THREE ESSAYS ON EVALUATION AND MEASUREMENT

IN DEVELOPING COUNTRIES

By

Mario González Flores

Submitted to the

Faculty of the College of Arts & Sciences

of American University

in Partial Fulfillment of

the Requirements for the Degree of

Doctor in Philosophy

In

Economics

Chair: Paul Winters, Ph.D.

Mieke Meurs, Ph.D.

Jessica Todd (USDA, ERS), Ph.D.

Dean of the College of Arts and Sciences

April 22, 2014

Date

2014

American University

Washington, D.C. 20016

# DEDICATION

I dedicate this work to my mother, Minerva Flores Sánchez, and my father, Pedro González Amigón, who taught me, by example, the meaning and value of hard work, and who always helped me and encouraged me to get an education. I love and admire you both.

*Dedico este trabajo a mi madre, Minerva Flores Sánchez, y mi padre, Pedro González Amigón, quienes me enseñaron, con su ejemplo, el valor y significado de lo que es trabajar duro, y quienes siempre me apoyaron y alentaron a obtener una educación. Los amo y admiro a los dos.*

I also dedicate this work to my wife, Alison A. Pflepsen, who has supported me in this long and challenging process. Without your love and support, I would not have been able to do this; this was truly a team effort and the rewards of this work are all yours as much as they are mine.

Finally, I dedicate this work to my son or daughter who will be joining my wife and me sometime in June of this year. We can't wait to meet you! You also have been an inspiration (and a great incentive) to get this all done before you arrive!

THREE ESSAYS ON EVALUATION AND MEASUREMENT

IN DEVELOPING COUNTRIES

BY

Mario González Flores

ABSTRACT

The overall theme of this dissertation is on evaluation and measurement in developing countries. All three essays make a contribution to the literature—either with the use of a new empirical method for evaluation or a unique or new data set, or the evaluation of a tool or program not yet evaluated, or by evaluating an aspect within the evaluation literature usually neglected. The results of the three essays have policy relevance and can be used to design or modify anti-poverty programs geared towards the social and productive sectors so that they can have a greater impact on poverty.

Chapter 2 uses data from small scale potato farmers in Ecuador to examine the impact of the program *Plataformas de Concertación* on productivity growth. Using propensity score matching combined with a Stochastic Production Frontier model recently introduced by Greene (2010) that corrects for sample selection bias, we disaggregate the yield growth attributable to the program into technological change (TC) and technical efficiency (TE). While the results do not exhibit a clear indication of selection bias, the analysis does show that on average beneficiaries exhibit higher yields than control farmers given the same input levels, but lower TE with respect to their own frontiers. These results suggest that while the program raised the technology gap in favor of beneficiaries, it had a negative effect on TE on the short run. The latter finding is consistent with the notion that beneficiaries enjoyed a significant change in production techniques, but it is very likely that they were still in the "learning by doing" stages at

the time the data was collected. In fact, the results suggest a fast recovery on TE levels on the part of beneficiaries as time with project increased.

Chapter 3 uses data from Peru to evaluate the Progress out of Poverty Index (PPI), which is a low-cost and easy to use poverty scorecard developed by Mark Schreiner of Mircrofinance Risk Management, L.L.C for the Grameen Foundation. The PPI estimates the likelihood that a household has expenditure below a given poverty line. The scorecard is a practical way for anti-poverty programs or pro-poor NGOs to monitor poverty rates among their clients, track changes in poverty rates over time, and target services. The results of Chapter 3 show that the PPI is very accurate (in terms of bias and targeting accuracy) when it is applied at the national level and in urban areas. However, when it is applied to rural areas, the tool is not able to discriminate well between the poor and non-poor. After making several modifications to the PPI the results show that these modifications do not translate into significant improvements in accuracy (neither for bias nor for targeting accuracy). We conclude that the PPI might benefit from the creation of a separate scorecard for rural areas that is based on a set of indicators that might be more relevant for rural areas.

Chapter 4 uses data from an unconditional cash transfer program, the Child Grant Program (CGP), implemented as a Randomized Control Trial in Zambia, to assess the impact of the program on agricultural production, productive investments, and technical efficiency. We find that the program did not have an impact on the value of harvest or gross margins. However, the program had a positive impact on expenditures in seeds, the share of households that spend on crop production, including miscellaneous expenditures. Moreover, the increase in crop expenditures seems to be reflected in the amount of harvest sold for maize and rice, in the total value of sales, primarily driven by maize sales, and the share of households that sell part of their

harvest, primarily driven by sales of maize and rice. We find that the CGP had a positive and significant impact on total operated land (an increase of 0.23 ha.). The impact of the program on farm tools was very modest, although the impact on farm animals was greater: the CGP only increased the share of households that own a hammer (4%); while the program had a positive and significant impact on the number of cattle owned (increase of 0.44), goats (0.13), chickens (1.11), and ducks (0.2). To provide a more nuanced picture of the impact of program, we ran the same set of analyses across three household types based on differences in size of household labor, defined as household members 17 years old or older and younger than 65: (1) single-headed households; (2) nuclear households; (3) and surplus-labor households. The program impacted each of these types of households differently: single-headed households who are labor-constrained seem to have retrenched from agricultural activities, while nuclear and surplus-labor households have intensified and expanded their agricultural endeavors. Importantly, while nuclear households benefitted the most from the program in terms of the number of indicators that were positively impacted, surplus-labor households benefited the most in terms of the magnitudes of the benefits. The results show that the program had no impact on technical efficiency. The results of the ancillary equation for the inefficiency part of the error structure shows that education and access to credit can significantly improve TE levels.

ACKNOWLEDGMENTS

For Chapter 2, I thank Boris E. Bravo-Ureta, Daniel Solís, and Paul Winters for their guidance and collaboration. The data collection for this chapter was funded by the FAO-Netherlands Partnership Program and the FAO Norway Partnership Program, and was made possible through the support of FAO-ESA in Rome, farmers and leaders of CONPAPA in Ecuador, the International Potato Center and its Papa Andina Partnership Program, INIAP Ecuador, the Swiss Agency for Development and Cooperation and its FORTIPAPA Project, CESA-Ecuador, Fundación M.A.R.CO. and IEDECA. I also thank Daniel Solís and Heath Henderson for guidance on modeling the SPFA.

For Chapter 3, I thank Mark Schreiner from Microfinance Risk Management, L.L.C. for his invaluable help in replicating his methodology to create the PPI, as well as his very helpful suggestions to previous versions of this chapter. I also extend my thanks to Matt Walsh and Mary Jo Kochendorfer from the Grameen Foundation for comments on previous versions of this work, and for coordinating the completion of this work.

For Chapter 4, I thank Silvio Daidone Joshua Dewbre form the Food and Agriculture Organization of the United Nations (FAO) for their collaboration on variable creation. I would also like to thank UNICEF, the University of North Carolina (UNC), the American Institute of Research AIR, FAO, and the Government of Zambia for use of the data. I also thank David Seidenfeld from AIR for his help on interpreting some of the results of this paper, and Silvio Daidone from FAO for his guidance in modeling the SPFA for this chapter.

The completion of the Ph.D. would not have been possible without the support and guidance of the faculty in the Economics department at American University. I thank the following professors for sharing their knowledge, for providing guidance, and for always having

their offices open to my questions: Isaac, Park, Wisman, Bono, Meurs, Golan, Blecker, Winters, Willoughby, Starr, Broder, Feinberg, Callahan, Reynolds and Floro.

The completion of this Ph.D. would have not been possible without the help, love, and friendship of my classmates, who have now become part of my family: Demet Cimen-Gulsen, Tual Suan Tuang, Abdul-Farouk Nabourema, Horacio Diego Rivera, Du Huancheng, and Dongping Xie. I have no doubt I would not have been able to make it through without your help and friendship.

I also thank the members of my committee, Paul Winters, Mieke Meurs, and Jessica Todd. Your constructive critiques and suggestions to my work have significantly improved this dissertation. I have greatly benefited from your knowledge, expertise, suggestions, and kindness; and I thank you for that.

Finally, I have no words to express my gratitude to the chair of my committee, Paul Winters.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

# CHAPTER 1

# INTRODUCTION

This dissertation follows the three-essay format. The overall theme is on evaluation and measurement with a specific emphasis on developing countries. All three essays make a contribution to the literature—either with the use of a new empirical method for evaluation (Chapter 2) or a unique or new data set (Chapters 2 and 4), or the evaluation of a tool (poverty scorecard) (Chapter 3) or program (Chapter 4) not yet evaluated, or by evaluating an aspect within the evaluation literature usually neglected (Chapters 2 and 4). The results of the three essays have policy relevance and can be used to design or modify anti-poverty programs geared towards the social and productive sectors so that they can have a greater impact on poverty (ideally).

Chapter 2 uses data from small scale potato farmers in Ecuador to examine the impact of the program *Plataformas de Concertación* on productivity growth. Using propensity score matching combined with a Stochastic Production Frontier model recently introduced by Greene (2010) that corrects for sample selection bias, we disaggregate the yield growth attributable to the program into technological change (TC) and technical efficiency (TE). While the results do not exhibit a clear indication of selection bias, the analysis does show that on average beneficiaries exhibit higher yields than control farmers given the same input levels, but lower TE with respect to their own frontiers. These results suggest that while the program raised the technology gap in favor of beneficiaries, it had a negative effect on TE on the short run. The latter finding is consistent with the notion that beneficiaries enjoyed a significant change in production techniques, but it is very likely that they were still in the "learning by doing" stages at the time the data was collected. In fact, the results suggest a fast recovery on TE levels on the part of beneficiaries as time with project increased.

Chapter 3 uses data from Peru to evaluate the Progress out of Poverty Index (PPI), which is a low-cost and easy to use poverty scorecard developed by Mark Schreiner of Mircrofinance Risk Management, L.L.C. for the Grameen Foundation. The PPI estimates the likelihood that a household has expenditure below a given poverty line. The scorecard is a practical way for anti-poverty programs or pro-poor NGOs to monitor poverty rates among their clients, track changes in poverty rates over time, and target services. Up-to-date, there has not been a rigorous evaluation of the PPI. This paper aims to fill this gap in the literature. The results of Chapter 3 show that the PPI is very accurate (in terms of bias and targeting accuracy) when it is applied at the national level and in urban areas. However, when it is applied to rural areas, the tool is not able to discriminate well between the poor and non-poor. As such, the tool shows high levels of leakage in rural areas. After making several modifications to the PPI—including adding a rurality variable and adding regional variables, the results show that these modifications do not translate into significant improvements in accuracy (neither for bias nor for targeting accuracy). However, when we create two separate scorecards—one for rural and one for urban areas—there is an important improvement in terms of lower levels of leakage in rural areas and higher inclusion rates in urban areas. We conclude that the PPI might benefit from the creation of a separate scorecard for rural areas (which is where the vast majority of poor live) that is based on a set of indicators that might be more relevant for rural areas, such as having agricultural land, livestock, distance to a road, or primary occupation of the head of household. This might improve the PPI's ability to discriminate between the poor and non-poor in rural areas.

Chapter 4 uses data from an unconditional cash transfer program, the Child Grant Program (CGP), implemented as a Randomized Control Trial in Zambia. The objective of Chapter 4 was to assess the impact of the CGP on agricultural production, productive

investments, and technical efficiency. Using a difference-in-difference model for the entire sample, we find that the program did not have an impact on the value of harvest or gross margins. However, the program had a positive impact on expenditures in seeds, the share of households that spend on crop production, including miscellaneous expenditures. Moreover, the increase in crop expenditures seems to be reflected in the amount of harvest sold for maize and rice, in the total value of sales, primarily driven by maize sales, and the share of households that sell part of their harvest, primarily driven by sales of maize and rice. We find that the CGP had a positive and significant impact on total operated land (an increase of 0.23 ha.), primarily driven by increases in land devoted to maize (0.14 ha.), rice (0.06 ha.), sorghum (0.02), and beans (0.01). The impact of the program on farm tools was very modest, although the impact on farm animals was greater: the CGP only increased the share of households that own a hammer (4%); while the program had a positive and significant impact on the number of cattle owned (increase of 0.44), goats (0.13), chickens (1.11), and ducks (0.2).

In order to provide a more nuanced and clearer picture of the impact of the program, we ran the same set of analyses across three household types based on the differences in size of household labor, defined as household members 17 years old or older and younger than 65: (1) single-headed households; (2) nuclear households; (3) and surplus-labor households. The program impacted each of these types of households differently. For single-headed households, the program helped increase the share of households that spend on miscellaneous crop expenditures, it helped them increase the amount of maize sold, the total value of sales, primarily driven by maize, and it increased the share of households that sell part of their harvest. Moreover, the program increased their land devoted to sweet potatoes, sorghum, and other beans. Finally, the program helped single-headed households increase the number of cattle owned (by

0.28) and the number of chickens (by 1.1). On the other hand, the CGP had the greatest impact on nuclear households for the total number of indicators that were positively impacted (23), and in fact, the positive impacts of the program presented for the entire sample are driven primarily by the impact of the program on these nuclear households: for crop expenditures, rice sales, land expansion, and for investments in small farm animals. Nuclear households also increased their expenditures on hired labor. For surplus-labor households, the CGP had the greatest impact in terms of the magnitude of the benefits: while the program had a positive impact on seed expenditures for nuclear and surplus-labor households, the magnitude of the increase was greatest for the latter; surplus-labor households were the only ones able to translate increases in expenditures into greater yields; the impact of the program on maize sales—for Kg./ha and the value of maize sold—was greatest for surplus-labor households; and the impact of the CGP on investments in large- and medium-size farm animals was also greatest for surplus-labor households. The results show that the program had no impact on technical efficiency. The results of the ancillary equation for the inefficiency part of the error structure shows that education and access to credit can significantly improve TE levels.

We should note that beyond the important findings provided individually by each of these three essays, taken all together, this dissertation highlights two broader themes that have relevance to the field of impact evaluation and measurement, and which should help improve decision-making in the policy world.

First, each of these three essays brings together different strands of the literature, or highlights the importance of using complementary methodologies for evaluation and measurement, which results in more accurate, relevant, and complementary information. For instance, Chapter 2 brings together the impact evaluation and the stochastic frontier analysis

literature in order to look not only at the impact of a program on yields, but also on technical efficiency. The results provide a more complete picture of how a program can impact farmers. Even if a program leads to increases in yields, it is still possible to provide more benefits to farmers if additional training makes them more efficient and thereby bringing them closer to their output potential. Similarly, Chapter 3 highlighted how the developer of the PPI not only uses statistics, but also judgment to select the final set of 10 indicators that go into the final poverty scorecard. This process is complemented by field testing and focus groups to be sure the scorecard is appropriate (context), and that it has a good chance of being used (user friendly and clear). In this sense, one is not only concerned with statistical accuracy, which we can improve, perhaps, by creating a separate scorecard as suggested in this dissertation, but we are also concerned on the likelihood that the tool will be used. Moreover, Chapter 3 made suggestions on several complementary ways in which the results of the PPI can be presented, including by concentrating on the lower end of the distribution of scores, which is where most of the poor households are scored. These suggestions help provide a clearer and more relevant picture of the accuracy of the PPI, which can potentially help users better gauge the strengths and weaknesses of the PPI. Chapter 4 also highlighted the need to bring together the impact evaluation literature for social programs with the impact evaluation literature for productive programs, particularly in agriculture. In recent years, most of the impact evaluations of cash transfer programs have primarily focused on the impacts on human capital formation, particularly of the young. Yet, concentrating only on indicators for education and health may lead to missed opportunities to evaluate the impact of these programs on agricultural production and productive investments. Indeed, Chapter 4 found that, even though the program is intended primarily as a safety net program, the CGP actually had significant and important impacts on agricultural-related

variables, as well as on investments in farm animals. Thus, considering these sets of indicators in an evaluation provides a more complete picture of the total impact of social programs.

The second broader theme highlighted in this dissertation is related to the importance of looking beyond the average impact of a program, or the overall results of an analysis. In other words, it is important to undertake multiple analyses at different levels and with different sub-populations to identify the heterogeneity of impact, or the differential results across groups or sub-samples. This approach provided invaluable information of how different groups or sub-populations are impacted by a program or an analysis. For instance, Chapter 2 found that, on average, farmers participating in the *Plataformas* had lower technical efficiency scores than control households. However, when we looked at the impact of the program conditional on the number of years of participation, we found that farmers that had recently joined the program had the lowest levels of technical efficiency. On the other hand, farmers that had been in the program for at least two years had much higher technical efficiency scores, similar to control households. Moreover, in Chapter 3, we found that while the targeting accuracy of the PPI is very high in urban areas and when applied at the national level, the tool is not as good when applied to rural areas: the ROC curve shows that the PPI has very high levels of leakage in rural areas regardless of the scorecard used. In this sense, the tool is not as accurate at discriminating the poor from the non-poor in rural areas. Finally, while the results of Chapter 4 for the entire sample showed that the CGP had very modest impacts (in terms of the number of indicators positively impacted and the magnitude), when we ran the analysis by household-labor type, we found that the CGP may have kick started some form of structural transformation: single-headed households who are labor-constrained seem to have retrenched from agricultural activities, while nuclear and surplus-labor households have intensified and expanded their agricultural endeavors.

