, bypassing NGO intentions (e.g. loans as solution to seasonal food/grain shortages not intended asset building).

2.3.2.3 Cost efficiency

A more practical argument is, third, the question of cost efficiency: Alternative randomization designs such as randomized phase-ins require the management of a far more complex project. This requires better-skilled development workers and raises costs substantially – an argument why this design is e.g. not feasible for the IFSUP follow-up. Observational studies are especially efficient if reality by itself provides the exogeneity necessary for evaluation, as in the evaluation of disaster preparedness programs (Goldin et al. 2006: 11).

2.3.3 Natural experiments

An additional note of caution concerns the question of natural or quasi-experiments. These terms “denote a situation where real randomization was employed, without the intent of providing a randomized experiment” (DiNardo 2008). Humphreys & Weinstein (2009: 369) outline that experiments always have an element of control. The investigator seeks to maintain control over treatment assignment and where applicable, especially in case of lab experiments, control over the treatment itself. “[I]n natural experiments, the researcher controls neither but seeks to find assignment processes that create comparable treatment and control groups ‘naturally’ [–] in this sense, natural experiments are not experiments at all”. This is the reason why the treatment of natural experiments has to be approached with great caution: What might seem randomized on paper is in many instances affected by unobservable variables. Deaton (2010a: 446) cites the example of Miguel & Kremer (2004), a study which is both policy relevant and methodologically central. The authors assess the impact of large-scale deworming on the school level on absenteeism. But assignment of phase-in status was in these cases actually ascertained through the alphabet: Schools were “listed alphabetically and every third school was assigned to a given project group” (Miguel & Kremer 2004: 365). As Deaton (2010a: 446f.) argues, this method of alphabetization is just very different from a true random assignment. Whether or not the method of alphabetization is justified in this case, it has to be applied cautiously: It is e.g. possible that the alphabet might serve as, even if only unconscious, basis for the allocation of scarce resources as well. Still, natural experiments can be an ingenious way to the counterfactual, given a good case for their exogeneity.

2.4 *Non-experimental methods*

While randomization is theoretically the most comprehensive solution for evaluation, the outlined arguments indicate that experimental data is not feasible in many circumstances. Non-experimental³⁸ methods will be equally important, as already summarized in Figure 3 above.

In the following, approaches for successfully assuming conditional exogeneity, at least in changes over time, will be addressed with three popular methods: Matching directly exploits covariates ensuring conditional exogeneity. Similarly, in regression discontinuity designs (RDDs) exogeneity is established by looking at units identical in pre-treatment variables through features of the selection design. Higher-order difference approaches argue that changes in outcomes of nonparticipants reveal counterfactual changes in outcomes of participants, thus ensuring exogeneity.

The next three approaches circumvent the endogeneity in the data altogether: With instrumental variable (IV) approaches, the researcher aims at finding a variable uncorrelated with the outcome, but associated with participation, thus bypassing the endogeneity in placement. Bounds aim at dropping the exogeneity assumption by establishing an upper and lower estimation for the bias. Finally, sensitivity analysis and control function methods will be very shortly addressed. Sensitivity analysis assesses whether results change if the exogeneity assumption is slightly relaxed. Control function methods aim at modeling the unobservable variables causing endogeneity.

Combinations of these approaches provide as well important possibilities.

2.4.1 **Matching**

The central question for comparing different groups at a point in time is the degree of their comparability. This is the starting point for the long used (comp. e.g. Rubin 1974: 691, Thistlethwaite & Campbell 1960: 315) technique of matching, recently undergoing a revival (Sekhon 2009): Through matching, the researcher aims at comparing the outcomes Y_i^T of a unit receiving treatment with the outcomes Y_j^C of a unit as similar as possible to the treatment-receiving unit. Matching thus follows the idea to “mimic the effects of randomization, even though the treatments were not applied randomly” (Haughton & Haughton 2011: 247).

The most clear-cut matching approach would imply exact matching on a vector X of covariates – which is usually impossible in practice due to samples being too small or continuous covariates necessary for comparison. Subsequently, two broad approaches can be dis-

³⁸ Also termed methods for “data from observational studies” (Imbens & Wooldridge 2009: 7) or, as far as they mimic RCTs, “quasi-experimental evaluation” (Ravallion 2008: 3792) – confoundable with natural experiments.

tinguished that are commonly adopted: Propensity score matching³⁹ (PSM) as well as multivariate matching techniques. PSM uses the covariate information to ascertain the probability of project participation and matches outcomes on this information. Multivariate matching directly measures the distance between the covariate vectors for relating the closest target units with controls (Sekhon 2009: 497). Although straightforward in theory, the core problem with these methods is that “there is no consensus on how exactly matching ought to be done, how to measure the success of the matching procedure, and whether or not matching estimators are sufficiently robust to misspecification so as to be useful in practice” (Sekhon 2009: 489). The crucial question is whether all factors relevant for participation as well as outcomes are included in the covariates for matching, X . As Ravallion (2008: 3806f.) stresses, detailed knowledge on the program and the local structure is necessary. Qualitative work and theory thus are back in the picture, as “[n]o mechanical algorithm exists that automatically chooses sets of variables X that satisfies the identification conditions” (Smith & Todd 2005: 333).

To illustrate the point of the method, propensity scores will be used as example in the following. As outlined by Khandker et al. (2010b: 55-60) the propensity score is given by $P(X) = \Pr(T = 1 | X)$ with $0 < P(X) < 1$, measurable through a logit or probit model (Todd 2008: 3863). X is a vector of pre-treatment indicators⁴⁰ that affect both treatment status and outcome. Importantly, “[t]he values taken by [X] are assumed to be unaffected by whether unit i actually receives the program” (Ravallion 2008: 3806). Thus, feedback mechanisms would lead to biases. The vector X affects participation and outcome, although participation may not affect X .⁴¹ After all units i are ascertained a propensity score, in a second step a balancing test should be conducted on whether the distribution in target and control group of the propensity score as well as of the covariates is similar for different strata of estimated $\hat{P}(X) | G = g$, $g = \{1,2\}$ – this avoids misspecifications that can be very relevant in practice.⁴² If the case for a plausible propensity score construction can be established, a second condition is necessary: The “region of common support” (Khandker et al. 2010b: 56f.), where the propensity scores of participating and control units overlap, has to be sufficiently large.

³⁹ Introduced by Rosenbaum & Rubin (1983).

⁴⁰ Glynn & Quinn (2011) even argue for the cautioned usage of post-treatment variables if they can provide evidence on the selection bias.

⁴¹ To exemplify this assumption, in the case of IFSUP distance to the village center is an important determinant of participation and should therefore be included in X , but results will be biased if in response to treatment assignments households move closer to the village center, thus participation would have been affecting X .

⁴² Smith & Todd (2005: 333) summarize examples for implications of choosing different specifications for X .

For this, the sample of control units is preferably much larger than the sample of targeted units – one of the reasons why matching is a difficult approach for establishing causal inference in the IFSUP case. Observations that fall out of the common support need to be dropped. If these belong to the treatment group, a source of bias is reintroduced (Ravallion 2008: 3794). As Todd (2008: 3863) notes, “[t]he estimated treatment effect must then be defined conditionally on the region of overlap.”⁴³

Given the exogeneity assumption (and sufficient common support), Rosenbaum & Rubin (1983: 43) proved that “[a]t any value of a [propensity] score, the difference between the treatment and control means is an unbiased estimate of the [ATE]”. For this, in the last step participants and non-participants are matched and their mean outcomes over the overlap-region compared: Khandker et al. (2010b: 59-63) as well as Todd (2008: 3864-3869) provide an overview of the most commonly used methods of matching participants and controls with propensity scores. These methods differ by the number of control units used to match participants (1:1 vs. 1:n), by the closeness of propensity scores that is accepted for ‘good’ matches, and concerning the use of weights for fitting, e.g. weighted means in case of (1:n) matching.

For the application of matching, the quality of data on project uptake and outcomes is crucial, as Smith & Todd (2005: 347) argue from their comparison of different estimators. If there is a plausible case that the observables used really drive project take-up and that they are measured comparably within the control and target group, the quality of causal inference of matching techniques is no worse compared to RCTs.

2.4.2 Regression-discontinuity designs

A similar approach is used by regression discontinuity designs (RDDs). Comparability between target and control group is ascertained by assuming their similarity in expectation through features in the selection design: RDDs try to exploit the fact that an intervention is taken up only after a certain eligibility threshold is reached, with this eligibility threshold not affected by the treatment.⁴⁴ As formally presented by Ravallion (2008: 3812), with M_i representing the value of the eligibility variable for individual i and m being the threshold value of the eligibility criteria, this implies an assignment to the target group, $G_i = 1$ for $M_i \geq m$

⁴³ One then has to weigh whether other methods, such as regression discontinuity discussed below, might be preferable in addressing selection bias.

⁴⁴ As outlined in Khandker et al. (2010b: 110) a similar argumentation holds in the case of the exploitation of time as cut-off value: These designs, “pipeline comparisons”, use the fact that a program is intervening in multiple phases. Thus the eligible participants that have just not yet received treatment can serve as an approximation to the counterfactual.

and $G_i = 0$ for $M_i < m$. Based on the assumption that individuals barely above and barely below this threshold can ex-ante be expected to have had very similar results if exposed or not exposed to treatment, this discontinuity in treatment assignment allows for the construction of a credible control group and a causal effect of $D = E(Y^T | M = m + \varepsilon) - E(Y^C | M = m - \varepsilon)$ for a small $\varepsilon > 0$ ⁴⁵. Thus, simply the means of a small range of both control and target group members just around the threshold value are compared and by their difference, a local average treatment effect of the intervention for the range $m \pm \varepsilon$ is established.⁴⁶

First developed by Thistlethwaite & Campbell (1960) already in the 1960s, this approach seeks to overcome critiques of matching methods:

“[R]egression discontinuity analysis does not rely upon matching to equate experimental and control groups, hence it avoids the difficulties of [...] incomplete matching due to failure to identify and include all relevant antecedent characteristics in the matching process” (Thistlethwaite & Campbell 1960: 315).

The popularity of regression discontinuity has increased only since the 1990s in the social sciences as one of a number advances in econometrics (examples include Green et al. 2009, Lee 2008, Lemieux & Milligan 2008). As Imbens & Lemieux (2008: 618f.) summarize, especially one feature of the method is useful: M_i does not have to be a deterministic function of one or more covariates, but it suffices that the probability of being included as project beneficiary rises in order to estimate a causal effect.⁴⁷

Regression discontinuity has the great advantage of allowing the evaluator to establish an internally valid estimation of the local average treatment effect around the cutoff-value. This, however, is also the greatest weakness of the method: The results of a regression discontinuity design are not simply transferable to other regions of the function. To plausibly deduce the project effectiveness of a project over the whole scale of M_i , the assumption of a constant treatment effect is necessary, which requires good arguments and some theory. Still, in many instances, only the marginal effect of increasing eligibility might be of interest. For example, when faced with the question of expanding a program, a regression discontinuity design may be exactly analyzing the population of interest, those just above the current threshold of participation (Hahn et al. 2001: 207).

⁴⁵ That $\varepsilon > 0$ cannot be too large as then the regression discontinuity might not become apparent implies as well that the regression discontinuity design is only feasible for large N studies where enough observations just around the threshold value can be analysed.

⁴⁶ The method is based on just a few assumptions, especially that the expected value of the outcome variable without treatment is continuous at the cutoff-value (“continuity restriction”) (Hahn et al. 2001: 202). Assumptions such as the mentioned absence of spill-overs additionally have to be taken into consideration.

⁴⁷ This is then called “fuzzy” instead of “sharp” regression discontinuity design.

In the case of IFSUP, a RD approach is not feasible due to missing data. The project was designed for all ultra-poor but eligibility was additionally determined by geographical proximity to group centers (conditional on covariates⁴⁸) to be formed for project implementation. In principle, the usage of this geographical proximity as cutoff value (e.g. walking distance in minutes) for inclusion would allow for a valid regression discontinuity design. One would compare the difference in outcomes of households just above and below this threshold, with one large advantage: Other non-experimental methods for estimating a causal effect would rest on the assumption that walking distance to the group meeting place does not induce bias – or at least that this influence is negligible. But this assumption does not necessarily have to hold: A short walking distance implies a housing within a cluster of other ultra-poor.⁴⁹

A regression discontinuity approach could invalidate concerns for this bias. Furthermore, a comparison to impact estimates by other methods could be interpreted as a test for its presence. But as data on the proximity to the group meeting place was neither recorded in the baseline nor in the impact questionnaire, this approach is impossible to implement.

2.4.3 Higher-order difference methods

The above mentioned methods aimed at establishing the case for causal interference through a look at the difference in outcomes at one point in time $Y_i^T - Y_i^C$, referred to as single cross-sectional analysis. In contrast, higher-order differences validate the assumption of exogeneity through exploiting changes in outcomes. Thus, they rely on longitudinal data. Though panel data is not necessary, it allows for additional slackening of assumptions, as outlined below.⁵⁰

Higher-order difference methods assume that “unobserved heterogeneity in participation is present – but that such factors are time invariant” (Khandker et al. 2010b: 71). This assumption of time invariance is the crucial point in limiting this method, as will be discussed below, but their main advantage over single-difference methods as well. In practice, the approach is oftentimes used to assess the impact of policy changes in so-called natural- or quasi-experiments, where policy changes affecting the outcomes in one group can then be compared to the development in other groups (Wooldridge 2009: 453f.).

⁴⁸ Preference was given to minorities, poorer poverty stratum and the female headed of eligible households.

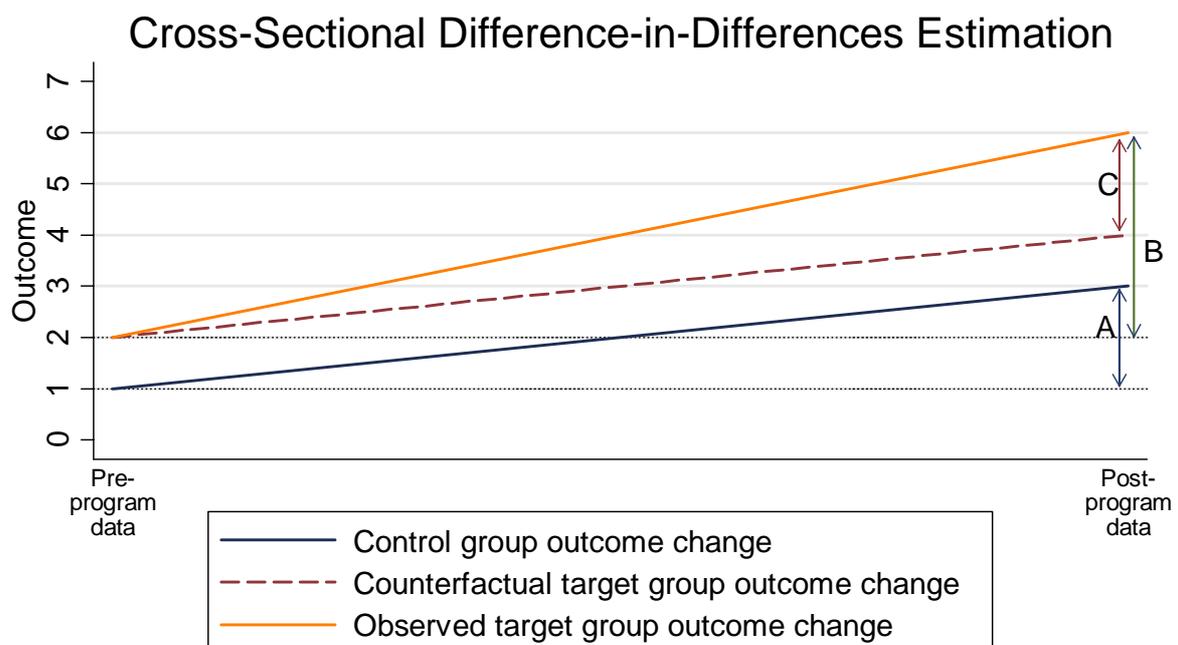
⁴⁹ Therefore, it is certainly possible that housing distance is correlated with outcomes depending on the predominant effects in either direction. Positive influences include e.g. short distance households with a higher likelihood of informal cooperation among ultra-poor household through some kind of risk-sharing arrangements, whereas longer distances indicate isolated ultra-poor livelihoods. Negative influences are equally possible. Longer walking distances could indicate a housing in a better-off neighborhood that provides quicker knowledge of employment opportunities or better access to financial capital.

⁵⁰ Comp. for a detailed discussion of the method Wooldridge (2010).

2.4.3.1 Pooling Cross-Sectional Data

The simplest case of higher-order difference methods is the comparison of changes between cross-section data-sets for control and target group in two time periods, the so called double-difference or difference-in-differences (DD) methodology: It gives, $ATE = E(Y_t^T - Y_t^C) - E(Y_{t-1}^T - Y_{t-1}^C)$, t indicating the post- and $t-1$ the pre-treatment period. Thus, not the control group outcome as such but the change in their outcome before and after project implementation is used as counterfactual. The condition $E(Y_i^C | G = 1) = E(Y_i^C | G = 0)$ of Chapter 2.2.2 for unbiasedness is thus relaxed to $E(\Delta Y^C | G = 1) = E(\Delta Y^C | G = 0)$ (Ravallion 2008: 3815). Figure 4 (on the basis of Khandker et al. 2010b: 75) exemplifies the set-up of a DD analysis: Between the pre- and post-intervention time period, observed mean outcomes changed from one to three for the control group, whilst the observed mean outcome for the target group changed from two to six. The different intersects of target and control group at the baseline period reveal that endogeneity in program placement is a problem. Therefore, a simple comparison between groups is not feasible. Yet, under the parallel trend assumption of counterfactual outcomes, the observed impact change for the target group (B) can be adjusted for the impact change observed in the control group (A). The impact as measured through DD is then given by the outcome change C, which is internally valid as far as the parallelism of counterfactual trends holds.

Figure 4: The principle set-up of difference-in-differences analysis



Own graph
The graph displays the construction of a counterfactual by the difference-in-differences estimation by assuming time-invariant selection bias

In practice, and as will be applied later for the IFSUP data in Chapter 3.2, a regression analysis is usually applied for estimating the DD estimator. This leads to an OLS estimation of $Y_i = \beta_0 + \beta_1 T_i + \beta_2 G_i + \beta_3 (T_i * G_i) + \text{covariates}_i + \varepsilon$, with $T_i = 1$ for the post-treatment period, $T_i = 0$ otherwise and $G_i = 1$ for project beneficiaries, $G_i = 0$ otherwise.⁵¹ The coefficient β_3 on the interaction of time period and treatment status will then give the DD estimator⁵², eventually conditional on additional controls (Wooldridge 2009: 453). As Khandker et al. (2010b: 73) highlight, DD will only be unbiased if both the regression model is specified correctly⁵³ and the correlations of the error term ε_i with all other variables in the equation are zero. Khandker et al. (2010b: 73, e.i.o.) especially emphasize the assumption $Cov(\varepsilon_{i,t}, (T_i * G_i)) = 0$, as this implies “the *parallel-trend* assumption [which] means that unobserved characteristics affecting program participation do not vary over time with treatment status”.⁵⁴

This parallel-trend assumption is the main obstacle to applying difference-in-differences in practice, which also applies to the IFSUP data. In data where e.g. attrition between the baseline and impact data occurred and the drop-outs cannot be expected to be randomly distributed, it is well possible to find arguments for different time trends in control and counterfactual target group outcomes. This is the case if the “changes over time are a function of initial conditions [of the surveyed population] that also influence program placement [or in this case the drop out process]” (Khandker et al. 2010b: 77). The DD impact estimator will consequently be biased. It is then important to figure out whether a mixture of approaches, such as the combination with matching methods (Todd 2008: 3869) provides a solution, or if it at least is possible to identify the direction of the bias.

⁵¹ This equation can include unlimited time periods, see Imbens & Wooldridge (2009: 68f.) for details.

⁵² It reveals the effect of the combined occurrence of a unit being in the post treatment period as well as in the target group compared to being in the baseline period as well as being in the control group. Thus, it indicates the effect of being in the target group on the outcome variable of interest over the time period change (Kam & Franzese 2007: 26).

⁵³ A linear-additive model is unproblematic for T=2 and no covariates. As soon as additional time periods are introduced or impacts of additional covariates analysed a linear model does not necessarily have the best fit. Potential problems of the functional form have to be kept in mind in every case, although as highlighted by Angrist & Pischke (2010: 11f.), good approximations to ATEs through linear models can be assumed in most cases.

⁵⁴ Additional assumptions concern the absence of serial correlation (or autocorrelation) of standard errors. A word of caution for interpreting DD results is voiced by Bertrand et al. (2004: 273f.): They argue that serial correlation is widespread in DD analysis and usually uncontrolled for, leading to overstated significance levels. They therefore caution the acceptance of DD estimates without tests for correlation and argue that the application of serial-correlation robust standard errors “should become standard practice in applied work”.

2.4.3.2 Panel Data Analysis

Additional attention deserves the presence of panel data: Panel data implies that data of all periods can be related to one single individual. This leads to additional possibilities in analyzing differences in outcomes, namely first differencing (FD), additionally with controls for lagged outcomes, as well as fixed effects (FE) models.

First differencing (Wooldridge 2009: Chapter 13.14) amounts to a regression linking individual and time specific data: $Y_{i,T} = \beta_0 + \beta_1 T + \beta_3(T * G_i) + \text{covariates}_{i,T} + \varepsilon_{i,T}$. But for the case without covariates, this can be reduced to a simple OLS regression of group status on the outcome changes, with $\Delta Y_i = \beta_1 + \beta_2 G_i + \varepsilon_i$ as Imbens & Wooldridge (2009: 70) point out. Analogous to the pooled cross-sectional DD estimator, this leads to an estimate of the treatment effect $\hat{\beta} = E(\Delta Y^T | G = 1) - E(\Delta Y^C | G = 0)$. The core difference is that due to the panel structure, the impact of the exact same individuals in these groups is subtracted before taking the mean. In these models, first differences are implicitly included for all fixed characteristics of the observed individuals as well. Oftentimes, concerns for unobservable biases inherent to the units of observation are a core limitation of an evaluation. These include motivation, skills, social networks, or other factors that can be assumed to influence both outcomes and selection but are not observable – and thus not controllable for mitigating selection bias. While both methods give similar estimates for data unaffected by e.g. attrition, in case bias through unobservables is a concern, first differencing allows to disregard this bias at least for the observed units. Concerning IFSUP, a reduced panel was constructed: Unobserved characteristics not changing over time are therefore no longer a source of bias in this panel. The question of course remains whether the results from the panel analysis are a valid approximation to the treatment effect for the overall group. One drawback of the method is that it may increase bias compared to cross-sectional DD if facing measurement if control variables are subject to measurement error: In this case “[d]ifferencing a poorly measured regressor reduces its variation relative to its correlation with the differenced error caused by classical measurement error, resulting in a potentially sizable bias” (Wooldridge 2009: 470).

Imbens & Wooldridge (2009: 70) describe how on top of the standard approach as outlined above, the researcher has the possibility to exploit the panel structure of the data by assuming exogeneity in program placement for lagged outcomes: This focus on lagged outcomes leads to an estimation of the change in outcomes between periods, $\Delta Y_i = \beta_1 + \beta_2 G_i + \beta_3 Y_{i,T=0} + \varepsilon$, with β_2 giving the causal effect of the intervention, whilst controlling for the baseline outcome level. As Imbens & Wooldridge (2009: 70) note,

“[w]hile it appears that the analysis based on unconfoundedness [exogeneity] is necessarily less restrictive because it allows a free coefficient in $Y_{i,0}$, this is not the case. The DID assumption implies that adjusting for lagged outcomes actually compromises the comparison because $Y_{i,0}$ may in fact be correlated with ε_i . In the end, the two approaches make fundamentally different assumptions. One needs to choose between them based on substantive knowledge ... [But] [a]s a practical matter, the DID approach appears less attractive than the unconfoundedness-based approach in the context of panel data. It is difficult to see how making treated and control units comparable on lagged outcomes will make the causal interpretation of their difference less credible, as suggested by the DID assumptions“.

This approach has the main advantage of having included a major time variant component that may have confounded standard difference-in-difference estimations.⁵⁵ On top, panel analysis allows to attach additional information about the unit of observation at baseline level and its changes to this estimation. In case more than two time periods are present, an expansion is possible through the inclusion of additional time period dummies (Wooldridge 2009: Chapter 13.15).

An alternative approach of utilizing the panel structure of the data would be a fixed effects model, which controls for unit-level unobservables as well. This fixed effects regression is as well called “within estimator” (Wooldridge 2009: 482) and can intuitively be interpreted as an inclusion of dummies for each observed unit. By this, it allows for a different intercept for each of the individuals. The regression thereby draws only on the variation of variables over time within each unit, by this circumventing any bias that could arise from unobservable but fixed characteristics. Fixed effects regressions are identical to first differencing for the simple case of two time periods. For more than two time periods, both are still similar with large N and T. But for “large N and small T, the choice between FE and FD hinges on the relative efficiency of the estimators and this is determined by the serial correlation of the idiosyncratic errors” (Wooldridge 2009: 487).⁵⁶

In the later application to the IFSUP case, both a cross-section and a panel difference-in-differences approach will be utilized; concerning the latter, a choice between FD and FE does not have to be taken due to the number of time periods $T=2$.

⁵⁵ As a panel approach by definition excludes the influence of time-constant variables, it prohibits the measurement of the effect of time-constant variables on the outcome of interest: “Even if we can obtain panel data, it does us little good if we are interested in the effect of a variable that does not change over time: first differencing or fixed effects estimation eliminates time-constant explanatory variables” (Wooldridge 2009: 506).

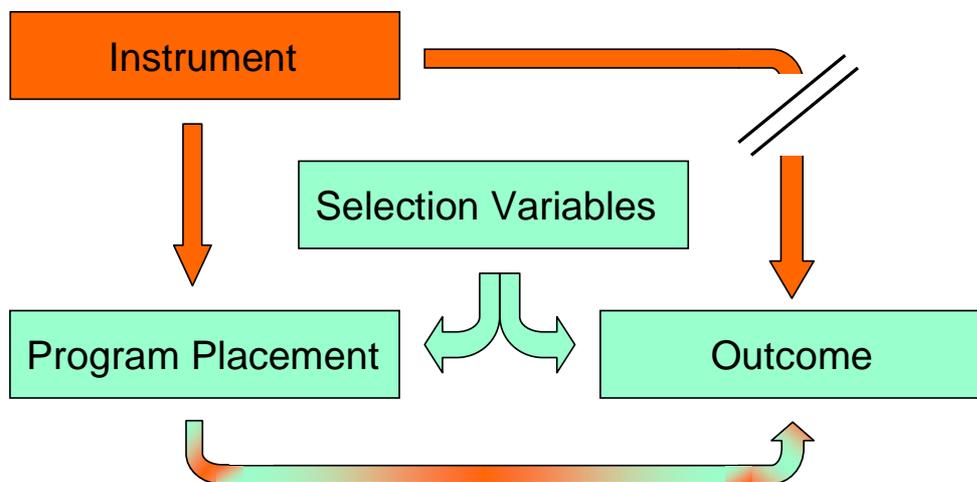
⁵⁶ See for a further comparison e.g. Angrist & Pischke (2009: Chapter 5.3). A lot of additional econometric literature is evolving around program evaluation with panel data and panel data analysis in general, Imbens & Wooldridge (2009: 72) discuss e.g. new ways of constructing artificial control groups in cases data on more than two groups is present, whilst Semykina & Wooldridge (2010: 380) develop further methods to correct for endogeneity arising from “time-specific variances and arbitrary serial dependence in idiosyncratic errors”.

2.4.4 Instrumental variables

While higher-order-differences allowed through focusing on changes a cancelling out of time and/or unit constant bias, the instrumental variable (IV) method is one of the methods that accept that unobserved heterogeneity is present, but bypasses the bias arising from it.

As exemplified in Figure 5, endogeneity in program placement leads to outcomes not only indirectly affected by selection variables through the individuals selected for the intervention but also directly. This would e.g. happen if target and control group differed by entrepreneurial ability in a microcredit intervention and field workers (unconsciously) chose, next to other criteria, ultra-poor they deemed more able for success. Measurement is then upwardly biased, as the target group higher counterfactual outcomes compared to the population they were drawn from. A proper instrument now has the attribute of excluding the direct relationship between selection variables and outcome. Then, only the direct effect of program placement on outcomes can be measured. But not all selection is based on this ‘ability’-factor. An instrument therefore has to be found that is correlated with selection but not with ability and/or outcomes.⁵⁷

Figure 5: The logical framework of the instrumental variable method



Own Graph

In a regression framework, these two attributes of an instrument Z are formally summarized by the exogeneity assumption demanding no relationship between the instrument and the error

⁵⁷ For the above example, this could be the variable ‘proximity to NGO headquarters’, if it can be plausibly argued that this proximity is positively influencing program placement but is neither correlated with ability nor outcomes; for this assumption, headquarter location more or less must have been selected at random. The case for this correlation to ability/outcomes not being present has to be plausibly established through argumentation, which might be problematic, e.g. as soon as headquarter location is related to infrastructure and proximity to infrastructure to better counterfactual outcomes in expectation. The instrument would then be endogenous in itself and possibly even aggravated the problem.

term $Cov(Z, \varepsilon) = 0$, as well as the relevance assumption demanding that variation in placement is related to Z , $Cov(G, Z) \neq 0$ ⁵⁸ (Wooldridge 2009: 508). The evaluation of the influence of group status G_i on outcomes Y_i is therefore conducted by regressing in two stages.

The first regression keeps hold of the exogenous part of the selection variables, the second one then estimates the impact of selection on the outcome variable of interest.⁵⁹ The validity of results rests on the credibility of these assumptions. The relevance assumption $Cov(G, Z) \neq 0$ is in principle testable, and thus is not only able to help detect misspecifications of instruments but as well caution the use of so-called ‘weak instruments’, referring to an insufficient relationship between instrument and placement that leads to imprecise results even in large samples (Khandker et al. 2010b: 91, Wooldridge 2009: 516). On the other hand, the exogeneity assumption $Cov(Z, \varepsilon) = 0$ is generally not testable (Wooldridge 2009: 508)⁶⁰: Its plausibility therefore has to be assessed in every single case. The relevance of qualitative data and of a proper understanding of the project context, possibly in combination with theoretical framework of the intervention pathway, comes back into the picture again.

Instrumental variables are an appealing approach to solving endogeneity. But several notes of caution are important:

First of all, as pointed out by Ravallion (2008: 3823f.) the IV approach implicitly includes the crucial assumption

“that impacts are homogeneous, in that outcomes respond identically across all units at given [levels of covariates] X [... Thus] the validity of causal inferences typically rests on ad hoc assumptions about the outcome regression, including its functional form. [E.g. propensity score matching], by contrast, is non-parametric in the outcome space.”

Secondly, Khandker et al. (2010b: 92f.) summarize that instruments usually do not capture all variation that leads to program participation. Thus, IV is in many cases only measuring local average treatment effects (LATE) that are valid for a subset of the population – similar to RDD. The LATE “is the average effect of treatment for the subset of persons induced by a change in the value of the instrument from Z_1 to Z_2 to receive the treatment” (Todd 2008: 3879). Latest developments initiated by Heckman & Vytlacil (2005), as summarized by Khandker et al. (2010b: 94), allow for an estimation of Marginal Treatment Effects. This es-

⁵⁸ Z can include multiple instruments. It can as well serve as instrument for more than one endogenous variable, as long as the regression contains “at least as many excluded exogenous variables as there are included endogenous explanatory variables” (Wooldridge 2009: 524).

⁵⁹ Hence, the IV approach is commonly referred to as the two-stages-least-squares approach (2SLS or TSLS) (Khandker et al. 2010b: 90), as two OLS models are fitted one after the other.

⁶⁰ Comp. Wooldridge (2009: 527-531) for the test of overidentifying restrictions that can in the case of multiple instruments aid in establishing a case for exogeneity.

timation tries to construct the overall impact function. Depending on the data available, it is possible to deduce ATE, LATE or TOT. Heckman & Vytlacil (2005: 669) by this aim to reconcile the structural and treatment effect approach to policy evaluation; they try to estimate and verify parts of a predefined structural equation.

Deaton (2010a: 429) raises a more fundamental concern about IV methods that is connected to the theory linkage in the marginal treatment effects approach: He argues that IV methods are oftentimes used to create analogies to experiments when the focus should rather lie on “the variation in [the impact estimator] that encapsulates the poverty reduction mechanisms that ought to be the main objects of our inquiry”⁶¹. He therefore insists that only a full understanding of the poverty reduction mechanisms at work, and therefore of the heterogeneity in place, allows for a credible adoption of IV methods – otherwise it is possible that numerous biases are contained in the instrument but not thought of. “The general lesson is once again the ultimate futility of trying to avoid thinking about how and why things work – if we do not do so we are left with undifferentiated heterogeneity that is likely to prevent consistent estimation of any parameter of interest” (Deaton 2010a: 432).

Two practical examples shed light on the question whether IV methodology could be a useful approach for the IFSUP case once the impact pathways are understood.

Khandker (1998: 205-207) uses program eligibility rules as instrument to solve endogeneity in program placement: In his evaluation, only households below a certain landholding-threshold were eligible for participation in a microcredit program. A dummy indicating landholdings below the threshold instrumented participation. To capture the unobserved effect of landholdings on outcomes, the quantity of land owned was additionally included in the impact regression. In principle, a similar approach could be used for assessing the IFSUP impact if the data were available: As discussed in the case of RDD, project participation was determined by distance from the group meeting points. If this data were present, a dummy indicating walking distance below a threshold could serve as instrument for inclusion. To control for differences in economic opportunities based on walking distance, the absolute walking distance could be included to the regression, analogous to the Khandker-evaluation.⁶²

⁶¹ He exemplifies his case by the (artificial) impact of railways on poverty reduction “Good candidates for [the instrument] might be indicators of whether the city has been designated by the Government of China as belonging to a special ‘infrastructure development area,’ or perhaps an earthquake that conveniently destroyed a selection of railway stations, or even the existence of river confluence near the city, since rivers were an early source of power, and railways served the power-based industries. I am making fun, but not much” (Deaton 2010a: 482).

⁶² Unfortunately, this data was not collected, which shows the importance of planning evaluations in detail. But even this approach would not be clear-cut: Meeting points are not only geographically defined e.g. by village

In the context of microcredit impact in India, Garikipati (2008: 2628) uses the variables “size of the respondent’s neighbourhood cluster [and belonging to the] dominant caste within the cluster” to instrument inclusion of households, exploiting that like in IFSUP 15 members are required to form a self-help group and similar cast-membership can be expected to enhance cooperation probability. Garikipati (2008: 2631) later uses these instruments to estimate the impact of group membership on vulnerability. But this approach has to be questioned as the direct relationship between the instruments and vulnerability cannot be expected to be zero: She assumes similar caste membership to enhance trust and thus establishes the case for her instrument. At the same time, however, it seems likely that similar caste membership influences directly one of her outcome variables, drought-related vulnerability, defined among others as “during the last drought, the household met all its food needs” (Garikipati 2008: 2626). This is certainly less probable if the household is not well connected in the village, as could be indicated by her instrument caste membership. In the IFSUP case, cluster-size, ethnicity, and religion could similarly serve as instruments for participation probability. But the model determining selection is not clear enough to exclude endogeneity issues.

2.4.5 Bounding the Endogenous Placement

In cases such as these, there may be no direct solution to the endogeneity present. However, it is possible to frame the problem as a problem of missing information. Ensuring a balance in the data would require adding data that is not available. Manski (comp. amongst others 1990, 1997, 2007) was among the first to propose an ingenious solution for this problem. It is often possible to narrow down the endogeneity problem by constructing bounds on the maximum bias present based on minimal assumptions. These bounds can lead to an interval that is informative to the overall treatment effect.⁶³ Manski (2007: 7) argues that this approach is much more reliable than a thrive for point estimates:

“Conventional practice has been to invoke assumptions strong enough to identify the exact value of this parameter. Even if these assumptions are implausible, they are defended as necessary for inference to proceed. In fact, identification [the possibility of inference if one had an infinitely large sample] is not an all-or-nothing proposition. We will see that weaker and more plausible assumptions often partially identify parameters, bounding them in informative ways.”

centres but as well depending on group members’ houses. It is unclear and cannot be confirmed due to missing data if participation was established by minimizing village distance or if other unobservables played a role.

⁶³ This bounding approach is taken up by various other researchers (comp. e.g. Balke & Pearl 1997, Lee 2002, 2009, Pearl 2009: Chapter 8.2). Comp. for one of the rare applications of bound applications (in this case Lee-bounds) to explaining income changes in South Africa Leibbrandt et al. (2010: esp. pp. 39-44) with an unbalanced panel.

Manski (2007: 183, e.i.o.) bases his results on minimal assumptions. Informative bounds can for example be derived under the assumption of monotonicity in treatment: The assumption of “*monotone treatment response* (MTR) [...] permits each unit of treatment to have a distinct response function [but] requires that all response functions be monotone.” It can be formalized for a dichotome setting by $T | G_i = 1 \geq T | G_i = 0 \Rightarrow y_i(T | G_i = 1) \geq y_i(T | G_i = 0)$: This is to say that a more (less) intense level of treatment T is related to a higher (lower) level of outcomes for a given individual i . As only one outcome is observable, this assumption cannot be tested; still it allows the potential of treatment invoking zero-results. Manski (2007: 186-189) describes how this allows a construction of bounds on the treatment effect $E(0) \leq E(Y^T - Y^C) \leq E(Y^{\max,T} - Y^{\min,C})$, where outcomes are bounded by a nil-effect and the maximum difference in possible responses. Although a very simple assumption, MTR can still be difficult to uphold in practice: On an *a priori* basis, an in-kind capital transfer cannot be expected to increase outcomes. It is plausible that investment in livestock is related to a risk of zero or negative outcomes in case the livestock dies. If additionally the household chooses to substitute outside labour, the monotonicity assumption could be violated in all cases where livestock generates zero or negative returns and counterfactual outcomes would be positive (e.g. due to work on the wage labour market). In the case of IFSUP, monotonicity can therefore not plausibly be upheld.

But other bounding approaches exist: As outlined in Manski (2007: 36-45), bounds for longitudinal data where some outcomes are missing can be constructed on the basis of three vectors of interest: The outcome variable, the relevant covariate vector including foremost treatment status as well as “a binary variable [z] indicating when outcomes are observed” (Manski 2007: 37). Although the author uses his setting for a general framework, it can easily be applied to the impact evaluation problem: Assume the problem that a random population subset is treated but non-random attrition occurs. Object of interest might be the probability $P(Y \in B | G)$ that the outcome of interest Y falls into a specified set B conditional on treatment status. As Manski (2007: 38) argues, from the Law of Total Probability follows that $P(Y \in B | G) = P(Y \in B | G, z = 1) * P(z = 1 | G) + P(Y \in B | G, z = 0) * P(z = 0 | G)$. Through randomization, information is not only provided on the equation parts $P(Y \in B | G, z = 1)$ that can be observed, but also on the probabilities $P(z = 1 | G)$ and $P(z = 0 | G)$ that are revealed in comparison with the baseline information, about which full data is available. Unobservable is $P(Y \in B | G, z = 0)$, the outcomes of the attritors in both groups. Given it is a probability, however, it must lie on the interval $[0,1]$. Sharp bounds for the worst-case scenario, in the

sense that they cannot be over- or underrun, thus can consequently be constructed by assuming either 0 or 1 for all missing outcomes:

$$P(Y \in B | G, z = 1) * P(z = 1 | G) \leq P(Y \in B | G) \leq P(Y \in B | G, z = 1) * P(z = 1 | G) + P(z = 0 | G).$$

This approach is useful in other settings as well: Manski (2007: 36-45) subsequently applies it e.g. to quantiles of $P(Y \in B | G)$, Imbens & Wooldridge (2009: 51-53) to the case where outcomes of interest lie within the interval [0,1]. Application is even possible when outcomes are normalized to the interval [0,1] as long as the assumption is plausible that the highest (lowest) observed outcomes provide a feasible mean upper (lower) bound for missing outcomes – the latter two estimations will be used in Chapter 3.2. Importantly, however, the information contained in the bounds depends on the width of the interval thus constructed, which in turn is depending on the amount of information missing.

Despite requiring no assumptions on the selection bias, Manski bounds have their great limitation in the information they contain: Their application is oftentimes not informative, as in the assessment of their usefulness by DiNardo et al. (2006: 26): “We do not display the Manski–Horowitz⁶⁴ bounds for our two bounded outcomes since these bounds are substantially larger than the Lee bounds we present and we can not rule out very large negative or positive treatment effects”. Analogously Lechner (1999: 23) notes in one of the few other applications that “without good knowledge of the relationship between potential outcomes and the selection/assignment process, it is very difficult to bound the treatment effects strictly away from zero.” Imbens & Wooldridge (2009: 51-53) and Duflo et al. (2008: 3944) consequently emphasize the applicability of Manski’s framework especially to a case with covariate information and knowledge on the direction of the bias to further sharpen the bounds above.

Overall, however, Manski (2007: 6-8) makes a strong case for constructively dealing with ambiguity: Researchers must accept that the social sciences seldom allow direct causal inference. Accepting ambiguous outcomes deriving from concurring hypothesis is certainly preferable to arbitrary choices based on e.g. the complexity of the hypothesis behind the results. In the case of bounding the effect, even if only loose bounds can be obtained with unrestrictive assumptions, this cannot be a case for not reporting them: One must “[focus] on the fact that [even a wide bound] establishes a domain of consensus about the value of [interest, regardless] of assumptions about the attrition process [...] Wide bounds reflect real uncertainties that cannot be washed away by assumptions lacking credibility.” (Manski 2007: 42).

⁶⁴ Referring to Horowitz & Manski (2000).