Importantly, while nuclear households benefitted the most from the program in terms of the number of indicators that were positively impacted, surplus-labor households benefited the most in terms of the magnitudes of the benefits.

Based on these two overarching themes highlighted in this dissertation, it seems clear that it is important to bring together different strands of the literature, to use multiple methodologies to undertake impact evaluations, and to look beyond the average impact by looking at the effect of a program on different groups or sub-samples. The application of these complementary alternatives will help decrease biases in the estimations, will help provide better measures of impact, will help provide a more complete and nuanced picture of the results and will better inform policy.

# CHAPTER 2

# THE IMPACT OF HIGH VALUE MARKETS ON SMALLHOLDER PRODUCTIVITY IN THE ECUADOREAN SIERRA: A STOCHASTIC PRODUCTION FRONTIER APPROACH CORRECTING FOR SELECTIVITY BIAS

## 2.1 Introduction

Agricultural projects often seek to improve productivity with the expectation that such improvements would lead to higher income and welfare among beneficiaries. Examples of interventions include the introduction of new seed varieties, the adoption of new farming techniques (such as integrated pest management or IPM), linking farmers to markets, better accounting practices, provision of extension services, farmer field schools (FFS), or a combination of various actions. The effective use of a newly adopted technology requires investing the time and effort to become acquainted with the new practices before the full benefits of adoption can be felt by the farmer. This may entail a process of trial and error during several agricultural cycles. While adopting new techniques or inputs can potentially lead to increases in production at the end of an agricultural cycle, this does not necessarily mean that the new procedures are being implemented in an efficient manner. This is particularly true for smallholders, who are typically characterized by having lower levels of education, living in isolated rural areas, and having limited access and exposure to information and markets. Thus, much of the innovative content and techniques can be quite foreign and may entail a challenging process for this type of farmers. Therefore, when evaluating the impact of an agricultural project, it is important to differentiate between indicators of technological change (TC) versus managerial performance (or technical efficiency, TE).

The economic impact evaluation literature has been growing in recent years, and this growth has been mainly focused on the social sectors where the indicators of impact tend to be more easily identifiable (Winters et al., 2011). Rigorous impact evaluations of agricultural projects have been relatively scarce and the evidence on the effectiveness of such projects in developing countries is mostly inconclusive (IDB, 2010; Del Carpio and Maredia, 2009). The relative scarcity of formal evaluations of agricultural projects is likely due to several reasons. First, agricultural projects are generally designed to increase output and therefore impact evaluations focus on production-based indicators, typically associated with TC. However, collecting this type of data can be challenging, beginning with the definition of the sample unit, since production is often linked to multiple plots but the decision-making process takes place at the household level. The challenge is greater when attempting to evaluate the impact of a project on different types of households, such as smallholders and large holders, who often have very distinct production systems (Winters et al., 2010).

Second, in analyzing agricultural production, the relationship between inputs and outputs or profitability is often examined through gross margins or total value product functions. Yet, presumably, agricultural projects have an impact not just on inputs and outputs, but also on how these inputs are used and combined. Whether these inputs are being used in an efficient manner to obtain the maximum possible levels of output needs to be considered in an evaluation (Winters et al., 2010). Yet, this is rarely done since, as noted, most project evaluations focus on TC indicators. While such focus allows the researcher to identify impact on different components of production, it does not provide any information on whether farmers made the right use of the available inputs and technology at their disposal, i.e., managerial performance is ignored.

Given these difficulties, combining Stochastic Production Frontier Analysis (SPFA) with impact evaluation methodologies provides a useful avenue for measuring the productivity impact of agricultural projects. SPFA is a widely used econometric technique that estimates the 'best practice' relationship between inputs and output of the farm households in the sample. In addition, SPFA can help identify the levels of efficiency (or inefficiency). Therefore, this approach makes it possible to quantify the potential to increase agricultural output without the need for additional inputs or new technology (Coelli et al., 2005).

*Papa Andina*, the focus of this paper, is a partnership that worked to address rural poverty in the Andean highlands by fostering innovation and market development for potatoes. The approach recognizes that while agricultural research is a main driver of TC and agricultural development in addressing rural poverty, this research needs to be linked to practical improvements in value chains that are important to smallholders (Horton et al., 2011). A key program within *Papa Andina* is the *Plataformas de Concertación*, hereafter *Plataformas.* This program offered a space for public and private sector partnerships where diverse actors—including farmers, potato processors, supermarkets, national research institutes, universities and non-governmental organizations—could work together to innovate and link small-scale potato producers to commercial interests. *Plataformas* offered a mechanism not just to support agricultural research in the field, through new varieties and different mechanisms to enhance production and marketing, but also served as an experiment in institutional innovation. The question is whether this approach is an effective mechanism to increase farmer production and efficiency and this constitutes the overall goal of this paper.

In implementing impact evaluations of development projects, several researchers have promoted the use of randomized experiments (Duflo et al., 2008). However, it is often the case

that experimental designs are costly and difficult to implement; thus, one needs to rely on non-experimental methods (Barrett and Carter, 2010). One common non-experimental approach to assessing project impact is propensity score matching (PSM), which alleviates biases stemming from observable variables (World Bank, 2006). However, in projects where beneficiaries self-select, unobservable variables (e.g., managerial ability) can also be a source of bias. If panel data are available, fixed effects estimators along with PSM can be used to deal with the problem, provided that the unobservables are time invariant (Angrist and Pischke, 2009). Thus, the generation of a counterfactual along with the mitigation of biases from observables and unobservables can be addressed in non-experimental designs as long as one has data on treatment and control groups at both the baseline and the endline. Recent applications of this methodology to agricultural projects include the work of Bravo-Ureta et al. (2011) in Honduras and Cerdán-Infantes et al. (2008) in Argentina.

A challenge frequently encountered in the field is that analysts and/or policy makers might be interested in measures of impact even when baseline data is not available. In such a situation, which is the case for *Plataformas*, one needs to rely on cross-sectional data along with suitable matching procedures and other econometric techniques, such as instrumental variables, in order to obtain the desired impact measures (Cavatassi et al., 2011a). In this paper, we are particularly interested in separating the effect of TC and TE on farm productivity. To achieve our goals we make use of cross-sectional data collected after the project was underway. A number of steps were taken to ensure that the data on treatment and control groups would have been very similar at the baseline, but it is likely that selection issues still remain. Therefore, we address possible self-selection in a stochastic frontier context using the model recently introduced by

Greene (2010) and adapted to the evaluation of development programs by Bravo-Ureta et al. (2012).

In sum, a key feature of this paper is bridging SPFA with impact evaluation methods. Development projects often have a major component intended to improve decision-making and managerial ability along with the transfer of technologies designed to increase output. Thus, for such projects, SPFA methodologies are ideally suited to decompose productivity growth into technological and managerial components; however, these methodologies have hardly been applied for this purpose. A major reason for the absence of such applications is likely to be the challenges posed by selectivity bias, which is a common feature in development projects. In this fashion, this paper adds to a very limited but emerging literature that combines SPFA modeling with impact evaluation techniques.

The remainder of the paper is structured as follows. Section 2.2 provides an explanation of *Plataformas*, and a description of the data is provided in Section 2.3. Section 2.4 presents the analytical framework for analyzing TC and TE and the closely related empirical strategy. Section 2.5 provides the results and Section 2.6 concludes.

*Plataformas de Concertacion*[1]

The *Plataformas* are multi-stakeholder alliances, which bring farmers together with a range of agricultural support service providers, including INIAP (*Instituto Nacional Autónomo de Investigaciones Agropecuarias*), local NGOs, researchers, universities, and local governments. The *Plataformas* pay special attention to expanding the direct participation of low-income farmers in high-value producer chains by providing them with new technologies, by promoting their organizational skills and social capital, and by involving them in a "value chain

[1] More information on the different aspects and activities of *Plataformas* can be found in Cavatassi et al. (2009).

vision" of production and commercialization that directly links them to the final output markets, thus circumventing intermediaries (Cavatassi et al., 2009). As noted by Devaux et al. (2009, p. 36), "this facilitates knowledge sharing, social learning and capacity building, leading to improvements in small farmer productivity and the quality of potatoes supplied to market." The overall objective of the *Plataformas* is then to "reduce poverty and increase food security, through increasing yields and profits of potato smallholders" (Pico, 2006, as cited in Cavatassi et al., 2009, p. 8).

A central component of *Plataformas* was the training provided at FFS, which involved lectures and applied exercises in experimental plots. FFS emphasized improved production technologies and IPM techniques aimed at enhancing the quality and quantity of production. Farmers were taught techniques to manage soil, seeds (renewing and stock selection), insects (Andean weevil and tuber moths), diseases (late blight), and agrochemicals (insecticides, fungicides, pesticides, and fertilizers) efficiently. An important element under the IPM approach was to reconcile improvements in production with the use of these techniques while preserving the environment and protecting human health. The training also exposed farmers to the quality requirements of high-value markets.

Previous evaluations have found positive and significant impacts of the program on technical indicators, such as yields, profits, and gross margins (Cavatassi et al., 2011a; Cavatassi et al., 2011b). However, an issue that has not been evaluated is whether the program had an effect on managerial performance or TE, i.e., are beneficiary farmers producing closer to their best practice production frontier than non-beneficiaries? Given that all participants in the *Plataformas* received the same comprehensive 'package' of 'treatments', one would expect that the program would have a positive effect on farmers' TE. On the other hand, if much of this new

information and technologies were too complex to process, comprehend, and master right away, then one would expect that while farmers might attain higher output, TE might remain about the same or might even decrease. The lack of improvement or even a lower TE is plausible if farmers had exposure to the program for only a short period and thus not enough time may have elapsed for them to become proficient with the new methods. Hence, the actual effect of the *Plataformas* on the TE of beneficiaries is an empirical question.

**2.2 Data**

The data used in this paper comes from "*La Nueva Economía Agrícola*" household and community level surveys implemented by the Food and Agriculture Organization (FAO) in collaboration with the International Potato Center (IPC). The *Plataformas* were active in the Ecuadorean provinces of Chimborazo and Tungurahua and surveys were administered in these provinces from June to August of 2007, and they encompass information for the year prior to the survey. Data was collected at the plot, household and community levels.

To ensure a robust evaluation of the impact of the program, the data collection was done in multiple steps and the counterfactual was designed with a great deal of care. First, communities participating in the *Plataformas* program (treated communities) were identified in each of the two provinces and information on these was obtained. Second, using population and agricultural census data, the treated communities and a set of potential control communities that did not participate in the program but who had similar geographic, agro-ecological and socio-demographic characteristics were identified. This provided a list of all possible treatment and control communities to be included in the survey. Third, using PSM, control communities that were most comparable to treated communities from a statistical standpoint were determined. Fourth, the resulting list of potential control communities was discussed with key local organizations that had a central role in *Plataformas* to determine if they were indeed comparable

to the treated communities. Some of the key characteristics considered were production and agro-ecological similarities and levels of community and farmer organization. Thus, the selection of treatment and control communities that resulted from the application of PSM was fine-tuned by local agronomists and leaders of organizations with local knowledge. Through this process, the most appropriate control communities were identified. Further, treated communities for which a suitable control community could not be found were excluded from the sample (three in total).

Once the communities for inclusion in the sample were determined, lists of households from treatment and control communities were obtained by *Plataformas* coordinators and community leaders to randomly select those to be included in the final sample. The lists from treated communities included households who participated in the program and those that did not. The final sample includes a total of 35 communities (18 treatment and 17 controls) and contains 1,007 households that were randomly selected from control communities and among participants and non-participants in treated communities. Given the focus of this paper, the analysis is restricted to farms that had at least one full potato production cycle (from planting to harvesting) and this final sample consists of 495 households. A test was conducted to check for systematic differences between treated and control farms in their probabilities of having had a complete production cycle and no differences were found.

The data used in this paper makes it possible to create alternative counterfactuals including: 1) non-beneficiaries living in treated communities (non-participants); 2) non-beneficiaries living in non-treated communities (non-eligible); and 3) a combination of 1 and 2. There are two concerns with option 1. First, non-participation is an explicit choice made by those producers and that choice might reflect that they are different from participants, which could lead

to self-selection bias, i.e., the estimates may reflect fundamental differences between the two groups rather than the impact of the program. Second, since non-participants live in close proximity to beneficiaries they may obtain indirect benefits from the program (spillover effects). Thus, using non-participants as a control group could be problematic. Nevertheless, this is a potentially useful group because their observable characteristics are likely to be very similar to those of participants since they live in the same communities. Further, previous analysis of these data indicates that there is little evidence of spillover effects (Cavatassi et al., 2009).

The non-eligible farmers can be considered as a "pure" counterfactual since spillover effects are unlikely. Importantly, given that the program was not offered in these communities, non-eligible farmers are not subject to possible self-selection bias, although they may exhibit program placement bias if the process used to select control communities is imperfect (Baker, 2000). As explained below, the final sample uses option 3 including 340 households who reside in treated communities (171 beneficiaries and 169 non-participants), and 155 in non-treated communities (non-eligible).

The survey has multiple modules containing several variables well suited for the prediction of participation into the program, such as information on land (number of plots, soil quality), socio-demographic variables (household composition and head of household characteristics), welfare and assets, social capital, and community-level variables. Table 2.1 presents descriptive statistics for the entire sample (pooled) and a test of difference in means between beneficiaries, non-participants and non-eligible as well as all control farmers together (non-participants and non-eligible).[2] Table 2.1 shows that, as expected, non-participants (column V) have very similar characteristics as beneficiaries (column II) — the two groups exhibit statistically significant differences for only 4 out of 28 variables.

[2] The variables included are theoretically exogenous, i.e., they were not affected by *Plataformas*.

Table 2.1: Description of the Data

| Variables (units) | I Pooled | II Benef. | III All controls | Test[a] | IV Non-elig. | Test | V Non-part. | Test |
|---|---|---|---|---|---|---|---|---|
| Altitude (meters) | 3475 | 3435 | 3497 | * | 3497 | | 3496 | |
| Land Owned (has.) | 2.77 | 2.75 | 2.78 | | 3.33 | | 2.27 | |
| Owned Plots (#) | 3.04 | 3.37 | 2.86 | *** | 3.11 | | 2.63 | *** |
| Black Soil (%) | 81% | 77% | 83% | ** | 86% | ** | 81% | |
| Flat Land (%) | 39% | 38% | 39% | | 37% | | 40% | |
| Irrigated Land (%) | 60% | 57% | 62% | | 63% | | 62% | |
| Family Size | 4.71 | 4.85 | 4.64 | | 4.66 | | 4.62 | |
| Max Educ. In HH | 7.85 | 8.44 | 7.53 | *** | 7.86 | | 7.24 | *** |
| % of Labor Force Male | 48% | 47% | 48% | | 48% | | 48% | |
| Dependency Share | 0.29 | 0.29 | 0.29 | | 0.28 | | 0.30 | |
| Credit Constrained (1,0) | 18% | 16% | 19% | | 24% | * | 15% | |
| Average Educ. Of Head | 4.91 | 5.05 | 4.83 | | 4.59 | | 5.05 | |
| Indigenous Head | 60% | 56% | 62% | | 70% | ** | 56% | |
| Female Head (1,0) | 11% | 12% | 10% | | 10% | | 11% | |
| Single Head (1,0) | 12% | 11% | 13% | | 14% | | 9% | |
| Age of Head | 41.7 | 42.0 | 41.5 | | 43.3 | | 39.9 | |
| House (1,0) | 87% | 83% | 89% | * | 88% | | 89% | * |
| Concrete/brick House (1,0) | 87% | 82% | 90% | ** | 89% | * | 91% | ** |
| Refrigerator (1,0) | 16% | 12% | 17% | | 18% | | 17% | |
| Access to Water System (1,0) | 94% | 93% | 95% | | 97% | | 93% | |
| Sewage (1,0) | 5% | 5% | 6% | | 5% | | 7% | |
| Big Farm Animals (#) | 5.67 | 5.32 | 5.43 | | 6.35 | | 5.28 | |
| Ag. Ass. Membership 5 yrs.+ (1,0) | 8% | 6% | 9% | | 9% | | 8% | |
| Non-Ag. Ass. Membership 5 yrs.+ (1,0) | 66% | 63% | 68% | | 70% | | 66% | |
| Bus in Community (1,0) | 46% | 43% | 47% | | 45% | | 50% | |
| Elementary School (1,0) | 87% | 88% | 87% | | 85% | | 89% | |
| Distance to Closest City (km) | 29.68 | 27.36 | 30.90 | ** | 36.54 | *** | 25.73 | |
| Chimborazo (1,0) | 50% | 50% | 50% | | 45% | | 41% | |
| ***N*** | 495 | 171 | 324 | | 155 | | 169 | |

[a] Tests are for differences in means with respect to treated farmers
* $p<.10$; ** $p<.05$; *** $p<.01$

The only category in which the two groups differ is for welfare variables: on average, non-participants have a higher probability of living in a house, rather than a hut, and have a higher probability of living in a brick house which gives an indication that they may be marginally

better off. The comparison between non-eligible (column IV) farmers and beneficiaries indicates that these two groups are also very similar in terms of land-related, social capital, welfare, and community variables, but they do differ slightly in socio-demographic characteristics. As expected, when the two possible control groups are pooled (column III), their differences with beneficiaries fall between what was described for columns IV and V. The two groups differ in 7 out of 28 variables. In sum, the descriptive statistics show that there are relatively minor differences between beneficiary and non-participant and non-eligible farmers. This gives an indication that the rigorous process of identifying control communities prior to data collection was quite successful although not perfect — a fact considered in the empirical strategy discussed below.

The data set also contains information on the value of potato production and input variables, which are used in the SPFA, and are presented in table 2.2 on a per hectare basis along with a test of difference in means. Looking at the value of total output per hectare, as expected given the emphasis of the project, the test of means shows that beneficiary farmers have a significantly higher value; and this is true when doing the comparison against non-participants, non-eligible farmers, or the pooled group of controls although the magnitudes vary. Table 2.2 also shows that beneficiary farmers, on average, spend more per hectare on labor, on seeds and on other inputs, and have a higher likelihood of hiring paid labor.

Table 2.2: Descriptive Statistics for Inputs and Output used in the SPF Models[d]

| Variables | Pooled | Benef. | All cont. | Test[a] | Non-elig. | Test | Non-partic. | Test |
|---|---|---|---|---|---|---|---|---|
| Total Output ($/ha.) | 5.72 | 6.17 | 5.48 | *** | 5.36 | *** | 5.59 | *** |
| Total Expenditures on Labor ($/ha.)[b] | 5.81 | 5.99 | 5.72 | ** | 5.60 | *** | 5.83 | |
| Total Expenditures on Seeds ($/ha.) | 4.22 | 4.71 | 3.96 | *** | 3.85 | *** | 4.05 | *** |
| Total Expenditures on O. Inputs ($/ha.)[c] | 5.11 | 5.41 | 4.96 | *** | 4.87 | *** | 5.03 | *** |
| Hired Labor (1,0) | 0.58 | 0.65 | 0.53 | *** | 0.61 | | 0.46 | *** |
| ***N*** | 495 | 171 | 324 | | 155 | | 169 | |

[a] Tests are for differences in means with respect to beneficiary farmers

[b] Includes Family, Hired, and Minga (community) Labor

[c] Includes expenditures on insecticides, fungicides (preventative and curative), fertilizer (organic and chemical), tractor, and animal draft.

[d] Values are in natural logarithm

* $p<.10$; ** $p<.05$; *** $p<.01$

## 2.3 Analytical Framework and Empirical Approach

Two critical components of productivity growth are TC and TE. The former captures "jumps" in the production function stemming from the application of improved practices that come from research and development efforts, whereas the latter can be interpreted as a relative measure of managerial ability for a given technology. In this paper, these two effects are disentangled making use of the Stochastic Production Frontier (SPF) framework in order to determine if the *Plataformas* had an impact on each component of productivity. Identifying program impact on TC and TE requires ensuring that selection, particularly self-selection, does not bias the estimates. Although the process of creating treatment and control groups, already discussed, attempted to minimize bias, this could still arise from differences in observable and unobservable characteristics of treatment and control households. To address this issue, we follow the framework presented by Bravo-Ureta et al. (2012), where PSM is used to mitigate biases stemming from observable variables when selecting the counterfactual/control groups and then bias from unobservables is addressed using SPF with sample selection.

The SPF model incorporates a composed error structure where a two sided symmetric term captures standard random variability and a one sided component captures inefficiency. In general terms, the model used in this study can be expressed as:

$$\boldsymbol{Y_{ij} = f(X, T_D) + v_{ij} - u_{ij}} \qquad \textbf{(1)}$$

where $\boldsymbol{Y_{ij}}$ is the yield (value of output per hectare) of the $i^{th}$ farmer, $\boldsymbol{X}$ is a vector of inputs per hectare, the variable $\boldsymbol{T_D}$ is a farm specific dummy variable that captures the effect of the new technology (i.e., innovations coming from participation in the *Plataformas*), $\boldsymbol{v_{ij}}$ is the two sided error term, and $\boldsymbol{u_{ij}}$ is the one sided error that captures efficiency. The subscript *j* is equal to *B* for beneficiaries or *C* for control farmers. The key technology effect that we are interested in identifying relates to the impact of participating in the *Plataformas* on potato yields. Two null hypotheses ($H_0$) will be tested:

1) The parameter of $\boldsymbol{T_D} = 0$; and
2) Mean $TE_B$ = Mean $TE_C$ or equivalently Mean $\boldsymbol{u_{iB}}$ = Mean $\boldsymbol{u_{iC}}$

Failure to reject the first null hypothesis indicates that there is no difference in yields, *ceteris paribus*, between control and treatment stemming from TC. Failure to reject the second hypothesis signifies that managerial ability, as measured by mean TE scores, across the two groups is the same. In this context, Fried et al. (2008) argue that although managerial ability is unobservable, it can be inferred from a ranking of TE scores derived from a "best practice" production frontier, which is the concept adopted here.

The empirical strategy to test these hypotheses involves several steps. The first step is to use PSM to construct a counterfactual group of farmers that have time-invariant characteristics similar to farmers that participated in the project (beneficiaries). The PSM procedure uses a Probit model to calculate the predicted probability of treatment based on observable

characteristics. These probabilities, or propensity scores, are then used to match similar households in the treatment group with those from the control group. The matching procedure applied here is the 1-to-1 nearest neighbor (NN) with replacement criterion using a caliper width of 0.001, which is more rigorous than the $0.25\sigma_P$ method suggested in the literature (Guo and Fraser, 2010).[3] While other papers rely on the 1-to-1 NN without replacement, arguing that it has the most intuitive interpretation of all alternatives available (e.g., Bravo-Ureta et al., 2011), this paper opts for 1-to-1 NN with replacement since it provides a better quality in the matching and it is more likely to decrease biases, although it adds variance in the estimations (Abadie and Imbens, 2002; Caliendo and Kopeining, 2008).[4] From this step, we obtain treatment and control groups with a similar range of observable characteristics.

The second step involves the estimation of SPF models. Conventional SPF models are estimated using the combined sample of beneficiary and control farmers (pooled data) as well as for each group separately. This is done to test whether beneficiary farmers display a different technology than the control group. If no technological difference is found then a single frontier combining all farmers, treated and control, is the more desirable option. At this stage, evaluating alternative functional forms for the SPF is advisable and the Cobb-Douglas (CD) was tested against the Translog (TL), which are the two most commonly used in efficiency studies (Bravo-Ureta et al., 2007). The results of maximum likelihood ratio tests were mixed. The linear parameters for the CD and TL were very similar in magnitude, and the coefficients for the

[3] This means using a caliper width for the nearest neighbor that is less or equal to a quarter of one standard deviation ($0.25\sigma_P$) of the estimated propensity scores of the sample.

[4] The alternative for 1-to-1 matching NN without replacement was also used and the results are very similar.

quadratic and interaction terms in the TL where, in most of the cases, not statistically significant. Thus, the Cobb-Douglas (CD) functional form was selected for the analysis.

To control for possible biases based on unobserved characteristics we implement the method recently introduced by Greene (2010). The model assumes that the unobserved characteristics in the selection equation are correlated with the noise in the stochastic frontier model (i.e., the term $v_i$ in equation 2).[5] Greene (2010) frames his model by noting that Heckman's (1979) original sample selection approach was developed for linear models and is not applicable for non-linear cases such as the SPF.[6] Thus, Greene proceeds to develop a selection approach for the SPF, which can be expressed as:

$$\textbf{Sample selection:} \quad d_i = 1[\alpha' \boldsymbol{z}_i + w_i > 0], w_i \sim N[0,1]$$

$$\textbf{SPF:} \quad y_i = \boldsymbol{\beta}' \boldsymbol{x}_i + \varepsilon_i, \varepsilon_i \sim N[0, \sigma_\varepsilon^2] \qquad \textbf{(2)}$$

$(y_i, \boldsymbol{x}_i)$ is observed only when $d_i = 1$

$$\textbf{Error structure:} \quad \varepsilon_i = v_i - u_i$$

$$u_i = |\sigma_u U_i| = \sigma_u |U_i|, \text{where } U_i \sim N[0,1]$$

$$v_i = \sigma_v V_i, \text{where } V_i \sim N[0,1]$$

$$(w_i, v_i) \sim N_2[(0,1), (1, \rho, \sigma_v, \sigma_V^2)]$$

[5] The rationale for the underlying assumption that the bias arises from possible correlation between unobservables in the production frontier with unobservables in the selection equation is that one "… might expect that observations are not selected into the sample based on their being inefficient to begin with" (Greene 2010, p. 23). As pointed out by one of the reviewers, this assumption is potentially a weakness of the approach but the implications for the empirical results and subsequent analyses are not evident. This is an issue that clearly merits additional methodological and empirical work.

[6] It is worth noting that several papers have applied the Heckman correction method to control for self-selection on unobservables in SPF studies. For instance, Sipiläinen and Lansink (2005) use a distance frontier model to analyze TE for Finnish organic and conventional dairy farms, while Solís et al. (2007) analyze TE for farmers in El Salvador who used different levels of adoption of soil conservation. Wollni and Brümmer (2012) examined productive efficiency for specialty and conventional coffee farmers in Costa Rica, and Rahman et al. (2009) analyzed production efficiency for a sample of rice producers in Thailand using the Greene (2010) method. However, these last two papers do not use matching techniques to control for differences based on observed characteristics. As far as we know, only Bravo-Ureta et al. (2012) use PSM along with Greene's (2010) self-selection model to identify the productivity impact of a development project.

where *d* is a binary variable that takes the value of 1 for beneficiaries and 0 for control farmers, *y* is output, **z** is a vector of explanatory variables in the sample selection equation, **x** is a vector of inputs in the production frontier, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameters to be estimated, and the error structure corresponds to that in the stochastic frontier model. In this model, the parameter $\rho$ captures the presence or absence of selectivity bias.

The log likelihood for the model in (2) is formed by integrating out the unobserved $|U_i|$ and then maximizing with respect to the unknown parameters. Thus,

$$\log L(\beta,\sigma_u,\sigma_v,\alpha,\rho) =$$

$$\sum_{i=1}^{N}\log\int_{|U_i|} f(y_i \mid \mathbf{x}_i,\mathbf{z}_i,d_i,|U_i|)p(|U_i|)d\,|U_i| \qquad (3)$$

The integral in (3) is not known so it has to be approximated. To simplify the estimation, Greene (2010) uses a two-step approach. The single equation MLE of $\alpha$ in the Probit equation in (2) is consistent but inefficient. The estimation of the parameters of the SPF does not require the reestimation of $\alpha$ so estimates for the latter are taken as given in the simulated log likelihood. Greene (2010) indicates that standard errors are adjusted based on the Murphy and Topel (2002) correction in a manner similar to what is done in Heckman (1979).

Greene (2010) goes on to argue that the non-selected observations (i.e., when $d_i = 0$) do not contribute information about the parameters to the simulated log likelihood and thus the function to be maximized becomes:

$$logL_{S,C}(\boldsymbol{\beta},\sigma_u,\sigma_v,\rho) = \sum_{d_{i=1}} log\ 1/R \sum_{r=1}^{R}[exp\big(-(1/2)((y_i - \boldsymbol{\beta'x_i} + \sigma_u|U_{ir}|)^2/\sigma_v^2)\big)/\sigma_v\sqrt{2\pi}\times$$

$$\boldsymbol{\Phi}\big(\rho((y_i - \boldsymbol{\beta'x_i} + \sigma_u|U_{ir}|)/\sigma_\varepsilon + a_i)/\sqrt{1-\rho^2}\big)]\ , \quad (4)$$

where $a_i = \hat{\alpha}' z_i$. Model parameters are estimated using the BFGS approach and asymptotic standard errors are obtained using the BHHH estimator. For full details on the model and its estimation see Greene (2010).

The estimation also requires modeling the farmer selection into the project. This selection process can be captured by a criterion function, which is assumed to be associated with exogenous household socio-economic variables, and can be expressed as:

$$B_i = \alpha_0 + \sum_{j=1}^{6} \alpha_j \boldsymbol{Z}_{ji} + w_i \quad (5)$$

where $B$ is a binary variable capturing the $i^{th}$ farmer's participation in the project ($B$=1 for beneficiary farmers, and 0 for control farmers); $\boldsymbol{Z}$ is a vector of exogenous variables; $\boldsymbol{\alpha}$ are the parameters to be estimated and $w$ is the error term distributed as $N(0, \sigma^2)$.

Within this framework, the predictor of TE can be obtained as the expectation of $u_i$ conditional on the composed error term $\varepsilon_i$ following Jondrow et al. (1982). Again, a full description of the estimation of the TE scores can be found in Greene (2010).

**2.4 Results**

Table 2.3 reports the results of Probit models on participation in the *Plataformas* using data on all controls (I), the non-eligible (II) and non-participants (III). For each model, marginal effects calculated at the sample mean are reported. The models accurately predict 69.9%, 68.7%, and 62.7% of outcomes, respectively. As expected, these results are consistent with those in table 2.1. The Probit results are used to calculate propensity scores for the treatment and control groups. Figure 2.1 shows the density estimates of the distribution of propensity scores for each group, along with the areas with and without common support. The scores obtained are almost entirely in the area of common support, suggesting that the different controls consistently represent a reasonable counterfactual for the treated group. Indeed, there are very few

observations off common support: 7 using all controls; 9 using non-eligibles; and 13 using non-participants.[7]

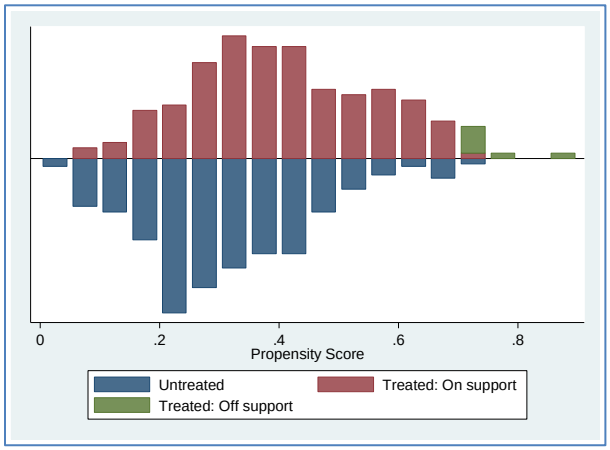
Table 2.3. Probit on *Plataformas* Participation (Marginal Effects)

| | (I) All controls | (II) Non-eligible | (III) Non-partic. |
|---|---|---|---|
| ***Land*** | | | |
| Land owned (ha) | -0.00325 | -0.00643 | 0.00769 |
| Owned plots (#) | 0.0365** | 0.0359** | 0.0365** |
| Black soil (%) | -0.126* | -0.196** | -0.112 |
| Flat land (%) | -0.0127 | 0.0416 | -0.0219 |
| Irrigated land (%) | -0.0728 | -0.0936 | -0.102 |
| ***Socio-demographic*** | | | |
| Family size | 0.00246 | 0.00598 | -0.00566 |
| Max educ. in HH | 0.0186** | 0.00722 | 0.0355*** |
| % of labor force male | -0.0506 | -0.0745 | -0.0578 |
| Dependency share | 0.0336 | -0.0220 | 0.135 |
| Credit constrained (1,0) | -0.0580 | -0.155** | 0.0244 |
| Indigenous head (1,0) | -0.0571 | -0.103 | -0.0233 |
| Female head (1,0) | -0.0387 | -0.0848 | -0.0276 |
| Age of head | -0.00104 | -0.00533* | 0.00113 |
| ***Welfare*** | | | |
| House (1,0) | -0.0461 | -0.0696 | -0.0405 |
| Concrete/brick house (1,0) | -0.158** | -0.142 | -0.187** |
| Refrigerator (1,0) | -0.148** | -0.204** | -0.143 |
| Access to water system (1,0) | -0.0993 | -0.313** | 0.000203 |
| Sewage (1,0) | -0.0916 | -0.0727 | -0.136 |
| Big farm animals (#) | -0.00228 | -0.00533 | 0.00223 |
| ***Social capital*** | | | |
| Ag. Ass. Membership 5 yrs.+ (1,0) | -0.0779 | -0.149 | -0.0629 |
| Non-Ag. Ass. Membership 5 yrs.+ (1,0) | -0.0800 | -0.0490 | -0.102 |
| ***Community variables*** | | | |
| Bus in community (1,0) | -0.160** | -0.177** | -0.147* |
| Elementary school (1,0) | 0.0458 | 0.158* | -0.0179 |
| Distance to closest city (km) | -0.00537** | -0.0114*** | 0.000951 |
| Chimborazo (1,0) | 0.0780 | 0.0534 | 0.153* |
| ***N*** | 495 | 326 | 340 |

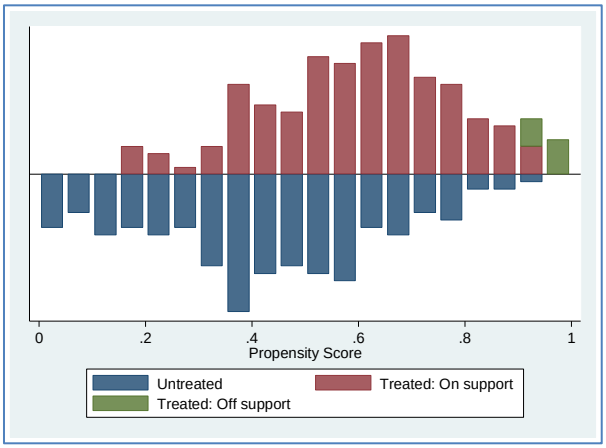
* $p<.10$; ** $p<.05$; *** $p<.01$

[7] Table A.1 in Appendix A reports the punctual test of means for the pooled sample following Leuven and Sianesi (2003).