2.4.6 Sensitivity analysis

While the presented methods above were either assuming a solution for the bias, a circumvention or bounds for it, sensitivity analysis deals with uncertainty on endogeneity by a stepwise relaxation of the endogeneity assumption – “violations of unconfoundedness are interpreted as evidence of the presence of unobserved covariates that are correlated, both with the potential outcomes and with the treatment indicator” (Imbens & Wooldridge 2009: 53). As Imbens & Wooldridge (2009: 53-56) outline, the degree of bias in outcomes (developed in Rosenbaum & Rubin 1983) or in the endogeneity of placement (developed in Rosenbaum 1995) is assessed that still allows a effect in the direction of interest. To arrive at sensible conclusions, the researcher has to choose hard limits to the size of a possible unobserved effect or the influence of unobservable on selection relative to the effect of observed variables – without these limits, no information can be extracted. DiPrete & Gangl (2004: 302) assess the usefulness of the method for bias in matching estimators and call for a routine application to increase confidence on estimates.

2.4.7 Control function methods

Lastly, if endogeneity is recognized as a (potential) problem, the researcher can aim at modeling the selection process through a control function. The approach goes back at least to Heckman (1979), who in viewing the omitted variables problem as a specification problem addressed omitted variables through their estimation: Then, “estimated values of the omitted variables can be used as regressors so that it is possible to estimate the behavioral functions of interest by simple methods” (Heckman 1979: 153). Estimation of these unobservables is based on variation in observed units. In the application to program evaluation, this “allow[s] selection into the program to be based on time varying unobservable variables at the expense of stronger functional form assumptions needed to secure identification” (Todd 2008: 2874). Todd (2008: 2874-2877) outlines further that an estimation of control functions is related to matching but constructs scores based on observables as well as estimated unobservables.

2.4.8 Combination of approaches

The above discussion treated one method after another: In practice, however, these methods must not be seen as substitutes but rather as complements. In his famous critique of observational studies on the effects of employment and training programs, LaLonde (1986: 617) compared the faring of experimental and non-experimental measurements and warned of non-experimental techniques that possibly “contain large and unknown biases resulting from

specification errors”; the consequence was an advice to make use of randomization. Even careful matching is no general solution to this problem, as Smith & Todd (2005) are able to show; but the reliance of observational methods can be increased through their combination. Smith & Todd (2005: 347) call for a case-by-case application of methods and a combination of approaches if necessary: Difference-in-differences matching is e.g. one promising solution to biases from data with different sources, as changes are looked upon. As Todd (2008) further notes, “difference-in-difference matching methods are more reliable than [single] cross-sectional matching methods, particularly when treatments and controls are mismatching geographically or in terms of the survey instrument.” Even the applicability of randomization may be increased through a combination of approaches: Attenuation bias⁶⁵ might be a concern where only the intention to project take-up can be randomized but not actual take-up. Random assignment can still be used as instrumental variable for measuring “intention-to-treat” effects (Khandker et al. 2010b: 42f., 89). Last but not least, bounds are not only a sensible approach if faced with biased randomized studies, but just as much in combination with observational studies, such as the bounds on a biased IV estimator by Hotz et al. (1997) concerning the effect of teenage childbearing on educational outcomes.

2.5 General Drawbacks of the Outlined Approaches

2.5.1 Spill-Overs as Confounding Factor

The discussion so far assumed a separable treatment for the intervention group and excluded the possibility of reciprocal interference between treatment and control group. Due to the embeddedness of quantitative research in complex social systems, however, feedback mechanisms between both groups are a major obstacle for measuring the causal effect of interventions, since they are potentially confounding results. Even with random assignment, the argument of thus deriving the ATE rests on the assumption of the independence of treatment group and non-treatment group, called “stable unit treatment value assumption (SUTVA) [...] SUTVA is a complicated assumption that is all too often ignored” (Sekhon 2009: 472). Even Manski (2007: 129) with his cautious approach to inference only in passing mentions how

“[t]he notation $y_i(\cdot)$ also supposes that treatment response is individualistic. That is, the outcome experienced by person j depends only on the treatment that this person receives, not on the treatments received by other members of the population. This assumption will be maintained throughout our analysis.”

⁶⁵ Attenuation bias is induced by a measurement error where e.g. “program participation varies more than it actually does. This overestimation in the variance of [treatment status] leads naturally to an underestimation of its coefficient b . This is called attenuation bias because this bias attenuates the estimated regression coefficient” (Baker 2000: 57).

For development interventions, SUTVA is difficult to uphold in many instances, as spill-overs are, intentionally or not, built in side effects of interventions: Concerning IFSUP, the project aims e.g. explicitly at changing the socio-political context the target group as well as control group ultra-poor are living in. Under these conditions, $E(Y_i^c | G = 0)$ is certainly influenced by the project. Ravallion (2008: 3796-3798) highlights an additional source of spill-over bias: the displacement of other project plans through external interventions. External interventions, e.g. a social protection initiative such as IFSUP, can replace government efforts that would have been undertaken counterfactually. If government resources for ‘target group ultra-poor’ are subsequently redirected towards ‘control group ultra-poor’, a downward bias is introduced in the estimates. Analysis must take this into account and try to construct arguments for the severity of the bias (e.g. whether it could be neglected or whether spill-overs could cancel out) or the direction to obtain at least an idea of over- or underestimation of the project effect.

But spill-overs can as well be introduced very subtly to a program e.g. by autonomous decisions of the project staff, as is the case for the Mexican government’s conditional cash transfer program PROGRESA, much hailed for its randomization approach and its carefully planned impact assessment strategy (comp. for a short overview Khandker et al. 2010b: 10). Yet, a mere impact analysis based on the assumption of random treatment assignment returned a zero impact. Only after taking into consideration that resources were redirected to malnourished children not covered by the program and redistributed among intended beneficiaries by local staff based on their own assessment of need, the project impact could be adequately estimated (Behrman & Hoddinott 2005: 554f.). This case emphasizes the need for detailed knowledge of the operational procedures of the actual project implementation for the ability to take these factors into account.⁶⁶ Ravallion (2008: 3790) for this explicitly introduces the concept of an “average treatment effect on the untreated” (TOU).

Sometimes, ingenious ways of program design provide solutions: Miguel & Kremer’s (2004: 208f.) approach to the confounding influence of spill-overs with treatments and measurements at the individual level (leading to positive TOU and a negative bias in measured ATE) was the introduction of target and control areas. Geographical distance thus allowed them to rule out spill-over effects and a comparison of the estimates even made the argument possible that spill-overs alone are so substantial that they justify the program. Methodically

⁶⁶ Additionally, in case these disruptions of the project routines are noted, the researcher needs some luck so that the unobserved heterogeneity can be controlled for in the data. In the PROGRESA case, Behrman & Hoddinott (2005: 547) found a way out through “child fixed-effects estimates that control for unobserved heterogeneity that is correlated with access to the supplement”.

speaking, in this context spill-overs would bias the results significantly towards zero. But even a group-level approach cannot be an overall solution, as the authors note:

“While group-level randomization can be used in other settings with externalities localized, either geographically or along some other dimension, such as the analysis of school vouchers or information transmission and technology diffusion, it cannot be used to estimate more global spillovers, such as those arising through general equilibrium price effects.”

They therefore argue for randomization at various levels, which is of course an approach not feasible in many instances and potentially largely inflating the costs of the intervention.

2.5.2 Issues of Heterogeneity

2.5.2.1 Heterogeneity in Impact Pathways

Connected with the spill-over theme is the question of heterogeneity in impact pathways. The intervention pathway is not revealed by a conventional ATE analysis, and thus neither the effect of various project components nor the effect of potential spill-overs within the control or target group through the program. As Rubin (1974: 700, FN 713) notes,

“[e]ven assuming a good estimate of the causal effect of E versus C, there remains the problem of determining which aspects of the treatments are responsible for the effect. Consider, for example, ‘expectancy’ effects in education [...] and the associated problems of deciding the relative causal effects of the content of programs and the implementation of programs.”

For IFSUP, this is a highly relevant question as the intervention used a threefold social protection approach intervening with the capital base of households, their vulnerability as well as the socio-political context, a set-up which in other contexts may only be replicable in part.

Duflo et al. (2008: 3902f., e.i.o.) stress for randomized experiments (but the point applies more general) that the ATE reveals the total project effect including the impact of all project effects on the variable of interest as opposed to partial derivatives

“*keeping everything else equal* [...] But the total derivative may not provide a measure of overall welfare effects. Again consider a policy of providing textbooks to students where parents may respond to the policy by reducing home purchases of textbooks in favor of some consumer good that is not in the educational production function. The total derivative of test scores or other educational outcome variables will not capture the benefits of this re-optimization. Under some assumptions, however, the partial derivative will provide an appropriate guide to the welfare impact of the input.”

An obvious solution to this problem would be a differentiation in treatments, by which sub-populations of the overall population receive subsets of the treatment – Chowdhury et al. (2009) provide an interesting example of this: They are concerned with the causes of migration constraints despite the presence of seasonal hunger periods amongst the poor in northern Bangladesh. They subsequently provide a differentiated treatment of information packages and cash-incentives to randomized subsets of the target population. Their results reveal savings- and borrowing constraints as prime limiting factors of migration, compared to other possible sources such as information asymmetries. But equally important is a focus on under-

standing the mechanisms by which programs work (Deaton 2010b). For this, methods beyond econometric impact evaluation are important: White (2006: 17) e.g. refers to the case of a qualitative evaluation, where the qualitative study

“identified the mistargeting of potential beneficiaries and the failure of mothers to put into practice the nutritional knowledge they acquired through counseling. A randomized approach might have only shown that the project didn’t work, and would not have helped explain the causal factors.”

2.5.2.2 Heterogeneity in Treatment Effects

The impact evaluation literature is traditionally focused on ATEs. Sophisticated methods are applied that allow for heterogeneity in treatment effects and are nevertheless able to extract a meaningful average treatment effect under minimal assumptions, such as in the RCT case. This approach, however, tends to suppress a concern that is often essential: Heterogeneity in treatment effects and its causes might just be what the object of interest is. It might be that “it is precisely the variation in [the impact estimator] that encapsulates the poverty reduction mechanisms that ought to be the main objects of our inquiry” (Deaton 2010a: 429). Thus, interest in treatment effects in many cases needs to go beyond mean analysis. Policy decision must e.g. additionally be based on distributional results on which ATE analysis says nothing. ATE analysis might thus lead to misinterpretation of effectiveness as the variability of the outcome and its reasons are not assessed: Although having the same ATE, a project where all participants gain slightly is radically different from a project where the status of half the participants worsened although the other half overcompensated this loss through their gains.

Kanbur (2001: 1086) asks how it is possible that in the development discourse “people [are continually] talking past each other, each side equally convinced that it has the truth, even when confronted with seemingly the same objective reality.” Just like in the above case, the problem rests on the level of aggregation. It is in this respect important to keep in mind that the method of impact analysis should not determine the levels of analysis possible. If the optimal method for clear-cut analysis of mean developments reveals nothing beyond the mean, this should not imply that beyond-mean evaluations are not to be undertaken – a switch in methods might be needed. Median effects or the fraction of the population positively affected are similarly important for policy decisions (Deaton 2010a: 439). This was as well highlighted by Rubin (1974: 690) in the original formulation of the counterfactual problem: He chose the mean not for its analytical strength (though this might be in many cases), but only because other estimators such as the median or midmean “lead to more complications when discussing properties of estimates under randomization”. Kanbur (2001: 1093) in this light calls for a look beyond these methodical limitations: “[T]he message is that explicitly taking into ac-

count these complications is more likely to shift the intellectual frontier than falling back yet again on conventional analysis.”

Additionally, even if analysis of the mean effect of a specific intervention convincingly reveals zero-results, “there is no implication about any specific dam [(the average effects of dam construction on poverty is used as example)], even one of the dams included in the study, yet it is always a specific dam that a policymaker has to approve” (Deaton 2010a: 441). If heterogeneity in responses is large, the mean analysis cannot produce a general guideline for the single-case examination.

The effort of learning about heterogeneity leads directly to subgroup analysis: The question then is whether based on different observable characteristics, failure or success of the intervention is upheld. Khandker et al. (2010b) argue for the application of graphical and descriptive analysis for a first understanding of potential heterogeneity. Besides descriptive analysis, two approaches are commonly used: Quintile analysis and interaction terms. For a quintile analysis, the data on the intervention of interest is separated into different distinct subpopulations defined through their levels of observational variables (comp. for an overview Imbens & Wooldridge 2009: 17f., Khandker et al. 2010b: 118-124). Carter (2008) for example uses information on pre-intervention asset levels to determine heterogeneous responses of the effect of natural disasters on asset growth.

For the interaction term approach, the characteristics by which treatment is supposed to be heterogeneous are entering a regression analysis via interaction effects, estimating the effect of the characteristics themselves on the outcome of interest as well as their interaction with treatment status. This therefore leads to estimations of regressions as in Deaton (2010a: 440): $Y_i = \beta_0 + \beta_1 G_i + \sum_j \delta_j * X_{i,j} + \sum_j \gamma_j * X_{i,j} * G_i + \text{further covariates} + \varepsilon_i$ (comp. for a detailed overview on the method Kam & Franzese 2007).

These approaches to heterogeneity are potentially very informative on the impact pathways and can lead to a better understanding of the intervention as well as improved policy in the future. Banerjee & Duflo (2010c: 26) for example estimate the differing impacts of microcredit on households with pre-existing business: They thereby can conclude that owning businesses changes the use of the loan but not the probability of taking it in the first place. Their analysis is inspired by theory on credit constraints according to which efficient markets should direct money to places with higher returns on investment – as these can be supposed to be higher in existing businesses without fixed costs, their results are intriguing.

Overall, it should be clear that at least tests for heterogeneous impacts should be conducted. Crump et al. (2008) develop for example tests to analyse whether, defined for a set of

covariates, treatment has zero or constant effects in all subgroups. Important is the call of Khandker et al. (2010a: 124) in this respect: To be able to sensibly measure heterogeneous impacts, one needs to carefully choose which data to gather on the individual as well as household and community level.⁶⁷ For these considerations alone, the link back to theory is crucial. But, as White (2009: 282) summarizes:

“Criticisms of reporting an average treatment effect should not be overstated. Heterogeneity matters, as does understanding the context in which a particular impact has occurred. But it will rarely be the case that the average treatment effect (usually both the treatment of the treated and the intention to treat) is not of interest. Indeed it is very likely to be the main parameter of interest. It would be misleading to report significance, or not, a particular subgroup if the average treatment effect had the opposite sign. Moreover the average treatment effect is the basis for cost effectiveness calculations.”

2.5.3 Data Mining Risks

The danger of using approaches such as interaction terms or non-linear specifications certainly rests in data mining, by which researchers can be tempted to specify their analysis just until they can report significant effects, as Rubin (1974: 700) notes:

“With or without random sampling or randomization, if an important prior variable is found that systematically differs in E and C trials or in the sample and target population, we are faced with either adjusting for it or not putting much faith in our estimate. However, we cannot adjust for any variable presented, because if we do, any estimate can be obtained.”

Brambor et al. (2006: 78) report for the context of general multiplicative interaction models that an incorrect application of interaction terms is flawing the reliability of articles in top journals of the political science: They find for example that

“[o]f the 101 articles that actually interpreted one or more of the constitutive terms, 63 interpreted them as unconditional or average effects. [...] We can say with absolute certainty that interpreting constitutive terms as unconditional or average effects is wrong. It is simply not appropriate to interpret the results of interaction models as if they came from a linear-additive model.”

The more general point of this analysis is that the case for causal inference is getting more complicated with additional regressors.⁶⁸ As Ravallion (2008: 3793) notes usually the “X’s enter in a linear-in-parameters form. This is commonly assumed in applied work, but it is an ad hoc assumption, which is rarely justified by anything more than computational convenience (which is rather lame these days).” Sims (2010: 66f.) backs this argumentation. He argues that standard errors, functional forms and the significance of results are intrinsically related: “Observing that robust standard errors are quite different from conventional ones, which do not cluster or account for heteroskedasticity⁶⁹, should be a signal to us that there is a

⁶⁷ The latter aspect of careful data gathering is related to the decision of the size of the sample as well: There exists an inherent trade-off between the costs of the intervention or at least its measurement and the effects that can be measured: The larger the sample size, the smaller the “minimum detectable effect size” and vice versa (Armendáriz & Morduch 2010: 302f., Duflo et al. 2008: 3918). Subgroup analysis inflates this trade-off.

⁶⁸ Besides, adding too many interaction terms can, besides reducing the degrees of freedom, lead to problems of multicollinearity (Khandker et al. 2010b: 117).

⁶⁹ Heteroskedasticity refers to the fact that variation in the error terms of an OLS regression is correlated amongst different strata of the population (Wooldridge 2010: 59-62).

great deal going on in the data that our linear model is missing” (2010: 66). As he outlines, a lot can be learned on the heterogeneity in-between the subjects of the observation from more careful approaches that are not too complicated to implement.⁷⁰ Other authors argue for a focus on more important aspects than linearity, such as Angrist & Pischke (2010: 11f.): “The linear models that constitute the workhorse of contemporary empirical practice usually turn out to be remarkably robust, a feature many applied researchers have long sensed and that econometric theory now does a better job of explaining”.

Some methods are especially sensitive to data mining risks: In RDD, the evaluator has the possibility to choose the bandwidth of $m \pm \varepsilon$, model specifications other than linear approximation around the cutoff-value as well as the inclusion of covariates which can strongly change the results of the analysis. The great effect of these modifications is reported by Green et al. (2009), who try to establish general rules especially for bandwidth selection. Further work in their line is required, testing the accuracy of RDDs in cases that are comparable to estimates from a classical randomized evaluation. This would bear the possibility to establish more knowledge on the rules of applying the designs, especially the bandwidth selection question. The “data mining” Deaton (2010a: 440f.) fears in respect to subpopulation analysis thus comes back into the picture: “In large-scale, expensive trials, a zero or very small result is unlikely to be welcomed, and there is likely to be considerable pressure to search for some subpopulation or some outcome that shows a more palatable result, if only to help justify the cost of the trial.” Bendavid (2011: 20) therefore makes an analogy to medical trials: “Generating significant findings through small tweaks of the research design and subtle changes to inclusion and exclusion criteria are unsavory practices, but the motivation behind them is unsurprising.” He therefore argues for the set-up of ex-ante-trial registration for experiments, just as in the medical field: In this, he has a strong point. Only the up-front declaration of research designs, including especially sub-group analysis, can prevent data-mining concerns. These data mining concerns must not prevent ex-post analysis of intriguing questions that originate from later research but can be added with a note of caution to results from the pre-declared research design. Glennerster & Kremer (2011a: 29) warn in this context that “[s]afeguards are important, but we don’t want the perfect to be the enemy of the good”.

⁷⁰ Sims (2010: 66) emphasizes the “use of generalized least squares [or] what statisticians call ‘mixed models’, in which conditional heteroskedasticity is modelled as arising from random variation in coefficients.” See Wooldridge (2010: Chapter 20.23.22, Chapter 22.23.24) for further details.

2.6 A Hierarchy of Methods?

The discussion of methods implicitly involved a hierarchy on the theoretical level: As presented, the most straightforward approach to the counterfactual and thus to internally valid estimates is achieved through RCTs. The other methods presented try to mimic this randomization or must accept that bias exists. But does this hierarchy hold in the applied work and beyond internal validity? Roodman & Morduch (2009: 41) in revisiting evidence on microfinance argue for a cautious approach:

“[E]xclusive reliance on one type of study is not optimal. But the present analysis suggests that for non-randomized studies to contribute to the study of causation in social systems where endogeneity is pervasive, the quality of the natural experiments must be very high. And it must be demonstrated.”

In the 1970s, however, Rubin (1974: 688) already saw a position of “extensive criticism of the use of nonrandomized studies to estimate causal effects of treatments [...] untenable [for the social sciences]”. Still, this debate is lively and ongoing: Especially authors centered around the influential MIT Poverty Action Lab (Ravallion 2009: 1) focus on the need of hard evidence to justify and improve development aid.⁷¹ This approach of methodological primacy is thus said to be more than another “‘big think’ fad [...] especially as it forces] to engage development problems where they play out” (Glennerster & Kremer 2011b: 13). At the same time, the ‘gold standard’ is extended beyond program evaluation to questions of behavioral research (Glennerster & Kremer 2011b) or even macro-level phenomena (Angrist & Pischke 2010: 16-20). While proving the working of development programs as well as their guiding theory on the ground is certainly a turn to be welcomed, it is left open by Glennerster & Kremer (2011a, 2011b) why RCTs should be best suited for this. Deaton (2006: 13f.) warns accordingly with the same metaphor for “the latest in a long string of development fads.”

2.6.1 Clear-Cut Research Designs

Angrist & Pischke (2010: 4) use RCTs as prime examples of what they call a credibility revolution in empirical work through clear-cut research designs – easily attainable through perfect randomization: “Such studies offer a powerful method for deriving results that are defensible both in the seminar room and in a legislative hearing.” The role of observational methods is in this light reduced to cases where time, money or practical reasons deter the implementation of RCTs.⁷² Various studies, popularized at least since LaLonde (1986), support the claim that

⁷¹ “Randomized trials [...] are the simplest and best way of assessing the impact of a program” (Banerjee 2006a: 8), comp. as well Banerjee (Banerjee 2006b).

⁷² Although the authors acknowledge that “the difference between a randomized trial and an observational study is one of degree. Indeed, we would be the first to admit that a well-done observational study can be more credible and persuasive than a poorly executed randomized trial.” (Angrist & Pischke 2010: 9)

RCTs are the way forward by providing evidence where for the same subjects and interventions RCT evaluations surpass observational methods in credibility.

It has to be acknowledged that the spread of RCTs has raised the standard for all research designs: Whatever method is used, the internal validity of the approach to the counterfactual has to be defended with RCT designs as benchmark. Angrist & Pischke (2010: 11) highlight this shift to the core of the evaluation problem – the question “whether the sources of variation in execution used by [...] statistical models justify a causal interpretation of [...] estimates”. They argue that the debate on causal inference was too often confounded by discussions on secondary methodological issues⁷³. In this understanding, a focus on research designs leads to a focus on internal validity.

In a second step, however, the danger of such an approach is a one-sided focus on settings in which RCTs are adoptable and straightforward internal validity is achievable. As outlined in Chapter 2.3.2, this is by far not everywhere. An over-focussing on randomization evidence and comparable research designs could go to the expense of knowledge-gathering on the effects of macro-level determinants for successful micro-level interventions (Goldin et al. 2006: 11), as e.g. tremendous problems of aid in fragile states where sustainable interventions on the ground rely on long term capacity building at the state level (comp. for a discussion of the task Schneckener 2004). Thus, for finding credible solutions to most empirical questions, Sims (2010: 61f.) argues that the complexity of reality requires to at least try “modeling the dynamic interactions [...] – something for which there is no push-button in Stata.”⁷⁴

2.6.2 RCTs Solving Second Tier Problems

Beyond internal validity arguments, the method is claimed to solve problems of a second tier. Duflo et al. (2008: 3908f.) e.g. assess the supremacy of RCTs in face of publication biases:

“[If] the true treatment effect is zero, but each nonexperimental technique yields an estimated treatment effect equal to the true effect plus an omitted variable bias term that is itself a normally distributed random variable with mean zero. Unfortunately, what appears in the published literature may not reflect typical results from plausible specifications, which would be centered on zero, but may instead be systematically biased.”

Nevertheless, preferring RCTs over non-randomized evaluations for this reason would only provide a superficial solution: The way forward must be the recognition that the publication of insignificant findings and replication studies alike is relevant – not to disqualify observational evidence for this reason.

⁷³ Such as the functional form of the impact regression or heteroskedasticity.

⁷⁴ He explicitly counters the logic of Angrist & Pischke (2010: 18-20) who call for shifting the focus from “theory-centric” to experimental evidence on the macro-level as well: Sims (2010: 64) emphasizes multivariate time series models and other complex evaluation methods.

Related to the previous chapter, their following argument is more convincing: “Some researchers may inappropriately mine various regression specifications for one that produces statistically significant results [even if] inadvertently”⁷⁵ (Duflo et al. 2008: 3909.); but here again, changing the method is going one step ahead of the straightforward solution: A specification guided by theory as well as rich on-the-ground-knowledge is less likely to be subject to data-mining-concerns, which brings the discussion back to the hypothetico-deductive method argued for by Cartwright (2007) and the call by Deaton (2010b: 3, e.i.o.) “to investigate *mechanisms* and to discover *why* projects work rather than *whether* they work”.⁷⁶

The debate is as well obscured by the overlap to other debates on aid effectiveness: Moore (2006) and Vásquez (2006) observe an aid industry centred around itself and unwilling to prove its effectiveness and thus call for RCTs as solution. However, their argument is only one for the evaluation of programs as such, not for randomization – even if RCTs force the project implementing agency clearly into an up-front consideration of the later-on evaluation.

2.6.3 Weakest Link Principle

As soon as estimates have proven to be internally valid, the question is how to apply results. Cartwright (2007: 18) subsequently lines out how there exists a real “trade-off between internal and external validity [...]. The RCT, with its vaunted rigor, takes us only a very small part of the way we need to go for practical knowledge.”⁷⁷ Banerjee et al. (2010a: 28f.) e.g. struggle to explain contradictory generalized results of their randomized intervention in India in comparison to results from Uganda (Björkman & Svensson 2009) on the success of programs initiating community participation for improved public good provision.⁷⁸

Concerning external validity, a RCT has to prove just as much as any other study how its effects are transferrable (comp. Chapter 4 for a detailed discussion). Similarly, as soon as one considers interest in heterogeneity of the impact or of certain impact pathways, interest in

⁷⁵ Inadvertence may occur if under uncertainty about the correct specification of the relationship between causes and outcome the researcher is unconsciously guided to accepting the specification providing the most robust results.

⁷⁶ Angrist & Pischke (2010: 12-16) would provide a perfect example for Deaton’s call: Citing “the best of today’s design based studies”, they explain how returns to schooling depend on state resources, how class size is negatively related to student achievements or how homicide rates are unrelated to the death penalty – what is the case is lined out, but not so much why. The danger of this approach is the mere focus on finding exogenous variation, not the understanding of the most pressing themes in development economics and beyond.

⁷⁷ This is not to say that this trade-off is not there with other methods, but just that RCTs are in no way methodologically superior in comparison: Duflo et al. (2008: 3953, FN 3926) as well argue, how RDD and IV methods are similarly focusing on parts of the evidence where generalization from can be a difficult case.

⁷⁸ They have to resort to general observations on differing context – “[t]here could be many reasons why it was possible to increase the involvement of citizens with the public sector in the Uganda case and not in UP, and it is difficult to tease them out” (Banerjee, et al. 2010a: 28) – and are left with the conclusion that participation is no panacea and context dependent.

the program at a larger scale, the transfer of results to other contexts, or factors that are not possible to randomize, randomization is one amongst many other methods. Cartwright's (2007: 14) weakest link principle comes back into the picture.

2.6.4 A Hierarchy beyond the Quantitative Approach?

Bamberger (2010: 617) goes even one step further and criticizes not only a supremacy of RCTs via other methods, but also an implicitly applied hierarchy of quantitative versus qualitative insights. He calls for a look beyond the econometric approach:

“None of the quantitatively oriented development participants in this debate [on the supremacy of RCTs], however, question the singularity of the econometric analysis of survey-based data as the core method of relevance for impact evaluations. [...] However, given that a central challenge in international development is that the decision makers (development economists included) are in the business of studying people separated from themselves by vast distances – social, economic, political and geographic – there is a strong case for using mixed methods both to help close this distance and to more accurately discern how outcomes (positive, negative, or indifferent) are obtained, and how any such outcomes vary over time and space (context).”

One should as well not forget that the understanding of development as such is a discipline that draws importantly on “historical and structural explanations for why poverty persists in some contexts more than in others, and many of these explanations can't be found through random evaluation or general experimental methods” (Bradhan 2011: 21). As Kanbur (2010: xvii) argues, detailed knowledge on what works has to be generated “using the full range of qualitative and quantitative methodologies from the social sciences.”

This does not necessarily imply that the very same researchers have to address their evaluation problem with multiple methods.⁷⁹ However, given the predominance of certain methods in different disciplines, an open dialogue is just as necessary as the recognition that different methods can contribute to different parts of the same picture. Intersections between disciplines, such as the field of development studies, play a pivotal role in this. Mutual stimulation can contribute to an increased variety of the usage of methods within the same discipline and thus to a richer understanding of the social processes surrounding questions of poverty and beyond.

⁷⁹ Though this is certainly possible as Lawson (2010) proves by assessing reasons for chronic poverty through combined econometric data evaluation and life history analysis. But as Deaton (2010b: 13) summarizes, “we make most progress when theory and empirical work are closely articulated, not necessarily in the same person, but at least when different people with different skills read and talk to one another. Good tests require deep understanding of models and the ability to manipulate them into delivering predictions that are not obvious and that are specific enough to the model to be informative about it. At the same time, good theories, or good modifications of existing theories, require theorists who are familiar with and pay attention to historical and empirical evidence.”

3 Evaluating a Pro-Ultra Poor Intervention: The Impact of IFSUP

“What determines poverty reduction? If we knew the answer to that, the poor would be millionaires.” (Kanbur 2010: xv)

Questions of poverty, poverty persistence and eradication are a central field of study for the social sciences, each field bringing in its own perspectives – be it discourses based on risks in neo-classical economics, rights-centered approaches in international law or needs-based arguments stemming from political philosophy (Munro 2008: 28-37). Poverty is a question to be addressed on all levels: Questions on the structures enabling and maintaining poverty on the macro level – such as discussions on economic liberalization or the social consequences of conflicts – have their say as does the generation of practical knowledge on what works in poverty reduction (Hulme & Lawson 2010: 2f.). The following second part of this thesis is concerned with linking impact evaluation to the latter: What works on the ground in poverty alleviation, in particular for ultra-poverty⁸⁰ alleviation?

In many developing countries, the reduction of the depth and severity of poverty remains a major challenge. Hulme & Lawson (2010: 2) highlight that “chronic and extreme poverty has spatial and social relational dimensions”: Regional pockets of ultra-poverty persist even in countries on track for MDG 1; ultra-poverty is additionally concentrated on special strata of the population such as ethnic minorities or refugees. Innovative mechanisms need to tackle these populations, as conventional strategies fall short of reaching these poorest sections of society or even use mechanisms that shut them out – as is the case with the ‘micro-credit revolution’⁸¹ (Amin et al. 2003, Dietzel 2006). Its regularly used group collateral, while it may serve its purpose in circumventing moral hazard and adverse selection problems widespread on financial markets for the poor (Amin et al. 2003, Armendáriz & Morduch 2010, Banerjee et al. 2003), implicitly excludes the ultra-poor: On group or village level, the group collateral system prevents their inclusion individually or in groups, as actual or perceived default rates inhibit their acceptance by other participants (Rudolph 2010: Chapter 5). This leads not only to an inefficient but also inequitable distribution of investment opportunities and in

⁸⁰ Many names are used to describe the bottom section of society in Bangladesh: the ultra-poor, the hardcore poor, extreme poor, poorest of the poor, the destitute, etc. The term ultra-poor in the present paper is related to the working definition of the NGO NETZ, focusing especially on calorie intake: Ultra poverty is defined by a threshold of 1800 kcal/day/person and (functional) landlessness, as described in detail in Appendix 2.

⁸¹ While the identification of capital market failures, widespread in developing countries (comp. for direct evidence e.g. Banerjee et al. 2003), as core obstacle for poverty reduction microcredit has been called for as win-win solution as discussed by Morduch (2000).

particular of the potential income gains thereof. The ability of the ultra-poor to increase their standard of living on their own is thus seriously constrained. Nonetheless it can be expected that the ultra-poor have worthwhile investment projects if given the capital base needed to get them underway.⁸²

As of today, only few projects focus on these long-term needs of the bottom section of society. Too often they are only addressed by food or emergency aid, if not left alone. But it is crucial that their vulnerability – “the likelihood of being in poverty in the future” (Barrientos & Hulme 2008: 4) – is addressed. While food aid is nonetheless necessary in many circumstances, a change in policy is needed, and could be emerging at present (see for an overview Barrientos & Hulme 2008), that would shift the focus beyond their current deprivation to their long-term prospects of freeing themselves of poverty. Organizations of larger, such as the Bangladesh Rural Advancement Committee (Halder & Mosley 2004, Matin & Hulme 2003), and smaller scale, such as NETZ (Dietzel 2003), are since decades at the forefront of developing related strategies in Bangladesh. Recent approaches focus on directly enhancing the productive potential of the poorest and “graduate” (Hashemi & De Montesquiou 2011: 1) them out of poverty by a combination of asset transfers, training and context intervention. In Bangladesh and beyond, several preliminary evaluations of pilot projects with such a framework are and have been undertaken (amongst others Banerjee et al. 2010b, Emran et al. 2009, Haseen 2006, Huda 2009, 2010, Mallick 2009, Matin et al. 2008, Sulaiman & Matin 2006). First lessons have been published (e.g. Hulme & Moore 2008, 2010), though partly without a counterfactual approach to evaluation (e.g. Ellis et al. 2009: 269-275).

The evaluation of the case of IFSUP can contribute to the understanding of this approach to ultra-poverty reduction.⁸³ The intervention classes in the broader field of social protection as a revived strategy for reaching the poorest households in developing countries⁸⁴, taking note of their especially deprived situation,

In the following Chapter 3.1, the IFSUP project will be classified as multidimensional poverty reduction scheme with a combination of promotional, preventive and transformative

⁸² Some innovative micro-credit projects address these issues by explicitly targeting the ultra-poor through innovative design features such as flexible pay-back options (see for a recent evaluation of this approach e.g. Khandker et al. 2010a).

⁸³ First limited knowledge on IFSUP was gathered by Rudolph (2010, 2011). As noted, this point of departure in analysis is described in Appendix 1. Additionally, Appendix 2 provides a general overview on the ideas behind the IFSUP approach and Appendix 3 an overview of the context of Bangladesh and its northern regions. Appendix 4 provides a map of the selected working areas for IFSUP.

⁸⁴ For a first overview to the field see Barrientos et al. (2010).

project elements. In this context, the concrete mechanisms are described that IFSUP applies for poverty alleviation. Chapter 3.2 will be concerned with the strategy for an internally valid estimation of the project impact. For this, Chapter 3.2.1 will describe the selection process of IFSUP beneficiary and control group members to assess the implications for a counterfactual approach to impact evaluation. Chapter 3.2.2 is concerned with, first, a description of the data present and second, the assessment of the quality of control group data as counterfactual. Chapter 3.2.3 to 3.2.5 will outline the description of the evaluation approach: Manski bounds will be used to control for attrition problems, difference-in-differences estimation as well as fixed-effect-estimation to assess the plausibility of *as-if-randomization*. Chapter 3.3 will describe the estimated impact⁸⁵ of IFSUP on three variable sets: A significantly positive influence is found for food security, expenditure data and socio-economic status of beneficiary households, as well as increased asset levels and reduced vulnerability (although outcomes for health-related variables are unclear). Chapter 3.3.4 summarizes findings for correlates and Chapter 3.4 concludes that an internal validity of results could plausibly be established. The discussion of the external validity of results and the link back to theory is kept for Chapter 4.

3.1 Theoretical Classification and Description of the IFSUP Approach: Promotion, Prevention and Transformation

The ultra-poor need special targeting for successfully supporting their livelihoods – new forms of social protection are emerging as means to achieve this. Following the conceptualization of Ellis (2009: 8), social protection includes all initiatives aiming at the poorest and pursuing the aim of reducing economic and social vulnerability by a protection of livelihoods, transfers with the goal of improved consumption and a reduction of social marginalization.⁸⁶

Adopting the typology of Sabates-Wheeler and Devereux (2008: 70, e.i.o.),

“social protection includes four categories of instruments: *provision measures*, which provide relief from deprivation; *preventive measures* which attempt to prevent deprivation; *promotive measures*, which aim to enhance incomes and capabilities; and *transformative measures*, which seek to address concerns of social justice and exclusion.”

⁸⁵ Estimation, table and most figure creation was done using STATA 11 (StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX, USA: StataCorp LP). Reference for Stata usage goes, beyond the useful Stata help files, to Kohler & Kreuter (2008), Cameron & Krivedi (2009) and Khandker et al. (2010b: Part II) additional to countless helpful comments especially on the ‘Statalist’ (available through <http://www.stata.com/statalist/>) and the ‘UCLA Stata FAQ’ (available through <http://www.ats.ucla.edu/stat/stata/faq/>). Data analysis beyond the code implemented in Stata 11 was conducted using *mdesc* (Medeiros & Blanchette 2011), the *bounds*-package (Beresteanu & Manski 2000), *matselrc* (Cox 2000) and *polychoricpca* (Kolenikov & Angeles 2004), table creation was aided by the *estout*-package (Jann 2005, 2007), figure creation by *grc1leg* (user written STATA program by Wiggins (without year)).

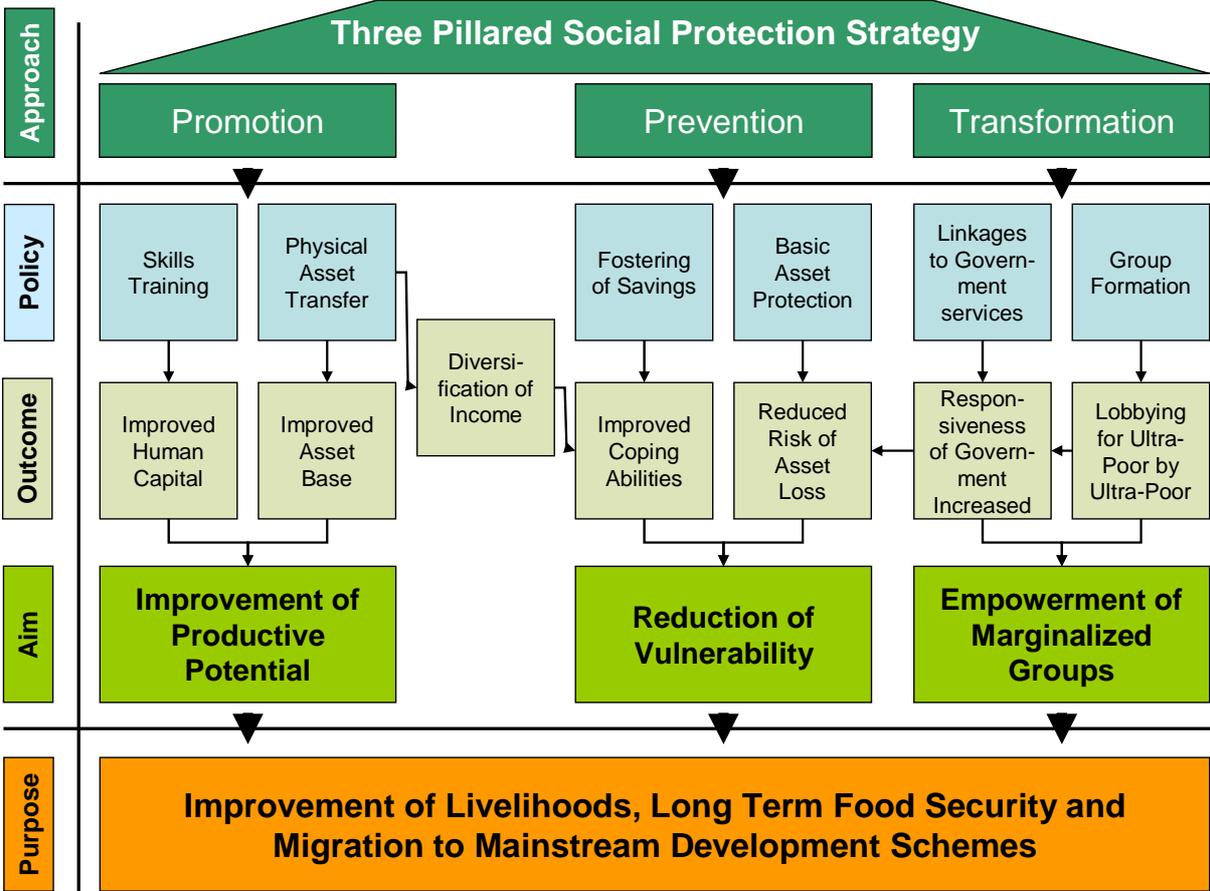
⁸⁶ Social protection can be this widely defined, although this somehow “straddles the conceptual divide between welfare payments and development policy” (Ellis et al. 2009).

The IFSUP project uses a threefold approach combining prevention, promotion and transformation, by this establishing a “springboard” for the poor with a “safety net” beneath (Sabates-Wheeler & Devereux 2008: 73). This combination of approaches is not uncontested: While major developing institutions acknowledge a focus on vulnerability and its *ex ante* reduction (ESCAP 2002: 37f.), a wide range of donors see no promotive elements in social protection interventions (World Bank 2000), even if some linkages are recognized (e.g. BMZ 2009: 13). A combination of approaches is necessary, though, to do the multidimensionality of poverty justice: Just as even emergency cash transfers can be shown to induce the building of asset stocks and contribute to livelihood protection and promotion at the same time (Devereux 2007: 56), productive asset transfers can be used as means to protect livelihoods by promoting income generation – with the advantage of explicitly raising the long term productive potential of households. Opposed to moderately poor households, where temporary safety-net measures allow them to mitigate a temporary crisis and bounce back to their original development trajectories, the chronically poor need protection against imminent deprivation simultaneous to support for their productive capacity (Farrington et al. 2004). It is in this light that Farrington et al. (2004) call for an incorporation of protection mechanisms in promotive action and vice versa, and that Sulaiman (2006: 1) concludes that an “intervention design for the poorest will have to be far more comprehensive including promotional, protective and transformative strategies to make a real dent on extreme poverty.”

Figure 6 conceptualizes the project approach and its envisioned impact trajectories. The project’s main component, the promotion of the respective household’s asset base, is combined with mechanisms aimed at preventing asset loss and building up liquid assets as well as a transformative intervention into the socio-political context the ultra-poor are living in. The overarching aim is to substantially improve the livelihoods of the ultra-poor households, especially by lifting their year round DCI above 2,122 kcal pc pd and migrating them to mainstream development schemes. This is addressed through the three interwoven pillars of the project with their respective subgoals of directly raising the productive potential of the poor, reducing their vulnerability to future shocks⁸⁷ and socially empowering them.

⁸⁷ The term ‘shock’ as used in this thesis refers to unexpected or with a low probability predictable events influencing the livelihoods of the poor positively (‘positive shock’ such as inclusion in an unexpected NGO scheme) or negatively (‘negative shock’ such as illness/accidents or environment related disasters) . See for more details and a critique of the term Sinha (2002: 187). In this thesis, the mention of mere ‘shocks’ is generally related to ‘negative shocks’, while ‘positive shocks’ are explicitly noted.

Figure 6: Conception of strategy and envisioned intervention pathway by the IFSUP project



Own Graph

With the first pillar on the left hand side of Figure 6, promotional aspects are included in the project. This includes primarily two policies, skill trainings and asset transfer. The skill trainings centre on the management of productive assets but additionally cover general livelihood aspects (such as disaster preparedness, health or basic literacy).⁸⁸ Concerning the management training, women are offered a range of courses related to Income Generating Activities (IGAs) they are interested in. Skill training by this is related to the second policy within the promotional pillar, the transfer of working capital to the households. This is the major development mean of the IFSUP project, leading to a substantial increase in the asset base of the ultra-poor with an overall transfer of about 11500 taka⁸⁹ per household for the purchase of at least two assets, including the subsequent running costs, in several tranches. Following baseline data, the transfer thus amounts to 75% of the average annual wage income of households

⁸⁸ These aspects additionally provide links to vulnerability through the aspect of risk coping ability: Functional literacy training is supposed to in general increase the confidence and market participation ability of the ultra-poor households, emergency preparedness includes e.g. strategies for saving livestock over flood periods.

⁸⁹ About 120 € at market exchange rates or 675 € at Purchasing Power Parity (PPP) adjusted exchange rates. PPP exchange rate: 1\$ = 12.7 BDT (World Bank 2006); interbank rate: 1 € = 93.6919 BDT, 1 € = 1,34187 \$ as of 24.09.2010.

(comp. Chapter 3.2.2.1). The women choose from a range of IGAs, in large majority livestock related (beef fattening, raising of heifer, poultry, etc.) but as well including small trade, small businesses (e.g. tea stalls), land purchase or services (e.g. rickshaw pulling but as well education as ‘para-veterinary’⁹⁰). Through these policies, the human capital and physical capital base of households is strengthened. The IGAs shall on the one hand establish a lasting and growing base of physical capital for the household and help in diversifying their income source, and on the other hand are supposed to quickly generate a steady income stream, contributing to the short term food security of households. 20% of this working capital received is supposed to be paid back, increasing the responsibility for the assets transferred and reducing incentive problems.⁹¹ By this, the repayment scheme familiarizes the households with the procedure of regular savings and conventional micro-credit, enhancing their financial competency. The 20% payback is additionally constituting a revolving credit fund accessible for households similar to the procedure in rotating savings and credit associations (Armendáriz & Morduch 2010: Chapter 3). Aim of these promotive policies is to improve the productive potential of households, lifting it to a level where they can self-sustain their livelihoods.

Concerning the second pillar, the preventive project elements, the project worked primarily with two policies (middle section of the figure above): A security fund of 1% of the working capital was established that serves as basic insurance against asset loss through natural disasters and diseases of livestock.⁹² A major second policy is the fostering of savings in small but regular installments. For this, the selected women form self-help groups of 10-20 women opening a collective bank account and generating a save storage place for their liquid capital. Outcome of this policy is hoped to be an increased coping ability with regard to shocks: A small stock of liquid assets enables households to cope with minor shocks without drawing on their productive potential. Additionally, the IGAs as such diversify household incomes and increase independence from previous occupations. Aim of these methods is a reduction in both dimensions of vulnerability: The security fund reduces the exposition of households to risks, the savings component their coping ability. The sustainability of the intervention and the potential for long term food security is by this hoped to be strengthened.

⁹⁰ ‘Para-veterinary’-training (the Bangla word ‘para’ standing for ‘village’) is one innovative design feature of IFSUP. Recognizing the problem of lacking access to veterinary service, some ultra-poor women were offered basic training as veterinaries, especially concerning livestock vaccinations. By this it was not only an innovative IGA introduced but access to veterinary services for all ultra-poor in the vicinity guaranteed.

⁹¹ But exceptions apply: To prevent the material destruction of families in case of failure of the respective IGA, women are not to repay their 20% share in case of asset loss and only parts of it in case of low productivity of the asset.

⁹² Loss of livestock is additionally prevented through the advancement of livestock vaccination e.g. through the mentioned ‘para-veterinaries’.

Lastly, the third project pillar introduces transformative project elements (right section of the figure above) to the intervention. Two policies are especially relevant: First of all, the project creates links to government offices, amongst others government veterinary services, and the PNGOs are lobbying for the inclusion of ultra-poor in existing social protection schemes of the GOB.⁹³ Secondly, the self-help groups form⁹⁴ an institutionalized special interest group of the ultra-poor⁹⁵, as they are combined in ultra-poor ‘federations’ on *upazila*⁹⁶ and even district level. These elements link the transformative pillar to vulnerability reduction through access to government resources. But foremost, they give the ultra-poor a voice and increase their political space. Outcomes of these transformative elements are subsequently hoped to be manifold: As empowerment device the group formation raises the awareness of the participating women through training, generating room for discussions and networking. Additionally, they increase the ability of ultra-poor for political lobbying. In this respect, the PNGOs together with the ultra-poor federations run public awareness raising campaigns and demonstration against their deprivation. This aims at transforming the socio-political context the IFSUP project is embedded in, to not only give the ultra-poor the means for sustaining their livelihoods but as well tackle the dependency relations that lock them in their deprivation in the first place. This is an aspect oftentimes neglected (Sabates-Wheeler & Devereux 2008) but important for project success: As Bastiaensen et al. (2005: 990) outline the political processes around poverty and capital markets have to be taken into consideration in poverty alleviation programs: The poor are deprived of opportunities for shaping the institutional landscape and thus cannot (co)determine the rules of the game they have to play. Neglecting these political forces can lead to “[poverty] interventions that converge with local elite networks with the effect that the [...] flow of funds destined to the poor ultimately reproduce[s] the local structures of poverty.” Thus, the third pillar supports the overall purpose of sustainable livelihoods for the ultra-poor.

Overall purpose of the program is the stabilization of ultra-poor livelihoods, their long term food security and the migration to mainstream development schemes such as microcredit in

⁹³ These include especially the GOB cash or food transfer social protection schemes ‘vulnerable group development (VGD)’ (with WFP as donor) and ‘vulnerable group feeding (VGF)’ that have been established to target ultra-poor but where corruption is regularly impeding their inclusion (comp. e.g. findings by TIB 2008: 44f.)

⁹⁴ Additionally, they serve as coordinative device between women and PNGOs and to facilitate savings.

⁹⁵ On top of the general marginalization of the poorest in Bangladeshi society, a majority of the target group belongs to especially excluded sections without political representation: female headed and *Adibasi* households.

⁹⁶ Administration in Bangladesh follows the levels national – division – district – *upazila* – union – *mouza*. The six *upazilas* where the project is implemented are Durgapur, Gangachara, Kalmakanda, Kaunia, Joypurhat and Panchibi (see Appendix 4 for an overview).

order to lift households permanently above the poverty line. The intervention can thus be classified as integrated poverty reduction program following a multidimensional understanding of poverty (Barrientos et al. 2010: 7).

3.2 Estimation Strategy of the Project Impact

As outlined in Chapter 2, the measurement of the project impacts hinges on the construction of a credible counterfactual. The degree and assumptions under which the control group in the IFSUP case constitutes such a counterfactual will be outlined in the following. For this, the selection process and similarities as well as dissimilarities of the two groups at baseline will be outlined in detail before describing a threefold impact measurement strategy, combining single difference with Manski bounds, difference-in-differences as well as within-estimation.

3.2.1 The Selection Process

Selection of beneficiaries was based on four steps: Analysis of statistical data, participatory collection of local knowledge, household level surveys and practical concerns.

As described by program managers, NETZ and the respective PNGOs chose three districts of the deprived northern part of Bangladesh as working areas and, based on data provided by the WFP and the BBS, subsequently selected the two poorest *upazilas* each, and within these the four poorest unions each. Based on their local knowledge and consultations with the respective local council (*Union Parishad*, UP), the poorest villages were singled out. Secondly, the 6000 poorest households (1000 per *upazila*, 250 per union) were selected for inclusion in the baseline survey.

For this step, the NETZ PNGOs used Participatory Rural Appraisal (PRA)⁹⁷, in particular transect walks, social mapping and wealth ranking: By physically traversing the transect of a village accompanied by local inhabitants, transect walks allow for a comprehensive understanding of the local characteristics of the village, the distribution of housing by different sub-groups of interest (in this case especially concerning religion, ethnicity and well-being), the usage of land and other resources, and possible constraints and opportunities of the locality. Following this first understanding of the local context, the PRA tool of social mapping is enriching the gathered information by facilitating the local community to draw maps of their community including all households: These allow to map e.g. the demographic structure, the distribution of household characteristics, the relative distribution of

⁹⁷ See for a call for a widespread adoption of these methods beyond the selection process to as well impact evaluation e.g. Chambers (2009, 1995). For an overview on PRA techniques see Rietbergen-McCracken & Narayan-Parker (1998: Section 2).

poverty and access to physical and social resources in the community. Last, the tools of wealth ranking is used to get a detailed statement on the relative wealth level of each household, based on a poverty concept self-constructed through the community – through the tool, local community members are asked to sort households on different piles based on perceived socio-economic status. The facilitator by this not only learns about the relative wealth level but as well about the aspects of poverty relevant for the community. In the case of IFSUP, wealth ranking was used to sort out the poorest 250 households⁹⁸ in the respective union that were to be included in the baseline survey. All tools are conducted with open and encouraged participation of all village members, in part by subgroups of community members based on e.g. interest in gender-differing information on access to village resources. The validity of results can be cross-checked by repeating the tools with different focus groups or by comparing PRA results with the later gathered baseline data.

All participants had at the time of the baseline survey a chance of being chosen for inclusion in the beneficiary group, determined by the PNGOs following targeting criteria as well as practical aspects such as group formation. After the collection of baseline data⁹⁹, selection of households into target and control group was conducted following envisaged amounts of at least 50% belonging to group II of ultra-poverty¹⁰⁰ and 40% to *Adibasi*¹⁰¹ ethnicity. Preference in selection was as well given to female headed households¹⁰², households with disabled member(s) and beggars. Group II poverty status was operationalized by landlessness, a monthly income below 400 taka, and a maximum of average two meals during at least nine months of the year. NGO involvement, especially coverage by microcredit was considered an exclusion criterion. Double-checks on the information in the baseline survey were conducted to verify rightful inclusion in the beneficiary group.

However, the selection process was, last but not least, strongly influenced by practical but in retrospect unobservable concerns as well: As the project implementation is based on the organization of households in groups, beyond these concerns preference was given to households with a walking distance below approximately 15 minutes to the group meeting points.

⁹⁸ As long as they were physically able to work as this is a core criterion for successful project implementation.

⁹⁹ Information was collected on the location of households, household members including sex, marital status and education, health information, house, land and asset ownership and value, savings and credit, NGO support, nutritional status, social awareness/status as well as in 2009 information on trainings and assets received.

¹⁰⁰ Ultra-poor are disaggregated by their DCI and capacity to work through the program. Group I ultra are incapable for physical work (e.g. disabled, elderly, orphans) and not targeted by the program. Group II and III differ by their degree of poverty with DCIs of below 1600 kcal/day for group II and below 1800 kcal/day for group III (comp. Appendix 2 for more detail).

¹⁰¹ *Adibasi* is used as umbrella term for the different indigenous minorities of especially India and Bangladesh.

¹⁰² These are households without primary male earner, mainly divorced, widowed or deserted women.

Preference was as well given to the comprehensive inclusion of ultra-poor neighbourhoods to minimize social tensions caused through side-by-side difference in program in- or exclusion of households with similar social status. Last but not least, with regard to efficiency-concerns for the travel time of project-fieldworkers, a tendency was built into the selection process that prefers villages clumped together and well-connected by a in tendency better infrastructure.¹⁰³

Overall, through this process four fifths of the 6000 interviewed were assigned to the beneficiary group, one fifth to the control group – stratification was done on union level, assigning 200 and respectively 50 households to beneficiary and control group.

3.2.2 Data on the Project Population¹⁰⁴ and Its Comparability over Groups and Time

The data gathered for the baseline survey allows a detailed description of the ultra-poor population in the selected unions¹⁰⁵, already hinting at constraints the households face. Given the selection process, the summary statistics are representative for the poorest strata of households¹⁰⁶ in each of the selected unions. As the question of endogenous program placement is pressing, divergences in patterns between target and control group are noted in the last columns. Data was gathered in two phases. In 2007, a baseline was conducted in six *upazilas* of the three districts in question; in 2009 a follow up study was conducted.¹⁰⁷ The survey was set up by NETZ in cooperation with the PNGOs as well as an external consultant and conducted by field workers of the respective PNGO.

Table 1 summarizes key characteristics of the sampled population. In its first three columns, it displays the mean, standard deviation and number of non-missing observations for the beneficiary group in 2007. These summary statistics will be discussed in Chapter 3.2.2.1. For the subsequent discussion of control-group deviations in Chapter 3.2.2.2, the following two columns depict the deviation of control group means as well as their significance as measured by a simple t-test of mean-equality, the following six columns the deviation of control group means on *upazila* level from respective beneficiary group means.

¹⁰³ Further hidden aspects of selection are of course possible: The data revealed for example that 140 households were already affiliated with a saving program of one of the implementing NGOs, Ashroi, in the district Joypurhat. These households constitute 7.7% of the beneficiary but only 4.3% of the control group which must not, but could hint to a bias in the selection – whether this bias would be correlated with outcomes is of course *a priori* as well unclear.

¹⁰⁴ This chapter draws in part on Rudolph (2010: Chapter 3) but was updated and extended.

¹⁰⁵ An analysis at union-level is complicated by erroneous and missing coding for the impact survey, therefore a view on upazila-level is taken where geographical subsamples are addressed.

¹⁰⁶ Households of group II and group III poverty status (comp. FN 98).

3.2.2.1 Descriptives of the Project Population at Baseline

Table 1 contains information on variables that were chosen for measurement of a causal project impact or as control variables in the impact regressions (comp. Chapter 3.3). The latter include the baseline distribution of key fixed household characteristics as well as the distribution of wealth, assets and overall well-being indicators. Detailed information on the definition of these variables and, if applicable, their construction can be found in Appendix 7: Variable Definitions and Construction Procedure.

Some of these variables are summarized indices based on principal component analysis. These indices serve as proxies for the underlying wealth structure determining the distribution of socio-economic status, housing quality and asset ownership in the sample (see Appendix 5: for a detailed description of the PCA methodology and the index construction in the case at hand). The absolute value of these indices is not informative as such, but it is in comparison between subjects and/or over time.

As presented in the first column of Table 1, the IFSUP intervention intervenes with a sampled population of 63% belonging to group II of ultra-poverty, the rest to group III; 50% are of *Adibasi* ethnicity, 22% of households are female headed. The sampled population of the six *upazilas* thus shows the characteristics of the selection process, as these figures are much different from their distribution in the basic population (BBS 2009a). The average household size amounts to 3.9 family members¹⁰⁸, which can be translated to 3.12 adult equivalents¹⁰⁹ and thus a support-gap of about 0.8 in an average family.

The severe poverty the families are living in is represented by the hunger they face with an average amount of 1.95 meals per day¹¹⁰ over the year. This is mirrored by expenditure figures with a yearly mean consumption of 17961 taka, translating into a yearly mean per

¹⁰⁸ This size seems at first sight surprisingly small compared to the 2005 national average of 5.0¹⁰⁸(BBS 2009a) but with female headed households (22% of the sample) being overrepresented in the sample (national average: 11 %, BBS 2009a), they draw the average household size down: Female headed families oftentimes are nuclear families¹⁰⁸, with the female household head being deserted or widowed. This is in line with findings of the Bangladesh Bureau of Statistics (BBS 2009a), reporting much smaller average household sizes for female (3.48) compared to male headed households (4.98) in Bangladesh.

¹⁰⁹ Adult equivalents figure calculation follows Townsend (1994: 554, FN 512) and are “for adult males, 1.0; for adult females, 0.9. For males and females aged 13-18, 0.94, and 0.83, respectively; for children aged 7-12, 0.67 regardless of gender; for children 4-6, 0.52; for toddlers 1-3, 0.32; and for infants 0.05.”

¹¹⁰ Households reported the amount of months with average meals intake of one, two or three for the year.

capita consumption of 4808 taka.¹¹¹ Mean pc consumption is thereby well below conservative¹¹² lower poverty line estimates for the region of around 5500 to 6000 taka.¹¹³

Additional to their direct poverty, households are highly vulnerable to shocks: Savings amount to a mean of 33 taka, mean credit to 315 taka. But these low figures are due to many families neither taking credit (n=3078) nor saving (n=3991), as indicated by the high standard deviations. Among the households who do save, mean savings amount to 409 taka, while among households taking credit, mean credit amounts to 637 taka.¹¹⁴ Vulnerability indicators for health outcomes stand at one average sick day per month. 50% of mothers receive medical checkups before and after birth, 70% of children basic immunization by the age of 1.5 years. While the first figure seems surprisingly large in comparison, immunization coverage is low given the widespread immunization schemes in Bangladesh.¹¹⁵

Concerning physical assets, households own mean tropical livestock units¹¹⁶ (TLU) of 0.27, as would be e.g. two goats and seven chickens, with a mean value of 1327 taka. Again, the high standard deviations point to the uneven distribution even among the ultra-poor. The same finding holds for the average reported values of productive (744 taka) and durable assets (243 taka). Last but not least, households are landless or functionally landless, as indicated by the mean size of productive land of 0.01 acres.

Finally, as reported in Rudolph (2010: Table 1) the maximum education of household head or its spouse is used as an indicator for the (formal) education level of the household. The marginalization of the ultra-poor concerning education is indicated by 22% of household heads without any formal education, 52% being only able to sign their name, 8% having dropped out of school after class one or two, 9% having dropped out after class 3-5 and only

¹¹¹ 192 € and respectively 51 € at market exchange rates or about 1053 € respectively 282 € Euro at PPP adjusted exchange rates, see FN 89 for calculating figures.

¹¹² Conservative is meant in the sense that these figures are based on readjustments based on a “Törnqvist price index [which tends to] *underestimate* the increase in price level” (Murgai & Zaidi 2004: 12, e.i.o.). As figures are derived from the 2000 Household Income and Expenditure Survey, taking inflation between 2000 and 2006 into consideration the lower poverty line since then should have further shifted upward.

¹¹³ The lower per capita income poverty line in 2000 is calculated to yearly 5456 taka for rural Rangpur division districts (in particular for rural Bogra, Rangpur and Dinajpur districts; Rangpur Division contains the project areas in Joypurhat district and Rangpur district) as well as 6012 taka for rural Dhaka division districts (in particular for Tural Faridpur, Tangail and Jamalpur districts; Dhaka Division includes the project areas of Netrakona district) (Murgai & Zaidi 2004: Table A7, recalculated).

¹¹⁴ Figures are for the overall sample.

¹¹⁵ Indicative for comparison are findings by the BBS & UNICEF (2007: 160, 178, 187) for the same year: For under-5 children, 89% in Rangpur, 94% in Joypurhat and 75% in Netrakona received full immunization. For pregnant women, 66% in Rangpur, 55% in Joypurhat and 48% in Netrakona received antenatal care.

¹¹⁶ TLUs are computed based on the proposal of Jahnke (1982: 9f.), standardizing one TLU to 250 kg liveweight. As suggested by Jahnke (1982: 9f.), weights of 0.7 for cattle, horse and buffaloes, 0.2 for pigs, 0.1 for sheep and goats and 0.01 for chicken, hens and ducks and were used.

10% having started or completed secondary education. Only 2% of households report being covered by any sort of GO/NGO intervention.¹¹⁷

¹¹⁷ Examples of assistance include old age allowance or food aid from the GOB/UN WFP, or health education through NGOs like World Vision or CARE Bangladesh

Table 1: Description of beneficiaries' baseline characteristics and control group differences

	mean		N	deviation		deviation of control to beneficiary group means on <i>upazila</i> -level					
	BG	sd		CG	N	Durgapur	Gangachara	Kalmakanda	Kaunia	Joypurhat	Panchibi
<i>fixed household characteristics</i>											
group II poverty status	0.63	0.48	4800	-0.07 ^{***}	1200	-0.03	-0.04	-0.11 ^{***}	-0.15 ^{***}	-0.04	-0.06
<i>Adibasi</i> ethnicity	0.50	0.50	4800	-0.01	1200	-0.10 ^{**}	0.03	-0.01	-0.01	0.00	0.00
female headed HH	0.22	0.41	4800	0.04 ^{***}	1200	0.00	0.07 ^{**}	0.10 ^{***}	-0.01	0.08 ^{***}	0.02
family size	3.95	1.49	4800	-0.23 ^{***}	1200	-0.07	-0.23 [*]	-0.59 ^{***}	-0.18	-0.26 ^{**}	-0.04
adult equival. HH size	3.12	1.11	4800	-0.12 ^{***}	1200	-0.04	-0.09	-0.36 ^{***}	-0.11	-0.18 ^{**}	0.03
<i>overall outcome variables</i>											
average meals pd pc	1.95	0.26	4723	0.01	1193	0.03 [*]	-0.02	0.01	0.03 [*]	0.00	0.02
SES (PCA)	-1.46	1.18	4778	0.01	1190	-0.11	-0.01	-0.37 ^{***}	0.16 [*]	0.12	0.26 ^{***}
sec. outcomes (PCA)	-1.00	1.09	4778	-0.04	1190	0.02	-0.03	-0.34 ^{***}	0.11	-0.11	0.10
yearly expenditure	17961	6130	4799	-590.28 ^{***}	1200	-512.06	-257.67	-2098 ^{***}	184.49	-84.53	-771.06
yearly expenditure pc	4808	1458	4799	450.54 ^{***}	1200	-42.69	258.16 ^{***}	126.14	204.70	1403 ^{***}	755.04 ^{***}
<i>vulnerability outcome variables</i>											
households' savings	32.88	138.47	3479	-0.04	861	1.13	1.38	0.00	-0.33	171.75 ^{**}	-55.98
households' credit	314.58	731.90	4800	-21.67	1200	-35.31	-75.94	-71.88	55.00	7.49	-9.36
housing quality (PCA)	-0.44	0.94	4778	0.05 [*]	1190	0.10	0.04	-0.12 [*]	0.20 ^{***}	-0.03	0.11 [*]
sick days per capita	0.96	1.99	4799	-0.26 ^{***}	1200	-0.87 ^{***}	-0.40 ^{***}	-0.15	-0.18	0.18	-0.14
pregnant mothers' age	29.88	7.97	4111	-0.63 [*]	589	0.08 ^{***}	0.28 ^{***}	-0.13 ^{***}	-0.05	0.06	-0.15 ^{***}
mothers checkup status	0.52	0.50	2040	0.17 ^{***}	879	-0.18 ^{***}	0.16 ^{***}	0.00	-0.10	0.03	-0.07
children's vaccination	0.70	0.46	1834	-0.03	394	.	.	0.62	-0.83	0.15	-0.17
<i>asset outcome variables</i>											
tropical livestock units	0.27	0.44	4799	0.05 ^{***}	1200	-0.01	0.01	-0.02	0.02	0.12 ^{***}	0.19 ^{***}
livestock value	1327	2248	4799	-76.65	1200	-828.14 ^{***}	-1113 ^{***}	-95.64	157.04	385.19 [*]	1036 ^{***}
product. assets (PCA)	-0.63	0.82	4799	0.04	1200	-0.10 [*]	0.06	-0.12 [*]	0.16 ^{**}	0.12 [*]	0.11
productive asset value	743.55	1197	4799	-122.34 ^{***}	1200	-369.97 ^{***}	-674.11 ^{***}	-91.54	181.77 ^{***}	77.88	142.17
durable assets (PCA)	-0.58	0.94	4799	-0.03	1200	-0.04	-0.06	-0.23 ^{***}	0.09	-0.03	0.08
durable assets value	243.32	342.63	4799	-1.68	1200	-19.94	6.96	-40.06 ^{**}	49.95 ^{**}	-19.49	12.49
productive land in acre	0.01	0.07	4799	0.01 ^{**}	1200	0.00	0.00	-0.01	-0.00	0.01	0.04 ^{***}
Observations	6000					1000	1000	1000	1000	999	1001

* p<0.10, ** p<0.05, *** p<0.01; The deviations of CG means and their significance at overall/*upazila* level displayed follow a t-test on mean equality assuming equal variance.

3.2.2.2 Differences between Control and Beneficiary Group at Baseline

Central for the estimation of a causal project impact is the problem of selection bias as discussed in Chapter 2.2. This problem is potentially very significant, as the logic of the selection process has not fully conformed to a conditionally randomized selection process. It is therefore no surprise that the groups of selected beneficiaries differ statistically significant on a couple of the indicators presented in Table 1. This is a serious concern for any impact estimation: Measurement of causal effects is likely to be confounded by pre-existing differences between beneficiary group and control group at baseline if these are related to project outcomes. Recalling from Chapter 2.2.2, a simple difference in mean observed outcomes between the groups, $D = E(Y_i^T | G = 1) - E(Y_i^C | G = 0)$ would lead to a bias $[E(Y_i^C | G = 1) - E(Y_i^C | G = 0)]$, representing expected differences of developments in both groups had the project not been implemented. Subsequently, it has to be assessed to what degree this bias is confounding results and/or by which strategies it can be circumvented.

3.2.2.2.1 Differences in Means on Overall and Upazila Level

Table 1 above depicts in columns five to twelve differences of control group means for the whole sample and on *upazila*-level (measured through a t-test of mean equality). As suggested by the number of significant differences (indicated by stars) for the overall sample alone (column four), the selection process is not likely to have approximated strict randomization.

The question though is, whether endogeneity can be expected to be structurally different for both groups. As presented in columns seven to twelve, directions of mean differences and their significance are similarly structured across the *upazilas* for most variables: A conclusive picture is on the one hand indicated for the fixed household characteristics section – which is self-evident given the partial selection of group status on observables. This difference will be controlled for in the impact regressions. On the other hand the variables ‘yearly expenditure’, ‘yearly expenditure per capita’, ‘sick days per capita’, ‘tropical livestock units’ and ‘productive land in acres’ show very similar patterns across the subsamples. Importantly, in these cases the bias is pointing in a direction that suggests an underestimation of the counterfactual, as exemplified by the finding that 7% of control group households belong the better-off poverty strata III.¹¹⁸ These figures indicate that among the 6000 identified ultra-poor households an on average poorer subsample was chosen for inclusion in the project.

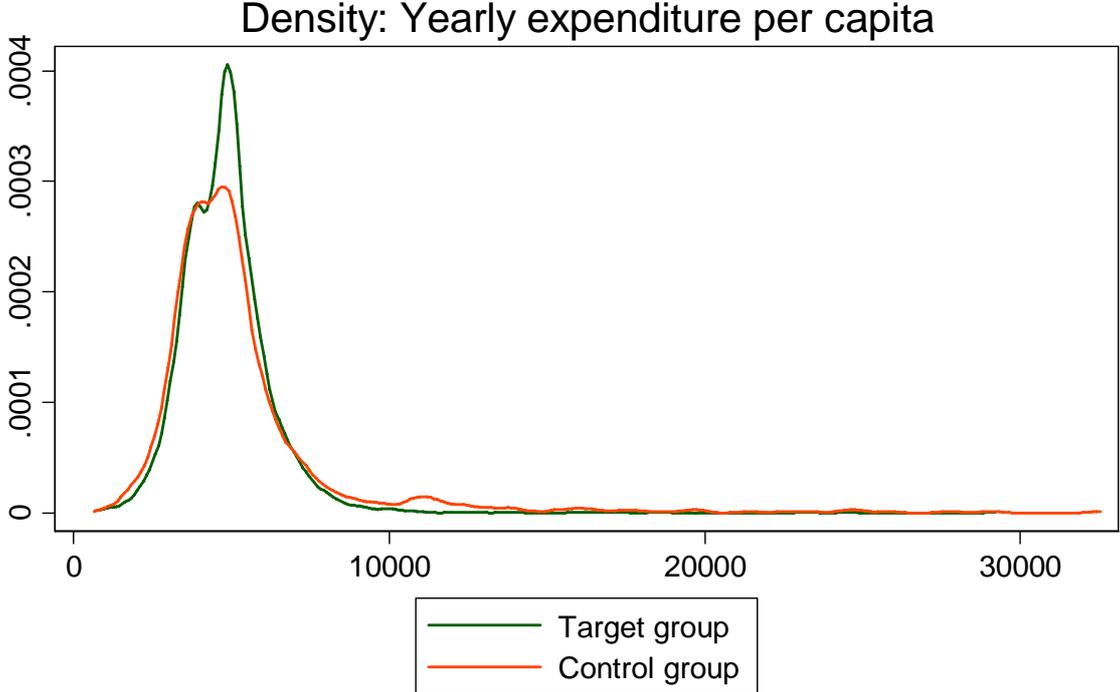
¹¹⁸ Subsequently, mean CG expenditure pc is 451 taka higher. Additionally, mean CG households have 0.26 less sick days per month, 0.05 more TLU and 0.01 more acres of productive land in economic use. Their by 590 taka lower overall household expenditure is a consequence of a significantly lower family size in the CG (mean difference 0.23 members or 0.12 adult equivalents, indicating a slightly better dependency ratio in mean CG HHs).

A note of caution concerns the over *upazilas* structurally dissimilar, but overall still significant differences for the variables ‘housing quality’, ‘pregnant mothers age’ and ‘mothers’ checkup-status’. These indicate further dissimilarities that would counteract the underestimation-argument above. Additionally, as the above analysis has only compared means, unrevealed structural differences could be hidden in the distribution of the above variables as analysed in the following chapter.

3.2.2.2.2 Differences in Distributions on Overall and Upazila Level

Figure 7 to Figure 11 summarize the density of the distribution of core variables. Figure 7, the overall distribution of yearly per capita expenditure of households, is very similarly distributed in control and target group. But as summarized in Figure 8, this is only the case for four of the selected *upazilas*, while the expenditure distribution for control group members in *upazilas* of Joypurhat District is, despite a similar maximum, broader compared to the target population. Still, the density of expenditure levels is for both groups very similarly distributed, reassuring a comparability of counterfactual outcomes for both groups.

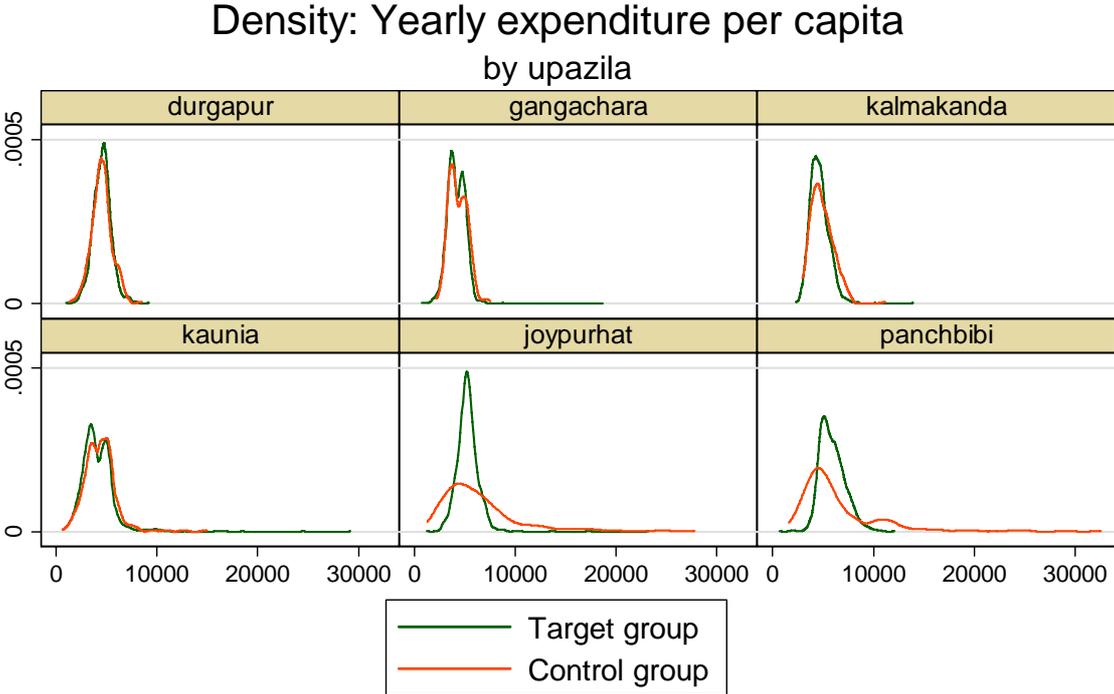
Figure 7: Density of yearly expenditure per capita



Own Graph
The graph shows the distribution of per capita yearly expenditure in BDT for households selected ("Target group") and not selected ("Control group") for IFSUP.

Importantly, the difference in shares of households belonging to group II poverty status does not suffice in explaining these average mean differences. T-tests on mean difference conditional on belonging to group II or III poverty status show significant differences between CG and BG in both cases. Results are available on request.

Figure 8: Density of average yearly expenditure per capita by upazila



Own Graph
The graph shows the distribution of per capita yearly expenditure in BDT for households selected ("Target group") and not selected ("Control group") for IFSUP.

But the economic potential of households is not only determined by their current income and expenditure status. Assets, especially physical and human capital, are important determinants in the long-term household production function. The socio-economic status (SES) index tries to capture these aspects (as depicted in Figure 9): The SES, based on principal component analysis, provides a comparative measurement for the underlying long-term wealth structure differentiating households, by this complementing the expenditure figures above¹¹⁹. A more direct insight in the distribution of human capital in the two group provides the distribution of the maximum education levels¹²⁰ of household head and, if present, its spouse (as reported in Figure 10). A more direct indicator for access to financial-productive capital as well as the vulnerability implied by the distribution of access to different sources of livelihood sustainment is provided in Figure 11 through a diversification index of income sources¹²¹ (DIS).

¹¹⁹ The SES-index is primarily based on the first principal component of physical assets (productive assets, durable consumer goods and housing) as well as human capital (maximum education levels). Comp. Appendix 5 for a detailed description of its construction.

¹²⁰ The maximum of reported education levels of household heads or their spouse is ranging from 1 to 9 (1: illiterate; 2: signature only; 3: basic reading/writing skills; 4: class 1-2 passed; 5: class 3-5 passed; 6: class 6-8 passed; 7: class 9-10 passed; 8: secondary school certificate obtained; 9: higher school certificate obtained).

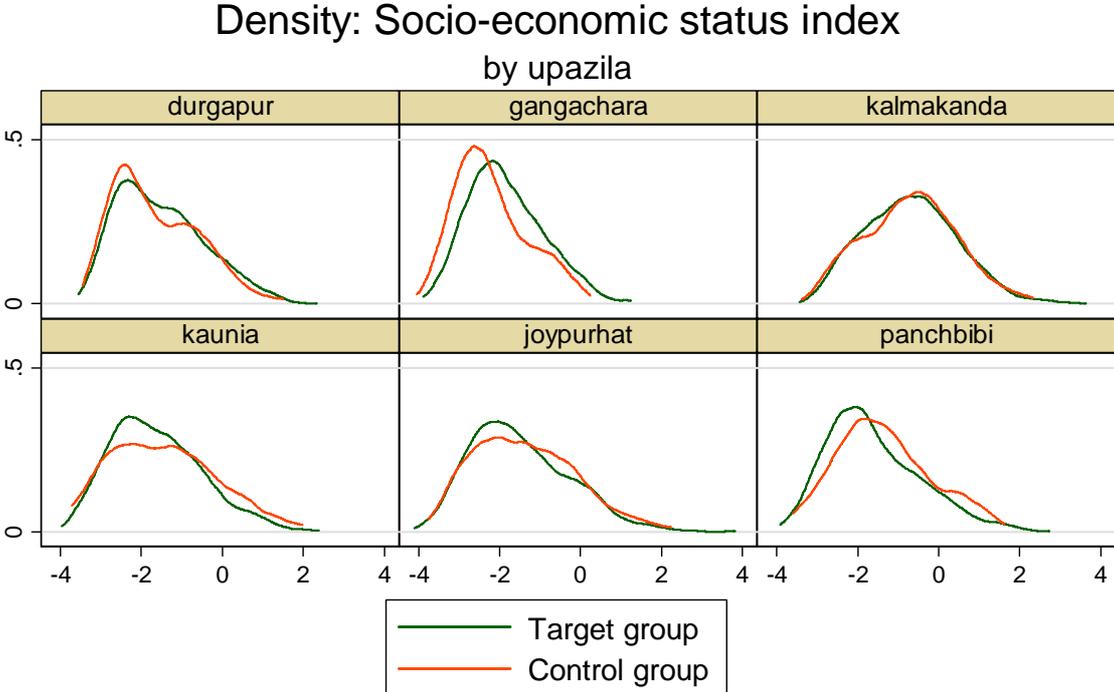
¹²¹ Index from 0-8 counting the number of income sub-classes of the household (land usage; livestock usage; fruit trees, bamboo clumps/vegetable gardens; productive assets of the type rickshaw/van, fishing net, boat, husking machine, small business capital; number of household members pursuing wage labour occupations).

All three figures confirm that control and target group seem to be selected from a comparable population. It has to be noted that the distribution of SES indicates that wealth structure differs between groups, but even more between *upazilas*.

Control group distributions are, besides Gangachara, shifted towards the right, backing the case for their structurally better position. Concerning the distribution of education levels and diversification of income sources, it is again Gangachara that stands out of a general picture of a credible counterfactual. This possible negative bias has to be assumed to be negligible in expectation

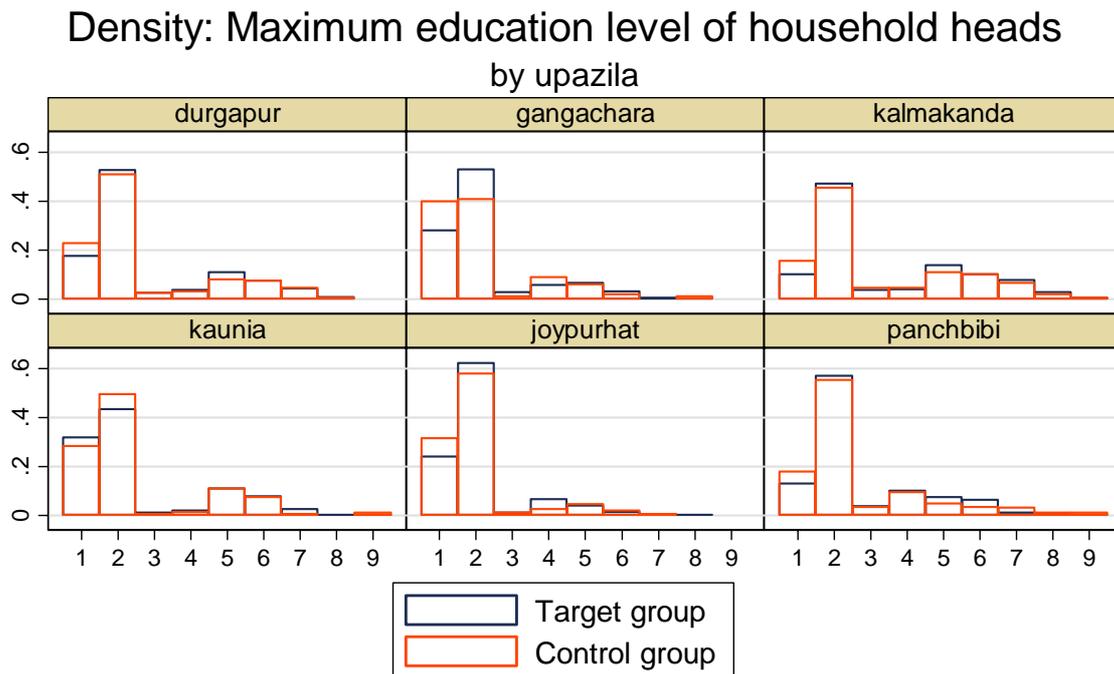
The unequal distribution for both groups in-between *upazilas* focuses attention to the fact that the project intervenes in three similar, but still structurally different regions of the country (comp. Appendix 3: Bangladesh and Its Northern Regions for more details). The estimation of a causal project effect has to take this into account by allowing for structural dissimilarities – cluster-robust standard errors on *upazila* level and *upazila*-dummies will be included in the impact regressions (comp. Chapter 3.2.5).

Figure 9: Density of socio-economic status by *upazila*



Own Graph
 The graph shows the distribution of the socio-economic status index for households selected ("Target group") and not selected ("Control group") for IFSUP. SES was through principal component analysis.

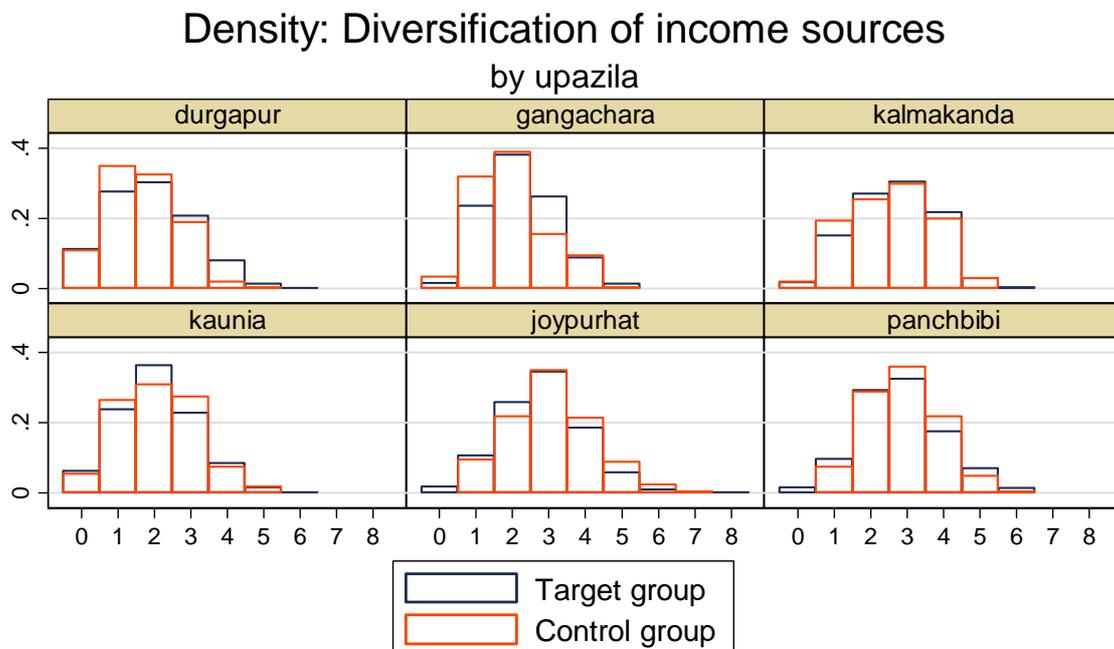
Figure 10: Distribution of the size of homestead by upazila



Own Graph

The graph depicts the histogram of the density of maximum education levels (1 "illiterate", 9 "higher secondary school") of household head and its spouse in households selected ("Target group") and not selected ("Control group") for IFSUP.