**Panel 1: All Controls**




**Panel 2: Non-eligible**




**Panel 3: Non-participants**

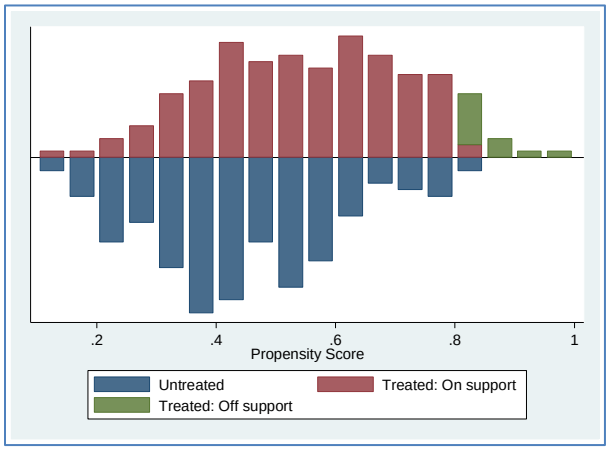



Figure 2.1: Common Support with Different Controls

Having identified the appropriate groups of control farmers, the next step is to determine if the conventional SPF should be run for the whole sample or if separate frontiers are necessary for beneficiary and control farmers. First, a pooled SPF is estimated that includes a binary variable for participation in *Plataformas.* Next, two separate SPF models, one for beneficiaries and a second one for control farmers, are estimated. The LR tests confirm that treated and control farmers display different technologies; thus, separate SPFs are preferable. These results are corroborated by the pooled models presented in tables 2.4 and 2.5 where the parameters for the variable Beneficiaries (beneficiary farmers) are positive and significant. Then, to correct for the

possible bias from unobservables, two separate SPF models are reestimated using Greene's (2010) selection correction framework.

The results of the SPF models are presented in table 2.4 for the unmatched samples, and in table 2.5 for the matched samples. All models presented in these two tables use the value of potatoes harvested per hectare (US$/ha.) as the dependent variable. Each table contains a total of nine models, divided into two groups, one denominated conventional SPF and the second sample selection corrected SPF. As expected, all estimated models present positive partial production elasticities; however, their magnitudes and statistical significance differ. In every model, expenditures on seeds make the greatest contribution to yields, followed by total expenditures on other inputs.[8] On the other hand, labor plays a minor role, particularly for beneficiaries and the non-eligible. The results on input expenditures are well in-line with those found by Kalirajan (1991) and Bravo-Ureta et al. (2012). The former argues that the cash required to buy inputs is one of the main production constraints for smallholders in developing countries. The mixed results for labor expenditures, where the coefficient is insignificant for beneficiaries and the non-eligible, are consistent with the Lewis (1954) model. This is the case because the farmers being studied have very limited amounts of land with a relatively abundant work force; thus, the marginal productivity of labor can be zero. In the case of beneficiaries, this might reflect the additional effort, in terms of labor used, put forth to ensure that the new technology is successful. On the other hand, the fact that the coefficient on labor is significant for the non-participants (in treated communities) may reflect their preference to diversify their labor allocation by using it outside potato production.

[8] This variable includes expenditures on insecticides, fungicides (preventative and curative), fertilizer (organic and chemical), tractors, and draft animals.

Table 2.4: Parameter Estimates for the Conventional and Sample Selection SPF Models: Unmatched Sample

| Variables | Conventional SPF | | | | | Sample Selection Corrected SPF | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Pooled** | **Benef [a]** | **All control** | **Non-eligible** | **Non-partic.** | **Benef [a]** | **All control** | **Non-eligible** | **Non-partic.** |
| β1 = Labor ($/ha.) | 0.160** | 0.063 | 0.180*** | 0.153 | 0.170* | 0.062 | 0.181*** | 0.147 | 0.188** |
| β2 = Seeds ($/ha.) | 0.560*** | 0.518*** | 0.599*** | 0.681*** | 0.550*** | 0.515*** | 0.599*** | 0.695*** | 0.532*** |
| β3 = O. Inputs ($/ha.) | 0.318*** | 0.436*** | 0.259*** | 0.197* | 0.341*** | 0.442*** | 0.257*** | 0.193** | 0.336*** |
| Hired Labor (0, 1) | 0.081 | -0.036 | 0.121* | 0.086 | 0.131 | -0.023 | 0.121* | 0.086 | 0.122 |
| Chimborazo | -0.416*** | -0.386* | -0.460*** | -0.404** | -0.604*** | -0.386** | -0.479*** | -0.426*** | -0.584*** |
| Beneficiaries | 0.140* | - | - | - | - | - | - | - | - |
| Controls in PC | -0.001 | - | - | - | - | - | - | - | - |
| Altitude | -0.001** | -0.001 | -0.001* | -0.001 | -0.001* | -0.001* | -0.001** | -0.001** | -0.001** |
| Constant | 2.671*** | 3.535*** | 2.434*** | 2.740*** | 2.663*** | 3.241** | 2.443*** | 2.864*** | 2.747*** |
| γ | 2.192*** | 2.658*** | 1.861*** | 1.667*** | 2.279*** | - | - | - | - |
| L. Likelihood | -441.745 | -173.378 | -259.873 | -130.035 | -125.636 | -316.064 | -316.064 | -244.745 | -242.025 |
| σ(u) | - | - | - | - | - | 0.967*** | 0.653*** | 0.646*** | 0.721*** |
| σ(v) | - | - | - | - | - | 0.385** | 0.388*** | 0.425*** | 0.355** |
| ρ(w,v) | - | - | - | - | - | 0.446 | -0.106 | -0.238 | -0.592 |
| ***N*** | 495 | 171 | 324 | 155 | 169 | 171 | 324 | 155 | 169 |

[a] Estimates from beneficiary sample using all controls for the matching procedure. The other two sets of estimates for beneficiaries show similar estimates and where omitted from this table due to space limitation.

* $p<.10$; ** $p<.05$; *** $p<.01$

Table 2.5: Parameter Estimates for the Conventional and Sample Selection SPF Models: Matched Sample

| Variables | Conventional SPF | | | | | Sample Selection Corrected SPF | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Pooled** | **Beneficiaries[a]** | **All controls** | **Non-eligible** | **Non-partic.** | **Beneficiaries [a]** | **All controls** | **Non-eligible** | **Non-partic.** |
| β1 = Labor ($/ha.) | 0.164*** | 0.074 | 0.180*** | 0.153 | 0.170* | 0.067 | 0.180*** | 0.147 | 0.194** |
| β2 = Seeds ($/ha.) | 0.550*** | 0.495*** | 0.599*** | 0.681*** | 0.550*** | 0.494*** | 0.609*** | 0.687*** | 0.527*** |
| β3 = Oth. Inputs ($/ha.) | 0.322*** | 0.447*** | 0.259*** | 0.197* | 0.341*** | 0.452*** | 0.247*** | 0.196** | 0.328*** |
| Hired Labor (0, 1) | 0.08 | -0.047 | 0.121* | 0.086 | 0.131 | -0.037 | 0.126** | 0.080 | 0.128 |
| Chimborazo | -0.4076*** | -0.367* | -0.460*** | -0.404** | -0.604*** | -0.343* | -0.48979*** | -0.421*** | -0.580*** |
| Beneficiaries | 0.153* | | | | | - | - | - | - |
| Controls in PC | -0.00056 | | | | | - | - | - | - |
| Altitude | -0.00032*** | -0.00045 | -0.0002* | -0.0003 | -0.0003* | -0.0004 | -0.0002*** | -0.0003** | -0.0003** |
| Constant | 2.636*** | 3.494*** | 2.434*** | 2.740*** | 2.663*** | 3.119** | 2.466*** | 2.804*** | 2.715*** |
| RTS | 1.04 | 1.03 | 1.04 | 1.03 | 1.06 | 1.01 | 1.04 | 1.03 | 1.05 |
| γ | 2.208*** | 2.644*** | 1.861*** | 1.667*** | 2.279*** | - | - | - | - |
| L. Likelihood | -434.999 | -166.324 | -259.872 | -130.035 | -125.636 | -343.465 | -396.058 | -236.020 | -239.961 |
| σ(u) | - | - | - | - | - | 1.0190*** | 0.621*** | 0.677*** | 0.642*** |
| σ(v) | - | - | - | - | - | 0.387* | 0.403*** | 0.405*** | 0.398*** |
| ρ(w,v) | - | - | - | - | - | 0.325 | -0.139 | -0.020 | -0.603 |
| ***N*** | 488 | 164 | 324 | 155 | 169 | 164 | 324 | 155 | 169 |

[a] Estimates from beneficiary sample using all controls for the matching procedure. The other two sets of estimates for beneficiaries show similar estimates and were omitted from this table due to space limitation

* $p<.10$; ** $p<.05$; *** $p<.01$

Given that expenditures on seeds and other inputs make the greatest contribution to farm production, one way to interpret these results is that since many of these farmers spend very little on inputs, the marginal returns to these expenditures are very high. The values for $\gamma$, also reported at the end of tables 2.4 and 2.5 are used to test whether inefficiency is statistically significant. As shown in the tables, the null hypothesis that $\gamma = 0$ is rejected, which reveals that technical inefficiency is indeed an important contributor to yield variability (Coelli et al., 2005).

Turning to the results obtained from the matching procedure along with Greene's (2010) sample-selection correction method, we see that these are very similar to those obtained with the conventional SPF method. The results of the selection equation are presented in the Appendix A table A.2, which show that only the parameters for the age variables are significant.

Moreover, when looking at the results for $\rho$ for the various models, there is no statistical support for selection bias. One implication of these results is that the rigorous process followed to define the counterfactual groups, including the PSM before estimating the SPF models, have eliminated significant biases from unobservables.

The results for average TE are presented in table 2.6 separately depending on the control group used along with tests for the difference in means.[9]

[9] Average TE is computed for each group of farmers with respect to their respective frontier and these means are then compared across groups. Therefore, these comparisons are not absolute but relative to each group's own frontier. In other words, it is not possible to say which group or which specific farmer exhibits higher or lower productivity for the overall sample. However, it is possible to investigate the effect of selectivity bias on TE within each group and then discuss how close each group, on average, is to its own frontier and thus make relative productivity/efficiency statements.

Table 2.6: Technical Efficiency Levels

| Control Group | Sample and Method | Pooled | Beneficiaries | All Controls | Test of Means |
|---|---|---|---|---|---|
| **All** | **Unmatched** | | | | |
| | Pooled | 0.57 | 0.55 | 0.57 | |
| | Separate | - | 0.51 | 0.61 | *** |
| | Sample Selection | - | 0.51 | 0.62 | *** |
| | **Matched** | | | | |
| | Pooled | 0.57 | 0.55 | 0.57 | |
| | Separate | - | 0.51 | 0.61 | *** |
| | Sample Selection | - | 0.50 | 0.61 | *** |
| **Non-eligible** | **Unmatched** | | | | |
| | Pooled | 0.55 | 0.54 | 0.55 | |
| | Separate | - | 0.51 | 0.61 | *** |
| | Sample Selection | - | 0.52 | 0.62 | *** |
| | **Matched** | | | | |
| | Pooled | 0.55 | 0.54 | 0.55 | |
| | Separate | - | 0.52 | 0.61 | *** |
| | Sample Selection | - | 0.51 | 0.61 | *** |
| **Non-partic.** | **Unmatched** | | | | |
| | Pooled | 0.55 | 0.54 | 0.57 | |
| | Separate | - | 0.51 | 0.60 | *** |
| | Sample Selection | - | 0.50 | 0.60 | *** |
| | **Matched** | | | | |
| | Pooled | 0.56 | 0.54 | 0.57 | |
| | Separate | - | 0.52 | 0.60 | *** |
| | Sample Selection | - | 0.51 | 0.57 | ** |

* $p<.10$; ** $p<.05$; *** $p<.01$

The average TE for the entire unmatched sample is slightly higher than those found in other similar studies for small-scale farmers in Central America (Bravo-Ureta et al., 2007). More specifically, in all pooled models, the difference in TE between beneficiaries and controls is negligible and non-significant. On the other hand, in all separate models beneficiaries exhibit lower levels of TE than control farmers independent of the control group used for comparison. Another point to highlight is that average TE decreases for beneficiaries when going from the

pooled to the separate model, and then to the separate with sample selection models, although the magnitude of the change is small. On the other hand, the opposite is true for most cases for the controls for which generally TE increases when going from the pooled to the separate, and from the separate to the sample correction models. In sum, the main result is that, independent of the group or method used for comparisons, beneficiary farmers exhibit lower levels of average TE. That is, beneficiary farmers tend to use their resources in a less efficient way than the control group with respect to their respective technologies.

To explore the issue of TE further, we focus on the matched sample and estimate TE levels for three different groups of beneficiaries based on the number of years they had participated in *Plataformas* at the time the data was collected: 1) those who had recently joined *Plataformas* (at least one year); 2) those who had been in the program for up to two years; and 3) those that had participated for three or more years.[10] We then conduct a test of means across groups by comparing the TE levels of farmers that had been in the *Plataformas* for up to three years against the other two groups, and against all controls. The results, presented in table 2.7, show that beneficiaries who had been in the program for only one year exhibit the lowest levels of average TE (0.49) compared to those who had been in the program for three years or more (0.57), and this difference is statistically significant. Similarly, beneficiaries who had been in the program for up to two years also exhibit lower levels of TE (0.50), although slightly higher than the first group. Finally, when comparing all controls against beneficiary farmers who had been in the program for three years or more, while there are differences between their levels of TE (0.61 versus 0.57), these are not statistically significant. One way to interpret these results is that farmers who had applied the new technology package for one year or less experienced a drop in

[10] We thank an anonymous reviewer for suggesting this additional analysis.

their efficiency, but TE levels started to pick up by the second year, and by the third year they got closer to the level exhibited by the counterfactual group.