Figure 11: Distribution of the diversification of income sources



Own Graph

The graph depicts the number of income sources available for households (outside wage labor occupations, land, livestock, home-production potential) of households selected ("Target group") and not selected ("Control group") for IFSUP, ranging from 0, (begging), to 8, (four household members with wage labor occupation, land and productive asset ownership, home-production of vegetable and animal products)

3.2.2.3 Single Difference as Conservative Impact Measurement

The general picture of mean similarities and distributions leads to a general conclusion: For many variables, differences between control group and beneficiary group seem negligible. The remainder is expected to overall indicate that households were selected for the program that had a structurally weaker status at baseline. This fits with the general approach of NETZ, aiming at a selection of the in comparison poorest households for their projects.

Continuing this argument would lead to the expectation that the control group has the potential for a better development in mean outcomes without the intervention over time. For this, it is assumed that outcome variables in 2009 ($t = 1$) are positively influenced by lagged outcomes in 2007 ($t = 0$) next to other factors, thus $Y_{i,t=1} = Y_{i,t=1}(Y_{i,t=0}^+, \dots)$. This could be a direct mechanism for some variables: More livestock or productive assets directly influence future incomes, thus asset accumulation in the same fields. For other variables, the process works intermediary: Better health is indicative of a better coping ability against shocks.

But this relationship is limited through non-linearities in these processes: As Carter (2007: 52-55) outlines, the asset base of households is crucial for movements out of and into poverty. Whereas shocks contribute to short term fluctuations in livelihood status, structural up- or downward trajectories are determined by the asset¹²² and skill levels of a household and the possibility to move to upward asset accumulation trajectories.¹²³ In this light, Carter (2007: 55-59) argues that different classes amongst the poor exist: Based on a choice of the poor between present and future consumption he models different livelihood strategies. The success of the ‘high potential strategy’ is depending not only on individual skill and the absence of shocks but, for individuals with intermediary skill level, as well a minimum capital base. Only if this capital base is crossed, upward trajectories out of poverty are possible.

Thus, one subsection of the poor is trapped in poverty for a lack of skill – those “*economically disabled* [...] who are inevitably in a poor, low-equilibrium trap” (Carter 2007: 56, e.i.o.) – , whereas an intermediate subsection of the poor might be trapped in poverty as their

¹²² Assets are here understood in a broad sense comprising the five kinds of capital (human, financial, social, productive, natural) additional to political resources and other assets (Moser 2007: 1,9).

¹²³ In this, the socio-political context is implicitly assumed as external to the model and fixed. By this, the socio-political context of material poverty is reduced to an intermediary factor that might influence the outlined trajectories but is no main determinant in the reasons for up- and downward mobility. But this context is certainly an additional factor of differentiation next to Carters skill level. Strategies chosen by the poor would then additionally depend on socio-political constraints by the system that importantly are not external to the system, but are influenced by the choices of the actors involved. Exemplarily, Radhakrishna et al. (2006: 148) point to the fact that poverty needs to be seen as more than income (or asset) poverty, as “an outcome of multiple deprivations, including attributes like powerlessness, alienation and lack of social justice. Income poverty provides only a simplified view of poverty and conceptualization of poverty should extend beyond what is captured by the money metric.”

initial endowments with skill and capital do not suffice to lift them above a threshold needed for an upward trajectory – “[t]he *multiple equilibrium poor* [...] will move to the high-potential strategy if they are not too far away from the needed minimum capital” (Carter 2007: 56, e.i.o.). Only the “*upward mobile* [...] are expected to always surmount a poor standard of living given a sufficiently long period of time” (Carter 2007: 56, e.i.o.).¹²⁴

This being, it can be expected that at an aggregate level future outcomes are positively influenced by initial asset levels, at least as long as the group in question contain a portion of “*upward-mobile*” poor. At worst, initial endowments do not matter and counterfactual outcomes amongst the ultra-poor are on aggregate identical in every case, as no matter what the initial situation is, a poverty trap exists.¹²⁵

A simple estimation strategy assuming as-if-randomization under these pretexts would assume that the beneficiary group counterfactual, $E(Y_i^C | G = 1)$, is underestimating the outcomes of the IFSUP intervention, given a structurally slightly better situation in the control group. With controls for the evident differences in fixed household characteristics X , the observed project effect $D = E(Y_i^T | G = 1, X) - E(Y_i^C | G = 0, X)$ would thus be a conservative measure of the real project effect, as the bias $[E(Y_i^C | G = 1, X) - E(Y_i^C | G = 0, X)]$ would be negative. Although the above assumption would allow for a clear-cut estimation strategy¹²⁶, it can only be upheld if the data presented for 2009 is covering the same individuals as the data for 2007. As mentioned, attrition for unknown reasons is making this assumption unrealistic. Additionally the above conclusion rests solely on observable factors. But as argued by Carter (2007: 56), the livelihood trajectories of the poor are not only depending on initial asset levels or the occurrence of shocks but crucially “individuals are heterogeneous in the sense that they have innate skill levels and abilities.” Especially the latter aspect renders doubt on the conservative-estimation argument and calls for difference-in-differences or fixed-effects strategies.

3.2.3 Estimation Strategies for the Project Impact

3.2.3.1 Manski-Bounds as Solution to Non-Random Attrition

Attrition refers to a process by which individuals scheduled for project inclusion or coverage by the control group survey dropped out of being surveyed. As long as the mechanism underlying attrition is random or unrelated to outcomes, this is of no concern. The only consequence would be a reduction in the statistical power of the estimation, due to the shrinking

¹²⁴ Comp. as well for empirical evidence on the matter Carter et al. (2008).

¹²⁵ Additionally, shocks confound the trajectories but these effects should cancel out on aggregate.

¹²⁶ As conducted in Rudolph (2010: Chapter 6).

sample size. But bias is introduced if this is not the case – if e.g. the poorest were not observed, migrating out of desperateness, outcomes are measured within a subgroup that is better off as the true group of reference; if the best-off households refused respond, e.g. unwilling to display their riches, impact is estimated with a reference group worse off than it truly is. Figure 12 displays the distribution of attrition households by group status and region: Attrition was a problem predominantly for the control group. Its size shrank by 21% from 1200 households in 2007 to 986 households in 2009. Importantly, attrition is very unevenly distributed across regions: While about 42% of control group members were not requested in Gangachara, respectively 31% in Kaunia and 22% in Durgapur, attrition is less a concern in Kalmakanda (11%) and negligible for Joypurhat and Panchibi (2% and 0% respectively).¹²⁷

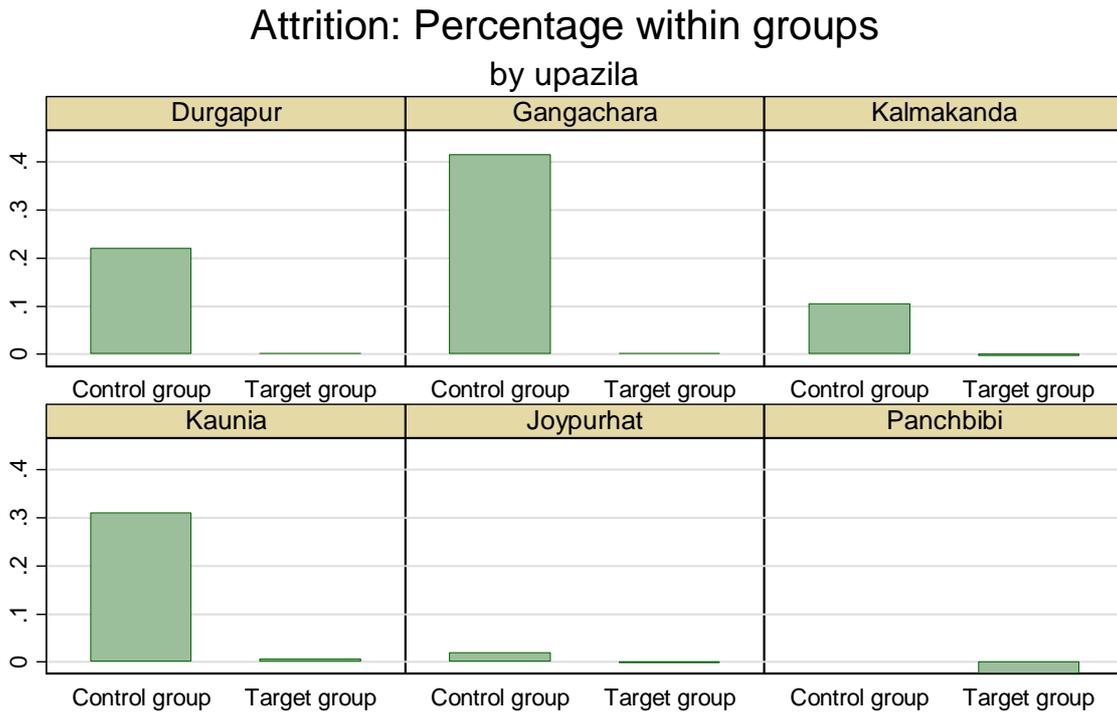
Project managers give no account for the reasons of these high attrition rates – they especially cannot provide estimations for the poverty strata that was most likely not resurveyed. The large difference in attrition rates is supposed to have a major reason in the non-acceptance of the selection procedure by some control group families. Tensions between families not included and the implementing PNGOs were reported. Thus, while some attrition might be uncorrelated with outcomes, a clear assumption cannot be taken.

That a cautious approach is not unwarranted is indicated in Figure 13: The distribution of fixed household characteristics¹²⁸ between groups and over time shows that the attrition process cannot be assumed to be random across covariates for the *upazilas* Durgapur, Gangachara and Kaunia. For Joypurhat and Panchibi, as expected by the data base, distributions remained similar over time and groups, whereas they are reasonably stable for Kalmakanda given the 11% attrition rate for the control group there.

¹²⁷ Attrition in the beneficiary group is amounting to a total of 74 households: “Among the 4803 women 90 dropped out from the action. A total of 23 died during the period. The others dropped out due to migration. 16 women were replaced from the same family. 15 women were replaced from the control group as in their families was none to become a group member. During phase out of the action the total number of members was 4744.” (NETZ 2010: 6). Additionally, 65 HHs split up, one HHs into three sub-households. All these are present in the impact survey, replacements were not separately recoded. They therefore can only be ignored for estimation it has to be noted though that they account for a total of 2.33% of the BG.

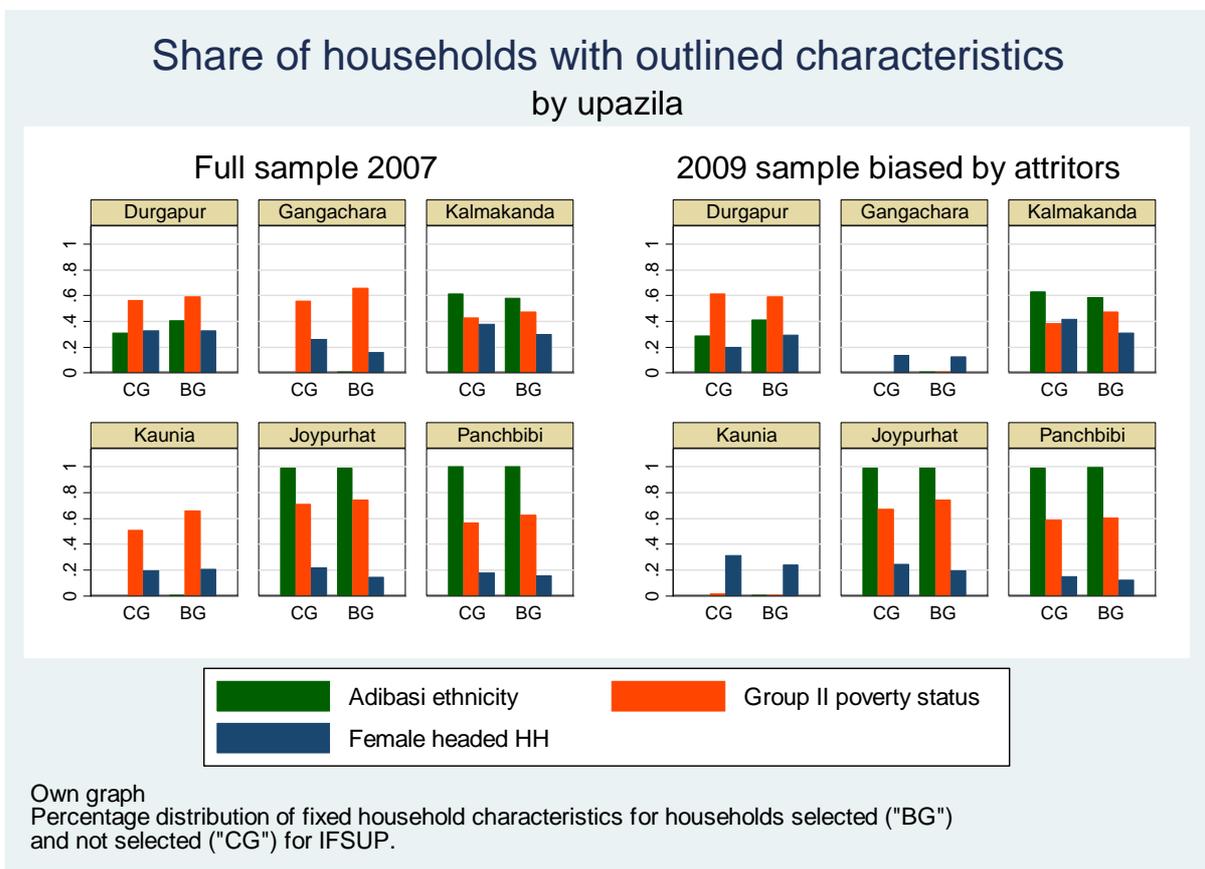
¹²⁸ The interpretation of the distribution of female headedness between groups and years is not clear-cut, as divorces and death of male household heads are not only possibly in-between the years but might as well be different between groups, with males in beneficiary group households more unlikely to leave their wives given her asset contribution to the household. Still, marriages are relatively stable in Bangladesh and divorce is a social stigma, therefore changes within two years are assumed to be negligible.

Figure 12: Attrition in control and target groups by upazila



Own graph
Percentage by which survey coverage shrank for the 2009 impact survey for households selected ("Target group") and not selected ("Control group") for IFSUP.

Figure 13: Difference in fixed household characteristics between control and target group households by upazila and in-between 2007 and 2009



Impact measurement in this context has to take the attrition process into account: Additional to a single difference estimator, Manski bounds as outlined in Chapter 2.4.5 will be constructed in order to narrow down possible bias in impact estimation. These bounds provide upper and lower limits of the treatment effect under the assumption of as-if-randomization at baseline.

An estimation of worst case bounds is implemented following a difference of means in 2009: Target group means of impact variables are straightforward, as target group outcomes are measured without attrition. For the control group, upper and lower bounds on the means are constructed. Attrition households are indicated by z ($z = 0$ if no outcome measured).

The lower bound for control group means is given by $E^{lower}(Y | G = 0) = E(Y | G = 0, z = 1) * P(z = 1 | G = 0) + E(Y^{min} | G = 0, z = 0) * P(z = 0 | G = 0)$

The upper bound for control group means is given by $E^{upper}(Y | G = 0) = E(Y | G = 0, z = 1) * P(z = 1 | G = 0) + E(Y^{max} | G = 0, z = 0) * P(z = 0 | G = 0)$.

As outcomes are rescaled to the interval $Y \in [0,1]$, $Y^{max} = 1$ and $Y^{min} = 0$ and thus $E(Y^{max} | G = 0, z = 0) = 1$ and $E(Y^{min} | G = 0, z = 0) = 0$.

This leads to the simplified form of the bounds for control group means, with $E^{lower}(Y | G = 0) = E(Y | G = 0, z = 1) * P(z = 1 | G = 0)$ for the lower bound and $E^{upper}(Y | G = 0) = E(Y | G = 0, z = 1) * P(z = 1 | G = 0) + P(z = 0 | G = 0)$ for the upper bound.

As beneficiary group means are observed without attrition, bounds on the impact are given by $D^{lower} = E(Y | G = 1) - E^{upper}(Y | G = 0)$ and $D^{upper} = E(Y | G = 1) - E^{lower}(Y | G = 0)$, reported following rescaling with the original variation of the variable of interest.

This bounding strategy is used similarly for naturally bounded as well as (partially) unbounded variables. In case a variable is naturally bounded (such as meals per day ranging from zero to three or dietary diversity score (DDS), ranging from zero to eight), these bounds are true worst case bounds of the treatment effect for the upper as well as lower impact measurement. As soon as these bounds are naturally unbounded¹²⁹, lower (upper) bounds are constructed under the assumption of missing outcome information being on the level of the lowest (highest) observable outcome variable.

If bounds lead to informative impacts on the treatment, which is the case when both D^{lower} and D^{upper} point in the same direction, 95% confidence intervals (CIs) are constructed using bootstrapping (comp. for an overview on the method Wooldridge 2010: 438-442). Due to external constraints, 50 bootstrap replications were used for the construction of CIs.¹³⁰

¹²⁹ Be it in one direction, such as expenditure or TLU, theoretically ranging from zero to infinity, or in both directions, such as PCA indices, covering negative as well as positive values.

¹³⁰ This in comparison small number (current standard are e.g. 100-500 replications in the medical sciences Pattengale et al. 2010: 337) could lead to a power loss of up to up to 2-10% depending on the sample size as

3.2.3.2 Difference-in-Differences as Solution to Time-Invariant Bias

Additional to the Manski bounds as a conservative approach for attrition with potential structural bias, a DD strategy can alleviate further doubts concerning endogeneity in placement. As a bounding approach is complex with longitudinal data, the DD estimation will exclude data from the high attrition *upazilas*: Although this reduces the overall number of observations and thus the statistical power of the estimation¹³¹ as well as the case for an external validity of results¹³², attrition can be excluded as a concern biasing results¹³³. This DD impact measurement using the two cross-sectional data-sets is implemented through an OLS regression with linear specification¹³⁴,

$$Y_i = \beta_0 + \beta_1 treatment + \beta_2 year + \beta_3 treatment \times year + X_i \alpha + Z_i \gamma + \varepsilon_i,$$

with Y_i as outcome variable of interest for household i ,

β_0 as constant,

$treatment \times year$ as interaction term for the actual project impact,

$treatment$ as a dummy variable indicating the status of project beneficiary or control group,

$year$ as a dummy variable indicating the year of the baseline (2007) or impact survey (2009),

X_i being a vector of geographical controls on *upazila* level,

Z_i as a variables-vector introducing the household characteristics ethnicity, group II poverty status in 2007, female-headedness (all as dummies) and household size (continuous variable).

3.2.3.3 Fixed Effects as Solution to Skill-Differences

The argument for as-if-randomization at baseline level was constructed following observable variables. Further difficulties could arise from unobservable factors biasing selection. In the asset framework discussed above, the future livelihood trajectories of the poor and poorest are distinguishable by three groups: Differences in distributions of those ‘economically disabled’ are of little concern, as group I ultra-poor are excluded from inclusion in either control or target group. But the ‘upward mobile’ as well as ‘multiple-equilibrium poor’ in the terminology of Carter (2007: 56) are differentiated by intrinsically unobservable characteristics such as

Monte Carlo studies of Davidson (2000: Figure 2) have shown. Due to computational power and time constraints this approach regardless had to be chosen.

¹³¹ Information on 4000 households in Durgapur, Gangachara, Kaunia and Kalmakanda, of which 210 are control group attrition households, is dropped.

¹³² Impact measurement of the implementation is now only measured on data from three *upazilas* in two different districts and two implementing organizations.

¹³³ Although by this it is assumed that 2% attrition for CG in Panchibi and 11% attrition for CG in Kaunia is negligible.

¹³⁴ It remains a question for further research, what insights different model specifications can generate. A first test with probit regressions, treating the impact indicator as dummy variable (3 meals per day as “1”, less than 3 as “0”), provided similarly consistent results.

entrepreneurial ability, motivation or social networks. As ability, skill, motivation and the like can plausibly be assumed to remain constant over time or at least over the covered two-year period, a panel approach can control for biases through these unobservables.

Data collection for IFSUP was conducted for the same individuals in the same vicinities for 2007 as well as 2009. In the data collection process, the linkages between 2007 and 2009 households through IDs was confounded, though, therefore no panel data set is present in the original dataset. Still, drawing on information present in the data, a clear link between 2007 and 2009 households could be established for part of the sample. Using information on the names of household head and spouse¹³⁵ as well as fixed household characteristics¹³⁶, data sets of 2007 were linked to data sets of 2009. However, it is problematic that the panel data set constructed with this procedure is significantly reduced in size: Overall 982 households can be linked from 2007 to 2009, of which 262 are control group and 720 beneficiary group members. As reported in Figure 14, attrition for the panel data-set is ranging for the beneficiary group from 79% (169 of 800 households remaining) in Gangachara to 89% (89 of 800) in Panchibi and for the control group from 68% (64 of 200) in Panchibi to 92% (16 of 200) in Durgapur. Despite this severe attrition, the distribution of fixed household characteristics is still comparable between the 2007 cross-section and the 2009 panel data, as reported in Figure 15, with the best comparability for Durgapur, Panchibi and Joypurhat data.

Reasons for this reduction are manifold: Most prominently, linkages failed where names of household heads and spouse were not recorded. Additionally, attrition households in the 2009 data sets subsequently dropped out of the panel data set as well. To be able to interpret the panel data, two assumptions have to be taken: First, the dropping of name information has to be randomly distributed concerning outcome variables. Additionally, both concerning the identification strategy as outlined in FN 136 and concerning the 2009 attrition rates, it is assumed that the households are kept in the sample without negative shocks occurring. This

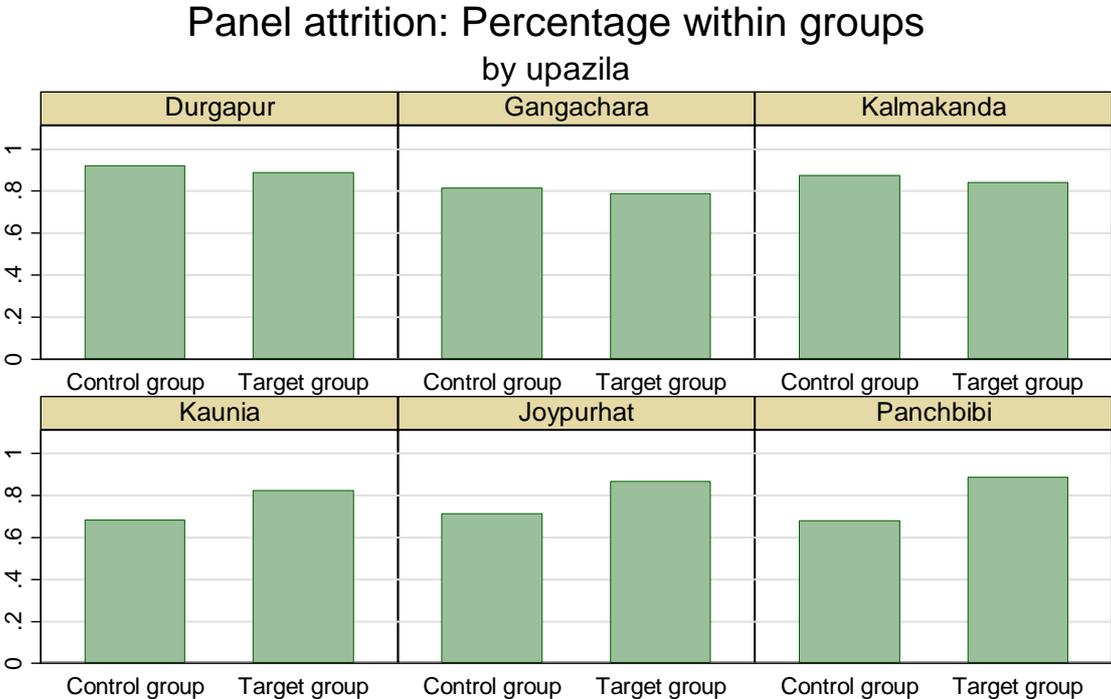
¹³⁵ Due to inconsistencies in the transcription of Bangla to Latin spelling, Stata's *soundex* command was used.

¹³⁶ Core characteristic for attribution were *upazila*, ethnicity and the *soundex* code of names for household head and spouse. Where these results were ambiguous, as multiple households with identical names and ethnicity in the same location were present, step by step the following characteristics were added to identify the links between these: Variables used were group status, ethnicity, female headedness, *upazila* of housing, education levels of household heads beyond signature-ability, family size corrected for children born in-between 2007 and 2009, home as well as homestead ownership corrected for home-construction/homestead-purchase in-between 2007 and 2009 as well as group II poverty status. Identity between these characteristics for 2007 and 2009 can only be assumed for female headed-households not remarrying, households without members dying, without household heads acquiring formal education, and households where group II poverty status 2007 was correctly reported in 2009. The identification strategy for households with multiple pairs following *upazila*, ethnicity and the *soundex* codes is thus potentially upward-biased as it prefers households without shocks. The identification strategy thus as well had to drop 2007 information on 2009 attrition households. The fact that Panchibi, Joypurhat and Kaunia are non-attrition *upazilas* is left out of consideration for the panel data due to the low number of observations that would be remaining in case of this further data reduction.

leads to an upward bias introduced in expectation for control group households relative to beneficiary households as only for them attrition is observed in 2009 and as they are more likely to face negative shocks due to not being included in the program. Impact would under these assumptions be measured against a relatively upward-biased control group and is thus a conservative approximation to the true causality.

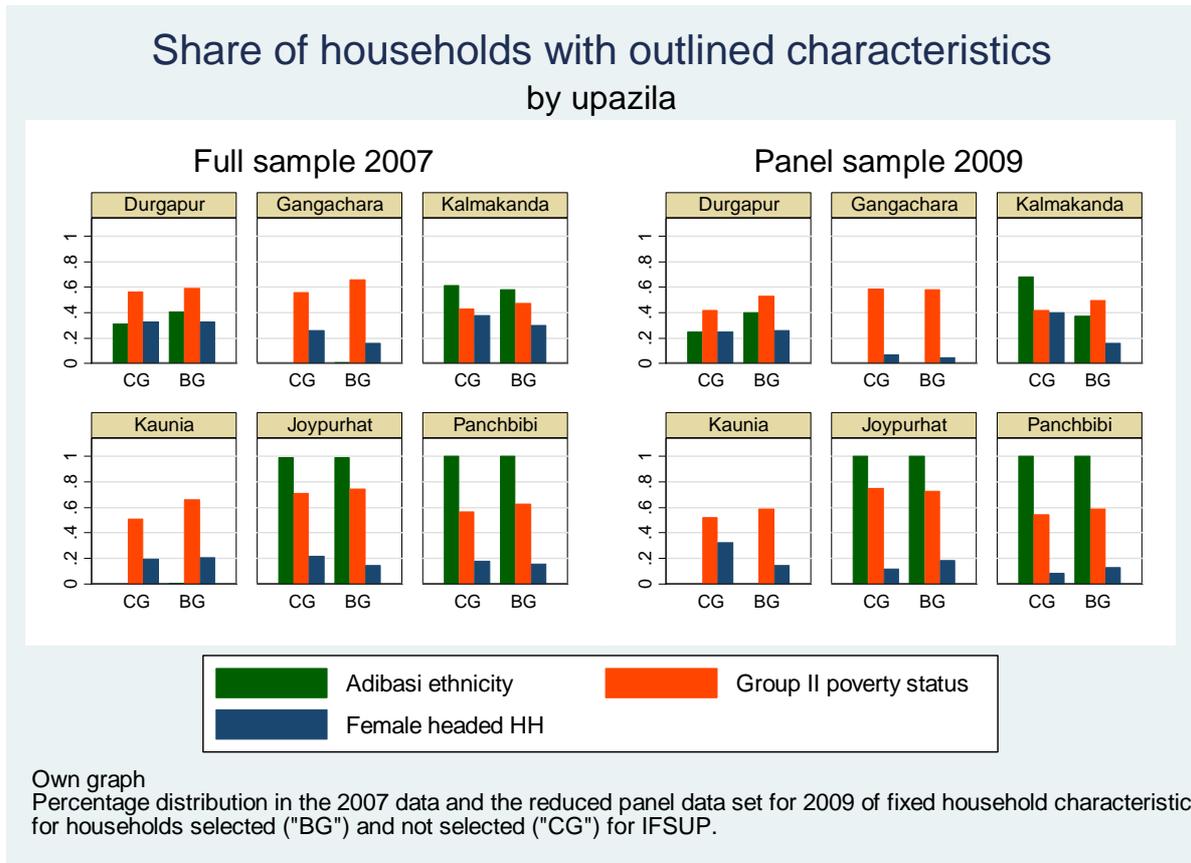
If this upward bias of control group means does not hold, impact measurement would be overstated. If the assumption of random non-collection of names does not hold, uninterpretable bias is introduced in estimates. For an assessment of these assumptions, point estimates of the SD, DD and within estimates have to be compared.

Figure 14: Attrition in control and target groups by upazila for the panel data-set



Own graph
 Percentage by which coverage shrank for the constructed panel data-set for households selected ("Target group") and not selected ("Control group") for IFSUP

Figure 15: Difference in fixed household characteristics between control and target group households by upazila and in-between the 2007 cross-section households and 2009 panel households after attrition.



Impact measurement is conducted following the estimation of a fixed-effects regression

$$(Y_{i,t=1} - Y_{i,t=0}) = B_i \delta + \beta_2 \text{year} + \beta_3 \text{treatment} \times \text{year} + (X_{i,t=1} \alpha - X_{i,t=0} \alpha) + (Z_{i,t=1} \gamma - Z_{i,t=0} \gamma) + \varepsilon_i,$$

with $(Y_{i,t=1} - Y_{i,t=0})$ as change in the outcome variable of interest for household i for 2007 ($t = 0$) as well as 2009 ($t = 1$),

B_i as vector of individual constants,

$\text{treatment} \times \text{year}$ as interaction term for the actual project impact,

year as a dummy variable indicating the year of the baseline (2007) or impact survey (2009),

X_i being, for each year, vectors of time-invariant household characteristics that cancel out due to the within-estimation approach, be they observable (from geographical location to ethnicity) or unobservable (from ability to social networks),

Z_i being, for each year, vectors of time-variant household characteristics such as HH size.¹³⁷

¹³⁷ Only included in the within-estimation of Appendix 6.

3.2.3.4 Specification Including Additional Controls (Appendix)

The above regressions include controls for fixed household characteristics influencing the selection process (ethnicity, female headedness, group II poverty status), household size (project implementation is conducted at household level, though many outcome variables are measured at individual level) as well as geographic locality (as as-if-randomization was conducted on union level).

Appendix 6: Further Analysis of Internal Validity in this respect reports different specifications of the impact estimation tables, introducing additional controls for education levels, outside wage labour income, further NGO/GO support. As these variables are likely to be, at least in part, a project impact themselves, they are included to the within estimation to be able to compare outcomes for the same individuals. Their inclusion serves the important purpose to warn against overturning of results or large deviations in point estimates or standard errors beyond a comparison of the DD, panel, and Manski bounds. Results will be noted in the main text where necessary.

Additional interest lies in heterogeneity of treatment effects and specific impact pathways. Results of Appendix 6: Further Analysis of Internal Validity give a first indication for this by including an interaction effect for IFSUP beneficiaries additionally receiving support by GOB social protection schemes.

3.2.4 Spill-Overs as Confounders

Estimation of treatment effects relies on the SUTVA assumption, implying no spill-overs. This assumption is very unlikely, especially given the transformative project pillar that aims to influence the socio-political context all ultra-poor in the respective area are living in. But it can be expected that spill-overs predominantly have a positive influence for ultra-poor control group households: First of all, through IFSUP employment opportunities are introduced that should relax pressure on the wage labour market. Secondly, IFSUP aimed at enhanced responsiveness for local government institutions in the just distribution of access to government social protection schemes. Third, IFSUP introduced not only knowledge on animal husbandry but as well access to livestock vaccination schemes in the regions that can be expected to spread beyond the project population.

This being, spill-overs are expected to introduce a downward-bias on estimated impacts.

3.2.5 Standard Error Concerns

Estimation of regressions always implies concerns for heteroscedasticity and, for time series data, serial correlation of error terms that would bias the estimation of t-values. Bertrand et al. (2004: 274f.) especially criticise the common practice of difference-in-differences estimations without taking these considerations into concern:

“because of serial correlation, conventional DD standard errors may grossly understate the standard deviation of the estimated treatment effects [...] computing standard errors that are robust to serial correlation appears relatively easy to implement in most cases, it should become standard practice in applied work”

Similarly, Stock & Watson (2008) recently brought the aspect to the agenda that usual robust standard errors controlling for heteroskedasticity can be inconsistent in case of fixed-effects panel analysis, at least if time periods exceed two.

Potential problems could be present for the IFSUP data, especially as the selection of control group and beneficiary group participants was implemented on union level. These potential problems are circumvented using *upazila*-level fixed effects¹³⁸ and cluster-robust standard errors on *upazila* level¹³⁹.

It has to be noted, though, that current discussions in the literature caution against the application of cluster-robust standard errors for the case of few groups (G) with a large number of within-group units. As Wooldridge (2010: 884) comments, “we should not expect good properties of the cluster-robust inference with small groups and very large group sizes when cluster effects are left in the error term.” For the evaluation in this thesis, cluster effects are therefore made explicit in the impact regression. But even if the cluster effects are excluded from the error term, “[w]ith small G , inference based on cluster-robust statistics could be very conservative when it need not be” (Wooldridge & Imbens 2007: 8).¹⁴⁰ Wooldridge & Imbens (2007: 9) therefore recommend for this case “whether or not we leave cluster (unobserved) effects in the error term, there are good reasons not to rely on cluster-robust inference” and propose the introduction of fixed effects together with, eventually, adjusting for heteroscedasticity by a robust regression.

¹³⁸ Oriented on the methodological approach of Banerjee et al. (2010b) in the estimation of effects of a randomized asset transfer project.

¹³⁹ Data on unions of housing is missing for part of the sample and cannot be easily linked for other parts due to different coding.

¹⁴⁰ A replication of the estimates using robust but unclustered standard errors resulted in more significant results (tables available on request). Further analysis concerning the necessity of clustering could draw on a test developed by Kdezi (2003: 12f.) for its presence.

3.3 *Impact Estimation*

The impact of IFSUP is measured for several bundles of variables while controlling for fixed household characteristics and geographical location.¹⁴¹ These are to give an indication for the impact of the various project pillars as well as the overall project effect. Chapter 3.3.1 provides an overview and discussion of variables indicative for the overall project effect. Chapter 3.3.2 reports impacts on asset holdings and Chapter 3.3.3 vulnerability related indicators. Chapter 3.3.4 will summarize the discussion on internal validity by drawing on findings from the covariates. Variables indicative for the transformative project pillar of IFSUP are not reportable at present, as clear-cut impact variables are missing¹⁴², difficult to compare over time¹⁴³ and weigh, although recently estimation strategies for this problem have been proposed, providing pathways for further analysis.¹⁴⁴

Indicators selected for impact measurement include recalls on food consumption, expenditure figures, summary indicators and PCA indices (comp. Appendix 5: PCA Analysis for the construction procedure). Impact measurement with the PCA indices is conducted given concerns for potential measurement error in the IFSUP data: Meals per day data as well as overall expenditure figures for 2009 were collected with a twelve month recall period. Expenditure figures for food expenditure and collected food value were unsystematically collected with a 30 day, seven day or one day recall. These recall periods are rather unusual compared to other approaches for expenditure figures and food security estimates¹⁴⁵, and problematic especially given the already persisting concerns with shorter recall periods: These

“methods rely heavily upon the memory of respondents leading to substantial measurement error even when people are asked to recall what they ate the day before, as in the case of the 24-hour recall [...T]these methodological challenges can lead to an unacceptably high measurement error, especially when interviewers are not fully trained and standardized against each other and the lead supervisor.” (Pérez-Escamilla & Segall-Corrêa 2008: 20).

¹⁴¹ Importantly, all impact estimates are robust to the exclusion of these control variables, both concerning significance and sign (results available on request).

¹⁴² Offenheiser & Jacobs (2006: 17) add, these might not even be quantitatively discernible, as is the case for diffuse measurement or a context specific intervention: Poverty reduction is not a technical issue, but centrally about challenging social injustice and addressing power imbalances as argues a rights based approach to poverty (Munro 2008: 28-37). “Advocacy projects [...] are context-sensitive and do not lend themselves to measurement by [RCTs. A]gencies funding advocacy projects therefore must rely on qualitative evaluation techniques and satisfy themselves with less tangible – though still meaningful – evidence.” (Offenheiser & Jacobs 2006: 17).

¹⁴³ It can additionally be expected that the effects of the transformative project aspects are, if any were present, lagging due to the inertia of changes in social and cultural systems; Fisman & Miguel (2007) e.g. provide evidence concerning the persistence of cultural attitudes to corruption.

¹⁴⁴ Banerjee et al. (2010c) apply an estimation strategy developed by Kling et al. (2007) in a similar context where they “test the null hypothesis of no effect [...] on ‘social outcomes’ against the alternative that [the intervention] improves social outcomes. [For this, they] construct an equally-weighted average z-score” (comp. as well Banerjee et al. 2010c: 7).

¹⁴⁵ For expenditure surveys, reference periods exceeding several months are already unusual (Pérez-Escamilla & Segall-Corrêa 2008: 17), for food intake 24-48 hour recall periods are common (Hatløy et al. 1998: 897).

Although twelve-month recall periods for food security impact measurement provide the important advantage of averaging seasonal fluctuations, especially of incorporating the locally occurring hunger season *monga*, their application requires a lot of care, since otherwise they are “prone to careless answers or to reporting their current situation as opposed to conditions during the prior twelve months” (Coates et al. 2003: 8, FN 5).

Additional measurement error is induced in the reported value of assets and livestock through the measurement and reporting system employed for data collection: Values are according to IFSUP project managers reported following the ‘best guess’ of the interviewee and interviewer. Furthermore, for the variables related to livestock, productive and non-productive asset values, interviewers did not separately code ‘zero’ and ‘missing’ in the impact survey. This information thus had to be combined as zero to be able to draw conclusions.¹⁴⁶ PCA indices on the other hand rely on assets generally visible to the interviewer, such as the number of rooms or the presence of durable goods. A comparison of estimates for indices and value figures is thus able to counter concerns for data validity.

On top of these measurement error problems, where indices based on criteria visible to the interviewer provide increased reliability, PCA indices aim at including dimensions of poverty “not captured by money metric measures” (Booyesen et al. 2008: 1113) – such as education levels, the quality of housing construction or ownership status of homestead.¹⁴⁷

Overall, these variable bundles based on different construction and measurement mechanisms is beyond being indicative for various project outcomes providing information on effectiveness through multiple complementing reference points in the data.

3.3.1 Overall project impact

Indicators for the overall project are selected to give an indication for the impact on overall food security and wealth of ultra-poor households. Variables include average meals per day per capita, the socio-economic status index (PCA), the index of secondary outcomes (PCA), yearly expenditure and yearly expenditure per capita, additionally impact on average food expenditure per capita per day and the value of collected food per capita per day.¹⁴⁸

IFSUP seems both over the range of presented variables and for the different specifications to have a very significant impact on the treated population, as reported in Table 2 (single

¹⁴⁶ Exemplarily, the amount of credit for the 982 control group households is reported for 396 households with a range of 110 to 30,000 taka, not distinguishing for the rest of the data between zero credit and households not willing or able to respond.

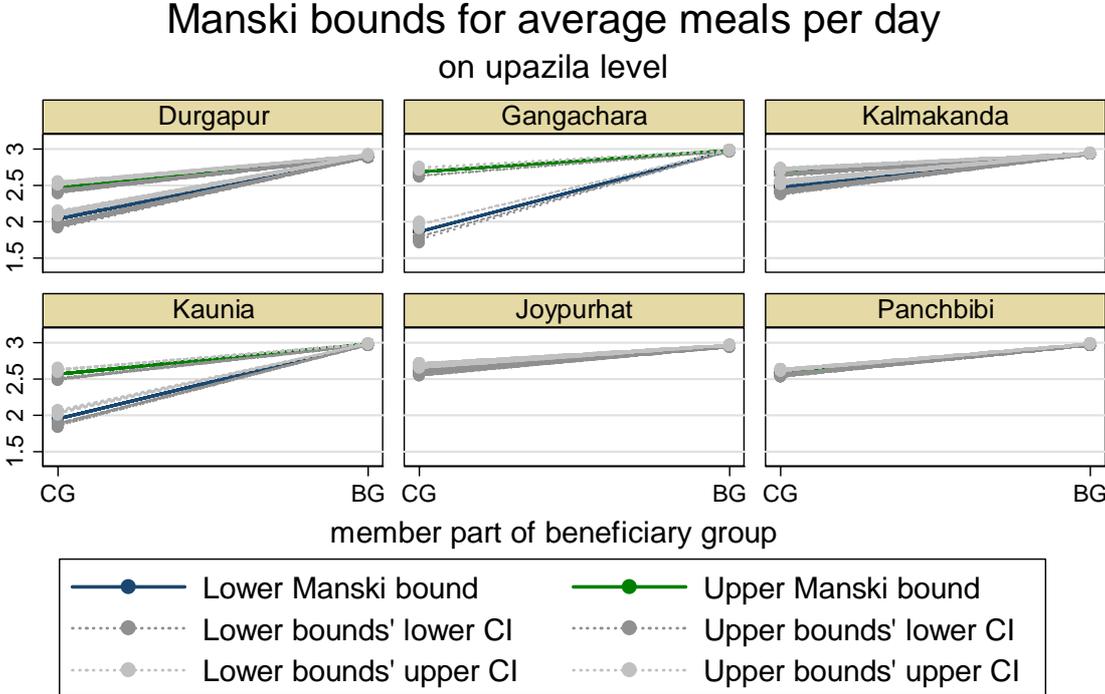
¹⁴⁷ Subsequently, PCA is applied to other evaluations of asset-based projects (comp. methodology in Banerjee et al. 2010b: 34).

¹⁴⁸ Effects for the latter two variables are not reported for the longitudinal impact estimations as baseline data for these variables is missing.

difference (SD) estimation with Manski bounds in the last two rows), Table 3 (DD estimation for non-attrition *upazilas*) and Table 4 (within-estimation for the reduced panel data set).

The intervention, as displayed in column 1 of the tables, raised the intake of average meals per day (mpd) from 0.328 (DD estimate) to 0.411 (within estimate) – results that are highly significant on the 1% level and similar to the Manski bounds of the effect $D^{mpd} \in [0.350, 0.700]$. These bounds are informative, as both lower and upper bounds have the same sign, and worst case bounds, as the outcome variable is naturally bounded (through the interval [0,3]). This figure indicates that the major purpose of the project, securing food security for ultra-poor households, was achieved. This figure highlights additionally that food insecurity can be overcome in a relatively short term of three years by an intervention such as IFSUP. The sharpness of bounds on the treatment effect for the overall sample reaffirms this finding. As displayed in Figure 16 bounds can be constructed conditional on attrition on *upazila* level as well. As outlined, upper and lower bounds for control group means are always lower than beneficiary group means for all cases with 95% confidence intervals. As expected, bounds on control group means are wider for *upazilas* with higher attrition rates.

Figure 16: Manski bounds and confidence intervals for average meals per day



Own graph
Manski bounds on upazila level with 95% confidence intervals (CI) following bootstrapping (50) for control group (CG) and beneficiary group (BG) of IFSUP.

The single variable analysis of the project effect for meals per day in Rudolph (2010) for single difference and Rudolph (2011) for DD estimates could raise concerns that the overall im-

impact on food security is a question of the choice of the indicator. This is notably the case as meals per day are calculated as average of reported months during the last year with approximately one, two or three meals per day for family members.

A comparison of impacts for other variables indicative for the overall project impact is therefore necessary. Evidence based on PCA indices as well as expenditure figures as independent variables reveal similarly significant results. The intervention is likely to have caused an increase in the SES index of 1.442 (DD estimate) to 1.962 (within estimate), again with informative, though very wide and on the 5% level insignificant bounds $D^{SES} \in [0.025, 2.371]$.¹⁴⁹ This increase of the SES by about 1.4 units can e.g. be represented by households owning a vegetable garden (0.73), a husking machine (0.31) and a tubewell (0.36) or households getting to own jewellery (0.63), a radio (0.40) and the homestead they live on (0.42).¹⁵⁰

The positive average effect of the project on the secondary outcomes status index (SOS) between 0.606 (DD estimate), though only significant on the 10% level, to 0.852 (within estimate), significant on the 1% level, is reassuring that the project raised the well-being of households beyond the direct transfer of assets through the households on education, home, homestead and durable asset related variables: An increase in secondary outcomes by one would be equivalent to the ownership of a radio (0.60) or the ownership of the house the respondent is living in (0.65). Manski bounds in this case are not informative, as the difference in means is ranging in the negative with $D^{SOS} \in [-0.910, 1.360]$. Attrition could thus cover an actual zero or negative effect of the project, at least for the *upazilas* where the DD estimation could not be implemented. But the positive and significant effect of the DD estimation reassures a positive impact at least for the *upazilas* Kalmakanda, Joypurhat and Panchibi.

Given the concerns for measurement error, it is comforting that the positive average project effect is visible not only through the asset-indices but as well in expenditure figures: Both yearly expenditure on household and on per capita level increased significantly (on the 5% and respectively 1% level for the single difference estimator, on the 10% level for the DD estimator and the 1% level for the within estimator) through the project, as reported in columns 4 and 5. The impact estimation for household expenditure ranges from 4532 taka (SD

¹⁴⁹ It has to be kept in mind though that these bounds are not worst case bounds, but upper (lower) bounds under the assumption of missing values having highest (lowest) observed values. It has to be noted additionally that factor loadings and subsequent PCA indices could only be calculated in ignoring attrition households. 95% confidence intervals constructed with bootstrapping (50) for the control group mean of SES leads to an upper Manski bound CI in-between 1.51-1.69 and 1.86-2.09, thus including the beneficiary group mean of 1.81.

¹⁵⁰ Estimation strategy analogous to Filmer & Pritchett (2001: 117) using the quotient of overall standard deviations and factor loadings as presented in Table 11 of Appendix 5 as weights for a change from zero to one in the dummy variables.

estimator) to 7143 taka (within estimator). This positive influence on expenditure is not driven by the baseline difference in household sizes, as indicated by the positive yearly pc expenditure, with estimates ranging from 960.1 (within estimator) to 1918.6 (DD estimator). That bounds are uninformative in the sense that they contain the possibility for a positive as well as negative impact for both variables is not unexpected, given the large range of expenditure figures.¹⁵¹ These point estimates still indicate that IFSUP households have been able to generate income streams that they are able to use: And these are quite significant in absolute terms. Using the DD figures and 2009 means as reported in the second last row as reference points, 15% of mean household expenditure and 25% of mean pc expenditure can be attributed to the project. To the extent that this measured expenditure can be interpreted as expression of permanent income levels (see for a summary of the inference problems involved Hsiang-Ke 2003), IFSUP with a one time intervention primarily aimed at the productive potential of ultra-poor households¹⁵² increased the well-being of household substantially. But for certainty on these conclusions, repeated measures after a longer time-span are necessary.

Finally, and only relying on data from the impact survey, IFSUP seems to have caused, significant on the 5% level, an increase in daily per capita food expenditure of 0.864 taka, which is not surprising, but as well an increase in the value of collected food by 0.355 taka. This latter finding was expected to be negative, but as the figure not only includes begging but as well food self-collected in nature, respondents could have reported the value of food consumed from home gardening as well. As home gardening was explicitly supported, a positive effect is then plausible.

Last but not least, the positive estimates of the panel data set allow with certainty the conclusion that at least for the individuals of the reduced panel data set the project effect was significant and positive in the indicators measured. Additionally, as all point estimates for the DD estimation are in fact lower (except per capita expenditure) than their counterparts in the SD scenario, this could indicate that attrition is biasing results downward and is indeed a concern for the data.

¹⁵¹ It has to be noted though that upper bounds are worst case bounds assuming zero expenditure for attrition households, leading to a maximum project effect of 9936.4 taka for yearly HH expenditure and 2105.3 taka for yearly expenditure pc.

¹⁵² The high correlation between current expenditure and expected permanent income can be theoretically expected given Friedman's (esp. 1957) permanent income hypothesis (comp. as well the application to SES indices in developing countries in Howe et al. 2008: 4).

Table 2: Difference estimation with Manski bounds for the overall project impact

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	average meals per day per average family member	socioeconomic status index (PCA)	secondary out- comes index (PCA)	yearly HH ex- penditure	yearly expendi- ture pc	average food expenditure pc pd	value of col- lected food pc pd
member part of beneficiary group	0.404 ^{***} (7.33)	1.450 ^{***} (10.31)	0.645 ^{***} (7.60)	4532.8 ^{**} (4.02)	1211.3 ^{***} (4.76)	0.864 ^{**} (3.15)	0.355 ^{**} (3.08)
female headed HH	-0.0422 [*] (-2.47)	-0.406 ^{***} (-6.27)	-0.241 ^{***} (-9.54)	-1774.3 [*] (-2.14)	-146.2 (-0.57)	-0.255 ^{**} (-2.87)	-0.0167 (-0.15)
<i>Adibasi</i> ethnicity	-0.0286 (-1.09)	0.447 ^{***} (9.00)	0.349 ^{***} (5.28)	1274.8 ^{***} (4.56)	233.7 (1.10)	0.507 ^{***} (9.89)	0.301 ^{**} (2.78)
group II poverty status	-0.0222 ^{***} (-6.20)	-0.386 [*] (-2.55)	-0.336 ^{**} (-2.97)	-1042.9 (-1.16)	-256.0 (-1.26)	-0.146 (-1.13)	-0.0946 (-1.46)
number of HH members	0.0146 [*] (2.12)	0.321 ^{***} (6.90)	0.236 ^{***} (8.33)	4005.4 ^{***} (21.17)	-856.1 ^{***} (-10.38)	0.906 ^{***} (27.01)	0.182 ^{***} (7.48)
Constant	2.453 ^{***} (45.21)	-1.069 ^{***} (-5.25)	-0.218 [*] (-2.03)	13582.6 ^{***} (13.62)	10955.8 ^{***} (29.43)	1.605 ^{***} (7.25)	0.685 ^{**} (3.42)
<i>upazila</i> dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R ²	0.375	0.339	0.295	0.351	0.234	0.489	0.171
N	5211	5273	5273	2217	2217	5362	5369
mean for BG	2.960	1.810	1.190	34374.3	7777.9	6.080	1.680
upper Manski bound	0.700	2.371	1.360	9936.4	2105.3	2.080	0.700
lower Manski bound	0.350	0.025	-0.910	-11638.3	-4188.8	-1.280	-2.830

Single difference estimation of outcome variable in the heading on project participation and the indicated controls. Regression includes *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. The last two rows show upper and lower Manski bounds for the difference of outcome variable means assuming as-if-randomization at baseline (N=6014 of which for N=213 CG members outcomes are missing due to attrition) for all households in one cluster; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Difference-in-Differences estimation for the overall project impact

	(1) average meals per day per average family member	(2) socioeconomic status index (PCA)	(3) secondary outcomes index (PCA)	(4) yearly household expenditure	(5) yearly expenditure per capita
project impact	0.328** (7.02)	1.442*** (10.72)	0.606* (3.76)	5221.0* (4.05)	1918.6* (2.98)
member part of beneficiary group	-0.00343 (-0.27)	-0.165 (-2.15)	-0.0174 (-0.36)	-159.4 (-0.34)	-635.4 (-2.17)
year 2009	0.673*** (13.24)	1.402*** (11.99)	1.239** (8.97)	7627.2* (3.02)	1128.4 (1.50)
<i>Adibasi</i> ethnicity	-0.0250* (-3.42)	0.546*** (16.97)	0.466*** (29.53)	298.0 (2.16)	-226.6 (-1.82)
female headed HH	-0.0280 (-0.88)	-0.224 (-2.13)	-0.142* (-3.07)	-418.8 (-1.26)	440.8 (1.50)
group II poverty status	-0.00999 (-0.42)	-0.345* (-3.13)	-0.238* (-4.16)	-1607.2 (-2.01)	-492.5 (-2.56)
number of HH mem- bers	0.00824 (1.38)	0.285** (7.66)	0.216** (9.81)	3288.9*** (31.42)	-698.1** (-6.55)
Constant	1.999*** (77.50)	-1.832*** (-22.38)	-1.128*** (-10.72)	6764.7*** (29.91)	8666.8*** (10.88)
<i>upazila</i> dummies	Yes	Yes	Yes	Yes	Yes
R ²	0.822	0.607	0.573	0.701	0.424
N	5816	5874	5874	4250	4250
mean 2009	2.900	1.359	0.852	33762.6	7676.8

Difference-in-differences estimation of the outcome variable in the heading on project participation and indicated controls. Impact estimation for *upazilas* Panchibi, Joypurhat and Kalmakanda. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Within estimation for the overall project impact

	(1) average meals per day per average family member	(2) socioeconomic status index (PCA)	(3) secondary outcomes index (PCA)	(4) yearly household expenditure	(5) yearly expenditure per capita
project impact	0.411*** (7.58)	1.964*** (13.21)	0.852*** (10.57)	7143.0*** (4.13)	960.1*** (4.31)
year 2009	0.622*** (14.01)	1.569*** (6.00)	1.557*** (7.48)	10552.1*** (5.26)	2126.6*** (7.38)
Constant	1.926*** (68.73)	-1.597*** (-12.75)	-1.153*** (-9.99)	17395.3*** (80.64)	5124.9*** (66.13)
N	1905	1926	1926	1387	1387
R ²	0.888	0.775	0.731	0.749	0.404
mean 2009	2.968	1.858	1.250	35698.8	8040.5

Within estimation on household level of the outcome variable in the heading on project participation for the restricted panel data set. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. The reported constant is an average over the FEs. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.3.2 Project Impact on Physical Asset Holdings

Table 5 reports the single difference with Manski bounds, Table 6 DD estimates and Table 7 within estimates for the variables tropical livestock units, value of livestock, productive assets index (PCA), value of productive assets, durable assets index (PCA), value of non-productive durable assets and the size of productive land in acre.

The selection of variables has two aims: First of all, it shall provide an indication for the impacts of the major project component, the distribution of assets. Secondly, it aims at answering the question whether IFSUP has been able to generate differences in the holding of non-productive assets already in the short time period. This would be an indicator, similar to the SOS, of an impact beyond the direct assets transferred.

The results confirm that households were able to better their productive potential due to the project:

As indicated in the first row for columns one and two in the tables below, beneficiary households hold significantly (1% for SD and within estimates, 5% for DD estimates) more livestock, both in units (ranging from 0.487 TLU for the SD and 0.687 TLU for the within estimate) and in value (ranging from 7254.9 taka to 7914.1 taka for the within estimate).

As indicated in columns three and four, beneficiary households have a significantly (on the 1% level for SD and within estimates and on the 5% level for DD estimates) increased position in asset holdings as measured by the first component as well as self-reported values. Measured impact ranging from 0.780 (SD estimate) to 1.082 (within estimate) for the PCA index, which is equivalent to households owning a tubewell (1.00) or a husking machine (1.00), and 2651.9 taka (DD estimate) to 3292.3 taka (within estimate) for the reported productive asset values.

As indicated in column five, the project significantly increased mean landholdings of households by 0.087 (SD estimation) to 0.101 (within estimation) acres, significant on the 1% (SD and within estimation) and 5% (DD estimation) level. This is a substantial increase given mean overall landholdings in 2009 of 0.105 acres, using the DD estimates.

Concerning livestock, productive assets as well as land holdings, the project therefore raised asset levels substantially throughout the three years. This is not surprising, given the transfers involved. The share of livestock related transfers as share of all assets distributed constitutes

about 70%¹⁵³. With a full livestock transfer equivalent to one cow (0.7 TLU), the measured SD impact of 0.487 TLU could be equivalent to exactly the transferred amount. With a negative spin, this could indicate that project beneficiaries did not build up livestock beyond the initial transfer¹⁵⁴. With a positive spin, beneficiary households' livestock asset base was maintained at least during the project implementation phase.

That beneficiary households could generate welfare beyond the face-value of the assets transferred could be indicated by the projects impact on durable assets in column 7, both in index and value measurement: Impact measurement ranges from 0.620 (SD estimate) to 0.768 (within estimate) for the PCA index, which would be more than compensated by households owning electronics (1.34) or jewellery (1.43), and from 968.3 (SD estimate) to 1137.9 (within estimate) for respective values. Importantly, though, this impact is only significant on the 5% level for the SD estimate and not at all for the DD estimate. On the other hand, Manski bounds, significant on the 5% level¹⁵⁵, for the durable assets PCA index are informative and positive with $D^{durables} \in [0.180, 0.980]$, although not worst case bounds as the index has neither a natural upper nor lower bound.

The findings on durable assets are therefore inconclusive, especially for the *upazilas* covered by the DD estimates. Positive within estimates allow at least a positive conclusion for part of the sample given control group households are an adequate counterfactual.

¹⁵³ Calculated following the asset distribution numbers in NETZ (2010: 30).

¹⁵⁴ In this respect it has to be kept in mind that 20% (10%) of the transfer had to be paid back if profits exceeded 7000 taka (lay between 5000 to 7000 taka).

¹⁵⁵ For the beneficiary group, the estimate of the mean of 0.73 has 95% CIs of maximally 0.68 to 0.77. For the control group, the estimate of the lower bound for the mean of -0.24 is surrounded by 95% CIs of maximally -0.33 to -0.15 and the estimate of the upper bound for the mean of 0.55 is surrounded by 95% CIs of maximally 0.43 to 0.66.

Table 5: Difference estimation with Manski bounds for physical asset-related variables

	(1) tropical live- stock units	(2) value of live- stock	(3) productive as- sets index (PCA)	(4) value of pro- ductive assets	(5) size of produc- tive land in acre	(6) durable assets index (PCA)	(7) value of non- productive du- rable assets
member part of beneficiary group	0.487*** (5.65)	7776.9*** (6.03)	0.780*** (7.84)	2998.5*** (5.23)	0.0867*** (7.10)	0.620** (3.95)	968.3** (3.79)
female headed HH	-0.0341 (-1.05)	52.15 (0.09)	-0.316*** (-4.48)	-373.7 (-1.44)	-0.0281 (-1.20)	-0.390*** (-14.19)	-470.7** (-3.37)
<i>Adibasi</i> ethnicity	0.211*** (4.81)	2924.8*** (8.63)	0.146** (3.77)	1207.7*** (5.13)	-0.0161 (-1.82)	-0.0653 (-0.96)	422.3*** (6.83)
group II poverty status	-0.102** (-2.62)	-1205.0* (-2.30)	-0.144 (-1.86)	-1102.2* (-2.51)	-0.0180 (-1.57)	-0.143 (-1.83)	-207.2* (-2.25)
number of HH members	0.0609*** (5.82)	741.6*** (5.73)	0.121*** (4.85)	445.6*** (9.01)	0.0227** (3.81)	0.0820*** (5.53)	178.2*** (13.19)
Constant	0.135** (3.34)	1326.4 (1.59)	-0.633** (-3.41)	1022.7** (2.70)	-0.0595** (-2.68)	-0.144 (-1.18)	330.0 (1.90)
<i>upazila</i> dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R ²	0.116	0.152	0.175	0.123	0.118	0.227	0.118
N	5373	5373	5373	5373	5373	5373	5373
mean for BG	0.958	12668.0	0.803	4851.5	0.162	0.731	2143.4
upper Manski bound	0.590	8712.6	1.110	3403.0	0.120	0.980	1260.0
lower Manski bound	-0.520	-24596.4	-0.520	-12620.8	-0.940	0.180	-5757.1

Single difference estimation of outcome variable in the heading on project participation and the indicated controls. Regression includes *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. The last two rows show upper and lower Manski bounds for the difference of outcome variable means assuming as-if-randomization at baseline (N=6014 of which for N=213 CG members outcomes are missing due to attrition) for all households in one cluster. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Difference-in-Differences estimation for physical asset-related variables

	(1) tropical live- stock units	(2) value of livestock	(3) productive assets index (PCA)	(4) value of pro- ductive assets	(5) size of produc- tive land in acre	(6) durable assets index (PCA)	(7) value of non- productive durable assets
project impact	0.527** (7.71)	7254.9** (5.50)	0.812** (5.28)	2651.9** (4.99)	0.103** (4.59)	0.659 (2.80)	1074.7 (2.36)
member part of beneficiary group	-0.114 (-2.10)	-151.2 (-0.22)	-0.110** (-4.64)	118.5 (0.44)	-0.0178 (-1.36)	-0.0191 (-0.55)	-25.65 (-2.70)
year 2009	0.127 (1.51)	4675.0** (4.74)	0.481* (3.35)	1455.3 (2.56)	0.00332 (0.58)	0.377 (1.55)	674.4** (8.48)
<i>Adibasi</i> ethnicity	0.190*** (22.67)	1990.8*** (12.81)	0.0683 (1.67)	509.0*** (12.22)	-0.0111 (-2.55)	0.0857*** (10.75)	315.9*** (10.33)
female headed HH	-0.0235 (-0.78)	131.4 (0.25)	-0.133 (-1.70)	-241.9* (-3.38)	-0.0153 (-1.74)	-0.265* (-4.26)	-263.2 (-2.67)
group II poverty status	-0.0712* (-3.57)	-603.5* (-3.75)	-0.159 (-2.07)	-618.8 (-1.93)	-0.0118 (-1.42)	-0.132 (-2.00)	-101.9 (-2.45)
number of HH members	0.0630* (3.81)	588.5* (4.12)	0.0906** (7.51)	297.0*** (24.22)	0.00826 (2.03)	0.0908** (8.70)	102.8*** (44.34)
Constant	0.0825 (1.60)	-390.7 (-0.37)	-0.764*** (-15.65)	670.3 (1.34)	0.00582 (0.84)	-0.552* (-3.54)	373.6 (1.97)
<i>upazila</i> dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R ²	0.227	0.374	0.320	0.249	0.113	0.221	0.236
N	5960	5960	5960	5960	5960	5960	5960
mean 2009	0.932	12536.6	0.478	4427.7	0.105	0.379	1853.6

Difference-in-differences estimation of the outcome variable in the heading on project participation and the indicated controls. Impact estimation for *upazilas* Panchibi, Joypurhat and Kalmakanda. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Within estimation for physical asset-related variables

	(1) tropical live- stock units	(2) value of livestock	(3) productive assets index (PCA)	(4) value of produc- tive assets	(5) size of productive land in acre	(6) durable assets index (PCA)	(7) value of non- productive durable assets
project impact	0.687*** (10.76)	7914.1*** (8.97)	1.082*** (7.54)	3292.3*** (13.74)	0.110*** (5.57)	0.768*** (4.17)	1137.9*** (4.15)
year 2009	0.0368 (0.62)	3203.9*** (5.83)	0.432** (2.74)	853.0** (2.90)	0.0424 (1.77)	0.698* (2.33)	960.5*** (5.56)
Constant	0.272*** (9.13)	1220.5** (2.60)	-0.625*** (-14.74)	679.8** (3.48)	0.0156 (0.82)	-0.588*** (-4.72)	225.1 (1.60)
N	1953	1953	1953	1953	1953	1953	1953
R ²	0.428	0.560	0.509	0.418	0.168	0.489	0.431
mean 2009	0.941	12125.4	0.855	4807.3	0.166	0.865	2314.5

Within estimation on household level of the outcome variable in the heading on project participation for the restricted panel data set. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. The reported constant is an average over the FE s. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.3.3 Project Impact on Vulnerability of Households

Below, for the variables savings, credit, index of income source diversification, housing quality index (PCA), dietary diversity score, sickness in days per capita, age of pregnant females, mothers' checkup status around birth and the vaccination status of children (>1.5 years) results are presented in Table 8 (SD estimates with Manski bounds), Table 9 (DD estimates) and Table 10 (within estimates).

Beyond the build-up of asset levels presented and the overall improved situation, these estimates give an indication for a reduced vulnerability of beneficiary households to shocks through the project.

Important indicator of decreased vulnerability – in its dimension to be able to “deal with risky events when they occur”¹⁵⁶ (Ellis et al. 2009: 23) – is the access of households to capital markets and the ability to build up savings. Savings are a new focal point of development practitioners and theorists alike and their potential to decrease vulnerability by providing both a buffer for shocks and the means for productive investment is increasingly recognized (Armendáriz & Morduch 2010: 147f.). Access to credit on the other hand has long been seen as major poverty alleviation mechanism, both by unleashing the productive potential of the poor (Yunus 2006) and by enabling the smoothing of consumption patterns throughout seasons and crises. That this might be a viable pathway even for the ultra-poor, who are in many contexts shown to be excluded from microcredit access (e.g. for northern Bangladesh Amin et al. 2003), has recently been argued for by Khandker et al. (2010a: 39), following the evaluation of a ultra-poor microcredit program as well in northern Bangladesh.

The impact of IFSUP on savings is positive, as expected by the savings-encouragement design of the program: Mean savings increased significantly (1% for SD and within, 5% level for DD estimates), from 607.3 taka (within estimate) to 723.5 taka (SD estimate). Though Manski bounds are not informative, this indicates an increased coping ability of beneficiary households – given the beneficiary group mean in 2009 as reported in the third-last mean of Table 8, the project was responsible for 87.5% of savings built up by beneficiary households. Concerning credit, the impact estimates tell a different story: Impact seems to be (for SD and within estimates significantly on the 5% and respectively 10% level, but not for DD estimates) negatively related to the IFSUP project. Mean credit levels sank by 771.5 for the DD and 917.6 for the SD estimate. Manski bounds are informative, with

¹⁵⁶ Opposed to the second dimension of being exposed to risks, unrelated to coping abilities.

$D^{credit} \in [-6028.3, -698.8]$, significant at the 5% level¹⁵⁷, and worst case bounds in the sense that impact must lie below -698.8 taka. But given the insignificance of the DD estimates, a time-variant or regional bias could be present in this case. Even then, the conclusion holds that IFSUP had a negative or negligible impact on credit.

A priori, it would have been expected that beneficiary group members would through their inclusion in the project increase their access to (informal) credit, through increased collateral (Banerjee & Duflo 2010: 62f.). It could additionally have been hypothesized that they would use this access to improve their businesses, given the expectation that their productive potential is underused. However it has to be kept in mind that credit is utilized for other purposes than productive investment as well: Banerjee et al. (2010c: 12) report from a randomized microcredit intervention on moderately urban poor in India that only around 50% of loans had a (partially) business related purpose. This indicates, as Banerjee & Duflo (2011: Chapter 9) argue, that access to credit is not enough for scaling-up businesses. These might just not be profitable enough after they have reached a certain small scale. Further external constraints and human psychology additionally play a major role in the decision of the poor to borrow. Recently, it has been proposed that the flourishing microcredit movement could be the expression of savings, rather than borrowing constraints, e.g. in following self-control problems that might even be higher the poorer individuals are (Banerjee & Mullainathan 2010: 43): As Banerjee & Duflo (2010: 74) put it, from this perspective credit “is not the only way (and possibly not the best way) to offer a commitment to saving”.

The negative to negligible impact of IFSUP on credit behaviour could be an expression of both these phenomena: On the one hand, it is plausible that IFSUP households have reached a certain level of productive asset ownership, from which further scaling up is difficult. Many of the ultra-poor households’ income generating activities are essentially home-base production. Scaling these up might require investments beyond the scope of the ultra-poor. On the other hand, the savings device of IFSUP provides households with means to save up large lump sums for the purchase of durable assets and the like, first evidence of which has been discussed in the previous chapter. Both reasons support the finding that IFSUP households, at least in a project phase-out situation, have reduced their borrowing. Given these findings, the question remains open whether IFSUP beneficiaries would have increased access to credit had they had wanted to.

¹⁵⁷ Mean credit levels of beneficiaries in taka at 183.05 and their CIs of a minimum 142.82 to 231.02 do neither overlap control group CIs around the lower bound of the mean (885.54), ranging from 732.3 to 1048.04 nor around the upper bound of the mean (6214.98), ranging from 5314.30 to 7193.35.

Beyond these questions of access to financial products, the risk coping ability of ultra-poor beneficiaries can be expected to have further increased, as indicated by column 3 in the tables below. Diversification of income sources is measured as simple index count of the different sources a household receives income from. This diversification index can be expected to be correlated with risk diversification. Nevertheless, the core aspect of diversification – “a mixture of activities that have net returns with negative correlations” (Rayhan & Grote 2010: 597) – cannot be directly captured due to data lacking on these correlations. Concerning the measured DIS, IFSUP increased the diversification score significantly (1% for SD and within, 5% level for DD estimates) from 1.093 (SD estimate) to 1.376 (within estimate). Manski bounds are informative and worst-case bounds with $D^{DIS} \in [0.520, 1.760]$, significant on the 5% level.¹⁵⁸ It is in this respect important to remember that IFSUP members on average received more than one IGA. All beneficiary households were e.g. given the means to start a vegetable garden additional to their main asset transferred. In this light, a diversification well above one would have been *ceteris paribus* expected. An impact ‘only’ around one seems to indicate that substitution effects are taking place for the households: IGA transfers lead to substitution away from outside occupations to the management of these IGAs.¹⁵⁹ This finding is highly relevant in another context as well, as it could be indicative of the argument for positive project spill-overs to control group members in Chapter 3.2.4.

Additionally, the increased coping ability of beneficiary households is as well indicated by improved housing facilities, measured by a PCA index. Significant on the 5%-level, housing quality increased by 0.139 (SD) to 0.340 (within estimate). Manski bounds are not informative. This increase is relatively small in absolute numbers, as a change in ownership status of house (1.26) or homestead (0.90) is related to much higher scores. Still, the accumulation of housing is a prime example of an increase in lasting well-being beyond the direct assets transferred by the project. In this respect, it is especially important that housing is the major prerequisite for escaping poverty, as is argued by Moser (2007: 6), drawing on findings from a panel study in Latin America: “[W]hile accumulating housing itself as an asset does not pull households out of poverty, it is a prerequisite for the assets that do.”

¹⁵⁸ CIs for the beneficiary group mean of 4.46 are ranging from minimally 4.40 to a maximum of 4.50, CIs for the lower bound of the control group mean of 2.70 range from minimally 2.56 to a maximum of 2.83, while CIs for the upper bound of the control group mean range from minimally 3.80 to 4.08.

¹⁵⁹ Investments in education are missing in this index, thus a full-time reduction in child labour would decrease the DIS score as well.

While the variables mentioned above tried to directly capture vulnerability aspects related to risk coping ability, the following discussion aims at finding indirect evidence on the health status of households. Health shocks are a prime reason for impoverishment; and poor health is a main determinant of being locked in ultra-poverty, as Kabeer (2010: 71) finds for rural Bangladesh.

In this respect, food variety as measured in the DDS is an important indicator of nutritional adequacy and a predictor of health outcomes, as summarized in Hodgson et al. (1994: 143-145): Higher food variety is related to height and weight for age measures of children as well as to the prevalence of heart and sugar related diseases and even to total mortality, though partially drawing on much more detailed measures of diversity. IFSUP achieved a significant (on the 1% for SD, 5% for DD and 10% level for within estimates) and positive impact of 0.810 (within estimate) to 0.956 (SD estimate) on dietary diversity, with informative and worst-case Manski bounds of $D^{DDS} \in [0.690, 2.110]$, significant on the 5% level.¹⁶⁰ This increase by at least one food variety could be related to the relatively cost effective promotion of home gardening by trainings and the distribution of seeds, which calls for an inclusion of this project component in related projects as well.

The impact of IFSUP on other health indicators, especially per capita sick days, age of pregnant mothers, the participation of mothers in ante- and postnatal checkups and vaccination status of children is inconclusive. SD estimates indicate a positive influence on sick days (which might as well be due to household members permitting themselves to stay at home), higher vaccination rates for children (with informative Manski bounds significant on the 5% level¹⁶¹) and, counter-intuitively, a reduction in the mean age of pregnant females (which might indicate that more children are born in the beneficiary group). These results are inconsistent with (insignificant) within and DD estimates. This could have various reasons: A first might be the severely reduced sample size for the latter three variables; a second difficulty in measurement (checkup itself might e.g. not be the variable of interest but checkup by qualified health personnel); a third that effects on health indicators have not played out yet. In this respect, follow up studies are necessary to get a clearer picture.

¹⁶⁰ CIs for beneficiary means of 7.32 range from minimally 7.28 to maximally 7.36; CIs for the lower bound of control group means (5.22) range from 5.00 to 5.44, CIs for the upper bound of control group means (6.64) from 6.54 to 6.75.

¹⁶¹ CIs for beneficiary means of 0.97 range from minimally 0.95 to maximally 0.98; CIs for the lower bound of control group means (0.71) range from 0.65 to 0.77, CIs for the upper bound of control group means (0.89) range from 0.82 to 0.93.

Table 8: Difference estimation with Manski bounds for vulnerability related variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	amount of households' (in)formal savings	amount of households' (in)formal credit	index of income source diversification	housing quality index (PCA)	dietary diversity score	amount of days lost due to sickness pc	age of pregnant females	mothers' checkup around birth	vaccination status of children (>1.5 years)
member part of BG	723.5*** (19.51)	-917.6** (-3.01)	1.093*** (7.23)	0.139** (3.00)	0.923*** (6.51)	0.199** (2.95)	-2.773** (-2.58)	0.111 (1.29)	0.0973** (2.65)
female headed HH	29.80 (0.98)	-49.05 (-1.62)	-0.239** (-2.58)	-0.0151 (-0.29)	0.00203 (0.02)	0.105** (3.20)	-0.598 (-0.60)	-0.0788 (-0.87)	0.0106 (1.15)
<i>Adibasi</i> ethnicity	71.90** (3.90)	-15.04 (-0.54)	0.365*** (10.37)	0.426*** (9.62)	-0.124 (-0.72)	-0.0219 (-1.06)	-0.737 (-1.83)	0.124 (0.69)	-0.00236 (-0.24)
group II poverty status	-12.20 (-0.54)	-66.32* (-2.42)	-0.113 (-1.39)	-0.406*** (-4.37)	0.0768 (1.17)	0.00834 (0.23)	0.428 (0.56)	0.000410 (0.01)	-0.0147 (-1.43)
number of HH members	17.72 (1.21)	66.60** (3.29)	0.234*** (8.28)	0.186*** (7.74)	0.0473* (2.12)	-0.0837*** (-4.37)	1.504*** (6.09)	0.00686 (0.18)	-0.00179 (-0.39)
Constant	34.75 (0.39)	954.0*** (5.36)	1.973*** (11.62)	-0.326** (-2.62)	6.512*** (34.37)	0.584*** (-9.05)	22.42*** (29.34)	0.212 (1.10)	0.900*** (22.51)
<i>upazila</i> dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R ²	0.0612	0.0779	0.268	0.157	0.261	0.0504	0.191	0.221	0.0417
N	5373	5369	5392	5284	5373	5373	435	464	1407
mean for BG	836.4	183.8	4.460	0.490	7.320	0.540	27.72	0.490	0.970
upper Manski	748.8	-698.8	1.760	0.870	2.110	0.220	0.140	0.150	0.270
lower Manski	-3024.8	-6028.3	0.520	-2.860	0.690	-4.500	-5.730	-0.0300	0.0900

Single difference estimation of outcome variable in the heading on project participation and the indicated controls. Regression includes *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. The last two rows show upper and lower Manski bounds for the difference of outcome variable means assuming as-if-randomization at baseline (N=6014 of which for N=213 CG members outcomes are missing due to attrition) for all households in one cluster. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Difference-in-Differences estimation with for vulnerability related variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	amount of households' (in)formal savings	amount of households' (in)formal credit	index of diversification of income sources	housing quality index (PCA)	dietary diversity score	amount of days lost due to sickness pc	age of pregnant females	mothers' checkup around birth	vaccination status of children (>1.5 years)
project impact	697.5** (9.28)	-771.5 (-1.55)	1.013** (6.13)	0.153** (5.87)	0.810** (4.61)	0.0866 (0.33)	-3.138 (-1.47)	0.163 (0.81)	0.0834 (1.26)
member part of BG	-4.998 (-0.16)	20.25 (1.04)	-0.106 (-0.90)	-0.0486 (-1.58)	0.0565 (1.12)	0.155 (1.00)	0.112 (0.64)	-0.156 (-0.88)	-0.0350 (-0.51)
year 2009	-15.14 (-0.08)	566.7 (1.20)	0.777 (2.13)	0.615*** (14.74)	1.673 (1.50)	-0.319 (-0.97)	1.156 (0.51)	-0.328 (-0.93)	0.141*** (12.00)
<i>Adibasi</i> ethnicity	28.59* (3.80)	-17.59 (-2.24)	0.583*** (16.98)	0.425*** (20.22)	-0.00413 (-0.09)	-0.0800* (-4.08)	1.507** (4.87)	0.0398 (2.90)	0.0188** (5.02)
female headed HH	30.06 (0.97)	18.93* (3.51)	-0.211 (-1.50)	0.0534 (1.78)	-0.162 (-1.26)	0.225* (3.23)	1.999 (2.32)	-0.0593 (-2.01)	0.0384 (2.01)
group II poverty status	-13.01 (-0.79)	-43.39* (-4.02)	-0.217 (-2.14)	-0.258*** (-13.30)	0.0324 (0.43)	0.00574 (0.09)	-0.718 (-1.68)	-0.0494 (-1.24)	-0.0131 (-0.59)
number of HH members	24.76 (1.11)	48.75 (1.35)	0.239** (9.08)	0.143*** (21.39)	0.0300 (1.79)	-0.116 (-2.64)	0.399 (2.54)	0.000438 (0.04)	-0.00961 (-1.76)
Constant	-45.53 (-0.82)	183.5 (0.94)	1.716*** (10.20)	-0.901*** (-12.86)	6.424** (9.14)	0.814 (2.41)	27.86*** (33.02)	0.853** (5.43)	0.793*** (34.44)
<i>upazila</i> dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R ²	0.266	0.0719	0.425	0.251	0.618	0.0458	0.0310	0.132	0.0957
N	4300	5958	5978	5879	5960	5960	2658	1393	1642
mean 2009	707.7	226.5	4.405	0.237	7.165	0.533	28.70	0.553	0.980

Difference-in-differences estimation of the outcome variable in the heading on project participation and indicated controls. Impact estimation for *upazilas* Panchibi, Joypurhat and Kalmakanda. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Within estimation for vulnerability related variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	amount of households' (in)formal savings	amount of households' (in)formal credit	index of diversification of income sources	housing quality index (PCA)	dietary diversity score	amount of days lost due to sickness pc	age of pregnant females	mothers' checkup around birth	vaccination status of children (>1.5 years)
project impact	607.3 ^{***} (4.18)	-774.7 [*] (-2.05)	1.376 ^{***} (9.62)	0.340 ^{**} (3.83)	0.956 [*] (2.03)	-0.0756 (-1.01)	0.232 (0.11)	0.0789 (0.72)	0.0800 (0.48)
year 2009	111.9 (0.81)	674.1 (1.74)	0.830 ^{**} (3.89)	0.682 ^{***} (13.26)	2.349 ^{**} (4.02)	-0.419 [*] (-2.56)	0.800 (0.40)	0 (0.00)	0.286 [*] (2.07)
Constant	31.32 (0.61)	306.8 ^{***} (5.41)	2.309 ^{***} (30.74)	-0.475 ^{***} (-11.02)	3.954 ^{***} (9.89)	0.967 ^{***} (11.95)	26.87 ^{***} (83.52)	0.431 ^{***} (5.58)	0.624 ^{***} (40.70)
N	1694	1953	1954	1931	1953	1953	160	102	697
R ²	0.410	0.0542	0.636	0.326	0.648	0.0394	0.0155	0.0125	0.311
mean 2009	758.6	199.9	4.409	0.525	7.297	0.539	27.66	0.526	0.981

Within estimation on household level of the outcome variable in the heading on project participation for the restricted panel data set. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. The reported constant is an average over the FE s. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.3.4 Findings for Correlates and Summary Statistics

The validity of the arguments in the discussion above is supported by the structure of the estimates for the correlates of the overall project impact.

Most importantly, for the DD estimates of Table 3, Table 6 and Table 9, second row, includes a measure for the effect of inclusion in the project at baseline (measured by the variable ‘member part of beneficiary group’). This inclusion is not significant for any variable but PCA asset index and in this case and most others¹⁶² negatively related to the outcomes. This reassures the argument that control group outcomes are, conditional on the correlates included, identical or even upwardly biased in comparison to the true beneficiary group counterfactual and can thus indeed be interpreted as point of reference for the project impact.

The plausibility of the data is further strengthened by findings for the vector of covariates included in Table 2-Table 10, leading to a conclusive picture over estimation strategies and variables selected:¹⁶³ Household size is overall significantly and positively linked to the variables presented.¹⁶⁴ This is in line with theory, suggesting that larger households can better diversify and spread risk (Ellis 2000: 14f.) and that economies of scale are relevant on the household level (Townsend 1994: 560). Female headedness is negatively associated with outcomes¹⁶⁵, in line with findings that female headedness is a main determinant of being chronically poor (Rahman et al. 2009: Chapter 1), at least if controlling for additional demographics (Meenakshi & Ray 2002: 558). Additionally, the negative correlation of group II poverty status in 2007 with the outcome variables¹⁶⁶ is indicative for some sort of continuing persistence in the poverty levels perceived, in line with poverty trap arguments (Carter 2007, Dasgupta 1997, Wood 2003). Last but not least, ethnicity is positively related to outcomes:¹⁶⁷ This is surprising, given the expectation that these ethnic minorities are especially discriminated against (comp. findings from India and Bangladesh in this direction by Lehmann 2006, Meenakshi & Ray 2002), but in line with findings of Miah (2008: 12f.) reporting no special discrimination in marketing (besides pig rearing) for *Adibasis* amongst the group of ultra-poor. It is therefore well possible that *Adibasi* ethnicity is a determinant of ultra-poverty but

¹⁶² Besides productive asset value, DDS, credit and mothers’ age; higher sick days are in the expected direction.

¹⁶³ Though health related variables show partially inconclusive patterns, in line with the hypothesis that data quality may be unclear in their case.

¹⁶⁴ The relation is insignificant for MDP, productive land size, credit, savings and health related variables in the DD estimates and savings and health-related variables in the SD estimates. A negative correlation is estimated for expenditure per capita, vaccination status and, as expected, sick days.

¹⁶⁵ Besides pc expenditure in the DD estimation, livestock value. Health and financial market indicators are inconclusive, which is plausible given the special role women is attributed in family care and foresight (Armendáriz & Morduch 2010: 233).

¹⁶⁶ Besides inconclusive findings for health related indicators.

¹⁶⁷ Besides MPD, credit, land size and inconclusive for health related variables.

not associated with a lower status amongst the ultra-poor; further research in this direction would be necessary to confirm this hypothesis.

Problematic is the interpretation of the positive year effect: A priori, as ultra-poor households are theoretically expected to be in a poverty trap, zero or only slight average gains would be expected over the years. Exogenous reasons such as a recovery period following natural disasters – in Bangladesh severe inundation in the whole country in 2004 and in its northwestern provinces in 2005 (Rayhan & Grote 2010: 597f.) – could explain these results as well as data inherent problems, such as measurement error under the assumption that households had incentives to underreport their wealth for the baseline survey. The continued existence of measurement error for PCA based indices would be counterintuitive, though. An analysis of additional time periods is necessary to find an answer to this puzzle, especially whether this upward trend for the control group will hold. Given the data available, it has to be assumed that a possible bias is constant over both groups.