Table 2.7: Technical Efficiency Levels by Years of Participation in *Plataformas*

| Length of Time in Plataformas | | | | | | |
|---|---|---|---|---|---|---|
| **3 or more yrs.** | **At least 1 yr.** | ***Test*** [a] | **At least 2 yrs.** | ***Test*** [a] | **All Controls** | ***Test*** † |
| 0.57 | 0.49 | ** | 0.50 | ** | 0.61 | |

[a] Tests are for differences in means with respect to treated farmers with 3+ years in the *Plataformas*
** $p<.05$

The results of table 2.6 show that on average beneficiary farmers are less efficient than control farmers, when compared to their own frontiers, and now we are interested in examining which of the two groups (beneficiaries versus controls) has higher value of yields after controlling for biases from observed and unobserved characteristics. These comparisons, shown in table 2.8 along with a test of the difference in means, are based on the average predicted frontier output obtained for beneficiaries and controls, using their respective frontiers, after applying both PSM and Greene's (2010) selection correction.[11]

Table. 2.8. Predicted Frontier Value of Output per Hectare after Sample Bias Correction at the Mean of the Data

| **Sample** | **Beneficiaries** [b] | **All Controls** | ***Test*** [a] | **Non-eligible** | ***Test*** | **Non-particip.** | ***Test*** |
|---|---|---|---|---|---|---|---|
| **Unmatched** | 1689 | 940 | *** | 946 | *** | 1185 | *** |
| **Matched** | 1738 | 954 | *** | 859 | *** | 856 | *** |

[a] Tests are for differences in means with respect to treated farmers
[b] Average frontier output for the 3 matched samples
*** $p<.01$

These results show that, on average, beneficiary farmers are significantly more productive than control farmers and this finding is consistent for all comparisons made. Thus, an

[11] The results of the matched conventional SPFA model are very similar. We present results using the matched sample with the sample correction since it is more rigorous than the conventional model.

important implication of these results is that participation in *Plataformas* has contributed to a technological advantage for participants *vis a vis* non-participants—that is, the program induced TC. This finding is compatible with previous studies that have looked at the impact of *Plataformas* on farm output using very different conceptual frameworks (Cavatassi et al., 2011a; Cavatassi et al., 2011b).

Overall, the results lead to the rejection of both of the null hypotheses of this study: (1) equality of yields between beneficiary and control farmers; and (2) equality in TE for both groups. However, the rejection of the second hypothesis is opposite to what was anticipated. More precisely, beneficiaries of *Plataformas* have significantly higher yields, but are significantly less efficient. Therefore, beneficiaries operate on a higher production frontier than controls due to the technology transferred by the program; but, on average, they are further away from their frontier compared to controls. However, TE levels rise with the length of participation in the *Plataformas*. An initial lower level of TE for beneficiaries is generally consistent with Schultz (1975) who postulates that traditional farmers, when faced with a production shock, such as the introduction of a new technology, could experience an initial drop in productivity.

## 2.5 Conclusions

While there is growing interest in conducting impact evaluations in the agricultural sector, most of the related literature focuses on output indicators, such as increases in yields, total value of product, profits, gross margins, etc., while attention to managerial performance is rare. Focusing on TE, as an indicator of managerial performance, can be a significant addition to impact evaluation studies of agricultural projects. This paper uses data from small-scale potato farmers in Ecuador to disentangle TE measures from TC indicators by combining impact evaluation tools with the SPF approach. A matched group of beneficiaries and control farmers is generated using PSM techniques to correct for biases based on observed characteristics. In

addition, this paper deals with possible self-selection arising from unobserved characteristics using Greene's (2010) selection correction method for stochastic frontiers. The analysis presented in this study, does not exhibit clear indication of selection biases. For instance, the PSM only dropped, on average, 10 out of 495 observations,[12] while Greene's (2010) method does not reveal clear evidence of selection bias from unobservables. These findings suggest that the rigorous process implemented in data collection and in constructing the counterfactual groups was able to mitigate the presence of biases from both observable and unobservable variables.

The analysis presented in this paper suggests that while beneficiary farmers have significantly improved their performance in terms of TC indicators, their levels of TE are lower than those of control farmers. Thus, while *Plataformas* had a positive and significant impact on inducing higher yields, the results suggest that the implementation of the new technology package had a cost in terms of lower managerial performance (TE), at least in the short run.

It is important to put these results in context. Given that *Plataformas* beneficiaries were taught an array of new procedures to implement throughout the production cycle (starting with selecting the right type of seeds, turning the land, applying IPM techniques, etc.), it is very likely that beneficiaries had to spend additional effort to master their newly acquired skills. Moreover, at the time of the interview, 30% of farmers had been with *Plataformas* for only one year, another 30% had been in the program for two years, and the average number of years for the entire sample of beneficiaries was 2.3 years. This means that for about one third of the beneficiaries, their most recent harvest was the first time they had a chance to implement their newly acquired technology. Thus, a significant share of beneficiaries had limited exposure and not enough time to fully incorporate their new know how from learning by doing. Indeed, our

[12] The models drop 7 using all controls; 9 using non-treated communities; and 13 using treated communities only.

results do show losses in TE but only during the first years of participation, while there is evidence that TE starts to pick up the longer farmers are exposed to the new technology package.

These results are in line with economic theory (Schultz, 1975), and are compatible with those reported in related studies that have looked at TE for farmers that have recently entered into new niche markets. For instance, Sipiläinen and Lansink (2005) found that organic dairy farms experienced a decrease in TE when they first converted into organic production. Moreover, these authors estimated that TE started to increase only after 6 years from the switch and concluded that learning by doing as well as experience acquired over time are likely to be important factors in the efficiency of organic production.

Our results do suggest the need to conduct follow up evaluations after a project is completed to address longer term impacts. Such evaluations would make it possible to assess the sustainability of the project and also to examine if the managerial performance of beneficiaries improves as a result of the project compared to that of the control group. If the TE of beneficiaries continues to lag, in relation to the controls, it would be important to understand why that might be the case.

# CHAPTER 3

# EVALUATION OF THE PROGRESS OUT OF POVERTY INDEX (PPI)

## 3.1 Overview

The elimination of poverty is an important goal for developed and developing countries and for the international development community in general. Stewart, Laderchi and Saith (2007) note that in official discourse—for instance, by the World Bank and major funders—almost every policy is currently assessed in relation to its impact on poverty. In fact, since early in this century there has been a worldwide heightened commitment to reach the poorest of the poor. For instance, the United Nations created the Millennium Declaration in 2000 with the ultimate goal of halving the share of world's people whose income is less than one dollar a day between 1990 and 2015 (Maes and Vekaria 2008). Likewise, the same year and in response to advocacy campaigns, the U.S. Congress passed the Microenterprise for Self-Reliance Act, which mandates that one-half of all U.S. Agency for International Development (USAID) microenterprise funds must benefit "very poor" people, defined as those living on less than US$1 a day or those among the bottom 50 percent of people living below a specific country's poverty line (Maes 2006). These efforts have gained further commitments by the microfinance industry. The Microcredit Summit Campaign, for example, has adapted two major goals to be achieved by 2015: reaching 175 million of the world's poorest families, especially women; and ensuring that 100 million families rise above the US$1-per day (Harris 2007).

Yet, the success of antipoverty programs, in general, and the success in reaching the poorest of the poor, is contingent on first successfully identifying those that are poor.

In the context of limited financial resources and with the goal of accurately identifying those most in need, antipoverty programs have to rely on some form of targeting mechanism to move the program closer to the ideal targeting solution of providing benefits only to the poor and

to all of the poor (Besley and Kanbur 1993). Due to incomplete information on household income or wellbeing, these programs usually identify potential beneficiaries through a combination of targeting procedures, such as means tests (Besley 1989), proxy means tests (Ahmed and Bouis, 2002), geographical targeting (Park, Wang, and Wu, 2002), community-based targeting (Yamauchi 2010), or self-selection (Besley and Coate 1988; Dreze 1986; Alderman 1987, as cited in Besley and Kanbur 1993).

While governments, major development banks and agencies may have the resources to undertake one or a combination of the above mentioned targeting methods to identify the poor (or their target population), this can be costly. Zeller et al. (2006) explain that "there is a lack at present of low-cost and reliable methods for assessing whether policy reaches the poorest" (p.446). This can be particularly problematic for pro-poor development partners in the global south, especially small NGOs, who lack the funds and institutional capacity to undertake such technical and costly endeavors. In fact, not having a means to measure poverty outreach may be costly from the funding side since most pro-poor development partners typically rely on donor funding which may be contingent on whether they are truly reaching the poor.

Given this scenario, there has been a move towards improving poverty assessment tools (Imp-Act 2003), as well as a move to develop low-cost, simple, and reliable methods to identify the poor (SEEP 2008). The "collective" shift to improve or develop new poverty assessment tools is exemplified by several projects undertaken by major development and research entities in the fields of microfinance and development. For instance, to satisfy the congressional requirement mentioned above, USAID contracted the IRIS center at the University of Maryland "to develop simple, low-cost quantitative poverty assessment tools for measuring the prevalence of extreme poverty among clients of microfinance and microenterprise programs" (SEEP 2008,

p. 184). Concurrent to the development of IRIS's PAT, the Grameen Foundation (GF) undertook a similar initiative to create a simple poverty assessment tool with funding from the Consultative Group to Assist the Poor (CGAP) and the Ford Foundation (SEEP 2008). This tool, called Progress Out of Poverty Index (PPI), was developed by Mark Schreiner of Mircrofinance Risk Management, L.L.C. Similarly, the International Food Policy Research Institute (IFPRI), with the technical and financial support from CGAP, developed an operational tool, called PAT, designed to assess the poverty level of project beneficiaries in relation to the general population in an intervention area in order to provide transparency of poverty focus (Zeller, et al. 2001).

The industry is dominated by IRIS-USAID's PAT and the Grameen Bank's PPI, since they both provide absolute measures of poverty outreach, whereas IFPRI-CGAP's PAT provides relative measures of poverty outreach, which cannot be used for comparisons. The Grameen Foundation's PPI, and IRIS-USAID's PAT are the two poverty assessment tools more widely used around the world. The Grameen Foundation's PPIs, for instance, are available for over 50 countries, representing poverty measurement tools for the countries that are home to 90 percent of people in the world who fall under $1.25/day 2005 PPP (Progress Out of Poverty, 2014).[13] Similarly, IRIS's PATs are available for over 30 countries, covering Asia, Europe and Eurasia, Latin America and the Caribbean, the Middle East and North Africa, and Sub-Saharan Africa.[14]

Given that the PPI is the most widely used poverty assessment tool around the world, the objective of this paper is to evaluate the PPI in terms of its levels of bias, precision, and targeting accuracy. This will be done using Peru as a case study with data from the 2007 National

[13] Specific information on these countries and the PPI can be found here: http://www.progressoutofpoverty.org/ppi-country

[14] Specific information on these countries and PATs can be found here: http://www.povertytools.org/tools.html Two of USAID's PATs (for the Philippines and Haiti) were developed in consultation with the Grameen Foundation.

Household Survey on Living Standards and Poverty (ENAHO for its Spanish acronym). As far as this author is aware, up-to-date there has not been a rigorous quantitative independent study that assesses the accuracy of the PPI.[15] Importantly, and as noted by Boucher, Summerlin, and Martinez (2010), it is highly desirable to have a poverty assessment tool that has as "little bias as possible while having a high degree of precision" (p. 12). As a very important first step, this paper will try to replicate the results presented in Schreiner (2009) for the PPI for Peru. This paper argues that this first step is essential before looking at any measures of accuracy, precision, or targeting. Moreover, this paper will provide some suggestions on alternative ways to measure bias and targeting accuracy, which it is argued, might be more informative and relevant than those presented in the PPI documentation. Finally, this paper will provide some suggestions on how to, perhaps, improve the tool's accuracy and relevance by accounting for rurality and other regional variables.

With this in mind, the remainder of the paper is structured as follows. Section 3.2 describes the process of creating the PPI. Section 3.3 presents the empirical approach to evaluate the PPI. Section 3.4 describes the replication process and provides results of this exercise. Section 3.5 introduces suggestions on alternative ways to assess the accuracy of the PPI. Sections 3.6 through 3.9 make the bulk of the evaluation of the PPI. In Section 3.6, the original PPI is applied to urban and rural areas separately to see if the tool is more accurate in one setting over the other. Section.3.7 makes minor but important modifications to the original PPI in an attempt to make it more accurate and relevant for its intended use. This is done by first incorporating a variable to capture rurality, and then by adding regional variables. Section 3.8 goes a step further and it creates two new scorecards—one for rural areas and one for urban

[15] While there has been a report by Boucher et al., (2010) that compares IRIS-USAID's PAT and the PPI in terms of accuracy and user satisfaction, this is a qualitative comparison based on very few observations.

areas. All along and whenever relevant, the four new suggested ways to evaluate the tool, as described in section 3.5, will be used throughout this section. Section 3.9 provides a discussion of the evaluation results, it discusses lessons learned, makes recommendations, and concludes.

### 3.2 PPI: Construction, Calibration, and Validation

This section provides a general description of the PPI in terms of the construction of the poverty scorecard, the association of scores with poverty likelihoods, and how the tool measures accuracy, with particular reference to the documentation for Peru's 2007 PPI. Key features of the PPI and the more relevant for this paper, as well as important innovations introduced by the developer of the PPI will also be highlighted. It should be noted that the PPI for every country is different, and its construction and evaluation is very well and extensively documented at Progress Out of Poverty, and Microfinance Risk Management, L.L.C.'s websites.[16] A big share of this section is closely based on Schreiner (2009, p. 1, pp. 29-34; and 2012, pp. 19-50), unless other sources are cited.

The PPI evaluated in this paper uses Peru's 2007 ENAHO, which is representative at the national, rural, urban, departmental, and regional level, to construct a poverty scorecard that estimates the likelihood that a household has expenditures below a given poverty line.[17] The scorecard uses 10 indicators that field workers can easily and quickly collect and verify. The poverty scorecard can be used by pro-poor programs to monitor poverty rates at a point in time, track changes in poverty rates over time, and target services.

[16] For more information on the PPI, see the extensive documentation at Progress Out of Poverty (http://www.progressoutofpoverty.org/) and http://www.microfinance.com/

[17] The PPI for Peru using ENAHO 2007 provides the results for eight different poverty lines: national poverty line, 150% and 200% of the national poverty, national food line, USAID's 'Extreme' poverty line, and international poverty lines using 2005 PPP for $1.25/day, $2.50/day, and $3.75/day.

The process of creating the PPI involves splitting the data into three random sub-samples[18]: 1. the construction sub-sample is used to select the indicators (explanatory variables used to predict poverty status) and points (the scores from the regression) to construct the poverty scorecard; 2. The calibration sub-sample is used for associating scores with poverty likelihoods; and 3. The validation sub-sample is used to assess the accuracy of the results.

*PPI: Construction*

Focusing on the construction sample, the scorecard is built using Peru's national poverty line and Logit regression. The PPI uses a set of more than 100 potential indicators that are highly predictive of poverty status.[19] These variables fall within the general areas of household composition, education, housing, and ownership of durables.[20]

One important innovation introduced by the developer of the PPI is that the selection of the variables relies not only on a statistical algorithm to select the final set of variables, but it also relies on judgment on the part of the researcher, as well as field testing and user reviews (POP, 2011a; 2011b). For the statistical part, this involves using an algorithm similar to the common $R^2$-based stepwise least-squares regression. The accuracy of each scorecard is measured using the concordance statistic (c-statistic) or area under the receiver operating characteristics curve (ROC), which is also referred to as area under the curve (AUC). This process starts with a one variable scorecard and it stops once the best 10 variable scorecard has been identified. The c-statistic, ROC, or AUC, is the most commonly used performance measure to indicate the discriminate ability of generalized linear regression models (Steyerbert et al.

[18] The new Peru PPI randomly splits the data into two sub-samples: 1. The first is for construction and calibration; and 2. The second one is for validation for measuring accuracy.

[19] The 2007 Peru PPI uses 150 potential indicators, while the 2010 Peru PPI uses 110.

[20] More recently, the new Peru PPI for 2010 also incorporates agricultural variables and receipt of social transfers.

2010) and it is widely used in the fields of medicine (Austin and Steyerberg 2012), finance, atmospheric science, machine learning (Gönen 2003), as well as credit scoring (Baesens et al. 2003; Blanco et al. 2013).

Austin and Steyerberg (2002) explain that the c-statistic is a unitless index that denotes the probability that a randomly selected subject who experienced a particular outcome (in this case a household being poor) will have a higher predicted probability of having the outcome occur as compared to a randomly selected subject who did not experience the event (in this case a household not being poor). Under this approach, we can calculate the c-statistic by taking all possible pairs of households consisting of one household who is below the poverty line and one household who is above the poverty line. The c-statistic, then, is the proportion of such pairs in which the household who is poor has a higher predicted probability of being poor. In this regard, the c-statistic can be used as a rank-order statistic for predictions against true outcomes (Steyerbert et al. 2010). The higher the c-statistic is, the higher the predictive ability of the scorecard.