Last but not least, the reported R^2 measures the explained variation in the data by the variables included in the regressions.¹⁶⁸ The absolute height of R^2 is of secondary interest. Given a sufficiently large N and under the assumption of as-if-randomization, an already small change in variation between households can be attributed to the project (Wooldridge 2009: 199f.). More interesting are changes in R^2 with a change in the variables for the model: Inclusion and exclusion of the correlate-vectors controlling for fixed household characteristics and in a second step regional dummies lead to the general and plausible picture that R^2 values for variables directly related to the program (e.g. TLU, productive assets, savings) change to a much lesser degree than secondary outcomes like housing quality or durables assets (results available on request).¹⁶⁹ For these secondary outcomes, the difference between groups is significantly influenced by the program; effects still have to play out, though.

3.4 Rounding up the Case for Internal Validity

The estimations above lead to a good case for internally highly valid results. It is of course possible, but not likely that additional factors confound the estimates. Appendix 6: Further Analysis of Internal Validity provides first insights to two further avenues of making the case for validity more robust: Interactions with government social protection schemes as well as bias through additional observable characteristics uncontrolled for.

¹⁶⁸ Partially, this can be misleading: The partially very high R^2 values for the within regression (e.g. around 70-80% in Table 4) are due to the inclusion of fixed effects for all individuals (Wooldridge 2009: 486).

¹⁶⁹ This is as well indicated by the share of point estimate measures in the variable mean: Point estimates for the socio-economic status amount to 80%, point estimates for the secondary outcomes index to 54% of the beneficiary group mean (measured for SD estimates, comp. Table 2).

Interactions of beneficiary group status and support by government social protection schemes is included into the impact regression in order to ascertain the finding that impacts are due to the asset-transfer, and not to transformative elements leading to inclusion of BG members in government programs through political lobbying. For these estimates, the general picture is a negative and insignificant relationship of the interaction of government program coverage with BG status. At the same time, the positive estimates for the project as such, as measured by the inclusion of households in the beneficiary group, is overall positive and remains significant for most variables.¹⁷⁰

Controls for the additional variables education level of household heads, wage labour income, household size and government support to a within estimates adds controls for changes in these variables for the same individuals to the impact estimation. The consistency of point estimates and significance supports the conclusion that IFSUP beneficiaries profited from the program even beyond changes in these covariates.

This additional evidence supports the picture that the IFSUP project had, within three years, a significant positive effect on food security, income poverty, asset levels both productive and non-productive as well as vulnerability. Health related variables are largely inconclusive.

IFSUP thus, in the context of northern Bangladesh and with the implementing organizations involved, proves to be a viable strategy in addressing the ultra-poor and improving their livelihoods. Three crucial questions are remaining in this respect:

First, the question concerning the sustainability of results: IFSUP aims at generating long-term pathways out of poverty. However, this can only be achieved by follow up studies with non-deteriorating data quality for both control and target group.

Second, the question concerning heterogeneity in impacts: Until now, it was addressed that IFSUP works; due to which impact pathways, however, remains unclear. Further analysis should differentiate between subgroups of the population in order to differentiate their treatment effects, as was started with the inclusion of interaction terms with coverage by government schemes.

Third, the question concerning the generalizability of effects remains: What would happen to IFSUP if the project were transferred to other contexts, in Bangladesh and even beyond, or scaled up? The following chapter will conclude the thesis with this question.

¹⁷⁰ These estimates are indicative in another aspect: The data is conform to the hypothesis that political lobbying led to the inclusion of the poorer strata among the beneficiary ultra-poor in GO schemes. Their lower status as such serves to explain the negative sign of estimates.

4 External Validity

“[T]o produce ‘useful knowledge’ beyond its local context, [an evaluation] must illustrate some general tendency, some effect that is the result of mechanism that is likely to apply more broadly” (Deaton 2010a: 448).

4.1 *Theoretical Perspective*

As long as only direct project evaluation is concerned, issues of external validity – that is questions of a transferability of the results obtained in the given context to other settings and/or a larger scale (Manski 2007: 26f.) – may not be of interest. The implementing organization is foremost concerned with generating knowledge about the effectiveness of their specific project, in order to be able to justify the costs to its stakeholders (be it individual or organizational donors or the general taxpayer) and to adapt their program if necessary. But beyond this scope, it is always the question of an application of the approach to other contexts and/or a larger scale – IFSUP was explicitly designed as a pilot project whose findings are to be disseminated for an adoption elsewhere. Moreover, the evaluation should always be inclined to answer questions of broader interest to society (Humphreys & Weinstein 2009). For this, impact evaluation has to solve not only questions of internal but as well questions of external validity – an important aspect too often left aside.¹⁷¹

As Cartwright (2007: 18) notes, “[t]he central question for external validity then is, ‘How do we come to be justified in the assumptions required for exporting a causal claim from the experimental to a target population?’.” Deaton (2010a: 450) makes a strong point of linking evidence to theory to ensure external validity:

“[R]unning RCTs to find out whether a project works is often defended on the grounds that the experimental project is like the policy that it might support. But the ‘like’ is typically argued by an appeal to similar circumstances, or a similar environment, arguments that depend entirely on observable variables. [...] In the end, there is no substitute for careful evaluation of the chain of evidence and reasoning by people who have the experience and expertise in the field. The demand that experiments be theory-driven is, of course, no guarantee of success, though the lack of it is close to a guarantee of failure.”

The hypothetico-deductive approach would try to falsify a general theory by harder and harder ‘acid tests’. Falsification of the theory leads to its adaption and in the process, factors relevant for external validity can step by step be adjusted for. This is in line with proponents of the experimental approach who argue in favor of a comparison of experiments in various

¹⁷¹ Concerning microcredit, Ault (2009) e.g. makes a strong case that for too long proponents have generalized ‘best practices’ from single evaluations that actually need to be adopted to local contexts – equivalently, the theoretical dependency of microcredit performance on macro context was ignored: “[A]ny comparison of [micro-finance institutions] that does not take into account the macroeconomic and macro-institutional environment [...] is incomplete” (Ahlin et al. 2011: 105). On the other hand, the evaluation of results generated on the field has at least an important advantage to evidence generated in laboratories: “[T]reatment occurs without the intervention of the researcher, through some incidental process” (Robinson et al. 2009: 344).

contexts.¹⁷² This expansion of knowledge is crucial for external validity, as theory alone is of no help when data is not available:

“Failures of a theory on the support are detectable; statisticians have developed methods of hypothesis testing for this purpose. Failures off the support are inherently not detectable [...] There is no objective way to distinguish among theories that ‘fit the data’ but imply different extrapolations” (Manski 2007: 28).

Without care for external validity, hypotheses from the comparison of projects can otherwise be misleading: According to the interpretation of Banerjee (2006a: 8)¹⁷³ RCTs e.g. revealed that school attendance can most cost-efficiently be boosted by deworming programs.¹⁷⁴ For this argument, he compares different interventions: deworming, conditional cash transfers (CCT), handing out school uniforms or free school meals. But each of these projects is adapted to a specific context and thus covers multiple dimensions. This invalidates a comparison on the single dimension ‘impact on school attendance’, while a comparison on all dimensions is prohibited by the need for local context adaption of programs.¹⁷⁵ And above all, who has looked into the question whether worms are obstacles to school attendance in Mexico, where evidence from CCTs has been gathered?

Duflo et al. (2008: 3952-3954) argue for a twofold distinction for assessing the external validity of programs: First, the degree to which a program is depending on the specific implementing NGDO is crucial: The effect of pilot programs could especially hinge on the fact that the quality of the implementing agency is high, positively influencing compliance rates and procedure adherence. Armendáriz & Morduch (2010: 307) argue that especially pilot RCT studies are usually very carefully set up: “Regionwide policies can seldom be implemented with the same level of care”. Added to this question must be concerns that success hinges on the relatively small scale of the operation: Beyond management capacity, this question relates to general equilibrium effects if projects are implemented on a larger scale. Second, they recommend to evaluate to what extent the sample used for obtaining results is specific,¹⁷⁶ and to what degree the implementation is restricted to carefully selected geographical

¹⁷² “Of course, even randomized trials are not perfect. Something that works in India may fail in Indonesia. Ideally, there should be multiple randomized trials in varying locations” (Banerjee 2006a: 8).

¹⁷³ Banerjee (Banerjee 2006a: 8) argues: “How costly is the resistance to knowledge? One way to get at this is to compare the cost-effectiveness of plausible alternative ways of achieving the same goal. Primary education, and particularly the question of how to get more children to attend primary school, provides a fine test case because a number of the standard strategies have been subject to randomized evaluations. The cheapest strategy for getting children to spend more time in school, by some distance, turns out to be giving them deworming medicine so that they are sick less often.”

¹⁷⁴ Drawing on the evaluation of Miguel & Kremer (2004).

¹⁷⁵ Goldin et al. (2006: 11) e.g. claim that the value of sustainability of an intervention is reduced by RCTs’, or more general quantitative methods’, focus on cost-benefit calculations, which comes down to the incomparability of one-dimensional impacts if projects necessarily include multiple facets due to context adaption.

¹⁷⁶ IFSUP results would for example *a priori* only be transferrable to other ultra-poor of group II and III that were targeted by the program. These effects can be found on an aggregate level as well: Gulli (1998: 25-29) for

sites. Evaluations could therefore try to randomize sites for achieving greater external validity, something that is “almost never done. Randomized evaluations are typically performed in ‘convenience’ samples, with specific populations” (Duflo et al. 2008: 3953). The solution can therefore only be “a combination of replications and theory that can help generalize the lessons from a particular program” (Duflo et al. 2008: 3953f.). Ensuring external validity is crucial, “[o]therwise, we will be gathering evidence, not knowledge” (Deaton 2006: 13).

4.2 Generalizations beyond IFSUP?

Through the above mentioned procedure, generalizations beyond IFSUP are possible to some extent. In the following, first the general question of a transferability of the approach to other NGOs will be addressed, followed by arguments of scale. It will be concluded that implementing agencies need a high management capacity. Beyond these concerns, the questions is whether the approach as such is valid in other contexts. Subsequently, the dependency on specific context is addressed and, in comparison with similar programs and theory, denied. IFSUP supports the case of asset-based approaches to poverty alleviation.

4.2.1 Transfer of the Approach to Other Implementing Organizations

A replication of the IFSUP strategy is in principle possible. NETZ and its PNGOs shared their approach in local and regional workshops with national and international donor agencies as well as the public. However, an online publication of the detailed final report of IFSUP has not taken place.¹⁷⁷ This is recommended for a further dissemination of the approach.

Furthermore, the question is crucial whether the positive example of IFSUP is related to the agencies involved. NETZ is carefully selecting partners with a long-standing experience in working with the poorest and most marginalized section of society. In this respect, it is likely that a transfer of the program hinges on finding suitable implementing organizations.¹⁷⁸ Nevertheless, it is reassuring that three different NGOs of different size¹⁷⁹ and experience in asset-transfer programs in different local contexts¹⁸⁰ succeeded with the implementation of

example reviews the evidence that financial sustainability and depth of outreach are negatively related for developmental microfinance institutions – she concludes that proponents arguing against this relation falsely generalize from too few data-points while proponents arguing for this trade-off have the problem of not being able to falsify that financially sustainable institutions could in fact reach the poorest, although the institutions under the respective study did not.

¹⁷⁷ Although NETZ (2010: 42) explicitly voiced “no objection to this report being published on [e.g. the] EuropeAid Co-operation Office website”.

¹⁷⁸ Differing attrition between *upazilas* could e.g. indicate the consequences of differing management capacity.

¹⁷⁹ With employees ranging from about 570 to 1800 and turnover from about 400,000 to 1.3 Mio € before project implementation (NETZ 2006: Part III).

¹⁸⁰ Compare for further details on differences between IFSUP regions Appendix 3.

IFSUP. Still, experienced organizations with high commitment are certainly necessary for a replication: As noted for similar programs, the

“approach demands a more compassionate work force compared to microfinance. The real challenge is creating such a compassionate work force and managing it with a focus on achieving results. This involves significant change and innovation in management” (Matin et al. 2008: 29).

4.2.2 Difficulties to Scaling Up

However good the reproducibility of the approach by other organizations may be, the question of the effectiveness of the approach at a larger scale is crucial, even within the very same context. Three aspects are important in this respect: First of all, management costs, secondly general equilibrium effects and third the capacity of implementing organizations.

Costs per unit of the IFSUP program amount to 245 € per household assisted with about 50% of the costs due to the asset transfer. This relates to 4082 households per million € invested. These high unit costs as such are an obstacle to large-scale implementation, especially as evidence on the large scale impact of these programs is missing. Taking the small scale of IFSUP into consideration, however, substantial economies of scale can be assumed in case of a broader coverage, driving unit costs down.

Concerning general equilibrium effects, impacts are unclear as well. Relating to the promotive project pillar, Udo et al. (1992: 50) e.g. argue that the seasonality in the availability of cattle feed is the major limiting factor in an expansion of cattle production for Bangladesh. This points to a serious limit in scaling up the project: A sufficient number of IGAs need to be found to support the livelihoods of the ultra-poor. This can most likely only be achieved if a redistribution of assets is implicitly involved in the project, e.g. by increasing the land purchase component in IFSUP. Additionally, the fact that spill-overs to other (ultra-)poor are in tendency positive has to be re-evaluated as soon as large-scale national transfers would e.g. influence livestock prices. While this has been termed a positive effect that “supports local livestock prices, and expands the sale opportunities” (Ellis et al. 2009: 273), it could well have a negative effect on the productivity of livestock investments by other ultra-poor. Concerning the preventive pillar, vulnerability can be addressed more sensible with a larger design, especially for insurance schemes risk will be spread more broadly.¹⁸¹ For the transformative pillar, backing by higher level authorities is certainly necessary for successful lobbying and obstacles given patronage elite structures in developing countries could well be expected in this respect – this links the approach to ‘good governance discourses’ and good governance as prerequisite for development (comp. Nanda 2006).

¹⁸¹ Comp. for the emergence of microinsurance schemes in the development discourse e.g. Cohen et al. (2005).

Related to the transformative project elements are, third, management problems that could provide obstacles to effectively scaling up. As Deaton (2010a: 448) cautions, “development projects that help a few villagers or a few villages may not attract the attention of corrupt public officials because it is not worth their while to undermine or exploit them, yet they would do so as soon as any attempt were made to scale up.” An additional concern relates to the adoptability of the program to local needs of the beneficiaries. Huda (2010: 40), relating to first experiences from scaling up for a similar program in India, reports how the importance of detail in selecting IGAs or in adopting skills training is crucial for project success but on the other hand difficult to achieve with a large scale government bureaucracy. Only by of the state, however, a large scale and national impact can be reached. Huda (2010: 41) therefore calls for a combined effort of NGDOs and state authorities:

“The state has the ability to effectively scale, while the NGO has a client-driven ethos. Perhaps the NGO can implement with resources provided by the state, or the government staff can be trained by the NGO on how to keep the varied needs of the extreme poor at the forefront of their strategy?”

Otherwise, the danger remains that the approach will end as one of the redistributive grant schemes criticised in the development discourse and replaced by a microfinance movement without the necessary outreach to the poorest (Matin et al. 2008: 29f.).

4.2.3 Specificity of the Sample and Transfer to Other Contexts

Crucial for a generalizability in principle, independent of the scale, is especially, whether results depend on the specificity of sample and context. But as data was gathered exactly under these specific circumstances, this question is by definition off the empirical support.

Still, the specific targeting design of IFSUP contains features that allow a good case for a transferability of results. IFSUP chose the poorest poverty strata fit for physical work in the on average poorest unions of the working area.¹⁸² Additionally, the project was implemented in *upazilas* with both a high and low percentage of ethnic minority groups (comp. Figure 13). As far as ‘average poverty’ of unions and *upazilas* is correlated to structural disadvantages concerning political space, infrastructure and environmental conditions, IFSUP estimates provide a bottom line for counterfactual impacts in other contexts of (northern) Bangladesh. Still, it has to be kept in mind that an application of the approach is only feasible for ultra-poor of group II and III. For those unable to be addressed by self-employment, the

¹⁸² Recall that IFSUP was implemented in three districts belonging to the poorest of the country (comp. Appendix 3). Within these districts the poorest *upazilas*, within which the poorest unions were selected. In these, an ultra-poor population group still fit for physical work was selected for coverage, of which on average a poorer subsection was singled out as project participants (comp. Chapter 3.2.1).

elderly, disabled and orphaned ultra-poor, must be targeted by other schemes, e.g. cash transfers (Schubert 2008)

As concerns the transferability of the approach beyond Bangladesh to another cultural and socio-political context, the case is not as clear-cut. Concerning microcredit, it is e.g. argued that its apparent success in Bangladesh hinges to a great extent on small travel distances due to the high population density in the country.¹⁸³ Similarly, the asset transfer strategy of IFSUP might need to be adopted, especially if group meetings are related to higher (opportunity) costs for both NGOs and households involved. A rigorous evaluation of the asset transfer approach in these contexts is necessary. First positive results are reported e.g. for Zimbabwe, though not as hard evidence (Ellis et al. 2009: 269-275). Without an estimation of counterfactual impacts, the conclusion that “very little is known more generally about the true success rates of livestock [and asset] transfer projects in Africa” (Ellis et al. 2009: 275) can only be replicated. Although the case for a transferability of the approach beyond the South Asian context is not yet rigorously studied, pilot programs are underway (Huda 2009).

4.2.4 Comparison with Other Programs

Preliminary evidence of these pilot programs seems to confirm the overall validity of the asset-based approach to ultra-poverty reduction: Ellis (2009: 273f.) concludes that “[t]he multiple roles of livestock in successful rural livelihoods in Sub-Saharan Africa are well known, and have been empirically verified in numerous studies. Livestock can provide the key to the successful construction of pathways out of poverty”. As reported in his evaluation of a Zimbabwean program, failures of the strategy were predominantly related to livestock diseases (reported similarly for other contexts by Huda 2009: 28f.). This highlights innovative features of IFSUP, such as the retraining of ultra-poor to veterinaries in order to enhance vaccination coverage while providing an innovative IGA.

For Bangladesh, the estimates for IFSUP are backed by findings from similar projects as well: Positive effects for BRAC’s TUP project are reported using change rankings from the community perspective (Sulaiman & Matin 2006: 23) as well as counterfactual outcomes (Hulme & Moore 2008, Matin et al. 2008: Chapter 5), leading to positive estimations for social, physical, human, financial and natural capital.¹⁸⁴ First positive results are as well reported from India (Huda 2010) and beyond for the experience of a pilot scheme testing the

¹⁸³ In this light, Pedrosa (2011: 298) finds that interest rates and the strictness of screening procedures increase with distance in Niger, reducing the feasibility of the approach to capital market constraints.

¹⁸⁴ Especially concerning per capita income, an increase to a DCI above 2100 kcal per day, food diversity social status, self-perceived health but not concerning investments in education of children.

effectiveness of the asset-based approach in various settings¹⁸⁵ (Huda 2009). Once rigorous evaluations have been concluded for these pilots (on the way e.g. by Banerjee et al. (2010b)), further lessons for the external validity of IFSUP can be drawn as well.

In perspective, the importance of the transformative project element must be highlighted: Hulme & Moore (2010: 160f.) emphasize that the inclusion of local elites in the process and institutionalization of ultra-poor organizations is highly relevant for improving the social position of ultra-poor and for ensuring lasting impacts. As they note, this “component is perhaps the most context specific and least transferable of the [design]” (Hulme & Moore 2010: 161). Given the analysis of mean impacts, IFSUP does not allow conclusions on its counterfactual impact without the transformative elements but findings in other contexts are highly enlightening in this respect: Chaudry (2010) reports how a one-time cash stipend to ultra-poor in Vietnam, similar in size with the IFSUP transfer, was used to a large extent for productive purposes and the building up of asset stocks. But the major reported drawback were “[a]ttempts to pressure beneficiaries and seize the project funds” (Chaudry 2010: 175), as the project lacked a successful transformative project component.

4.2.5 Relation to Theory

Overall, the theoretical approach behind IFSUP has to be brought back into the picture: Dowling & Chin-Fang (2009: 30) summarize that the “chronically poor [are] subject to multiple poverty traps” – related to education levels, physical capital, high vulnerability, poor natural environments, insufficient access to financial resources, isolation from markets and discrimination as well as political exclusion. In these circumstances, they cannot be expected to free themselves out of poverty. Economically, their productivity is so low that they need to be pushed to a level from where self-help is possible again (comp. for an approach to modelling these ‘smart subsidies’ Armendáriz & Morduch 2010: Chapter 10.15); socio-politically, their exclusion from society and dependence on exploitative patron-client relationships needs to be overturned (Ferguson et al. 2007, Wood 2003). These multiple deprivations not only prevent the poor from freeing themselves out of poverty, but as well prohibit mainstream development schemes in reaching them.¹⁸⁶ The “graduation model” (Hashemi & Umaira 2011) assumes that a multiple intervention, centred around an asset and skills-transfer to overcome the economic trap, and an intervention in the socio-political context to overcome the political and social discrimination are a viable solution to these problems.

¹⁸⁵ The approach centres on asset transfer but includes next to transformative aspects additionally a health component. Pilots are implemented in Peru, Honduras, Pakistan, Ethiopia, Haiti, Yemen, and India.

¹⁸⁶ Dowling & Chin Fang (2009: 64) note that in face of discriminatory practices by local elites “well-meaning government programs directed to help the poor can easily be undermined.”

Yet, as soon as the deprivation of the poor is found to be dependent on (additional) factors not covered by this model, e.g. low soil fertility or isolation from markets, it has to be adapted. While asset-building is termed crucial for escapes from poverty (Moser & Dani 2008), the potential for a sustainability of transfers might depend on context-related factors not accounted for in these situations. Analogously, Sen (2003: 527) emphasizes the combination of different escape routes from poverty for a sustainable transition above the poverty line. In his panel study from rural Bangladesh, those households succeeded that

“integrate various anti-poverty strategies, resulting in relatively high savings-investment and income growth rate [...] ‘structural’ factors related to the asset base of the household and market conditions were seen as the drivers of change” (Sen, Binayak 2003: 527).

4.3 Policy Lessons

Given the above argumentation, a general applicability of asset-based approaches is possible where managing capacity of implementing organizations and local determinants of poverty allows its application. For an implementation of the approach at a large scale, the case is not as clear-cut.¹⁸⁷ But especially for a larger scale it must be clear that replications of the program must be adapted to local circumstances. There is no ‘golden bullet’ approach to ultra-poverty reduction. This concerns not only the selection of feasible IGAs for ultra-poor households but also the transformative project aspects: Lobbying for ultra-poor must always depend on the space possible for these actions in the socio-political context. Other project components might have to be included in different contexts as well: A major point of departure in comparison with other asset transfer projects is e.g. the question of an additional inclusion of cash stipends¹⁸⁸ to relieve ultra-poor from concerns with current deprivation. This cash stipend can be a necessary project mean, but as IFSUP shows it is not necessary in every context and could just be a major driver of costs. At best, future intervention strategies differentiate treatment with and without additional stipends for only parts of the sample. Thereby, more can be learned about the contexts in which these stipends are necessary or where they are substitutable by quick-income generating IGAs.

¹⁸⁷ Although far from universal coverage, evidence from BRAC’s TUP with 300000 beneficiaries as of 2010 can contribute to this question (Hashemi & de Montesquieu 2011 @1).

¹⁸⁸ As e.g. in BRAC’s TUP program as reported by Hulme & Moore (2010: 153).

5 Conclusion

“[T]he best argument for the experimental approach: it spurs innovation by making it easy to see what works” (Banerjee 2006b: 19)

This thesis presented in its first part the general approach to the counterfactual in program evaluation: How is it possible to ensure causality in the data gathered on the ground? This question is at the heart of the exogeneity assumption necessary for sensible evaluations. While randomization is providing a direct pathway to this causality, it is not always the optimal estimation strategy. But randomized control trials provide a benchmark for clear-cut research designs to compare to. This benchmark increases the standards for evaluations and potentially the quality of the evidence gathered.

The second part of this thesis applied especially comparisons over time and bounds to solve possible endogeneity problems in the evaluation of an innovative approach to ultra-poverty reduction, based on an integrative social protection scheme centred on an asset transfer. While the point estimates for comparisons over time give a first indication of the project effect, validity is only added to them by a bounding approach to narrow down the endogeneity left in the data. With sensible assumptions both informative and worst-case bounds were constructed even for variables with only one natural bound. A look beyond point estimates can be suggested as a viable strategy for impact evaluation, where endogeneity is pervasive.

The estimations in this part led to a general conclusion: The entire data material considered indicates that the estimation strategy generated internally valid estimates. This concerns the estimates as such, their height, significance and comparability over the estimation strategies. The findings are backed by analogous effects of measurement with summary indices, PCA based indices and value or expenditure related data. An overall conclusive picture for covariates and summary statistics adds to a good case for causality in the estimation as do first findings for further covariates and interaction terms. While measurement error could still be a concern and heterogeneity in impacts needs to be evaluated further, IFSUP has to a large extent proven its ability to support a sustainable livelihood for poverty strata previously excluded from development cooperation or addressed by short-term aid schemes.

But does this mean the evaluation has gathered knowledge beyond the single evaluation? As outlined in the third part of this thesis, the case for external validity is, beyond the estimation strategy, hinging on arguments, comparisons and theory. The selection of sites for the project and first evidence from other locations establish a good case for a general applicability of asset-based approaches to other contexts. The question remains open, however, whether these projects can work on a larger scale, without an intensive adaptation to local livelihoods. Additionally, long-term evidence on IFSUP is necessary, since the major justification for the high unit costs of the project is the argument of long-term stability, or even an upward trajectory, of ultra-poor livelihoods.

Credible data must prove this assumption, a fact the stakeholders of development cooperation industry need to give more attention to. Carefully planned evaluation designs are a challenging task and their costs can be high. However, as Duflo & Kremer (2005: 225) stress, evaluations are “far cheaper than pursuing ineffective policies.”

By a general drive for effective program evaluation not only donors but as well the implementing organizations on the ground can increase their impact on poverty reduction beyond the current project and potentially to a global scale.

Bibliography

- AHLIN, C., J. LIN, and M. MAIO (2011): "Where Does Microfinance Flourish? Microfinance Institution Performance in Macroeconomic Context," *Journal of Development Economics*, 95: 2, 105-120.
- AMIN, S., A. S. RAI, and G. TOPA (2003): "Does Microcredit Reach the Poor and Vulnerable? Evidence from Northern Bangladesh," *Journal of Development Economics*, 70: 1, 59-73.
- ANGRIST, J. D., and J.-S. PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- ANGRIST, J. D., and J.-S. PISCHKE (2010): "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics," *Journal of Economic Perspectives*, 24: 2, 3-30.
- ARMENDÁRIZ, B., and J. MORDUCH (2010): *The Economics of Microfinance*. Cambridge, Mass.: MIT Press.
- AULT, J. K., and A. SPICER (2009): "Does One Size Fit All in Microfinance? New Directions for Academic Research," in *Moving Beyond Storytelling: Emerging Research in Microfinance*, ed. by T. A. Watkins, and K. M. Hicks. Bingley: Emerald, 271-284.
- BAKER, J. L. (2000): *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Washington, D.C.: World Bank.
- BALKE, A., and J. PEARL (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92: 439, 1171-1176.
- BAMBERGER, M., V. RAO, and M. WOOLCOCK (2010): "Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development," in *Handbook of Mixed Methods in Social & Behavioral Research*, ed. by A. Tashakkori, and C. Teddlie. Thousand Oaks, Calif.: SAGE Publications, 613-643.
- BANERJEE, A. V. (2002): "The Uses of Economic Theory: Against a Purely Positive Interpretation of Theoretical Results," *BREAD Working Paper*, 7, Cambridge, MA: MIT, Bureau for Research in Economic Analysis of Development, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=315942.
- BANERJEE, A. V. (2006a): "Making Aid Work: How to Fight Global Poverty - Effectively," *Boston Review*, 31: 4, 7-9.
- BANERJEE, A. V. (2006b): "Making Aid Work: The Best Argument for the Experimental Approach Is That It Spurs Innovation," *Boston Review*, 31: 4, 19-19.
- BANERJEE, A. V., R. BANERJI, E. DUFLO, R. GLENNERSTER, and S. KHEMANI (2010a): "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," *American Economic Journal: Economic Policy*, 2 1, 1-30.
- BANERJEE, A. V., and E. DUFLO (2010): "Giving Credit Where It Is Due," *Journal of Economic Perspectives*, 24: 3, 61-79.
- BANERJEE, A. V., and E. DUFLO (2011): *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: PublicAffairs.
- BANERJEE, A. V., E. DUFLO, R. CHATTOPADHYAY, and J. SHAPIRO (2010b): "Targeting the Hardcore-Poor: An Impact Assessment," Consultative Group to Assist the Poorest, <http://www.cgap.org/p/site/c/template.rc/1.26.12411/>, accessed 25.07.2011, cited with permission of the authors.
- BANERJEE, A. V., E. DUFLO, R. GLENNERSTER, and C. KINNAN (2010c): "The Miracle of Microfinance? Evidence from a Randomized Evaluation," MIT Department of Economics, <http://econ-www.mit.edu/faculty/eduflo/papers>, accessed 16.03.2011.
- BANERJEE, A. V., E. DUFLO, and K. MUNSHI (2003): "The (Mis)Allocation of Capital," *Journal of the European Economic Association*, 1: 2/3, 484-494.

- BANERJEE, A. V., and S. MULLAINATHAN (2010): "The Shape of Temptation: Implications for the Economic Lives of the Poor," *NBER Working Papers*, 15973: National Bureau of Economic Research, Inc, <http://ideas.repec.org/p/nbr/nberwo/15973.html>, accessed 27.11.2010.
- BARRIENTOS, A. (2011): "Social Protection and Poverty," *International Journal of Social Welfare*, 20: 3, 240–249.
- BARRIENTOS, A., and D. HULME (2008): "Social Protection for the Poor and Poorest: An Introduction," in *Social Protection for the Poor and Poorest: Concepts, Policies and Politics*, ed. by A. Barrientos, and D. Hulme. Basingstoke ; New York: Palgrave Macmillan, 3-24.
- BARRIENTOS, A., M. NIÑO-ZARAZÚA, and M. MAITROT (2010): "Social Assistance in Developing Countries Database - Version 5.0," Manchester: Brooks World Poverty Institute, <http://ssrn.com/abstract=1672090>, accessed 10.03.2011.
- BASTIAENSEN, J., T. DE HERDT, and B. D'EXELLE (2005): "Poverty Reduction as a Local Institutional Process," *World Development*, 33: 6, 979-993.
- BBS (2009a): *Gender Statistics of Bangladesh 2008*. Dhaka: Bangladesh Bureau of Statistics.
- BBS (2009b): "Updating Poverty Maps of Bangladesh," Dhaka: Bangladesh Bureau of Statistics, World Bank and UN World Food Program, <http://www.bbs.gov.bd/WebTestApplication/userfiles/Image/UpdatingPovertyMapsofBangladesh.pdf>, accessed 04.05.2011.
- BBS (2011a): "GDP of Bangladesh at 2007-08 to 2010-11," Dhaka: Bangladesh Bureau of Statistics, http://www.bbs.gov.bd/WebTestApplication/userfiles/Image/BBS/GDP_2011.pdf, accessed 26.11.2011.
- BBS (2011b): "Millenium Development Goals: Bangladesh Progress at a Glance," Dhaka: Bangladesh Bureau of Statistics http://www.bbs.gov.bd/WebTestApplication/userfiles/Image/Latest%20Statistics%20Release/Millenium_Development_Goals.pdf, accessed 22.11.2011.
- BBS (2011c): "Preliminary Report on Household Income & Expenditure Survey 2010," Dhaka: Bangladesh Bureau of Statistics, <http://www.bbs.gov.bd/WebTestApplication/userfiles/Image/HIES/HIES-PR.pdf>, accessed 22.11.2011.
- BBS, and UNICEF (2007): "Bangladesh Multiple Indicator Cluster Survey 2006 - Final Report," Dhaka: Bangladesh Bureau of Statistics; Unicef, http://www.unicef.org/bangladesh/2006-08_MICS_2006.Vol_II.FinalJuly08.pdf, accessed 15.12.2011.
- BEHNKE, J., J. HINTERMAIER, and L. RUDOLPH (2010): "Die Bedeutung Von Werten Für Verteilungsergebnisse Im Ultimatum- Und Diktatorspiel," in *Jahrbuch Für Handlungs- Und Entscheidungstheorie Band 6: Schwerpunkt Neuere Entwicklungen Des Konzepts Der Rationalität Und Ihre Anwendungen*, ed. by J. Behnke, T. Bräuninger, and S. Shikano. Wiesbaden: VS Verlag für Sozialwissenschaften., 165-192.
- BEHRMAN, J. R., and J. HODDINOTT (2005): "Programme Evaluation with Unobserved Heterogeneity and Selective Implementation: The Mexican PROGRESA Impact on Child Nutrition," *Oxford Bulletin of Economics and Statistics*, 67: 4, 547-569.
- BENDAVID, E. (2011): "Those Who Fail to Learn from Medical History ..." *Boston Review*, 36: 2, 19-20.
- BERESTEANU, A., and C. F. MANSKI (2000): "Bounds for STATA - Draft Version 1.0," Chicago: Northwestern University, http://faculty.wcas.northwestern.edu/~cfm754/bounds_stata.pdf, accessed 09.09.2011.
- BERTRAND, M., E. DUFLO, and S. MULLAINATHAN (2004): "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, 119: 1, 249-275.

- BJÖRKMÄN, M., and J. SVENSSON (2009): "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda," *Quarterly Journal of Economics*, 124: 2, 735-769.
- BLUNDELL, R., and M. COSTA DIAS (2000): "Evaluation Methods for Non-Experimental Data," *Fiscal Studies*, 21: 4, 427-468.
- BMZ (2009): "Sector Strategy on Social Protection," *Strategies*, 190, Berlin: Federal Ministry for Economic Cooperation and Development, http://www.bmz.de/en/publications/topics/social_security/konzept190.pdf, accessed 07.01.2011.
- BOOYSEN, F., S. VAN DER BERG, R. BURGER, M. V. MALTITZ, and G. D. RAND (2008): "Using an Asset Index to Assess Trends in Poverty in Seven Sub-Saharan African Countries," *World Development*, 36: 6, 1113-1130.
- BRADHAN, P. (2011): "Experimental Fad," *Boston Review*, 36: 2, 20-21.
- BRAMBOR, T., W. R. CLARK, and M. GOLDBER (2006): "Understanding Interaction Models: Improving Empirical Analyses," *Political Analysis*, 14: 1, 63-82.
- BRAUN, J. P. (2010): "Study on Conflicts in Adivasi Villages," Wetzlar: NETZ Partnership for Development and Justice, <http://www.bangladesch.org/pics/download/Study-on-conflict-in-Adivasi-villages.pdf>, accessed 11.10.2011.
- CAMERON, A. C., and P. K. TRIVEDI (2009): *Microeconometrics Using Stata*. College Station, Tex.: Stata.
- CARTER, M. R. (2007): "Learning from Asset Based Approaches to Poverty," in *Reducing Global Poverty : The Case for Asset Accumulation*, ed. by C. O. N. Moser. Washington, D.C.: Brookings Institution Press, 51-61.
- CARTER, M. R., P. D. LITTLE, T. MOGUES, and W. NEGATU (2008): "Poverty Traps and Natural Disaster in Ethiopia and Honduras," in *Social Protection for the Poor and Poorest: Concepts, Policies and Politics*, ed. by A. Barrientos, and D. Hulme. Basingstoke ; New York: Palgrave Macmillan, 85-120.
- CARTWRIGHT, N. (2007): "Are RCTs the Gold Standard?," *BioSocieties*, 2: 01, 11-20.
- CATTELL, R. B. (1966): "The Scree Test for the Number of Factors," *Multivariate Behavioral Research*, 1: 2, 245-276.
- CHAMBERS, R. (2009): "So That the Poor Count More: Using Participatory Methods for Impact Evaluation," *Journal of Development Effectiveness*, 1: 3, 243.
- CHAMBERS, R., U. KIRDAR, and L. SILK (1995): "Poverty and Livelihoods: Whose Reality Counts?," 1-16.
- CHAUDRY, P. (2010): "Unconditional Cash Transfer to the Very Poor in Viet Nam: Is It Enough to 'Just Give Them Cash'?", in *What Works for the Poorest? Poverty Reduction Programmes for the World's Extreme Poor*, ed. by D. Lawson, D. Hulme, I. Matin, and K. Moore. Rugby, Warwickshire: Practical Action Publ., 169-179.
- CHOWDHURY, S., A. M. MOBARAK, and G. BRYAN (2009): "Migrating Away from a Seasonal Famine: A Randomized Intervention in Bangladesh," *Human Development Research Papers*, 2009/41: United Nations Development Programme, http://hdr.undp.org/en/reports/global/hdr2009/papers/HDRP_2009_41.pdf, accessed 10.09.10.
- COATES, J., P. WEBB, and R. HOUSER (2003): "Measuring Food Insecurity: Going Beyond Indicators of Income and Anthropometry," Washington DC: Food and Nutrition Technical Assistance Project, Academy for Educational Development, http://www.fantaproject.org/downloads/pdfs/foodinsecurity_bangladesh03.pdf, accessed 04.06.2011.
- COHEN, M., M. J. MCCORD, J. SEBSTAD (2005): "Reducing Vulnerability: Demand for and Supply of Microinsurance in East Africa", *Journal of International Development*, 17: 3, 319-325.

- COX, N. J. (2000): "matselrc: Stata Module for Selection of Rows and/or Columns from Matrix," Durham: University of Durham, <http://www.stata.com/users/vwiggins>, accessed 14.12.2011.
- CRUMP, R. K., O. K. MITNIK, G. IMBENS, and V. J. HOTZ (2008): "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90: 3, 389-405.
- DASGUPTA, P. (1997): "Nutritional Status, the Capacity for Work, and Poverty Traps," *Journal of Econometrics*, 77: 1, 5-37.
- DAVIDSON, R., and J. G. MACKINNON (2000): "Bootstrap Tests: How Many Bootstraps?," *Econometric Reviews*, 19: 1, 55-68.
- DEATON, A. (2006): "Making Aid Work: Evidence-Based Aid Must Not Become the Latest in a Long String of Development Fads," *Boston Review*, 31: 4, 13-14.
- DEATON, A. (2010a): "Instruments, Randomization, and Learning About Development," *Journal of Economic Literature*, 48: 2, 424-455.
- DEATON, A. (2010b): "Understanding the Mechanisms of Economic Development," *Journal of Economic Perspectives*, 24: 3, 3-16.
- DEVEREUX, S. (2001): "Sen's Entitlement Approach: Critiques and Counter-Critiques," *Oxford Development Studies*, 29: 3, 245.
- DEVEREUX, S. (2007): "The Impact of Droughts and Floods on Food Security and Policy Options to Alleviate Negative Effects," *Agricultural Economics*, 37: S1, 47-58.
- DIETZEL, P. (2003): "Hilfreich, praktisch, gut? Entwicklungsarbeit für die extrem arme Bevölkerung," *NETZ*, 2003: 3, 3-10.
- DIETZEL, P. (2006): "Ein Leben lang genug Reis," in *Solidarität die Ankommt! Ziel-Effiziente Mittelverwendung in der Entwicklungszusammenarbeit*, ed. by Global Marshall Plan Initiative. Hamburg: Global Marshall Plan Initiative, 219-249.
- DINARDO, J. (2006): "Constructive Proposals for Dealing with Attrition: An Empirical Example," *NBER Working Paper*, http://www.nber.org/public_html/confer/2006/si2006/ls/dinardo.pdf, accessed 07.08.2011.
- DINARDO, J. (2008): "Natural Experiments and Quasi-Natural Experiments," in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf, and L. E. Blume. Basingstoke: Palgrave Macmillan.
- DIPRETE, T. A., and M. GANGL (2004): "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments," *Sociological Methodology*, 34: 1, 271-310.
- DOWLING, J. M., and C.-F. YAP (2009): "Chronic Poverty in Asia: Causes, Consequences, and Policies," Singapore ; Hackensack, N.J.: World Scientific.
- DUFLO, E., R. GLENNERSTER, and M. KREMER (2008): "Using Randomization in Development Economics Research: A Toolkit," in *Handbook of Development Economics*, ed. by T. P. Schultz, and J. A. Strauss. Amsterdam/Oxford: Elsevier, 3895-3962.
- DUFLO, E., and M. KREMER (2005): "Use of Randomization in the Evaluation of Development Effectiveness," in *Evaluating Development Effectiveness*, ed. by G. K. Pitman, O. N. Feinstein, and G. K. Ingram. New Brunswick, N.J.: Transaction Publishers, 205-231.
- EASTERLY, W. R. (2006): *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. Oxford ; New York: Oxford University Press.
- EC (2001): *Country Strategy Paper - Bangladesh - 2002-2006*. Brussels: European Commission - Directorate General External Relations.
- ELLIS, F. (2000): *Rural Livelihood and Diversity in Developing Countries*. Oxford: Oxford University Press.