As for the reliance on judgment on the part of the researcher, the researcher considers the likelihood of acceptance by the users (face validity, simplicity, and cost of collection), the likelihood of changing as poverty status changes, verifiability and susceptibility to strategic falsification, improvements in accuracy, and variety among indicators. As such, judgment and statistics are combined at every step of the process (Schreiner et al., 2004; Zeller 2004; cited in Schreiner (2006)). Finally, beyond the reliance on judgment, each new PPI is also tested in the field during the development stages to gauge accuracy of scorecard collection and ease of use (POP, 2011a; 2011b). Similarly, the PPI also undergoes a process of user reviews, where local users provide feedback on a draft version of the PPI scorecard (POP, 2011b).

Going back to the construction of the scorecard, the final step involves transforming the logit coefficients from the final 10 indicators into non-negative integers. Each of these 10 questions (or indicators) in the final scorecard, and respective possible values (categories) or answers, is associated with specific points. These points are then added to create the total score. The maximum possible score is 100, while the lowest possible score is 0. The lower the score the higher the likelihood that the household is poor, while the higher the score the higher the likelihood that the household is not poor.

*PPI: Calibration*

The second sample is used for calibration where scores are associated with poverty likelihoods. This involves scoring every household using the poverty scorecard and points created in the first sample, and then identifying the share of households with a given score who are below the poverty line. The poverty likelihood for each score, then, is obtained simply by dividing the number of poor households in each score by the number of total households with the same score.

*PPI: Validation*

The third sample is used to assess the accuracy of the scorecard. In general, this is done using three different measures: 1. Bias of poverty rates (difference between estimated and true poverty rates at a point in time), and precision (using confidence intervals) using the bootstrap; 2. Bias of changes in poverty rates (difference between estimated change and true change over time), and precision (using confidence intervals) using the bootstrap; and 3. Targeting accuracy.

The PPI uses a resampling method called bootstrap to measure how accurate are the estimates of households' poverty likelihoods obtained from the calibration sample. This method was introduced in this industry by the developer of the PPI, and it has now been adopted by

IRIS-USAID's PATs (IRIS-USAID 2011), as well as by other researchers (Johannsen 2006). Although this technique has been used in the for-profit field of credit-scoring, as well as an array of other fields, based on this review, it had not been used for poverty scorecards.

Bootstrapping was introduced by Efron (1979). Efron and Tibshirani (1986) note that a typical problem in applied statistics involves the estimation of an unknown parameter $\theta$, and that researchers primarily want to answer two main questions: (1) what estimator of $\hat{\theta}$ should be used? and (2) having chosen to use a particular $\hat{\theta}$, how accurate is it as an estimator of $\theta$. The bootstrap is a general methodology for answering the second question (Ibid.). Under this framework, the PPI has chosen the probability likelihoods as its preferred estimator $\hat{\theta}$ and it uses the bootstrap to measure how accurate it is as an estimator of poverty rates, i.e., $\theta$.

The bootstrap is a non-parametric approach that relies on the assumption that the sample at hand is representative of the population of interest (Guan 2003). In the general case of Peru, we have a population distribution $\mathcal{F}$ and we are interested in estimating the (parameter) poverty rate $\theta = \theta(\mathcal{F})$. We have a sample of data *F*, from the ENAHO 2007, and obtain the estimate $\hat{\theta} = \theta(F)$. Let *N* denote the number of observations in *F*. In the simplest case, the bootstrap proceeds as follows. Using a random seed, we draw a random sample $F_i$ of size *N* with replacement from *F* and compute $\hat{\theta} = \theta(F_i)$ once. We then repeat this many times, obtaining the set of estimates $\{\hat{\theta}_i\}_{i=1}^{i=B}$, where *B* is a large numbers, such as 1,000, which is typically the accepted number of repetitions used in the literature (Cassell 2007). Under this framework, the empirical distribution function $\hat{F}$ is a non-parametric estimate of the population distribution $\mathcal{F}$, and from the sample dataset, the desired statistic $\hat{\theta}$ is calculated as an empirical estimate of the true parameter $\theta$. Stine (1989) notes that making use of numerous samples drawn from the initial observations, the bootstrap requires fewer assumptions and offer greater accuracy and

insight than do standard methods in many problems. In order to gauge the precision of the estimates, one can also build confidence intervals from the empirical distribution (Stine 1989, Cassell 2007).

More specifically on the PPI, the bootstrap is applied 1,000 times where each household is scored in the validation sample. For each bootstrap sample and score, the difference between the true poverty likelihood and the estimated poverty likelihood is obtained and reported. This provides the measure of bias for all scores and across the 1,000 bootstrap samples; and the overall level of bias is obtained when averaging out all the biases across bootstrap samples and scores.[21] Precision is then measured using two-sided intervals containing the central 900, 950, and 990 differences between estimated and true poverty likelihoods.

The PPI for Peru 2007 reports accuracy by measuring bias (the difference between the true poverty likelihoods and the estimated poverty likelihoods) and precision (using confidence intervals) at a point in time out of sample (validation sample), and out of time (using data for 2005 and 2006). It also reports bias and precision in changes in poverty rates using independent samples (2007 and 2005; 2007 and 2006) and using a panel (2006 and 2005). This is done for eight different poverty lines.[22]

Finally, in terms of targeting accuracy, the PPI documentation provides six different measures of targeting accuracy: inclusion, undercoverage, leakage, exclusion, total accuracy, and the Balance Poverty Accuracy Criterion (BPAC), which is used by IRIS-USAID's PATs.

[21] See table 3.4.4.1 for a more intuitive interpretation of how bias is obtained: it is simply the average bias for the bias of every score after applying the bootstrap.

[22] The PPI reviewed in this paper provides results for eight different poverty lines: national poverty line, 150% and 200% of the national poverty, national food line, USAID's 'Extreme' poverty line, and international poverty lines using 2005 PPP for $1.25/day, $2.50/day, and $3.75/day. The new Peru PPI provides results for 15 poverty lines, including legacy poverty lines (the eight poverty lines noted in the previous sentence) and the poverty lines using new updated definitions.

Additionally, the documentation includes four measures using 20 different cut-off points: % of all households who are targeted; % of targeted who are poor; % of poor who are targeted; and a ratio of poor households targeted per non-poor household targeted.

### 3.3 Empirical Approach

In an attempt to evaluate the accuracy of the PPI for Peru using data from the ENAHO 2007, a vital first step is to replicate the results presented in Schreiner (2009). Note that there are various kinds of replications: pure replications and scientific replications (Hamermesh 2007; Burman, Reed, and Alm 2010). The former is concerned with confirming that the results reported by authors are independently verifiable (Vinod 2005; Dewald, Thursby, and Anderson 1986), while the latter is concerned with examining if previous results are reliable, robust, and stable, with the use of different data from a different population, and may involve extensions of a model (Hamermesh 2007; Vinod 2005; Burman et al., 2010).

The empirical approach undertaken in this paper might be considered to be a hybrid of these. In Section 3.4, we attempt to replicate the results in Schreiner (2009) (pure replication). The objective of this first step is to verify the original results and to see if we uncover any errors the PPI might have made. This first step is crucial, since in Sections 3.6 through 3.8 we try to modify the model slightly to see if the results from the original PPI are reliable and stable in the sense that bias and targeting accuracy are not significantly changed. In order to accomplish these slight modifications, the pure replication is crucial since we have to be sure we are not introducing our own errors in the evaluation.

The scientific part of the replication (although with a lower case "s") aims to answer three important and inter-related questions. First, following previous research that looks at accuracy in non-nationally representative sub-groups (see for example Tarozzi and Deaton (2007) for the case of Mexico, and Elbers, Lanjouw, and Leite (2008) for the case of Brazil), we

apply the PPI to rural and urban areas to see if targeting accuracy and bias differ across these two settings. This analysis is undertaken because poverty scorecards like the PPI report accuracy measures for nationally representative samples; however, and as noted by Schreiner (2014), in practice, poverty scorecard users do not take such samples. This set of analysis, then, aims to answer the question of whether sub-group accuracy falls significantly or slightly, and if it does fall significantly, then it will be important to investigate if the tool can be modified to correct for this.

The second and third question that this paper aims to answer for the scientific part of the replication is also related to looking at the PPI's levels of accuracy and bias when considering rural and regional variables. The second set of analysis tests if adding rural and regional variables might also improve targeting accuracy and bias, while the third set of analysis goes a sept further and creates two poverty scorecards with the same indicators for rural and urban areas but with different points to see if targeting accuracy and bias can be improved.

It should be noted, however, that previous researchers have addressed this issue of trying to segment scorecards by urban and rural settings, as described in the PPI documentation.[23] In fact, Schreiner tested for this for the case of India and Mexico (Schreiner, 2006b and 2005a), while Narayan and Yoshida (2005) do this for Sri Lanka, and Grosh and Baker (1995) do this in Jamaica (as cited in Schreiner (2009)). The overall results suggest that segmenting scorecards by urban and rural areas does not improve accuracy much (Ibid.).

Yet, this has not been done for the case of Peru. Moreover, and as noted by Schreiner (2014), no tests of bias have been undertaken in the literature; although Schreiner has done this "extensively for India in unpublished work" (p. 5). Thus, this paper adds to the literature in this regard.

[23] We thank Schreiner for highlighting this oversight in a previous version of this paper.

As discussed above, the PPI documentation provides an extensive coverage of various ways in which the accuracy of the PPI is measured. Needless to say, it would be a colossal task to evaluate the PPI in every single measure in which the PPI assesses its accuracy. In order to select the most important indicators to gauge the accuracy of the PPI, it will be useful to recall how the PPI is intended to be used in practice (Schreiner 2007, p. 1): "The poverty scorecard is a practical way for pro-poor programs in Peru to monitor poverty rates, track changes in poverty rates over time, and target services."

As such, this paper will look at bias at a point in time. The results arising from this measure of accuracy, in principle, should also apply when looking at the bias in measuring changes in poverty rates over time. In terms of targeting services, this paper will evaluate targeting accuracy with the "concentration curve" (ROC or AUC), which is more or less universally accepted as the best way to visualize targeting power (Schreiner 2014, Baulch 2002)[24]. It should be noted that earlier poverty-scoring papers by Schreiner did include a ROC curve (see for example Schreiner (2006 for the case of Bangladesh), but this was later dropped from the documentation because Figure 14 in Schreiner seemed simpler to users (Schreiner 2014).

From a statistical point of view, an ideal targeting exercise would be able discriminate or identify all of the poor, and by definition would also identify all the non-poor. This would minimize errors of exclusion (of the poor), while also minimizing leakage or minimizing errors of inclusion (of the non-poor) (Cornia and Stweart 1995; Legovini 1999).

[24] We thank Schreiner for making this suggestion. This paper initially used the Targeting cut-off and the % of poor who are targeted as a measure of targeting power. However, Schreiner (2014) pointed out that the ROC is the standard to visualize targeting power and that these measures were in fact already provided in the PPI documentation.

However, in the real world, this is not possible as there are always tradeoffs between these two types of errors depending on the cut-offs used for targeting: a higher cut-off has better inclusion of the poor (but greater leakage to the non-poor), while a lower cut-off has better exclusion of the non-poor (but higher undercoverage of the poor).

In the classification literature, these trade-offs are referred to as sensitivity (the probability of true prediction given a positive outcome), and specificity (the probability of false prediction given a negative outcome) (Hermansen (2008). The great appeal of the ROC is that it plots the area under the curve, which shows these trade-offs in a graphical representation for all possible cut-offs as shown in Figure 3.3.1.

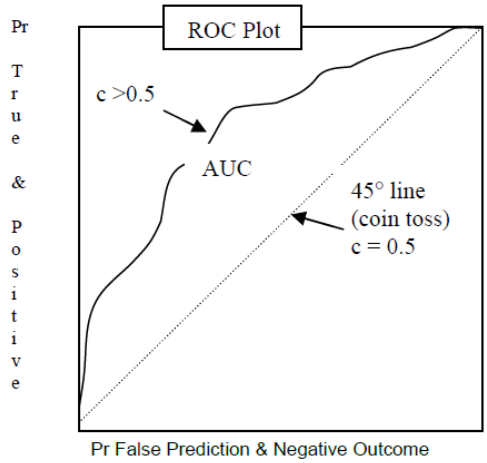



Figure 3.3.1. Illustration of a ROC Curve

*Source*: Hermansen (2008)

The diagonal line (45˚ line) represents pairs of sensitivity and specificity values that cancel each other out, which can be considered random targeting. Any point above the 45˚ line represents the probability of true predictions being higher than false predictions. As such, the greater the distance of a curve from the 45˚ line and the closer it is the north-west corner, the better the targeting accuracy of the predictive tool.

Importantly, unless we know the benefits and costs that the researcher attaches to the true and false predictions, i.e., how much we value inclusion, undercoverage, leakage, and exclusion, then no one point in the ROC plot represents a better choice than another. Yet, if every point in the ROC plot for a given model (call it A) lies above every point in the ROC plot of a second model (call it B), then model A dominates over B. In this regard, the ROC curves for alternative models in the same plot make it clear at which levels of sensitivity and specificity, one model dominates others (Hermansen 2008).

In this paper, the ROC is used for measuring the targeting power of each of model used in comparison with the original PPI applied to the national sample. However, this paper does not put any weights to the true and false predictions.

**3.4 Replication of Schreiner's (2009) PPI for Peru**

Thanks to the guidance provided by Schreiner, this paper successfully replicated the following key results: the exact values for every variable used in the creation of the poverty scorecard; the poverty scores (Figure 1 in Schreiner (2009)) for each question and respective values in the scorecard; the estimated poverty likelihoods associated with scores (Figures 4 and 5 in Schreiner (2009)); and the bootstrapped differences between estimated and true poverty likelihoods from the validation sample (Figure 7 in Schreiner (2009)). The confidence intervals we obtained for the bootstrapped differences from Figure 7, however, are slightly different: the replicate confidence intervals (for the 90-, 95-, and 99-percent CIs) are narrower. Perhaps this is due to the use of a different formula to obtain the confidence intervals.

There are three important reasons why the replication of these results is fundamental to the evaluation of the PPI. First, replicating the exact values of all the variables used in the creation of the scorecard gives credibility to this study, since this means we are not introducing our own errors in the analysis. Second, by having an identical dataset as the one used by the PPI

we can then move to replicate the poverty scorecard and respective scores. This is important because once we start off with the identical dataset and model used by the PPI, we can then move to make slight modifications to the scorecard—by accounting for rurality for example—to see if minor key changes can improve accuracy and relevance (in terms of bias and targeting accuracy). Finally, the process of replicating these results (exactly) gave us a thorough and deep understanding of the steps taken to create the poverty scorecard and the way accuracy is measured. A comprehensive understanding of these steps puts us in a better position to provide feedback on how the scorecard was created, and in how accuracy is being measured. Likewise, this also allows this study to provide suggestions on ways in which minor modifications might lead to more accurate and relevant results.

With this in mind, sub-section 3.4.1 describes how the replication of the data was done. Sub-section 3.4.2, provides the results of our replicates for the construction of the poverty score card, including the points for each question and respective values. Sub-section 3.4.3 provides the results of the replicates for the derivation of estimated poverty likelihoods associated with scores (calibration). Sub-section 3.4.4 provides the results of the replicates for the validation where we replicate the bootstrapped differences between estimated and true poverty likelihoods in order to capture the bias across scores and at the aggregate level (validation). Sub-section 3.4.5 provides a discussion of the key findings of the replications process.

### *3.4.1 Replication: General Description*

In attempting to replicate results of previous studies, researchers have encountered a host of problems (Dewald, Thursby, and Anderson, 1986; McCullough, McGeary, and Harrison, 2008; McCullough, 2007; Anderson, Greene, McCullough and Vinod, 2005). Differing results from an original study may arise due to differences in the variable creation (or transformation), coding (Herndon, Ash, and Pollin, 2013), the use of different data (or years used), the inappropriate use of weights (McCray, 2001; Herndon, et al., 2013), different software giving different results, 'bugs' in the software (Feldstein, 1974; Dewald et al., 1986), the lack of a clear code or program, inadequate documentation on variable creation and data sources, as well as human error. It is no surprise, then, that past comprehensive reviews in the economics literature show that very few studies have been successfully replicated, and in fact, replicable economic research is the exception rather than the rule (Anderson, et al., 2005; McCullough 2007).

Nevertheless, it has been argued, quite fittingly, that the confirmation of research findings through replication by peer researchers is a crucial part of the scientific method (Dewald, et al., 1986; Anderson, et al., 2005). While there might be some costs (on the part of the replicator) and risks (on the part of the researcher agreeing to have his/her work replicated) incurred in the process of replicating a study, Dewald, et al. (1986) and Anderson, et al. (2005) note some of the key benefits of doing so and conclude that the benefits certainly outweigh these costs. For example, the willingness of researchers to have their work replicated, and the successful replication of their work by others will improve the original author's reputation since they will be perceived as more transparent, while increasing the likelihood of being cited more frequently. Moreover, their work will be more likely used as a solid benchmark from where other researchers can build upon. A move to create a structure where more researchers are willing to

have their work replicated, while providing data, code, and/or documentation on how a study was undertaken will lead to a decrease in the frequency of errors and better economic research (Dewald, et al., 1986; Anderson et al., 2005).

With all this in mind, it might be fair to say that the most important step in the process of evaluating the PPI and its methodology is to first successfully replicate the results presented in Schreiner (2009). If we are able to reproduce the author's results, this will bring credibility not only to the author's findings, but to our own attempt to evaluate the PPI.

Recall that the poverty scorecard for Peru consists of 10 questions. Each of these questions, and respective possible values (categories) or answers, is associated with specific points. These points are then added to create the total score. The maximum possible score is 100, while the lowest possible score is 0. The lower the score the higher the likelihood that the household is poor, while the higher the score the higher the likelihood that the household is not poor.

Each of the 10 variables/questions in the scorecard was constructed to be ordinal in nature with the lowest value (0) capturing higher poverty levels or a higher likelihood of being poor, while the highest value is meant to capture less poverty, more wealth, or a lower probability of being poor. For example, Question 1 asks "How many household members are 17-years-old or younger?" There are 5 possible mutually exclusive values (or answers) the variable can take: 0 = Four or more; 1 = Three; 2 = Two; 3 = One; and 4 = None. The lowest value (0 = having four or more kids that are 17 or less) aims to capture households that are poorer with a higher dependency ratio, while the highest value (4 = no kids that are 17 years old or less) aims to capture the least poor households (lower dependency ratio) or a later stage in the life cycle. The same logic is applied in the construction of the 9 remaining variables.

Given that each variable and respective value (category or answer) yields the coefficients in the regression used in the construction of the scores, the accurate replication of these variables is crucial in the process of replicating the results. After several failed attempts, Schreiner provided invaluable guidance to finally replicating an identical dataset as the one he used. Guidance was provided on how to address "missing" values, and on the interpretation of how the education variable was to be created (Q. 2). Finally, we were provided with a data set that contained all the variables in the model so that we could verify that each of the variables we replicated were identical to the ones created by Schreiner (2009). Moreover, the dataset shared with us also identified the subsamples used in the construction of the scores, the calibration, and the validation. All of this significantly contributed to an accurate replication process by minimizing any possible errors that we may have introduced into the analysis. Table 3.4.1.1 summarizes the replication process.

Table 3.4.1.1: Summary of Replication Process for the 10 Questions in the Poverty Scorecard

| | I | II | III | IV |
|---|---|---|---|---|
| | Coded | Obs. Wrongly Coded | | Identical |
| | Correctly? | Replicate | PPI | After Corrected? |
| 1. How many household members are 17-years old or younger? | ✓ | 0 | 0 | ✓ |
| 2. What is the highest educational level that the female head/spouse completed? | × | 10, 270 | 0 | ✓ |
| 3. What is the main material of the floors? | × | 53 | 0 | ✓ |
| 4. What is the main material of the exterior walls? | × | 53 | 0 | ✓ |
| 5. Excluding bathrooms, kitchen, hallways, and garage, how many rooms does the residence have? | × | 53 | 0 | ✓ |
| 6. What fuel does the household most frequently use for cooking? | × | 53 | 1 | ✓^ |
| 7. Does the household have a refrigerator/freezer? | ✓ | 0 | 0 | ✓ |
| 8. How many color televisions does the household have? | × | 50 | 1 | ✓^ |
| 9. Does the household have a blender? | ✓ | 0 | 0 | ✓ |
| 10. Does the household have an iron? | ✓ | 0 | 0 | ✓ |

Notes: ^ = Can be considered as being "identical"; 1/18,934 coded differently does not change any result

Column I shows whether our original attempt at replicating the variables was successful or not: we were only successful in replicating 4 out of 10 variables identically. Column II shows the number of observations where we erred in the coding. In 5 out of the 6 wrongly coded variables, the error was relatively small: 53 out of 18,934 observations were coded incorrectly (less than a third of a percent). However, for question 2, the number of households that were

coded incorrectly was substantial: 10,270 out of 18,934 (54%). The primary error was in our interpretation of how to create the variable, which was taken from the poverty scorecard (Figure 1 in Schreiner 2009). Schreiner provided clarification on how to construct the variable.

On the other hand, the main reason the other variables were coded incorrectly was due to the number of households we used from the beginning of the variable creation. The original dataset contains 26, 527 households (*hogares*). However, only 18,934 households had a complete survey. This is summarized in table 3.4.1.2. Although Schreiner (2009) only uses the 18,934 households that completed the survey, he uses all 26,527 households to create the dwelling-related variables since multiple households (*hogares*) might live in the same dwelling (*vivienda*), and since information that applies to all the households, such as information on the main material of the dwelling, is only asked to the principal household and not to every household living in the dwelling (ENAHO 2007b). In other words, if a dwelling is home to 3 households, the enumerator asks questions pertaining to the dwelling only to the principal household, since the answer to these questions will be identical for the other 2 households living in the same dwelling.

Table 3.4.1.2: Breakdown of Survey Completion

Final Outcome of the Survey

| | Freq. | Percent |
|---|---|---|
| 1. Complete | 18,934 | 71.38 |
| 2. Incomplete | 3,270 | 12.33 |
| 3. Refused survey | 562 | 2.12 |
| 4. Not at home | 454 | 1.71 |
| 5. Dwelling is abandoned | 1,100 | 4.15 |
| 6. Other | 2,207 | 8.32 |
| | 26,527 | 100 |

*Source*: ENAHO, 2007 Data. (Author's translation)

Since we worked with only the 18,934 households, our initial variable creation had "missing" information on the dwelling for 53 households. Following Schreiner's guidance, we fetched this information from the "principal" household which did not have a complete survey

and which we initially dropped before creating the variables, but who had information on the dwelling.

The process of replicating these variables and the exchange with Schreiner was beneficial for two reasons. First, in the process of correcting our mistakes, we were able to identify if there were errors made by the PPI in the variable creation. It turns out that in the original PPI there were only two observations that were coded incorrectly, as shown in Column III of table 3.4.1.1. These were minor errors that in the overall analysis do not make any difference in the validity of the results presented in Schreiner (2009). Finally, after going over each observation that we coded incorrectly, we were much more confident that we were not introducing our own errors into the evaluation. After the corrections made in our code, column IV of table 3.4.1.1 shows that our final dataset is essentially identical to the one created by the PPI.

Before proceeding to the next section, it is important to briefly note that the entire process of creating the PPI as well as this evaluation only uses the 18,934 observations, while the original dataset contains information for 26,527 households, as noted. The results of the PPI and this evaluation are still valid since the documentation provided by National Institute for Information and Statistics (INEI for its Spanish acronym) explains that the sampling weights (*factor expansion*) for each household in the data set were re-calibrated in order to account for missing data, as well as for non-responses. As such, the sampling weights provided in the dataset have been adjusted so that the results of any analysis are taken to be representative at the national, rural, urban, departmental, and regional level (ENAHO 2007).

### *3.4.2 Replication: Construction of the Scorecard*

This section describes the steps taken to replicate the poverty scorecard, including the replication of points for each possible value of each question, as in Figure 1 of Schreiner (2009).

Having replicated (exactly) the variables used for the creation of the poverty scorecard, and having identified the sample for the construction of the poverty scorecard, the next step is to run the logit model using the construction sample with the dependent variable taking the value of 1 if the household is below the poverty line and 0 otherwise. This binary variable was already in the ENAHO dataset, and it is created from the "national" poverty line based on expenditures per capita. Note, however, that "national" is in quotes, since the actual poverty line is different across departments and across rural and urban areas since it is adjusted to control for price differences and demographic differences across departments and in rural and urban (INEI 2008). This is implied by Schreiner (2009, p. 10): "This paper focuses on the ―national poverty line, which adjusts the food line downwards for economies of size in the household (for example, because kitchen facilities are shared) and upwards to match the total food plus non-food expenditure observed for households who just meet their caloric needs (Instituto Nacional de Estadística e Información, 2006)". As for the right-hand side variables, these are the 10 questions in the scorecard.

Given that we are dealing with complex survey data that is representative at the national level (and other levels, see ENAHO (2007)), it is necessary to add the sampling weights in the regression. This was highlighted by Schreiner during our exchange, but it is not noted in the PPI documentation. Not including the sampling weights would provide a biased estimate of the coefficients in the regression model (Dargatz and Hill, 1996; Rodgers-Farmer and Davis, 2001). However, although the sampling design is done in multiple stages while incorporating stratification as well as clustering, given that we are interested in the coefficients of the

regression and not the standard errors, it is not necessary to also account for the sampling strategy when running the regression (Chromy and Abeyasekera, 2005).[25]

The results of the logit model using the sampling weights are presented in table 3.4.2.1. There are three important features of the methodology used to create the scorecard and scores that make it very intuitive to interpret, and which also capture the relationship between poverty and the variables and possible values, which should be highlighted. First, the possible values of each question are ordered such that the lowest value is associated with a higher likelihood of being poor, while the highest value is associated with a higher likelihood of being non-poor (or being less likely poor). As such, in the regression framework the lowest value (0) is used as the reference category. Thus, all other possible values (or categories) of the variable correspond to the probability of that value (or category) of being poor vis-à-vis the base or reference category. Second, the sign of all the coefficients are negative. Given the first and second points, the interpretation of the coefficients is that they are less likely to predict poverty in reference to the base category. In other words, the higher the negative value of the coefficient the lower the log odds are of being poor in relation to the base category. Third, given how the variables were created —for those that have more than one value or category—we can observe that the coefficients exhibit non-linearities (in most cases). In other words, for one variable the ability of one category of the variable to predicting poverty may increase non-linearly as the value of the

[25] We are not interested in the standard errors for two reasons. First, the final scorecard is selected based on the c-statistic, which measures the predictive capabilities of the scorecard, and which is not a function of the standard errors. Second, once the final scorecard has been selected, the PPI uses the values of the coefficients for all 10 variables in the model, independent on whether the coefficients are statistically significant or not. In this sense, we are only concerned with adding the sampling weights to make the results representative of the population. The only added benefit we would get if we applied the survey design into the regression is that the standard errors would change and the level of significance of the variables might change as well; however, the coefficients would be identical as the model that does not account for the survey design. Thus, we are not concerned with the standard errors since they play no role in the value of the coefficients of the logit, which are used to create the final scores/values.

category increases by one unit. (This will become clearer when we convert the coefficients to actual points.)

Table 3.4.2.1: Results of the Logit Using Sampling Weights

| Indicator | Ordinal Value/Category | Coeff. |
|---|---|---|
| 1. How many household members are 17-years old or younger? | A. Four or more | *BASE* |
| | B. Three | -0.69400 |
| | C. Two | -1.14825 |
| | D. One | -2.10048 |
| | E. None | -3.07247 |
| 2. What is the highest educational level that the female head/spouse completed? | A. None, pre-school, or kindergarten | *BASE* |
| | B. Grade school (incomplete) | -0.64333 |
| | C. Grade school (complete) | -0.94510 |
| | D. High School (incomplete) | -1.17008 |
| | E. High School (complete), non-university superior (incomplete) or no female head | -1.25957 |
| | F. Non-university superior (complete) or higher | -2.04062 |
| 3. What is the main material of the floors? | A. Earth, wood planks, other, or no residence | *BASE* |
| | B. Cement | -0.24241 |
| | D. Parquet, polished wood, linoleum, vinyl, tile, or similar | -1.95163 |
| 4. What is the main material of the exterior walls? | A. Adobe, mud or matting | *BASE* |
| | B. Wattle and daub, wood, brick or cement blocks, stone blocks with lime, other, or no residence | -0.32460 |
| 5. Excluding bathrooms, kitchen, hallways, and garage, how many rooms does the residence have? | A. One | *BASE* |
| | B. Two | -0.19666 |
| | C. Three, four or five | -0.68013 |
| | D. Six or more | -1.31926 |
| 6. What fuel does the household most frequently use for cooking? | A. Other | *BASE* |
| | B. Firewood, charcoal, or kerosene | -0.66799 |
| | C. Gas (LPG or natural) | -1.20480 |
| | D. Electricity or does not cook | -2.14362 |
| 7. Does the household have a refrigerator/freezer? | A. No | *BASE* |
| | B. Yes | -0.65676 |
| 8. How many color televisions does the household have? | A. None | *BASE* |
| | B. One | -0.33388 |
| | C. Two or more | -0.91292 |
| 9. Does the household have a blender? | A. No | *BASE* |
| | B. Yes | -0.43792 |
| 10. Does the household have an iron? | A. No | *BASE* |
| | B. Yes | -0.26491 |

*Source*: Author's estimation.

*Determining the "contribution" of each question (or variable) to predicting poverty*

In order to determine the points that each question is going to contribute to the poverty scorecard, the next step is to take the coefficient with the biggest absolute value for each variable

and to see their share of contribution out of the 10 questions. The contribution of each question to the scorecard is presented in table 3.4.2.2, below.

Table 3.4.2.2: The Contribution of Each Question/Variable to the Model

| Q# | Indicator | Coeff. |
|---|---|---|
| Q1. | # 17-year old or less | -3.07247 |
| Q2. | Education Fem Head/Spouse | -2.04062 |
| Q3. | Floor (mat.) | -1.95163 |
| Q4. | Wall (mat.) | -0.32460 |
| Q5. | # Rooms | -1.31926 |
| Q6. | Cooking Fuel | -2.14362 |
| Q7. | Refrigerator | -0.65676 |
| Q8. | # Color TV | -0.91292 |
| Q9. | Owns Blender | -0.43792 |
| Q10. | Owns Iron | -0.26491 |

*Source*: Author's estimation.

A more illustrative and intuitive representation of Table 3.4.2.2, is given in figure 3.4.2.1.[26]

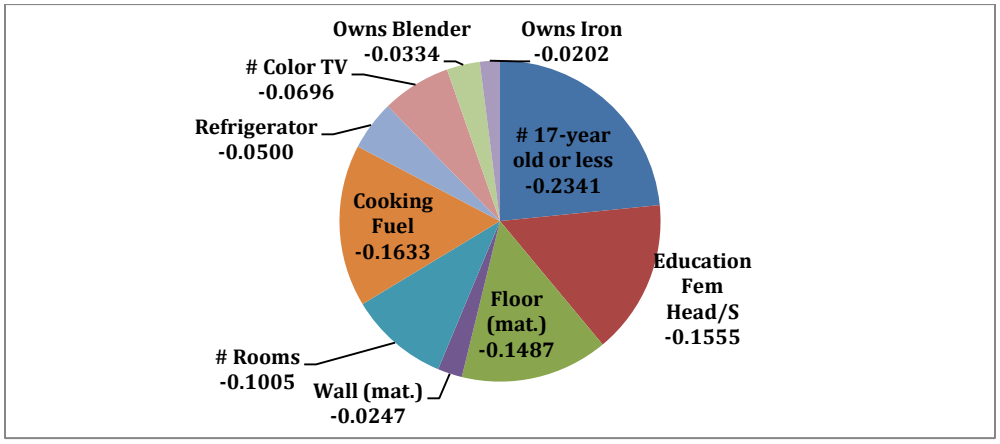



Figure 3.4.2.1. Breakdown of the Contribution of Each Question/Variable to the Model

*Note*: The percentages are derived from first adding all the coefficients in table 3.4.2.2, and then obtaining the share for each question. All percentages are negative since all the coefficients are also negative in table 3.4.2.2.

From figure 3.4.2.1 we can see that the demographic variable for the number of household members that are 17-years old or less makes the greatest contribution to the poverty

[26] If we were interested in measuring the relative importance of each indicator or response option for any indicator, Schreiner suggests "to measure scorecard's accuracy with-versus-without an indicator (or with-versus-without a response option). The larger the change in accuracy, the more important the item" (Schreiner 2014, p. 9).

scorecard, followed by the education of the female head, the type of fuel used for cooking, the type of material primarily used in the floor, and the total number of rooms in the household.