- ELLIS, F., S. DEVEREUX, and P. WHITE (2009): *Social Protection in Africa*. Cheltenham ; Northampton, MA: Edward Elgar.
- EMRAN, M. S., C. S. STEPHEN, and R. VIRGINIA (2009): "Assessing the Frontiers of Ultra-Poverty Reduction: Evidence from CFPR/TUP, an Innovative Program in Bangladesh," *IIEP Working Papers: The George Washington University, Institute for International Economic Policy*, <http://ideas.repec.org/p/gwi/wpaper/2009-06.html>, accessed 23.02.2011.
- ESCAP (2002): "Reducing Poverty and Promoting Social Protection," *Social policy paper*, 5, New York: United Nations Economic and Social Commission for Asia and the Pacific, <http://www.unescap.org/esid/psis/publications/spps/05/2168.pdf>, accessed 04.02.2011.
- FARRINGTON, J., R. SLATER, and R. HOLMES (2004): "Social Protection and Pro-Poor Agricultural Growth," *ODI Natural Resource Perspectives* London: Overseas Development Institute, <http://www.odi.org.uk/resources/docs/1664.pdf>, accessed 04.01.2011.
- FERGUSON, C., C. O. N. MOSER, and A. NORTON (2007): "Claiming Rights: Citizenship and the Politics of Asset Accumulation," in *Reducing Global Poverty: The Case for Asset Accumulation*, ed. by C. Moser. Washington, D.C.: Brookings Institution Press, 273-288.
- FILMER, D., and L. PRITCHETT (2001): "Estimating Wealth Effects without Expenditure Data - or Tears: An Application to Educational Enrollments in States of India," *Demography*, 38: 1, 115-132.
- FISMAN, R., and E. MIGUEL (2007): "Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets," *Journal of Political Economy*, 115: 6, 1020-1048.
- FOSTER, E. M., and L. BICKMAN (1996): "An Evaluator's Guide to Detecting Attrition Problems," *Evaluation Review*, 20: 6, 695-723.
- FRIEDMAN, M. (1957): *A Theory of the Consumption Function*. Princeton: Princeton University Press.
- GARIKIPATI, S. (2008): "The Impact of Lending to Women on Household Vulnerability and Women's Empowerment: Evidence from India," *World Development*, 36: 12, 2620-2642.
- GLENNERSTER, R., and M. KREMER (2011a): "Context Is Important, and Meticulous Experimentation Can Improve Our Understanding of It," *Boston Review*, 36: 2, 28-29.
- GLENNERSTER, R., and M. KREMER (2011b): "Small Changes Big Results," *Boston Review*, 36: 2, 12-17.
- GLYNN, A. N., and K. M. QUINN (2011): "Why Process Matters for Causal Inference," *Political Analysis*, 19: 3, 273-286.
- GOLDIN, I., F. H. ROGERS, and N. STERN (2006): "Making Aid Work: We Must Tackle Development Problems at the Level of the Economy as a Whole," *Boston Review*, 31: 4, 10-11.
- GREEN, D. P., T. Y. LEONG, H. L. KERN, A. S. GERBER, and C. W. LARIMER (2009): "Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks," *Political Analysis*, 17: 4, 400-417.
- GULLI, H. (1998): *Microfinance and Poverty : Questioning the Conventional Wisdom*. Washington, D.C.: Inter-American Development Bank.
- HAHN, J., P. TODD, and W. V. D. KLAUW (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69: 1, 201-209.
- HALDER, S. R., and P. MOSLEY (2004): "Working with the Ultra-Poor: Learning from BRAC Experiences," *Journal of International Development*, 16: 3, 387-406.

- HARPER, C., R. MARCUS, and K. MOORE (2003): "Enduring Poverty and the Conditions of Childhood: Lifecourse and Intergenerational Poverty Transmissions," *World Development*, 31: 3, 535-554.
- HASEEN, F. (2006): "Change in Food and Nutrient Consumption among the Ultra Poor: Is the CFPR/TUP Programme Making a Difference?," *CFPR-TUP Working Paper Series*, 11, <http://ideas.repec.org/p/ess/wpaper/id749.html>, accessed 27.11.2011.
- HASHEMI, S., and A. DE MONTESQUIOU (2011): "Reaching the Poorest: Lessons from the Graduation Model," *Focus Note*, 69: Consultative Group to Assist the Poor, <http://www.cgap.org/gm/document-1.9.50739/FN69.pdf>, accessed 27.11.2011.
- HASHEMI, S., and W. UMAIRA (2011): "New Pathways for the Poorest: The Graduation Model from BRAC," *CSP Research Report*, 10, Sussex: Institute for Development Studies, Centre for Social Protection, <http://socialprotectionasia.org/Conf-prgram-pdf/10-SPA-Final-Paper-No-10.pdf>, accessed 26.11.2011.
- HASHEMI, S., and A. DE MONTESQUIEU (2011): "Reaching the Poorest: Lessons from the Graduation Model," *CGAP Focus Note*, 69, Washington D.C.: Consultative Group to Assist the Poorest, <http://www.cgap.org/gm/document-1.9.50739/FN69.pdf>, accessed 26.11.2011.
- HATLØY, A., L. E. TORHEIM, and A. OSHAUG (1998): "Food Variety - a Good Indicator of Nutritional Adequacy of the Diet? A Case Study from an Urban Area in Mali, West Africa," *European Journal of Clinical Nutrition*, 52: 12, 891-898.
- HAUGHTON, D., and J. HAUGHTON (2011): *Living Standards Analytics: Development through the Lens of Household Survey Data*. New York: Springer.
- HAUSMAN, J. A., and D. A. WISE (1979): "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47: 2, 455-473.
- HECKMAN, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47: 1, 153-161.
- HECKMAN, J. J., and E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73: 3, 669-738.
- HODGSON, J. M., B. H. H. HSUHAGE, and M. L. WAHLQVIST (1994): "Food Variety as a Quantitative Descriptor of Food-Intake," *Ecology of Food and Nutrition*, 32: 3-4, 137-148.
- HOROWITZ, J. L., and C. F. MANSKI (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95: 449, 77-84.
- HOTZ, V. J., C. H. MULLIN, and S. G. SANDERS (1997): "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analysing the Effects of Teenage Childbearing," *The Review of Economic Studies*, 64: 4, 575-603.
- HOWE, L. D., J. R. HARGREAVES, and S. R. HUTTLY (2008): "Issues in the Construction of Wealth Indices for the Measurement of Socio-Economic Position in Low-Income Countries," *Emerg Themes Epidemiol*, 5: 1, 3-17.
- HSIANG-KE, C. (2003): "Milton Friedman and the Emergence of the Permanent Income Hypothesis," *History of Political Economy*, 35: 1, 77-104.
- HUDA, K. (2009): "Mid-Term (12 Month) Trickle up India TUP Process Evaluation - CGAP-Ford Foundation Graduation Pilot," New York: Trickle Up, <http://www.microfinancegateway.org/gm/document-1.9.41179/Final%20eval%20MAY252009%20formatted.pdf>, accessed 14.03.2010.
- HUDA, K. (2010): "Overcoming Extreme Poverty in India: Lessons Learnt from SKS," *IDS Bulletin*, 41: 4, 31-41.
- HULME, D., and D. LAWSON (2010): "What Works for the Poorest?," in *What Works for the Poorest? Poverty Reduction Programmes for the World's Extreme Poor*, ed. by D.

- Lawson, D. Hulme, I. Matin, and K. Moore. Rugby, Warwickshire: Practical Action Publ., 1-24.
- HULME, D., and K. MOORE (2008): "Assisting the Poorest in Bangladesh: Learning from Brac's 'Targeting the Ultra-Poor' Programme," in *Social Protection for the Poor and Poorest: Concepts, Policies and Politics*, ed. by A. Barrientos, and D. Hulme. Basingstoke ; New York: Palgrave Macmillan, 194-210.
- HULME, D., and K. MOORE (2010): "Assisting the Poorest in Bangladesh: Learning from Brac's 'Targeting the Ultra-Poor' Programme," in *What Works for the Poorest? Poverty Reduction Programmes for the World's Extreme Poor*, ed. by D. Lawson, D. Hulme, I. Matin, and K. Moore. Rugby, Warwickshire: Practical Action Publ., 149-168.
- HUMPHREYS, M., and J. M. WEINSTEIN (2009): "Field Experiments and the Political Economy of Development," *Annual Review of Political Science*, 12: 1, 367-378.
- IMBENS, G. W., and T. LEMIEUX (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142: 2, 615-635.
- IMBENS, G. W., and J. M. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47: 1, 5-86.
- JAHNKE, H. E. (1982): *Livestock Production Systems and Livestock Development in Tropical Africa*. Kiel: Wiss.-Verl. Vauk.
- JANN, B. (2005): "Making Regression Tables from Stored Estimates," *STATA Journal*, 5: 3, 288-308.
- JANN, B. (2007): "Making Regression Tables Simplified," *STATA Journal*, 7: 2, 227-244.
- JOLLIFFE, I. T. (2002): *Principal Component Analysis*. New York: Springer.
- KABEER, N. (2010): "Alternative Accounts of Chronic Disadvantage: Income Deficits Versus Food Security," in *What Works for the Poorest? Poverty Reduction Programmes for the World's Extreme Poor*, ed. by D. Lawson, D. Hulme, I. Matin, and K. Moore. Rugby, Warwickshire: Practical Action Publ., 59-78.
- KADIGI, R. M. J., N. S. Y. MDOE, and G. C. ASHIMOGO (2007): "Understanding Poverty through the Eyes of the Poor: The Case of Usangu Plains in Tanzania," *Physics & Chemistry of the Earth - Parts A/B/C*, 32: 15-18, 1330-1338.
- KAISER, H. F. (1960): "The Application of Electronic Computers to Factor Analysis," *Educational and Psychological Measurement*, 20: 1, 141-151.
- KAM, C. D., and R. J. J. FRANZESE (2007): *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor: The University of Michigan Press.
- KANBUR, R. (2001): "Economic Policy, Distribution and Poverty: The Nature of Disagreements," *World Development*, 29: 6, 1083-1094.
- KANBUR, R. (2010): "Foreword," in *What Works for the Poorest? Poverty Reduction Programmes for the World's Extreme Poor*, ed. by D. Lawson, D. Hulme, I. Matin, and K. Moore. Rugby, Warwickshire: Practical Action Publ., xv-xviii.
- KANBUR, R., and P. SHAFFER (2007): "Epistemology, Normative Theory and Poverty Analysis: Implications for Q-Squared in Practice," *World Development*, 35: 2, 183-196.
- KEZDI, G. (2003): "Robus Standard Error Estimation in Fixed-Effects Panel Models," Budapest: Budapest University of Economics, <http://129.3.20.41/eps/em/papers/0508/0508018.pdf>, accessed 30.12.2011.
- KHANDKER, S. R. (1998): *Fighting Poverty with Microcredit: Experience in Bangladesh*. New York: Oxford University Press.
- KHANDKER, S. R. (2005): "Microfinance and Poverty: Evidence Using Panel Data from Bangladesh," *World Bank Economic Review*, 19: 2, 263-286.
- KHANDKER, S. R., M. A. B. KHALILY, and H. A. SAMAD (2010a): "Seasonal and Extreme Poverty in Bangladesh: Evaluating an Ultra-Poor Microfinance Project," *Policy*

- Research Working Paper*, 5331: World Bank, <http://ssrn.com/paper=1620311>, accessed 29.08.2010.
- KHANDKER, S. R., G. B. KOOLWAL, and H. SAMAD (2010b): *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington, D.C.: World Bank.
- KLING, J. R., J. B. LIEBMAN, and L. F. KATZ (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75: 1, 83-119.
- KOHLER, U., and F. KREUTER (2008): *Datenanalyse mit Stata: Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung*. Munich: Oldenbourg.
- KOLENIKOV, S., and G. ANGELES (2004): "The Use of Discrete Data in Principal Component Analysis with Applications to Socio-Economic Indices," *CPC/MEASURE Working paper*, 85, Chapel Hill: University of North Carolina, <https://www.cpc.unc.edu/measure/publications/pdf/wp-04-85.pdf>, accessed 09.12.2011.
- KOLENIKOV, S., and G. ANGELES (2009): "Socioeconomic Status Measurement with Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer?," *Review of Income and Wealth*, 55: 1, 128-165.
- KONOLD, F., E. AHSAN, and B. NATH (2001): "Prospects and Limitations of Income Generating Activities (Igas) as an Instrument for Food Security for the Ultra-Poor - Volume I: Main Report," Dhaka: RESAL Bangladesh; GOPA unpublished.
- LALONDE, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 76: 4, 604-620.
- LAWSON, D. (2010): "A 'Q-Squared' Approach to Enhancing Our Understanding of the Chronically Poor," in *What Works for the Poorest? Poverty Reduction Programmes for the World's Extreme Poor*, ed. by D. Lawson, D. Hulme, I. Matin, and K. Moore. Rugby, Warwickshire: Practical Action Publ., 45-58.
- LECHNER, M. (1999): "Nonparametric Bounds on Employment and Income Effects of Continuous Vocational Training in East Germany," *Econometrics Journal*, 2: 1, 1-28.
- LEE, D. S. (2002): "Trimming for Bounds on Treatment Effects with Missing Outcomes," *NBER Technical Working Papers*, 277: National Bureau of Economic Research, <http://ideas.repec.org/p/nbr/nberte/0277.html>, accessed 22.11.2011.
- LEE, D. S. (2008): "Randomized Experiments from Non-Random Selection in U.S. House Elections," *Journal of Econometrics*, 142 2, 675-697.
- LEE, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 76: 3, 1071-1102.
- LEHMANN, H. (2006): "Eigene Sprache, eigene Schule - Die indigene Gemeinschaft der Oraon," *NETZ*, 2006: 4, 8-11.
- LEIBBRANDT, M., J. A. LEVINSOHN, and J. MCCRARY (2010): "Incomes in South Africa after the Fall of Apartheid," *Journal of Globalization and Development*, 1: 1, 1-60.
- LEMIEUX, T., and K. MILLIGAN (2008): "Incentive Effects of Social Assistance: A Regression Discontinuity Approach," *Journal of Econometrics*, 142: 2, 807-828.
- LI, X., C. GAN, and B. HU (2011): "The Welfare Impact of Microcredit on Rural Households in China," *Journal of Socio-Economics*, 40: 4, 404-411.
- MALLICK, D. (2009): "How Effective Is a Big Push to the Small? Evidence from a Quasi-Random Experiment," *MPRA Paper*, 22824, Munich: University Library of Munich, <http://ideas.repec.org/p/pramprapa/22824.html>.
- MANSKI, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80: 2, 319-323.
- MANSKI, C. F. (1997): "Monotone Treatment Response," *Econometrica*, 65: 6, 1311-1334.
- MANSKI, C. F. (2007): *Identification for Prediction and Decision*. Cambridge, Mass.: Harvard University Press.
- MATIN, I., and D. HULME (2003): "Programs for the Poorest: Learning from the IGVGD Program in Bangladesh," *World Development*, 31: 3, 647-666.

- MATIN, I., M. SULAIMAN, and M. RABBANI (2008): "Crafting a Graduation Pathway for the Ultra Poor: Lessons and Evidence from a BRAC Programme," *Chronic Poverty Research Centre Working Paper*, 109: BRAC Research and Evaluation Division, http://www.dfid.gov.uk/R4D//PDF/Outputs/ChronicPoverty_RC/109Matin_et_al.pdf, accessed 07.09.2011.
- MCKENZIE, D. (2005): "Measuring Inequality with Asset Indicators," *Journal of Population Economics*, 18: 2, 229-260.
- MCKERNAN, S.-M. (2002): "The Impact of Microcredit Programs on Self-Employment Profits: Do Noncredit Program Aspects Matter?," *The Review of Economics and Statistics*, 84: 1, 93-115.
- MEDEIROS, R. A., and D. BLANCHETTE (2011): "mdesc: Stata Module to Tabulate Prevalence of Missing Values," Boston: Boston College Department of Economics, <http://ideas.repec.org/c/boc/bocode/s457318.html>.
- MEENAKSHI, J. V., and R. RAY (2002): "Impact of Household Size and Family Composition on Poverty in Rural India," *Journal of Policy Modeling*, 24: 6, 539-559.
- MIAH, T. H. (2008): "Profitability and Risk Management of Income Generating Activities of Ultra-Poor – in Particular Adibasis under the Project of 'Income and Food Security for Ultra-Poor (IFSUP)'," Mymensingh: NETZ Partnership for Development and Justice, <http://www.netz-bangladesh.de/pics/download/IFSUP-Study-on-Profitability.pdf>, accessed 19.09.2010.
- MIGUEL, E., and M. KREMER (2004): "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72: 1, 159.
- MOORE, M. (2006): "Making Aid Work: The New Private Philanthropies Could Challenge the Existing Aid Business," *Boston Review*, 31: 4, 11-12.
- MORDUCH, J. (2000): "The Microfinance Schism," *World Development*, 28: 4, 617-629.
- MOSER, C. O. N. (2007): "Introduction," in *Reducing Global Poverty: The Case for Asset Accumulation*, ed. by C. O. N. Moser. Washington, D.C.: Brookings Institution Press, 1-14.
- MOSER, C. O. N., and A. A. DANI (2008): *Assets, Livelihoods, and Social Policy*. Washington, DC: World Bank.
- MOSER, C. O. N., and A. FELTON (2007): "The Construction of an Asset Index Measuring Asset Accumulation in Ecuador," *CPRC Working Paper*, 87, Washington DC: The Brookings Institution.
- MOYO, D. (2009): *Dead Aid: Why Aid Is Not Working and How There Is Another Way for Africa*. London ; New York: Allen Lane.
- MUNRO, L. T. (2008): "Risks, Needs and Rights: Compatible or Contradictory Bases for Social Protection," in *Social Protection for the Poor and Poorest: Concepts, Policies and Politics*, ed. by A. Barrientos, and D. Hulme. Basingstoke ; New York: Palgrave Macmillan, 27-46.
- MURGAI, R., and S. ZAIDI (2004): "Poverty Trends in Bangladesh During the Nineties," *SASPR Working Paper Series*, 2B: Poverty Reduction and Economic Management Unit, South Asia Region, The World Bank, http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2004/12/06/000012009_20041206122937/Rendered/PDF/308630PAPER0SA1Trends0in0Bangladesh.pdf, accessed 15.12.2011.
- NANDA, V. P. (2006): "The 'Good Governance' Concept Revisited", *Annals of the American Academy of Political and Social Science*, 603: 1, 269-283.
- NETZ (2006): "Food Security 2003 Call for Proposals - Income and Food Security for Ultra-Poor Project Application," *EuropeAid/123246/C/ACT/BD*, Wetzlar: NETZ Partnership for Development and Justice, unpublished.

- NETZ (2009a): "Annual Report 2009," Wetzlar; Dhaka: NETZ Partnership for Development and Justice, http://www.bangladesch.org/set.php?id=english&uid=annual_report, accessed 16.11.2011.
- NETZ (2009b): "Out of the Black Hole of Poverty - Lessons Learnt from the Project Income and Food Security for Ultra-Poor (IFSUP)," Wetzlar; Dhaka: NETZ Partnership for Development and Justice, <http://www.bangladesch.org/pics/download/IFSUP-Lessons-Learnt.pdf>, accessed 07.09.2010.
- NETZ (2010): "Final Narrative Report," Dhaka: NETZ Partnership for Development and Justice, unpublished.
- OECD-DAC (2010): "Evaluating Development Co-Operation: Summary of Key Norms and Standards," Paris: Organisation for Economic Co-operation and Development, <http://www.oecd.org/dataoecd/12/56/41612905.pdf>, accessed 30.07.2011.
- OECD (2005): "Paris Declaration on Aid Effectiveness – Ownership, Harmonisation, Alignment, Results and Mutual Accountability," *2nd High-Level Forum on Aid Effectiveness*, 2nd High-Level Forum on Aid Effectiveness, Paris, France.
- OFFENHEISER, R. C., AND D. JACOBS (2006): "Making Aid Work: The Global Poverty Challenge is Political as well as Technological," *Boston Review*, 31: 4, 17-18.
- PATTENGALE, N. D., M. ALIPOUR, O. R. BININDA-EMONDS, B. M. MORET, and A. STAMATAKIS (2010): "How Many Bootstrap Replicates Are Necessary?," *J Comput Biol*, 17: 3, 337-354.
- PEARL, J. (2009): *Causality : Models, Reasoning, and Inference*. Cambridge ; New York: Cambridge University Press.
- PEDROSA, J., and Q.-T. DO (2011): "Geographic Distance and Credit Market Access in Niger," *African Development Review*, 23: 3, 289-299.
- PÉREZ-ESCAMILLA, R., and A. M. SEGALL-CORRÊA (2008): "Food Insecurity Measurement and Indicators," *Revista de Nutrição*, 21, 15-26.
- PERRIN, B. (2011): "What Is a Results/Performance-Based Delivery System? An Invited Presentation to the European Parliament," *Evaluation*, 17: 4, 417-424.
- PITT, M. M., and S. R. KHANDKER (1998): "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?," *Journal of Political Economy*, 106: 5, 958-996.
- RADHAKRISHNA, R., K. HANUMANTHA RAO, C. RAVI, and B. SAMBI REDDY (2006): "Extreme and Chronic Poverty and Malnutrition in India: Incidence and Determinants," in *Chronic Poverty and Development Policy in India*, ed. by A. K. Mehta, and A. Shepherd. New Delhi ; Thousand Oaks, Calif.: Sage Publications, 148-167.
- RAHMAN, P. M. M., N. MATSUI, and Y. IKEMOTO (2009): *The Chronically Poor in Rural Bangladesh: Livelihood Constraints and Capabilities*. London: Routledge.
- RAVALLION, M. (2008): "Evaluating Anti-Poverty Programs," in *Handbook of Development Economics*, ed. by T. P. Schultz, and J. A. Strauss. Amsterdam/Oxford: Elsevier, 3787-3846.
- RAVALLION, M. (2009): "Should the Randomistas Rule?," *Economists' Voice*, 6: 2.
- RAYHAN, M. I., and U. GROTE (2010): "Crop Diversification to Mitigate Flood Vulnerability in Bangladesh: An Economic Approach " *Economics Bulletin*, 30: AccessEcon, <http://www.accessecon.com/Pubs/EB/2010/Volume30/EB-10-V30-I1-P54.pdf>, accessed 21.12.2011.
- RIETBERGEN-MCCRACKEN, J., and D. NARAYAN-PARKER (1998): *Participation and Social Assessment : Tools and Techniques*. Washington, D.C.: International Bank for Reconstruction and Development/World Bank.
- ROBINSON, G., J. E. McNULTY, and J. S. KRASNO (2009): "Observing the Counterfactual? The Search for Political Experiments in Nature," *Political Analysis*, 17: 4, 341–357.

- ROODMAN, D., and J. MORDUCH (2009): "The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence," *CGD Working Paper*, 174, Washington D.C.: Center for Global Development, http://www.cgdev.org/files/1422302_file_Roodman_Morduch_Bangladesh.pdf, accessed 28.08.2010.
- ROSENBAUM, P. R. (1995): *Observational Studies*. New York: Springer.
- ROSENBAUM, P. R., and D. B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70: 1, 41-55.
- RUBIN, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66: 5, 688-701.
- RUDOLPH, L. (2010): *Subsidizing Asset Transfer Towards the Ultra-Poor - Evidence from Bangladesh*. Munich: Ludwig-Maximilians-University, Faculty of Economics, Bachelor Thesis.
- RUDOLPH, L. (2011): "Increasing Food Security by Asset Transfer - Evidence from a Pro-Ultra-Poor-Intervention in Bangladesh," *First International Conference on International Relations and Development*, First International Conference on International Relations and Development, 19.-20.05.2011, Bangkok, Thailand: Thammasat University, http://www.icird.org/2011/files/Papers/ICIRD2011_Lukas%20Rudolph.pdf, accessed 01.06.2011.
- SABATES-WHEELER, R., and S. DEVEREUX (2008): "Transformative Social Protection: The Currency of Social Justice," in *Social Protection for the Poor and Poorest: Concepts, Policies and Politics*, ed. by A. Barrientos, and D. Hulme. Basingstoke ; New York: Palgrave Macmillan, 64-84.
- SCHNECKENER, U. (2004): *States at Risk: Fragile Staaten als Sicherheits- und Entwicklungsproblem*. Berlin: SWP.
- SCHUBERT, B. (2008): "Protecting the Poorest with Cash Transfers in Low Income Countries," in *Social Protection for the Poor and Poorest: Concepts, Policies and Politics*, ed. by A. Barrientos, and D. Hulme. Basingstoke ; New York: Palgrave Macmillan, 211-223.
- SEGERS, K., J. DESSEIN, P. DEVELTERE, S. HAGBERG, G. HAYLEMARIAM, M. HAILE, and J. DECKERS (2010): "The Role of Farmers and Informal Institutions in Microcredit Programs in Tigray, Northern Ethiopia," *Perspectives on Global Development & Technology*, 9: 3/4, 520-544.
- SEKHON, J. S. (2009): "Opiates for the Matches: Matching Methods for Causal Inference," *Annual Review of Political Science*, 12: 1, 487-508.
- SEMYKINA, A., and J. M. WOOLDRIDGE (2010): "Estimating Panel Data Models in the Presence of Endogeneity and Selection," *Journal of Econometrics*, 157: 2, 375-380.
- SEN, A. (1981): *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford ; New York: Oxford University Press/Clarendon Press.
- SEN, B. (2003): "Drivers of Escape and Descent: Changing Household Fortunes in Rural Bangladesh," *World Development*, 31: 3, 513-534.
- SHAFFER, P. (2002): "Poverty Naturalized: Implications for Gender," *Feminist Economics*, 8: 3, 55-75.
- SIMS, C. A. (2010): "But Economics Is Not an Experimental Science," *Journal of Economic Perspectives*, 24: 2, 59-68.
- SINHA, S., M. LIPTON, and S. YAQUB (2002): "Poverty and "Damaging Fluctuations": How Do They Relate?," *Journal of Asian & African Studies*, 37: 2, 186-243.
- SMITH, J. A., and P. E. TODD (2005): "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125: 1-2, 305-353.
- STOCK, J. H., and M. W. WATSON (2007): *Introduction to Econometrics*. Boston, Mass. ; London: Pearson/Addison Wesley.
- STOCK, J. H., and M. W. WATSON (2008): "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression," *Econometrica*, 76: 1, 155-174.

- SULAIMAN, M., and I. MATIN (2006): "Using Change Rankings to Understand Poverty Dynamics: Examining the Impact of CFPR/TUP from Community Perspective," *CFPR/TUP Working Paper Series*, 14, <http://ideas.repec.org/p/ess/wpaper/id651.html>, accessed 20.11.2011.
- SUPPES, P. (1970): *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Co.
- TEDESCHI, G. A., and D. S. KARLAN (2010): "Cross-Sectional Impact Analysis: Bias from Dropouts," *Perspectives on Global Development and Technology*, 9, 270-291.
- THISTLETHWAITE, D. L., and D. T. CAMPBELL (1960): "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51: 6, 309-317.
- TIB (2008): "National Household Survey 2007 on Corruption in Bangladesh," Dhaka: Transparency International Bangladesh, <http://www.ti-bangladesh.org/research/HHSurvey07full180608.pdf>, accessed 08.01.2011.
- TODD, P. E. (2008): "Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated," in *Handbook of Development Economics*, ed. by T. P. Schultz, and J. A. Strauss. Amsterdam/Oxford: Elsevier, 3847-3894.
- TOWNSEND, R. M. (1994): "Risk and Insurance in Village India," *Econometrica*, 62: 3, 539-591.
- UDO, H., C. HERMANS, and F. DAWOOD (1992): "Seasonality of Cattle Feed Sources in Pabna, Bangladesh," *Tropical Animal Health and Production*, 24: 1, 50-56.
- UN (2000): "United Nations Millenium Declaration," *Resolutions adopted by the General Assembly*, A/RES/55/2, New York: United Nations General Assembly.
- UN (2011): "World Statistics Pocketbook 2010: Least Developed Countries," New York: United Nations, UN-OHRLLS, <http://www.unohrlls.org/UserFiles/File/LDC%20Pocketbook2010-%20final.pdf>, accessed 26.11.2011.
- VÁSQUEZ, I. (2006): "Making Aid Work: There Is a Huge Gap between the Expert Consensus and the Political Push for More Aid," *Boston Review*, 31: 4, 12-12.
- VYAS, S., and L. KUMARANAYAKE (2006): "Constructing Socio-Economic Status Indices: How to Use Principal Components Analysis," *Health Policy and Planning*, 21: 6, 459-468.
- WATKINS, T. A., and K. M. HICKS (2009): "Moving Beyond Storytelling: Emerging Research in Microfinance," Bingley: Emerald.
- WHITE, H. (2006): "Making Aid Work: Technical Rigor Must Not Take Precedence over Other Kinds of Valuable Lessons," *Boston Review*, 31: 4, 16-17.
- WHITE, H. (2009): "Theory-Based Impact Evaluation: Principles and Practice," *Journal of Development Effectiveness*, 1: 3, 271.
- WIGGINS, V. (without year): "grc1leg: Graph Combine One Legend," <http://www.stata.com/users/vwiggins>, accessed 14.12.2011.
- WOOD, G. (2003): "Staying Secure, Staying Poor: The 'Faustian Bargain,'" *World Development*, 31: 3, 455-471.
- WOOLDRIDGE, J. M. (2009): *Introductory Econometrics: A Modern Approach*. Mason, OH: South Western, Cengage Learning.
- WOOLDRIDGE, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass.: MIT Press.
- WOOLDRIDGE, J. M., and G. IMBENS (2007): "What's New in Econometrics? Session 8: Cluster and Stratified Sampling," *Sumer Institute Mini Course, July 31 - August 1*, Cambridge, MA, : National Bureau of Economic Research, http://www.nber.org/WNE/lect_8_cluster.pdf, accessed 27.12.2011.
- WORLD BANK (2000): *Balancing Protection and Opportunity : A Strategy for Social Protection in Transition Economies*. Washington, DC: World Bank.

- WORLD BANK (2006): "Table 4.14: Exchange Rates and Prices," World Bank - World Development Indicators, http://devdata.worldbank.org/wdi2006/contents/Table4_14.htm, accessed 26.09.2010.
- WORLD BANK (2010): "Bangladesh Country Assistance Strategy: 2011-2014," Dhaka: Bangladesh Country Management Unit, World Bank, <http://siteresources.worldbank.org/BANGLADESHEXTN/Resources/295759-1271081222839/6958908-1284576442742/BDCASFinal.pdf>, accessed 23.09.2010.
- YUNUS, M. (2006): "Nobel Lecture," *The Nobel Peace Prize 2006* Stockholm, http://nobelprize.org/nobel_prizes/peace/laureates/2006/yunus-lecture.html, accessed 13.10.2010.
- ZUG, S. (2006a): "Monga - Seasonal Food Insecurity in Bangladesh - Bringing the Information Together," *The Journal of Social Studies*, 111: 4, 21-39.
- ZUG, S. (2006b): "Monga - Seasonal Food Insecurity in Bangladesh - Understanding the Problem and Strategies to Combat It " Sundarganj, Dhaka, Bochum: NETZ Partnership for Development and Justice, http://www.bangladesch.org/pics/download/Final_Report_Monga_Sebastian_Zug.pdf, accessed 13.04.2010.
- ZWANE, A. P., J. ZINMAN, E. VAN DUSEN, W. PARIENTE, C. NULL, E. MIGUEL, M. KREMER, D. S. KARLAN, R. HORNBECK, X. GINÉ, E. DUFLO, F. DEVOTO, B. CREPON, and A. V. BANERJEE (2011): "Being Surveyed Can Change Later Behavior and Related Parameter Estimates," *Proceedings of the National Academy of Sciences*, 108: 5, 1821-1826.

Appendix 1: The Point of Departure in IFSUP Analysis

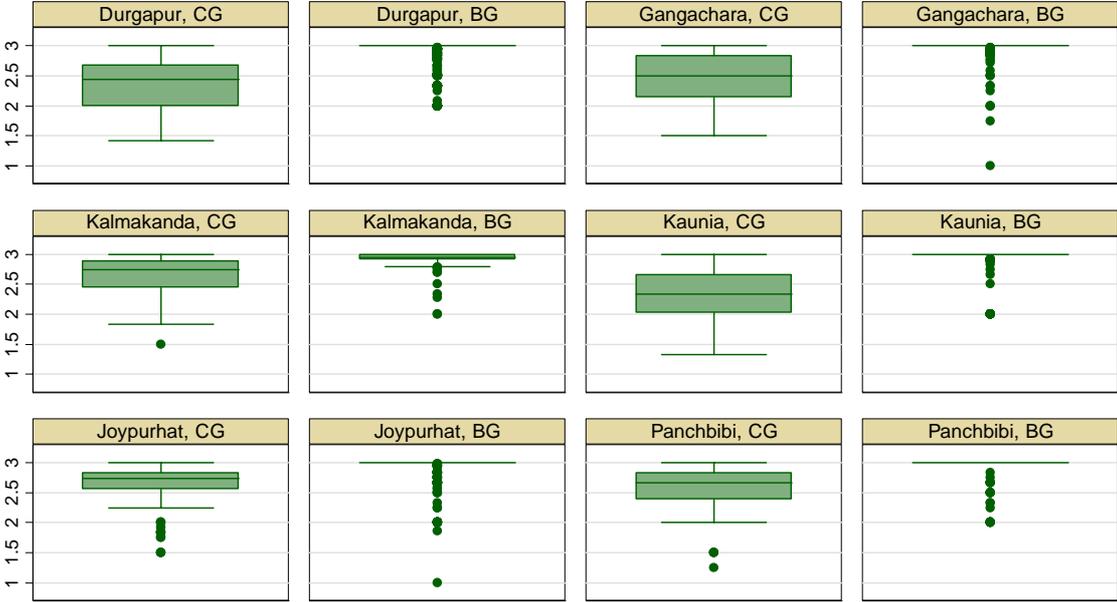
The impact analysis in this thesis is an extension of a first evaluation in Rudolph (2010: Chapter 6, 2011). Rudolph (2010: Chapter 6) assumed that an overall indication of project success should be given by the impact of the project on overall food security – first evidence using a single-difference equation on the impact data-set and treating the project as quasi-randomized experiment with random attrition indicated positive impacts of about 0.44 average meals per day. Using baseline and impact data and thus a difference-in-differences approach with two cross-sections, Rudolph (2011: 15) equally estimated an impact of about 0.45 average meals per day.¹⁸⁹ These results are estimated under the assumption of quasi-randomized project implementation and random attrition, although the difference-in-differences approach allowed for time-invariant unobserved heterogeneity. Both approaches controlled for observable characteristic that could have influenced project placement, namely district of the intervention, household size, ethnicity, female headedness of households, education of household heads, wage income and coverage by other GO/NGO programs. The results are significant on the 1% level and robust to the in- or exclusion of these controls. While food security measurement through recalls on meal intake is used in the literature (comp. e.g. Banerjee et al. 2010b: Table 9), the censoring of the data complicates the interpretation of the analysis: Figure 17 summarizes the distribution of average meals per day for control group and beneficiary group ultra-poor on *upazila* level, as used as outcome variable in Rudolph (2010) for single difference impact measurement. Both control group and beneficiary group distributions are very similar across *upazilas*: While control group distributions are centring on an average 2.3 to 2.8 meals per day, the IFSUP intervention seemed to be successful at lifting about 85% of the beneficiary group ultra-poor to food security with three meals per day throughout the year in all *upazilas*.

But the natural upper bound of the indicator prohibits insights into the real quantitative difference between both groups (this is the question whether project participants would be in a position to afford even four, five or six meals per day). Additionally, heterogeneity in impacts cannot be assessed with this indicator, as wealth differences among various subgroups are hidden behind the censoring at three meals per day and the low variability of the impact indicator for target group members. Analysis in Rudolph (2011: Table 3, 4) revealed only a first indication for a reduced project effect for target group participants with *Adibasi* ethnicity.

¹⁸⁹ This similarity in baseline distributions is the reason for the quantitative similarity of the results in Rudolph (2010) and Rudolph (2011) and was interpreted as indication for as-if-randomization at baseline.

Figure 17: Censoring and low variability in the indicator meals per day thus far used for analysis

Box plots for average meals per day on upazila level



Own graph
 Box plot distributions of average meals per day per average family member on upazila level for control group (CG) and beneficiary group (BG) of IFSUP.

Foremost, further analysis is necessary to counter concerns that the positive impact is related to hidden bias not evident in the data so far, both related to attrition and endogeneity in program placement. Secondly, unbounded indicators can give additional indications on the size of the effect of IFSUP. Third, theory suggested the project approach as viable option for ultra-poverty reduction. But only a research design investigating a range of aspects relevant for ultra-poverty livelihoods is able to comprehensively link the evidence back to theory.

For analysis beyond the scope of this thesis, impact variables with higher variability additionally allow for estimations of heterogeneity in impact.

Appendix 2: The IFSUP Project Intervention

The IFSUP project¹⁹⁰ aims at generating sustainable livelihoods for 4800 ultra-poor households with approximately 19000 family members in six *upazilas*¹⁹¹ of three Bangladeshi districts¹⁹² (Rangpur, Joypurhat and Netrakona), intending to lift about 10% of the ultra-poor population in these *upazilas* over the upper poverty line of 2122 kcal per day per capita. In each district, a different PNGO¹⁹³ of NETZ is responsible for the implementation. *Upazilas* with high incidence of ultra-poverty were selected. The project explicitly targets the senior women of households¹⁹⁴ and pays attention to the inclusion of local ethnic minorities, termed *Adibasis*, who are to make up at least 40% of the project beneficiaries. The ultra-poor are disaggregated into subgroups – group I, II and III. The main distinction between these groups lies in their capacity to work. Group I, the most marginalized ultra-poor, not embedded in supporting family structures and not capable of physical labor (e.g. elderly, physically or mentally ill, orphans), are not targeted by the project, as they cannot respond to self-employment schemes. Group II and III differ mainly by their degree of poverty. In the project, this distinction is operationalized by defining group II as households with a food intake of less than 1600 kcal/day operationalized as a maximum of two meals in minimum nine months, a monthly income of less than 400 taka per capita, being completely without land and/or assets. Group III is characterized by a food intake of less than 1800 kcal/day operationalized as a maximum of two meals in minimum six months, a monthly income of less than 500 taka per capita and being functionally landless.

Skepticisms prevail concerning self-employment schemes especially for the group II ultra-poor: The European Commission sees income generating activities (IGAs) by group II possible only “[s]poradic by highly motivated and capable beneficiaries” (EC 2001: Annex 5)¹⁹⁵, following a study highlighting exclusion and self-exclusion of ultra-poor households

¹⁹⁰ This appendix draws to a large part on Rudolph (2010: Chapter 2.1).

¹⁹¹ *Upazila* is an administrative unit in Bangladesh comparable to a county.

¹⁹² IFSUP is implemented in the context of one of the least developed countries of the world, and within Bangladesh the most deprived part of the country, its northern and north-western Region. Appendix 4 provides a map and Appendix 3 describes this context in detail.

¹⁹³ These are ‘ASHRAI’ in Rangpur, ‘Jagorani Chakra Foundation (JCF)’ in Joypurhat and ‘Sabalamby Unnayan Samity (SUS)’ in Netrakona.

¹⁹⁴ Implicitly, the approach assumes that women are more responsible in caring for social outcomes. This is in line with a large part of the literature. Behrman & Hoddinott (2005: 552) e.g. summarize how studies on “poor populations have concluded that larger shares of resources that go to mothers are directed towards child health and nutrition than of resources directed to fathers”.

¹⁹⁵ Although the direct transfer of assets seems not to have been taken into account as program design feature.

leading to low entrepreneurial prospects requiring intensive up front training and motivation (Konold et al. 2001: Chapter 7.3).¹⁹⁶

The IFSUP approach still centres on this approach to ultra-poverty reduction. Project-instruments include the boosting of self-employment through asset transfers to a level preventing hunger and distress. At the same time the project aims to reduce their vulnerability to an extent that a future fallback into poverty becomes unlikely. This approach is based on the identification of four major problems these ultra-poor are facing: They are suffering from hunger at least 6-9 months per year. Their state is vulnerable to an extent that they are at high risk of being locked in their situation and would slide back even if faced with a positive shock. To the social and political aspects of their vulnerability – social exclusion is assumed especially for female-headed and indigenous households, especially as faced with a government largely unable to include their needs to its agenda – adds the risk of various natural disasters and unfavorable macroeconomic conditions, as well as market forces.¹⁹⁷ This leaves many households incapable to invest in human or physical capital and prevents emancipation from their status: Thus, Rahman et al. (2009: Chapter 1.2) find in their assessment of chronic poverty in Bangladesh that a large majority of the chronically poor (72%) inherited their status over generations and are likely to pass it on (comp. for a detailed overview on reasons Harper et al. 2003). Sinha et al. (2002: 186), using the concept of ‘damaging fluctuations’¹⁹⁸, similarly note how this vulnerability is likely to influence the ultra-poor:

“Such fluctuations cause immediate damage and may trigger responses leading to chronic poverty or intergenerational continuities of poverty. Moreover, just the possibility of [damaging fluctuations], even if they do not occur, may generate more risk-averse behavior, which hampers growth and (because it is likely to be more common among poorer people) increases inequality.”

These poor are thus expected to be trapped in poverty, as noted in similar contexts (Carter et al. 2008).

¹⁹⁶ As the authors subsequently conclude the “promotion of IGAs represent a valuable instrument in food security for the ultra-poor within determined types of [...] interventions [but] are not an appropriate approach for all ultra-poor. Our impression is that no more than 50% of the ultra-poor are able [...] to run successful and sustainable IGAs” (Konold et al. 2001: 37).

¹⁹⁷ Much noted is especially their inability to get access to credit or savings facilities (Armendáriz & Morduch 2010, Banerjee et al. 2003, Rudolph 2010).

¹⁹⁸ The commonly used concepts of ‘risks’ and/or ‘shocks’ are subsets of ‘damaging fluctuations’, as the latter include as well largely or wholly predictable changes, e.g. seasons, the (ultra-) poor are vulnerable to. Additionally, the authors argue, the concept of ‘damaging fluctuations’ brings to light four central dimensions of shocks usually unnoted for in the literature: the source, the stress it imposes on systems, the resulting strain on persons, and the resulting damage.

Appendix 3: Bangladesh and Its Northern Regions¹⁹⁹

The IFSUP project intervenes in the context of one of the least developed countries (UN 2011) of the world, with a per capita income of US\$ 755²⁰⁰ for 2010/11 (BBS 2011a: 7).

Still, Bangladesh experienced considerable economic and social progress in the last decade. Supported by steady GDP growth rates around 6.0%, poverty levels²⁰¹ declined from 57% in 1990 to 31.5% in 2010 and main health and education indicators improved: Life expectancy is up from 54.4 years in 1990 to projected 69 years for females and 66.5 years for males for the 2010-2015 period. Infant mortality per 1000 births declined from 92 in 1990 to projected 27.5 for the 2010-2015 period. Primary education indicators improved to almost universal gross enrolment with 98% in 2008, while closing the significant gender gap that had existed. While being among the most densely populated areas in the world with 964 inhabitants per square kilometer and population growth remaining a concern, it is still down from about 2% in the 1990s to 1.3% in 2011.

While relative figures might indicate improvements, overall poverty levels are still high. Convincing progress is lacking for a range of other variables as well: Secondary education levels need to be improved, with the primary and secondary gross enrolment ratio being at 65.5 for females and 60.8 for males. Additionally, the quality of education remains poor, as indicated by primary drop out rates of 48%. As the World Bank (2010: 7) notes, “[l]earning outcomes are generally poor, especially among low income and disadvantaged groups.” Though down from 574 death per 100000 live births, the reduction in maternal mortality is high in regional comparative and stalling, with 320 deaths in 2001 and 348 in 2008. A major problem is as well malnourishment: Taking weight as well as height for age as indicators leads to estimates of malnourishment prevalence for about 46-48% of children. The poorest are especially affected: “Chronic malnutrition is pervasive across all socioeconomic strata, at 56 % of all children among the poorest and 32 % among the wealthiest quintiles” (World Bank 2010: 6). Amongst the 31.5% of the population still living in poverty, about 17.6% are considered as being ultra-poor²⁰², with a large rural-urban divide of 21.1% for rural and 7.6% for urban areas. They receive few aid by the government: Although social protection themes

¹⁹⁹ This appendix draws on data published by the United Nations Statistics Division, the World Bank and the Bangladesh Bureau of Statistics (BBS 2011b, 2011c: 11, 2011d: 2f., United Nations 2011: 5, World Bank 2010a: 62, 2010b: 5-7). It builds in part on the description in Rudolph (2010: Chapter 2.2) but is updated and extended.

²⁰⁰ The figure is at current prices and preliminary.

²⁰¹ As defined by a direct calorie intake (DCI) of less than 2,122 kcal/day.

²⁰² As defined by a DCI of less than 1,805 kcal/day

are in place, “coverage is low, targeting is weak and government’s planning and delivery capacity needs significant strengthening” (World Bank 2010: 38). Additionally, large parts of the population live close to the poverty line and are highly vulnerable to being impoverished by small shocks. In this respect, the susceptibility of Bangladesh to natural disasters with its dense population and 80% of its area consisting of flood plains is especially noteworthy.

Overall, economic progress is not reaching the bottom section of society as fast as it should and inequality is stalling even according to government figures as the BBS (2011c: 3) reports:

“Incomes accruing to households belonging to Decile-1 to Decile-5 are recorded at 2.00 percent, 3.22 percent, 4.10 percent, 5.00 percent and 6.01 percent respectively at national level in 2010. These shares have not changed relative to 2005. These five deciles continue to share only 20.33 percent of total income, although they comprise 50 percent of the population.”

Rural northern and northwestern Bangladesh, where the project is implemented²⁰³, is an especially deprived part of the country: Politically neglected, prone by natural disasters and recurrent seasonal food crisis. For the latest region-specific data from 2005, the poverty head-count ratio is at about 50% compared to the then 40% nationwide, with ultra-poverty at 43% compared to 25% nationwide (BBS 2009b: 2, 2011b: 2, Chowdhury et al. 2009: 1, FN 1).²⁰⁴ The region is prone by natural disasters, with river floods being a concern especially for the project areas in Rangpur along the Teesta River and flash floods for the project region in Ne-trakona (BBS 2009b: 10). The special political problems of the region become evident in the targeting failures of GOB safety net programs: The poverty incidence of the population not targeted by these government programs is e.g. in Rangpur at 37.6% compared to 27.5% nationwide (BBS 2011c: 5). The weak industrialization in the northwest of the country leads to lacking off-farm employment in the lean seasons which in turn drives a seasonal pattern of chronic food shortages, called *monga*, prevalent especially in the Rangpur and to a lesser extent in the Joypurhat area of the project (Zug 2006b). *Monga* is leading to a vicious cycle of second-best coping strategies and increased vulnerability for the next hunger season (Zug 2006a: 30-32, 2006b: Chapter 27). Wood (2003: 468) in this respect refers to a “Faustian bargain” where the poor and poorest, faced with life threatening distress, are forced to discount the future in favor of short term security “whatever the longer term costs”. IFSUP is intended to overturn this context.

²⁰³ See Appendix 4 for a map.

²⁰⁴ The deprivation of the north-western region is recognized by other donors besides NETZ as well: The World Bank (2010: 61) e.g. notes, that government efforts to improve completion rates for primary education failed – rates remained “very low” – in the northwestern region and especially in the local *monga*-prone areas.

Appendix 4: IFSUP Working Area in Bangladesh

Figure 18: Map of Bangladesh with Target Districts (Grey) and Upazilas of the IFSUP working area



Picture source: NETZ (2009b: 2).

Appendix 5: PCA Analysis

Index construction for socio-economic status (SES), secondary outcomes (SOS), productive and durable assets as well as housing quality in this thesis draws on principal component analysis (PCA) (see for a general overview Jolliffe 2002). This approach is based on the idea that the ownership of different asset types is correlated with each other based on some underlying latent variable. This latent variable is unobservable, but “manifests itself through ownership of the different assets” (Moser & Felton 2007: 3). Concerning socio-economic status this unobserved variable can be interpreted as general wealth level of households and extracted with PCA from various indicators – from education over housing quality to asset ownership. As Moser (2007: 3f.) outlines, PCA is technically similar to regression analysis, just that “residuals are measured against all of the variables”. PCA is directed at maximizing the variance of the different possible linear combinations by the variables involved, to which the solution is given by the eigenvalues and eigenvectors of the covariance matrix of these variables (Kolenikov & Angeles 2009: 133). It is especially appealing that interpretation is straightforward: The higher the loading of a variable, the more information this variable contains about the occurrence of other variables, in positive or negative direction depending on the coefficient sign. The intuition behind the approach then is to find the, at best single, sometimes few components “that underlie all of the structure in (covariance of) the data” (Kolenikov & Angeles 2009: 134) and aggregate their information to an index.

There is a clear distinction in approaches between choosing several as compared to choosing only the first principle component for index construction.

When choosing several components to be used for the explanation of variance in the sample, the researcher has to choose between different possibilities: One way is the inclusion of principal components so that their combined explained variance reaches more than a cut-off point of about 70-90%. The concrete cut-off is flexible, based on the interest in the variance structure remaining and the impracticability of including too many principal components (Jolliffe 2002: 112f.). Very popular throughout the literature is the usage of the ‘Kaiser criterion’, which would imply the usage of all principal components with an eigenvalue greater than one – intuitively, this leads to a usage of principal components which explain more than only their own variance.²⁰⁵ Another popular method, despite being criticized for its subjec-

²⁰⁵ Interesting enough, Kaiser (1960: 145) himself speaks only of a reliability criterion: “[F]or a principal component analysis to have positive Kuder-Richardson reliability, it is necessary and sufficient that the associated eigenvalue be greater than one”. What he cites as prime reason for the Kaiser criterion should make the re-

tiveness, is the analysis of scree-plots, following the proposal of Cattell (1966). In this, plots of eigenvalue against component numbers are examined, choosing the component as cut-off where the graph “becomes fairly constant for several subsequent values” (Jolliffe 2002: 117). Jolliffe (2002: 117-130) discusses further methods of choosing the optimal number of principal components, although “it still remains true that attempts to construct rules having more sound statistical foundations seem, at present, to offer little advantage over the simpler rules in most circumstances” (Jolliffe 2002: 133).

Recently, and especially for the construction of SES and asset indices, only the first principle component is used (as e.g. proposed by Filmer & Pritchett 2001, McKenzie 2005, Vyas & Kumaranayake 2006), although “[i]t is also possible to use the sum of a number of eigenvectors, based on some criteria” (Moser & Felton 2007: 3, FN 4). But as has been shown, the first component can plausibly be assumed to be a measure of economic wealth, with further components adding substructures to this interpretation (Vyas & Kumaranayake 2006: 463). As outlined in the literature, results from indices based on the first principle component are strikingly robust to the inclusion of more dimensions (see the tests conducted by e.g. Filmer & Pritchett 2001, McKenzie 2005). This thesis will follow this approach.

A second point of discussion concerns the inclusion of discrete, including ordinal, variables in PCA. This is technically not straightforward as PCA is classically assuming multivariate normal data input (Kolenikov & Angeles 2009: 134). As Jolliffe (2002: 134) stresses, “[h]owever, the basic objective of PCA – to summarize most of the ‘variation’ that is present in the original set of p variables using a smaller number of derived variables – can be achieved regardless of the nature of the original variables.” Much debate surrounds the best practice for the inclusion of discrete data. While Filmer & Pritchett (2001) advocated the splitting up of ordinal data into binary variables and Booyesen et al. (2008: 1114) the use of “multiple correspondence analysis”, more recently polychoric PCA as developed by Kolenikov & Angeles (2004) emerged as method of choice (Moser & Felton 2007: 5).

But importantly, as Kolenikov & Angeles (2009: 161f.) advice and test in Monte Carlo simulations, using discrete data with PCA is a feasible approach and due to computational power concerns even preferable in many instances:

„The gain [of polychoric PCA] is only related to more accurate estimation of the proportion of explained variance that other methods tend to underestimate [...] If there is a reliable and well established ordering of categories, the ordinal PCA should be used. However, if the proportion of explained variance is of importance, the polychoric method should be used.”

searcher cautious of its application: The “most important viewpoint [is that eigenvalues greater than one lead in practice to] the number of factors which practicing psychologists were able to interpret.”

This approach was subsequently adopted for the construction of asset indices for the IFSUP data. Variables used for the respective indices are presented in Table 11, columns one to five, and range from the maximum education level in households, the quality of construction material for houses and the amount of land owned by households to dummies indicating the ownership of productive and non-productive assets. Variables were measured and included on household level, not adjusting for family sizes, as benefits of these variables (e.g. construction material, durable assets owned) are available on household level as well (comp. the similar argument in Filmer & Pritchett 2001: 120). The measured index thus has a “focus on inequality across households” (McKenzie 2005: 233).

Straightforward criteria for the in- or exclusion of variables to indices have not been found. As Vyas & Kumaranayake (2006: 467) summarize “studies using asset-based indices appear to have relied on the ‘face validity’ of the variables included, i.e. they appear to capture household wealth.” The construction in this thesis including as many variables as present in the data indicative for asset ownership and quality, though thorough tests on the robustness of the indices to the in- or exclusion of variables are beyond the scope of this thesis. Subsequently, the case for project effectiveness is not exclusively build upon these indices, but on a combined interpretation with other variables, in order to increase the validity of conclusions.

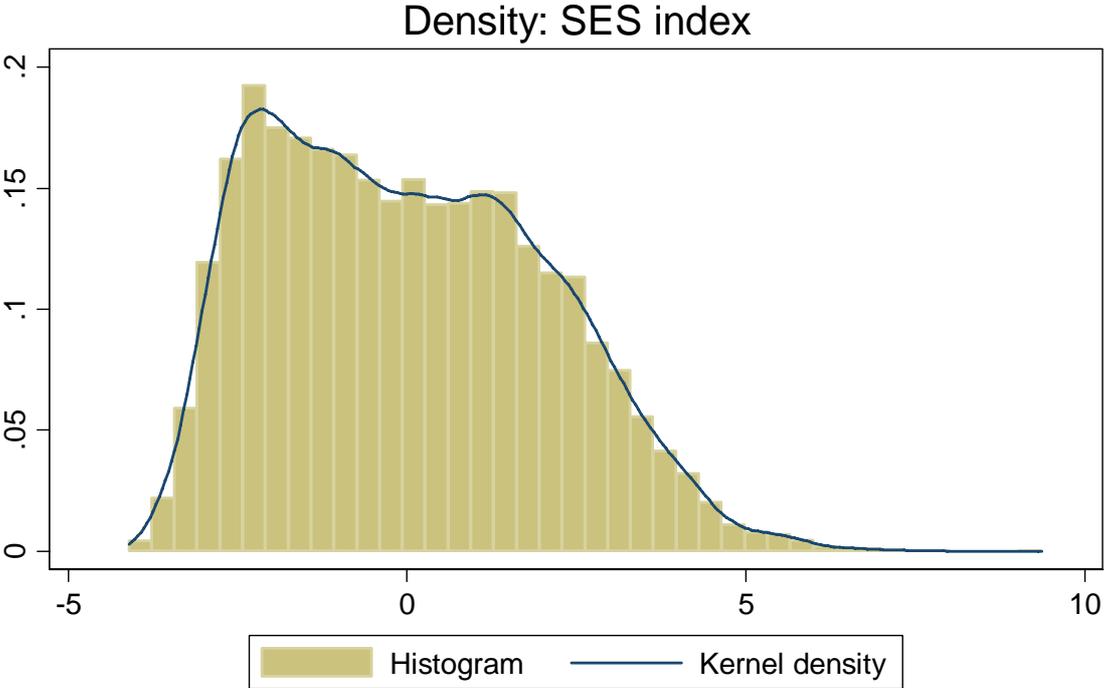
As proposed by McKenzie (2005: 234), in order to be able to compare over time and communities, the calculation of factor scores was computed in a pooled data set for both cross-sections, “since it is likely that the weights on the asset variables which most explain variation among households in the current cross-section may differ from the weights which explained the cross-sectional variation 10 [3] years before.” For the purpose of comparison, certain sets of variables had to be aggregated (chairs and tables were e.g. separately reported in baseline data, but aggregated in the impact survey) and coded as dummy variables indicating ownership. As dummy variables are the commonly used data input for PCA asset index construction (comp. e.g. the approach of Filmer & Pritchett 2001, McKenzie 2005), this is unproblematic besides the data on quantities lost in the process.

In principle, the recommendations of Vyas & Kumaranayake (2006) were followed in index construction. Crucial question is whether the indices following the first component are really a measurement of wealth (for the SES), wealth besides direct project outcomes (for the SOS) and respective asset endowments (for asset and housing quality indices). For this, the factor loadings presented in Table 11 are an important indicator: The positive association of

factor scores with the indices provides a first reassurance. The negative association of ‘further interior’ is plausible as well, given it is indicative of owning low quality furniture substituting the presence of especially chairs/tables. Owning ‘further interior’ can then be correlated to other variables indicative of a lower SES/SOS. Additionally, the eigenvalue of the first component, at least for the SES and SOS, as well as the variance associated is in line with indices constructed in similar conditions (comp. the summary of Vyas & Kumaranayake 2006: 463).

Last but not least, clumping and truncation as potential confounders (McKenzie 2005: 235) seem not to be of concern at least for SES, SOS, and housing quality, as a detailed look at histograms and densities of the estimates reveals. As presented for the distribution of scores for the SES index in Figure 19, scores are smoothly distributed, giving no hint to clumping and truncation and indicating that the variables chosen suffice for adequately distinguishing the wealth of households. The two maxima evident in the data are due to a positive shift in the distribution between 2007 and 2009. This pattern is similar with the secondary outcomes and housing quality indices, although the latter reveals various spikes. For productive and durable assets indices, truncation and clumping could be problematic, which is mainly due to the few variables included (detailed results available on request). Their results therefore have to be interpreted cautiously and only in complementing other asset indicators.

Figure 19: Distribution of scores for the SES index



Own graph
 The figure displays the histogram and kernel density of the distribution of scores constructed from the first principle component of the variables presented below for both cross-sections and all households.

Table 11: Factor loadings and descriptive statistics for variables involved in PCA index construction

	factor loadings for first component of PCA analysis					descriptive statistics					
	(1) SES	(2) secondary outcomes	(3) housing quality	(4) product. assets	(5) durable assets	2007 mean	sd	2009 mean	sd	total mean	sd
highest education level in HH	0.191	0.318				4.013	1.834	4.593	1.722	4.296	1.804
size of homestead (decimal)	0.155	0.276	0.473			4.352	3.285	4.918	3.671	4.628	3.490
ownership homestead	0.204	0.297	0.435			0.190	0.393	0.576	0.494	0.378	0.485
number of rooms	0.240	0.316	0.526			1.118	0.323	1.370	0.572	1.241	0.478
ownership house	0.084	0.130	0.253			0.944	0.231	0.973	0.163	0.958	0.201
quality of roof material	0.063	0.060	0.224			4.898	2.067	5.232	1.801	5.061	1.949
quality of wall material	0.110	0.212	0.443			2.910	1.552	3.293	1.670	3.097	1.622
dietary diversity score	0.302	0.438				4.300	2.262	7.167	1.158	5.698	2.308
productive land size (decimal)	0.220					1.574	7.344	14.466	31.095	7.860	23.251
tropical livestock units	0.302					0.283	0.450	0.874	0.697	0.571	0.654
index of diversification of in- come sources	0.404					2.415	1.199	4.286	1.191	3.328	1.518
ownership rickshaw/van	0.187			0.456		0.026	0.160	0.190	0.392	0.106	0.308
ownership trees/bamboo/ vege- table garden	0.364			0.582		0.278	0.448	0.852	0.355	0.558	0.497
ownership net	0.184			0.449		0.029	0.168	0.156	0.363	0.091	0.288
ownership boat	0.044			0.193		0.002	0.041	0.007	0.084	0.004	0.065
ownership tubewell	0.166			0.457		0.222	0.416	0.382	0.486	0.300	0.458
ownership husking machine	0.023			0.075		0.007	0.086	0.004	0.062	0.006	0.075
ownership electronics	0.108	0.162			0.362	0.072	0.258	0.087	0.282	0.079	0.270
ownership jewels	0.289	0.398			0.649	0.086	0.280	0.507	0.500	0.291	0.454
ownership chair/table	0.310	0.439			0.639	0.321	0.467	0.869	0.337	0.589	0.492
ownership further interior	-0.024	-0.031			0.199	0.801	0.399	0.497	0.500	0.653	0.476
eigenvalue of first component	4.203	2.394	1.396	1.443	1.436						
Variance associated	0.200	0.200	0.233	0.241	0.359						
N	11649	11649	11663	11785	11785	5969		5680		11649	

Explained proportions are approximations for SES, SOS and housing quality due discrete variables included (Kolenikov & Angeles 2009: 161). For these discrete variables (education, roof/wall quality, diversification index) mean/sd are not directly interpretable. Other variables besides TLU, land/homestead-size and rooms are dummies.

Appendix 6: Further Analysis of Internal Validity

Below, further evidence on the validity of conclusions derived and causality of the impact is presented, on the one hand concerning interactions with coverage of GOB programs, on the other hand concerning the inclusion of additional controls for education levels, outside wage labour income and education of household heads to the impact estimation.

These analysis are driven by concerns for additional bias in the data: One possible confounder could be related to the fact that IFSUP could have lead to a greater coverage of beneficiary households by GOB programs, as lobbying for inclusion of ultra-poor in GOB programs was explicit aim of the transformative project component. Outcomes could then be due not to the combined IFSUP approach as such (with its costly asset transfer component), but only to its transformative pillar. In this respect, Table 12 provides SD estimates, Table 13 DD estimates and Table 14 within estimates for the impact regression on overall outcomes with additional controls for the coverage by government programs and the interaction of beneficiary group status with this coverage (for SD and within estimates respectively the interaction of beneficiary status with GOB coverage and year 2009).

The inclusion of these interaction terms serves the purpose of singling out the impact of IFSUP apart from influences accruing from a possible increased coverage by GOB programs. As indicated in the tables below, the conclusions of the main part of this thesis hold: The variable support by NGOs/GOB²⁰⁶ is generally positive but not significant for most outcome measures.²⁰⁷ The interaction between NGO/GOB coverage and year for the within and DD estimates is, besides SES, negatively related to estimates, but again not significant for most variables²⁰⁸.

Most important for the effect of IFSUP are the findings for the interaction term beneficiary group with NGO/GOB support and, for DD and within data, year 2009: For these estimates, the general picture is again a negative relationship.²⁰⁹ This is to say that project effects are in tendency diminished (though insignificant for most variables²¹⁰) for the IFSUP households additionally included in GOB schemes.

²⁰⁶ The variable 'support by NGOs/GOB' is constructed as dummy variable indicating for baseline and impact data the coverage of households by any kind of NGO/GOB program besides IFSUP.

²⁰⁷ Support as such is significantly positively related only to SES for the within and DD estimation and meals per day and pc expenditure for the DD estimates

²⁰⁸ Support in 2009 is significantly negatively related only for SES and SOS for the DD and meals per day for the within estimates.

²⁰⁹ Besides food expenditure for the SD estimate and household expenditure for the within estimate.

²¹⁰ Besides SES and SOS for the within and value of collected food for SD estimates.

At the same time, the positive estimates for the project as measured by the inclusion of households in the beneficiary group is overall positive and remains significant for all variables but yearly pc expenditure (within estimates) and expenditure figures (DD estimates).²¹¹

The estimates for beneficiary ultra-poor additionally covered by NGO/GOB-programs are indicative in another aspect: As the transformative project elements aimed at local government-lobbying with the purpose of rightful coverage of GOB schemes, it is likely that the poorer strata among the beneficiary ultra-poor were included in GOB schemes. While they additionally benefit from GOB programs, their lower status as such serves to explain the negative sign of estimates.

Lastly, Table 15 presents findings for additional covariates in the restricted panel data set, again for the variables related to the overall project impact. The general conclusions still hold when controlling the impact of IFSUP for government support, income by outside wage labour, maximum education levels of household heads and household size. Impacts are still positive and highly significant on the 1% level (but for pc expenditure, significant on the 5% level). Additionally the impact size is comparable, though in general slightly lower (but for pc expenditure) to estimates in the reduced controls-model as estimated in Table 4.

The correlation of these additional covariates is generally conclusive as well: Outside wage labour is positively associated with outcomes (but for SOS) and significant for education related variables. Higher education levels to the reference category 'illiterate' are in general positively related to outcomes, though especially for expenditure estimates inconclusive. But clear-cut findings were not expected in this case as the absolute number of ultra-poor with higher education levels is very small.²¹² Additionally, as in the SD and DD estimations of Chapter 3.3, household size is positively related to outcomes besides pc expenditure.

As the within estimates control for changes in control variables for the same individuals, the conclusion can be drawn that IFSUP beneficiaries profited from the program even beyond changes in education level, household size and government support and even if outside wage labour was substituted away from.²¹³

²¹¹ Robust regressions instead of clustered regressions lead to overall significant positive impacts. Point estimates are by this not affected; therefore the general conclusions above can only be strengthened if robust standard errors were superior to clustered standard errors (comp. the discussion in 3.2.5).

²¹² Absolute number for maximum education of households are as follows: illiterate 289; can sign 1092; can read and write 48, class 1-2 passed 91; class 3-5 passed 221; class 6-8 passed 136; class 9-10 passed 53; secondary school certificate obtained 8; higher school certificate obtained 3.

²¹³ At least concerning the individuals in the reduced panel data set.

Table 12: Difference estimation for the overall project impact conditional on GOB/NGO support of BG by year

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	average meals per day per capita	average food expenditure per capita per day	value of col- lected food pc per day	socioeconomic status index (PCA)	secondary out- comes index (PCA)	yearly house- hold expendi- ture	yearly expendi- ture per capita
HH part of benefici- ary group	0.440 ^{***} (6.78)	0.798 ^{**} (2.68)	0.405 ^{***} (4.91)	1.447 ^{***} (15.08)	0.689 ^{**} (7.91)	4921.2 ^{**} (2.77)	1359.7 ^{**} (3.14)
support of BG by NGOs/GOB	-0.109 (-1.63)	0.0281 (0.16)	-0.141 ^{**} (-3.00)	-0.132 (-1.31)	-0.123 (-1.72)	-447.7 (-0.28)	-185.3 (-0.46)
support by NGOs/GOB	0.113 (1.83)	0.116 (0.58)	0.157 (1.89)	0.313 (1.73)	0.177 (1.42)	715.2 (0.50)	144.9 (0.41)
female headed HH	-0.0423 ^{**} (-2.61)	-0.270 ^{**} (-2.98)	-0.0254 (-0.22)	-0.416 ^{***} (-7.69)	-0.245 ^{***} (-9.68)	-1801.4 [*] (-2.15)	-148.9 (-0.58)
<i>Adibasi</i> ethnicity	-0.0310 (-1.23)	0.482 ^{***} (11.76)	0.289 ^{**} (2.91)	0.421 ^{***} (6.23)	0.341 ^{***} (4.44)	1210.3 ^{**} (3.83)	229.3 (1.03)
group II poverty status	-0.0186 ^{***} (-7.32)	-0.147 (-1.17)	-0.0925 (-1.32)	-0.391 ^{**} (-2.62)	-0.341 ^{**} (-3.13)	-1015.5 (-1.17)	-253.6 (-1.27)
number of HH members	0.0150 [*] (2.19)	0.906 ^{***} (27.12)	0.181 ^{***} (7.57)	0.320 ^{***} (6.96)	0.236 ^{***} (8.41)	3991.5 ^{***} (21.97)	-857.1 ^{***} (-10.56)
Constant	2.414 ^{***} (39.47)	1.589 ^{***} (6.04)	0.637 ^{**} (3.04)	-1.159 ^{***} (-5.10)	-0.279 ^{**} (-2.58)	13097.9 ^{***} (10.81)	10856.1 ^{***} (31.93)
<i>upazila</i> dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R ²	0.374	0.490	0.171	0.342	0.298	0.352	0.237
N	5210	5362	5369	5273	5273	2218	2218

Single difference estimation of the outcome variable in the heading on project participation and the indicated controls. Regression includes *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Difference-in-Differences estimation for the overall project impact conditional on GOB/NGO support of BG by year

	(1) average mpd per av. family member	(2) socioeconomic status index (PCA)	(3) secondary outcomes index (PCA)	(4) yearly household expenditure	(5) yearly expenditure per capita
project impact	0.338*** (13.22)	1.505*** (10.06)	0.683* (3.33)	5888.1 (2.39)	2132.8 (2.47)
HH part of beneficiary group	-0.00557 (-0.48)	-0.172 (-2.31)	-0.0211 (-0.49)	-180.0 (-0.37)	-640.9 (-2.17)
year 2009	0.633** (15.22)	1.306** (5.43)	1.131*** (14.45)	7170.5 (1.97)	1034.8 (1.03)
support of BG by NGOs/GOB in 2009	-0.0578 (-2.04)	-0.150* (-3.30)	-0.173 (-2.76)	-808.3 (-0.33)	-262.9 (-0.47)
support by NGOs/GOB in 2009	-0.0777 (-1.31)	-0.463* (-2.98)	-0.116** (-5.27)	-1062.8 (-0.49)	-354.2 (-0.73)
support by NGOs/GOB	0.180*** (42.57)	0.676* (3.63)	0.340 (2.07)	1705.0 (2.36)	487.9*** (11.55)
Constant	2.002*** (77.51)	-1.824*** (-20.79)	-1.119*** (-12.49)	6768.2*** (24.69)	8660.3*** (10.90)
fixed HH characteristics (ethnicity, group II poverty, HH size, female headed)	Yes	Yes	Yes	Yes	Yes
<i>upazila</i> dummies	Yes	Yes	Yes	Yes	Yes
R ²	0.824	0.610	0.575	0.703	0.425
N	5816	5874	5874	4249	4249
mean 2009	2.900	1.358	0.851	33774.7	7680.7

Difference-in-differences estimation of outcome variable in heading on project participation and the indicated controls. Impact estimation only for *upazilas* Panchibi, Joypurhat and Kalmakanda. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 14: Within estimation for the overall project impact conditional on GOB/NGO support of BG by year

	(1) average meals per day per average fam- ily member	(2) socioeconomic status index (PCA)	(3) secondary outcomes index (PCA)	(4) yearly household expenditure	(5) yearly expenditure per capita
project impact	0.433*** (11.31)	2.273*** (9.74)	1.013*** (12.49)	6592.0** (2.67)	926.4 (1.90)
year 2009	0.643*** (14.28)	1.326*** (4.28)	1.432*** (6.73)	10793.8*** (4.54)	2049.3*** (4.83)
support by the NGOs/GOB for BG in 2009	-0.0374 (-0.44)	-0.671** (-2.66)	-0.337* (-2.30)	1168.2 (0.52)	-95.53 (-0.12)
support by the NGOs/GOB in 2009	-0.205** (-3.08)	0.379 (0.72)	-0.0431 (-0.15)	-1475.1 (-0.23)	554.2 (0.34)
support by the NGOs/GOB	0.182* (2.34)	0.205 (0.57)	0.347 (1.25)	909.1 (0.14)	-335.0 (-0.23)
Constant	1.922*** (69.35)	-1.601*** (-12.96)	-1.160*** (-10.12)	17372.8*** (68.12)	5133.5*** (51.65)
N	1904	1926	1926	1389	1389
R ²	0.888	0.778	0.734	0.750	0.402
mean 2009	2.967	1.863	1.257	35731.9	8029.5

Within estimation of the outcome variable in the heading on project participation for the restricted panel data set with household fixed effects. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 15: Within estimation with additional control in the panel data set on overall outcome related variables

	(1) av. mpd per av. HH member	(2) SES (PCA)	(3) SOS (PCA)	(4) yearly HH expenditure	(5) yearly expenditure pc
project impact	0.394*** (6.32)	1.717*** (16.23)	0.634*** (18.29)	5537.8*** (4.18)	1590.5*** (3.93)
year 2009	0.644*** (17.85)	1.507*** (6.10)	1.476*** (7.46)	11680.5*** (6.59)	2429.6*** (10.26)
support by the NGOs/GOB	-0.0370 (-0.65)	0.0812 (1.09)	0.0742 (1.35)	-659.1 (-0.69)	-30.93 (-0.10)
total outside wage labor income	0.000000769 (0.45)	0.00000447 (0.78)	-0.00000292 (-0.89)	0.280*** (13.89)	0.0706*** (5.39)
education 2: signature only;	0.0293 (0.90)	0.272* (2.28)	0.256** (2.70)	567.6 (1.03)	-201.7 (-0.74)
education 3: basic reading/writing skills;	0.0277 (0.49)	0.158 (0.70)	0.114 (0.68)	5847.8** (3.18)	1332.9* (2.24)
education 4: class 1-2 passed;	0.0119 (0.32)	0.343* (2.56)	0.450** (3.06)	-873.7 (-0.58)	-716.1 (-1.49)
education 5: class 3-5 passed	-0.00110 (-0.02)	0.725** (3.32)	0.654** (3.62)	819.1 (0.52)	-746.9 (-1.06)
education 6: class 6-8 passed	0.0537 (1.20)	0.728** (3.55)	0.830*** (6.32)	1913.3 (1.11)	307.8 (0.39)
education 7: class 9-10 passed	0.0830 (0.84)	0.993** (3.28)	1.192*** (4.16)	-1037.3 (-0.60)	-1028.5 (-0.93)
education 8: secondary school certificate obtained;	-0.0663 (-0.48)	0.752 (1.47)	0.918** (2.78)	180.2 (0.11)	-967.8 (-1.04)
education 9: higher school certificate obtained	0.106 (0.90)	-0.253 (-0.57)	0.292 (0.52)	-9143.0*** (-4.12)	-3219.7** (-3.19)
HH size	0.0225* (2.46)	0.321*** (4.98)	0.269*** (13.88)	1605.9*** (5.00)	-1390.9*** (-6.15)
Constant	1.811*** (49.52)	-3.146*** (-13.25)	-2.397*** (-16.02)	6349.7*** (4.17)	9590.3*** (11.76)
N	1892	1916	1916	1385	1385
R ²	0.887	0.800	0.765	0.817	0.598
mean 2009	2.967	1.863	1.257	35731.9	8029.5

Within estimation of the outcome variable in the heading on project participation for the restricted panel data set with household fixed effects. Regressions include *upazila*-level fixed effects. In parentheses t-statistics from *upazila*-cluster-robust standard errors. Reference category for HH heads max. educ. is 'illiterate'. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix 7: Variable Definitions and Construction Procedure²¹⁴

Household Characteristics and Regression Controls

<i>Adibasi</i> ethnicity	Household self-classified as belonging to <i>Adibasi</i> ethnicity, an umbrella term for the indigenous population of Bangladesh (and as well parts of India) with many individual sub-groups and tribes.
Adult equivalent HH size	Adult equivalents give an indicator for demographic composition of households. Aggregation follows Townsend (1994: 554, FN 512): “for adult males, 1.0; for adult females, 0.9. For males and females aged 13-18, 0.94, and 0.83, respectively; for children aged 7-12, 0.67 regardless of gender; for children 4-6, 0.52; for toddlers 1-3, 0.32; and for infants 0.05.” As specific age information is available only in the panel data set, adult equivalent household size can be used only for analysis of baseline data.
Female headed HH	Household classified as female-headed, i.e. without primary male earner (widowed, divorced, deserted or with disabled husband)
Group II Poverty Status	Classification of a HH as belonging to group II is defined by NETZ as HH with a food intake of less than 1600 kcal/day, operationalized as a maximum of two meals in minimum nine months, a monthly income of less than 400 BDT per capita and being completely without land and/or assets. In comparison, group III is characterized by a food intake of less than 1,800 kcal/day, operationalized by a maximum of two meals in minimum six months, a monthly income of less than 500 BDT per capita and being functionally landless (<0.01 acres).
Highest education level	Highest level of education reported in the household ranging from one to nine (1: illiterate; 2: signature only; 3: basic reading/writing skills; 4: class 1-2 passed; 5: class 3-5 passed; 6: class 6-8 passed; 7: class 9-10 passed; 8: secondary school certificate obtained; 9: higher school certificate obtained).
Household /family size	Size of the household. A household is for IFSUP defined as the sum of individuals eating from the same cooking.
Maximum education level of household heads/spouse	The maximum of reported education levels of household heads or their spouse ranging from one to nine (values as above).

²¹⁴ Variables are alphabetically ordered within their respective categories as introduced in Table 1. Construction procedure is outlined where appropriate. Expenditure figures between the years are not corrected for inflation given the insecurity on the adequate points of reference (indicative for the problem is e.g. Figure 2 of Chapter 2.2, though only reporting national rice prices). Generally, ‘missing’ was treated as zero for some variables in the impact survey, as zero was partially not separately coded by interviewers.

Upazila Household found living in the *upazila* in question (Durgapur, Gangachara, Kalmakanda, Kaunia, Joypurhat and Panchibi); an *upazila* is the fourth level of administration in Bangladesh (administration levels are division – district – *upazila* – union – *mouza*) comparable to a county.

Overall Outcome Indicators

Average meals per day per average family member	For the baseline, households reported the amount of months with one, two or three meals over the last twelve months from the point of interviewing. For the impact survey, households reported these figures by male, female and underage family members – these figures were averaged over household members and aggregated for comparison.
Collected food value pd	Self-reported value of the household’s food (variables rice, pulse, vegetable, fruit, meat, fish, oil, egg, fuel) collected in nature or by begging, unsystematically reported for the last month, week or day, recalculated to daily figures (only available in the impact data set).
Food expenditure pd	Self-reported value of household’s food (see above for variables) expenditure unsystematically reported for the last month, week or day, recalculated to daily figures (only available in the impact data set).
Socio-economic status (PCA)	The SES-index is primarily based on the first principal component of physical assets (productive assets, durable consumer goods and housing) and human capital (maximum education levels). Comp. Appendix 5: PCA Analysis for a detailed description on construction and variables included.
Secondary outcomes (PCA)	The status of secondary outcomes (SOS) is based on variables not explicitly influenced by IFSUP, primarily the first principle component of physical non-productive assets (housing, durable consumer goods) and human capital (maximum education levels). Comp. Appendix 5: PCA Analysis for a detailed description on construction and variables included.
Yearly expenditure	Expenditure in the respective areas (food, health, clothing, housing, education, others) is reported yearly expenditure of households partly calculated from more detailed accounts (e.g. health treatment cost), partly reported as such. For the baseline, only an aggregate figures is available.
Yearly expenditure pc	Per capita yearly expenditure is the family size average of household’s yearly expenditure.

Asset Outcome Variables

Durable assets (PCA)	The durable asset index is based on the first principle component of HH ownership indicators for the variables chair/bench/table/bedstead, TV/mobile, jewellery and safe (for baseline additionally radio, dress-stands, blankets, clocks and others, besides the mentioned TV/mobile and safe). Partially, variables were grouped to asset complexes for comparability between data-sets. Comp. Appendix 5: PCA Analysis for a detailed description on construction and variables included.
Durable assets value	Sum of reported value of all (as above) non-productive assets in taka.
Livestock value	Sum of the self-reported value of livestock (subcategories as in TLU) owned in taka.
Productive assets (PCA)	The productive asset index is based on the first principle component of HH ownership indicators for the variables fruit trees, vegetable gardens, bamboo clumps, tubewells, fishnets, boats, rickshaws/vans, husking machines, irrigation pumps and, for baseline, cots. Partially, variables were grouped to asset complexes for comparability between data-sets. Comp. Appendix 5: PCA Analysis for a detailed description on construction and variables included.
Productive asset value	Sum of the self-reported value of all (as above) productive assets in taka.
Productive land in acre	Acre of land self-reported as being in economic use of the household (status self-owned, share-cropping, leasing or mortgaged out and status share-cropping, leasing or mortgaged in). Original accords are given in decimal which are equivalent to 1/100 th of an acre and recalculated in impact estimation tables.
Tropical livestock units	TLUs are computed based on the proposal of Jahnke (1982: 9f.), standardizing one TLU to 250 kg liveweight. As suggested by Jahnke (1982: 9f.), weights of 0.7 for cattle, horse and buffaloes, 0.2 for pigs, 0.1 for sheep and goats and 0.01 for chickens, hens and ducks were used. For the category 'other' value was used as conversion factor with 9164 taka (mean cow value) as 0.7 TLU. TLU represent all livestock in economic use by the household, thus as well e.g. share rearing arrangements.

Vulnerability Indicators

Children's vaccination status	Vaccination status is counted as complete if all relevant vaccines were received by a child age of 18 months (otherwise as incomplete) for households reporting children of age 18-60 months.
Credit	Amount of taka taken as credit from informal sources (money-lender, neighbours, self help group, advance selling of labour, land mortgages) as reported by household. For the baseline survey, credit constitutes the mean value of reported credit ranges.
Dietary diversity score	Dietary diversity score, theoretically ranging from zero to eight, practically from one to eight, is a simple index aggregating the reported consumption of several food groups (rice, pulse, vegetables, egg, fish, meat, fruits, oil). Dietary diversity is supposed to capture the nutritional adequacy of the diet, as proposed by Hatløy et al. (1998: 892) for Malawian children, though substituting the category milk by oil and green leaves by pulse as available in the IFSUP data.
Diversification Index (DIS)	Index ranging from zero to eight counting the number of income sub-classes of the household: Land usage; livestock usage; fruit trees, bamboo clumps or vegetable gardens used; productive assets of the type rickshaw/van, fishing net, boat, husking machine used, small business capital; number of household members pursuing wage labour occupations (maximum in the data: 4 [at baseline]). Begging is not counted as separate income source.
Housing quality (PCA)	The housing quality index is based on the first principle component of the number of rooms owned by the household, the quality of materials for the walls as well as the quality of materials for the roof (concrete and brick, tin, jute stick, bamboo, clay, clay/straw mixture, straw; in descending order; order derived by mean reported value of rooms by construction type), the size of the homestead (in decimal) and ownership of the homestead (self-owned vs. rented). Comp. Appendix 5: PCA Analysis for a detailed description on construction and variables included.
Pregnant mothers' age	Age of the pregnant mother in the household. In case of several pregnancies occurring at the same time, it contains the minimum of the respective pregnant mothers' ages.
Mothers' checkup status	Coded 1 if pregnant mothers in the households received both antenatal and postnatal checkup of any kind.
Savings	Amount of taka in savings accounts as reported by households.
Sick days per capita	Number of self-reported days household members fell sick per month averaged by household size. Figures were inquired as yearly data for the impact survey and subsequently recalculated.

Eigenständigkeitserklärung / Declaration of Academic Honesty

Hiermit versichere ich, dass ich die vorliegende Hausarbeit selbständig und ohne fremde Hilfe angefertigt, alle benutzten Quellen und Hilfsmittel angegeben und Zitate als solche kenntlich gemacht habe.

Ich versichere ferner, dass ich die Arbeit weder für eine Prüfung an einer weiteren Hochschule noch für eine staatliche Prüfung eingereicht habe.

München, den 05.01.2012

Lukas Rudolph

Curriculum Vitae

Geboren am 05.09.1984 in Bayreuth
Kinder: Timo David Rudolph (*20.11.2010)

Studium und Schule

Seit Oktober 2006 Magisterstudium der Politischen Wissenschaft mit den
Nebenfächern Recht und Interkulturelle Kommunikation

10/2008 – 10/2010 Studium der Volkswirtschaftslehre (B.Sc.) an der LMU
München

Juli 2005 Abitur am Peutingen Gymnasium Augsburg

Arbeitserfahrung

03/2011 – 07/2011 Wissenschaftliche Hilfskraft am Lehrstuhl für Internationale
Wirtschaftsbeziehungen an der LMU München

06/2010 – 03/2011 Praktikum in der Inhouse Consulting der Allianz Deutschland
AG

08/2008 – 10/2008 Praktikum bei der InWEnt Internationale Weiterbildung und
Entwicklung gGmbH

07/2005 – 08/2005 Praktikum bei der Augsburger Allgemeinen Zeitung

Auslandserfahrung

09/2009 – 10/2009 Französischsprachkurs mit Eurocentres, La Rochelle,
Frankreich

10/2005 – 09/2006 Freiwilliges Soziales Jahr in Bangladesch mit der deutschen
Nichtregierungsorganisation NETZ e.V.

08/2002 – 08/2003 „High School Year“ in Westfield, Massachusetts, USA, über das
Deutsche Youth for Understanding Komitee e.V.

Soziales Engagement

Seit Januar 2011 Vorstand, Schatzmeister und Regionalleiter Süd- und Ostasien
von IFAIR – Young Initiative on Foreign Affairs and
International Relations e.V.

Seit Oktober 2006 Engagement in der Bildungsarbeit von NETZ e.V.

09/2003 – 08/2004 Schülersprecher am Peutingen Gymnasium Augsburg

10/1999 – 03/2005 Jugendleiter im Verband Katholische Studierende Jugend

Stipendien

Seit September 2008 Stipendiat der Studienstiftung des deutschen Volkes

Publikationen

Mai 2010 „Die Bedeutung von Werten für Verteilungsergebnisse im
Ultimatum- und Diktatorspiel“, in: Behnke, J. et al. (Hrsgb.):
„Jahrbuch für Handlungs- und Entscheidungstheorie – Band 6“,
VS Verlag (mit Behnke, J./ J. Hintermaier).