Once we identified the maximum contribution to the scorecard for each question, the next step is to determine the contribution of the value(s) (or categories) of each question. This is done in relation to the base reference (contribution of 0) and in relation to the category with the highest value (and coefficient), as shown in table 3.4.2.2 and figure 3.4.2.1. This is summarized in table 3.4.2.3.

Table 3.4.2.3: Obtaining Contribution of Each Question and Respective Answer to the Model

| | | A | B | C | D | E |
|---|---|---|---|---|---|---|
| **Indicator** | **Ordinal Value/Category** | **Coeff.** | **Max. Coeff.** | **A/B** | **B's Contribution** | **C*D** |
| 1. How many household members are 17-years old or younger? | A. Four or more | *BASE* | -3.07247 | 0 | 23.41 | 0 |
| | B. Three | -0.69400 | -3.07247 | 0.22588 | 23.41 | 5 |
| | C. Two | -1.14825 | -3.07247 | 0.37372 | 23.41 | 9 |
| | D. One | -2.10048 | -3.07247 | 0.68365 | 23.41 | 16 |
| | E. None | -3.07247 | -3.07247 | 1.00000 | 23.41 | **23** |
| 2. What is the highest educational level that the female head/spouse completed? | A. None, pre-school, or kindergarten | *BASE* | -2.04062 | 0 | 15.55 | 0 |
| | B. Grade school (incomplete) | -0.64333 | -2.04062 | 0.31526 | 15.55 | 5 |
| | C. Grade school (complete) | -0.94510 | -2.04062 | 0.46314 | 15.55 | 7 |
| | D. High School (incomplete) | -1.17008 | -2.04062 | 0.57339 | 15.55 | 9 |
| | E. High School (complete), non-university superior (incomplete) or no female head | -1.25957 | -2.04062 | 0.61725 | 15.55 | 10 |
| | F. Non-university superior (complete) or higher | -2.04062 | -2.04062 | 1.00000 | 15.55 | 16 |
| 3. What is the main material of the floors? | A. Earth, wood planks, other, or no residence | *BASE* | -1.95163 | 0 | 14.87 | 0 |
| | B. Cement | -0.24241 | -1.95163 | 0.12421 | 14.87 | 2 |
| | D. Parquet, polished wood, linoleum, vinyl, tile, or similar | -1.95163 | -1.95163 | 1.00000 | 14.87 | 15 |
| 4. What is the main material of the exterior walls? | A. Adobe, mud or matting | *BASE* | -0.32460 | 0 | 2.47 | 0 |
| | B. Wattle and daub, wood, brick or cement blocks, stone blocks with lime, other, or no residence | -0.32460 | -0.32460 | 1.00000 | 2.47 | 2 |
| 5. Excluding bathrooms, kitchen, hallways, and garage, how many rooms does the residence have? | A. One | *BASE* | -1.31926 | 0 | 10.05 | 0 |
| | B. Two | -0.19666 | -1.31926 | 0.14907 | 10.05 | 1 |
| | C. Three, four or five | -0.68013 | -1.31926 | 0.51554 | 10.05 | 5 |
| | D. Six or more | -1.31926 | -1.31926 | 1.00000 | 10.05 | 10 |
| 6. What fuel does the household most frequently use for cooking? | A. Other | *BASE* | -2.14362 | 0 | 16.33 | 0 |
| | B. Firewood, charcoal, or kerosene | -0.66799 | -2.14362 | 0.31162 | 16.33 | 5 |
| | C. Gas (LPG or natural) | -1.20480 | -2.14362 | 0.56204 | 16.33 | 9 |
| | D. Electricity or does not cook | -2.14362 | -2.14362 | 1.00000 | 16.33 | 16 |
| 7. Does the household have a refrigerator/freezer? | A. No | *BASE* | -0.65676 | 0 | 5 | 0 |
| | B. Yes | -0.65676 | -0.65676 | 1.00000 | 5 | 5 |
| 8. How many color televisions does the household have? | A. None | *BASE* | -0.91292 | 0 | 6.96 | 0 |
| | B. One | -0.33388 | -0.91292 | 0.36572 | 6.96 | 3 |
| | C. Two or more | -0.91292 | -0.91292 | 1.00000 | 6.96 | 7 |
| 9. Does the household have a blender? | A. No | *BASE* | -0.43792 | 0 | 3.34 | 0 |
| | B. Yes | -0.43792 | -0.43792 | 1.00000 | 3.34 | 3 |
| 10. Does the household have an iron? | A. No | *BASE* | -0.26491 | 0 | 2.02 | 0 |
| | B. Yes | -0.26491 | -0.26491 | 1.00000 | 2.02 | 2 |

Table 3.4.2.3 summarizes the entire process to replicate the scorecard and the points for each question and respective categories. Column A provides the coefficients from the

regression, column B identifies the coefficient with the highest (absolute) value (least associated with poverty), column C gives the share of the contribution of each value of each question in the model (A/B) in relation to the highest possible value (B), column D presents the contribution of each question to the model (obtained from figure 3.4.2.1), and column E gives the product of C and D, which gives the final points for each value of each question.

Table 3.4.2.4, compares the poverty scorecard created by Schreiner (2009) with our replicate scorecard, and it shows that the original scores and the replicates are identical, with the exception of one value for question 1. Perhaps this is due to rounding differences, which might be necessary in order to obtain a maximum score of 100.[27] As noted, the lower the score, the higher the likelihood of being poor and the higher the score, the less is the likelihood of being poor.

Now that we have replicated the points, the non-linearities noted above can be more easily identified. For example, looking at Question 2 for the highest level of education, we see that for each additional level completed or attended, the points increase in a non-linear fashion: from 0 to 5, from 5 to 7, from 7 to 9, from 9 to 10, and from 10 to 16. Similar non-linear patterns can be seen for all the questions that have more than one possible value.

Having successfully replicated the poverty scorecard, let us turn to the replication of the calibration—associating scores with poverty likelihoods.

[27] Schreiner (2014, p. 11) confirmed that he added a point to option E of Question 1 so that the maximum score is 100. He further notes that "[A]dding it here (a more-or-less rare response option, and the one with the most points) has a smaller effect on accuracy than adding it elsewhere." Clearly, this minor modification does not affect any of the results.

Table 3.4.2.4: Replicate of Figure 1 in Schreiner (2009): A Simple Poverty Scorecard for Peru

| Indicator | Value | PPI 2007 Points | Replicate Points |
|---|---|---|---|
| 1. How many household members are 17-years old or younger? | A. Four or more | 0 | 0 |
| | B. Three | 5 | 5 |
| | C. Two | 9 | 9 |
| | D. One | 16 | 16 |
| | E. None | 24 | 23 |
| 2. What is the highest educational level that the female head/spouse completed? | A. None, pre-school, or kindergarten | 0 | 0 |
| | B. Grade school (incomplete) | 5 | 5 |
| | C. Grade school (complete) | 7 | 7 |
| | D. High School (incomplete) | 9 | 9 |
| | E. High School (complete), non-university superior (incomplete) or no female head | 10 | 10 |
| | F. Non-university superior (complete) or higher | 16 | 16 |
| 3. What is the main material of the floors? | A. Earth, wood planks, other, or no residence | 0 | 0 |
| | B. Cement | 2 | 2 |
| | D. Parquet, polished wood, linoleum, vinyl, tile, or similar | 15 | 15 |
| 4. What is the main material of the exterior walls? | A. Adobe, mud or matting | 0 | 0 |
| | B. Wattle and daub, wood, brick or cement blocks, stone blocks with lime, other, or no residence | 2 | 2 |
| 5. Excluding bathrooms, kitchen, hallways, and garage, how many rooms does the residence have? | A. One | 0 | 0 |
| | B. Two | 1 | 1 |
| | C. Three, four or five | 5 | 5 |
| | D. Six or more | 10 | 10 |
| 6. What fuel does the household most frequently use for cooking? | A. Other | 0 | 0 |
| | B. Firewood, charcoal, or kerosene | 5 | 5 |
| | C. Gas (LPG or natural) | 9 | 9 |
| | D. Electricity or does not cook | 16 | 16 |
| 7. Does the household have a refrigerator/freezer? | A. No | 0 | 0 |
| | B. Yes | 5 | 5 |
| 8. How many color televisions does the household have? | A. None | 0 | 0 |
| | B. One | 3 | 3 |
| | C. Two or more | 7 | 7 |
| 9. Does the household have a blender? | A. No | 0 | 0 |
| | B. Yes | 3 | 3 |
| 10. Does the household have an iron? | A. No | 0 | 0 |
| | B. Yes | 2 | 2 |
| | **Total Score:** | | |
| | **Max Score** | **100** | **99** |
| | **Min Score** | **0** | **0** |

Only score that does not agree with original PPI scores.

### *3.4.3 Replicating Calibration: Associating Scores with Poverty Likelihood Estimates*

This section describes the steps taken to replicate the calibration process for the association of scores with poverty likelihoods, as in figure 4 and 5 in Schreiner (2009).

Having replicated (exactly) the points for each question in the poverty scorecard, and having identified the sample for the calibration, the next step is to associate scores with poverty likelihoods. This involves scoring each household in the calibration sample and noting for each score the percentage of households that are poor. As with the regression framework, the sampling weights are needed to identify the share of households that are poor for any given score.

Table 3.4.3.1 provides the results of the replication of the association of scores with poverty likelihoods—panel A provides the replicates of this paper and panel B provides the results from figure 5 in Schreiner (2009). We can see that the replicates and the original estimates of poverty likelihoods are essentially identical with some very minor differences in the thousandths place for three scores that is most likely due to rounding. Schreiner (2009) normalizes the poverty likelihood estimates so that all households sum to 100,000, but this is simply to illustrate the approach implemented by the PPI.[28] On the actual estimate of the poverty likelihoods, Schreiner (2009) does not normalize, but even if normalizing was done the results should be the same.

There are three important things to highlight from the results of table 3.4.3.1. First, the results of the replicates confirm we have recreated the method accurately. Second, given the methodology used by the PPI, poverty likelihoods are a decreasing function of the scores, as expected. The lowest possible score is 0, which is associated with a poverty likelihood estimate of 100%. On the other extreme, the maximum possible score is 100, which is associated with a poverty likelihood estimate of 0. Finally, based on the data, there are no poor households with scores greater than 74, while the vast majority of poor households have a score of 30-34 or less.

[28] In an exchange with Schreiner, he notes that this is done in order to illustrate the method without the need to explain the sampling weights and in order to make it easier to understand.

This provides an indication that although we are interested in identifying the poverty rates and bias for the entire sample, it might be more relevant to look at these measures with a particular focus on the lower end of the scores, since it is plausible that few to none of microfinance beneficiaries might be found in the upper tail of scores. Let us now turn to the replication of the validation.

Table 3.4.3.1: Replicate of Figure 5 in Schreiner (2009): Derivation of Estimated Poverty Likelihoods Associated with Scores

| | **REPLICATE *** | | | **PPI ^** | | |
|---|---|---|---|---|---|---|
| | **A** | | | **B** | | |
| **Score *i*** | **HS below poverty line** | **All HHs at score** | **Poverty likelihood (estimated, %)** | **HS below poverty line** | **All HHs at score** | **Poverty likelihood (estimated, %)** |
| 0-4 | 2439.25 | 2439.25 | 1.000 | 181 | 181 | 1.000 |
| 5-9 | 23810.7 | 24719 | 0.963 | 1368 | 1420 | 0.963 |
| 10-14 | 70525.6 | 78408.6 | 0.899 | 3258 | 3622 | 0.900 |
| 15-19 | 107042 | 124625 | 0.859 | 5193 | 6046 | 0.859 |
| 20-24 | 104673 | 136966 | 0.764 | 5892 | 7710 | 0.764 |
| 25-29 | 89705 | 140282 | 0.639 | 5039 | 7879 | 0.640 |
| 30-34 | 90018.4 | 176321 | 0.511 | 5047 | 9882 | 0.511 |
| 35-39 | 63915.2 | 172544 | 0.370 | 3331 | 8999 | 0.370 |
| 40-44 | 44570.9 | 191160 | 0.233 | 2328 | 9982 | 0.233 |
| 45-49 | 25800.9 | 155241 | 0.166 | 1374 | 8265 | 0.166 |
| 50-54 | 13069.6 | 168705 | 0.077 | 641 | 8267 | 0.078 |
| 55-59 | 5258.31 | 124623 | 0.042 | 319 | 7570 | 0.042 |
| 60-64 | 1341.9 | 97790.8 | 0.014 | 73 | 5362 | 0.014 |
| 65-69 | 0 | 86718.4 | 0.000 | 0 | 4487 | 0.000 |
| 70-74 | 905.01 | 71352.6 | 0.013 | 37 | 2950 | 0.013 |
| 75-79 | 0 | 43151.1 | 0.000 | 0 | 2751 | 0.000 |
| 80-84 | 0 | 41376.2 | 0.000 | 0 | 2468 | 0.000 |
| 85-89 | 0 | 27086.2 | 0.000 | 0 | 1730 | 0.000 |
| 90-94 | 0 | 11367 | 0.000 | 0 | 382 | 0.000 |
| 95-100 | 0 | 1844.34 | 0.000 | 0 | 47 | 0.000 |
| **Totals** | 643,075.77 | 1,876,720.49 | **0.343** | 34,081 | 100,000 | **0.341** |

^ Figure 5 in Schreiner (2009) normalizes to 100,000 to simplify the illustration of this process.

* Uses the same calibration sample as Schreiner (2009).

### *3.4.4 Replicating Validation with a Focus on Bias*

This section describes the steps taken to replicate the validation of scores with a focus on bias, i.e., the difference between estimated and the true poverty likelihoods.

Following the steps described in Schreiner (2009, p. 32), and using the same validation sample, we score each household, then for each score we identify the share of households that are poor and we do this using the bootstrap method 1,000 times with $n$ = 16, 384. For the purpose of replicating the results as closely as possible as the original, we use the same sample used by Schreiner (2009), the same seed in the random generator, and the same size.[29]

After attempting and failing to do this in Stata 11, Schreiner provided guidance on how to do this in SAS. The main contribution in this regard was in the way to apply the sampling weights. Recall that the data comes from a complex sampling design where each household is associated with a sampling weight that is the inverse of the probability of inclusion in the sample. Thus, the sampling weights tell us how many households each household in the data set it represents. Thus, when doing the resampling (bootstrap), each household in the data has a probability of being selected into the sample proportional to the sampling weight. This assures an accurate resampling and an accurate representation of the true population.

Let us now turn to the results of the replicates for the bias, which is shown in table 3.4.4.1. Panel A compares the original results of the PPI for the difference between estimated and true poverty likelihoods for the 1,000 bootstraps and the replicates. The replicates are essentially identical to the original results with some very minor differences. This result corroborates the accuracy of the PPI when applied to the validation sample with an overall average bias of about a third of a percent (0.33). Yet, as noted by Schreiner (2009) given that the

[29] Although the validation sample has 6,287 observations, Schreiner uses $n$ = 16,384 for the bootstrapping. We use the same number of observations so that the replication is as identical as possible as the original results. We did the same analysis using $n$ = 6,287 and the results were very similar.

2007 validation sample "is representative of the same population as the data that was used to construct the scorecard and all the data comes from the same time frame, the scorecard estimators are unbiased and these differences are due to sampling variation…" (p. 6). Even though this is the case, the very small level of bias gives added credibility and validity to the methodology developed by Schreiner (2009). Given that the replicates of the bias are essentially identical to the original results, we continue to feel confident that we have replicated the methodology as accurate as it is possible.

Table 3.4.4.1: Replicate of Figure 7 in Schreiner (2009): Bootstrap Differences between Estimated and True Poverty Likelihoods, and Confidence Intervals *

| | A | | B | | C | | | D | |
|---|---|---|---|---|---|---|---|---|---|
| | Diff. | | 90% NCI (+/-) | | 95% NCI (+/-) | | | 99% NCI (+/-) | |
| **Score *i*** | **PPI** | **Replicate** | **PPI** | **Replicate †** | **PPI** | **Replicate †** | **Replicate SAS ^** | **PPI** | **Replicate †** |
| 0-4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5-9 | 0.5 | 0.5 | 2.2 | 2.0 | 2.6 | 2.4 | 2.4 | 3.3 | 3.1 |
| 10-14 | -5.4 | -5.4 | 3.4 | 1.4 | 3.5 | 1.6 | 1.7 | 3.7 | 2.2 |
| 15-19 | 6.9 | 7.0 | 2.5 | 2.4 | 2.8 | 2.9 | 2.8 | 3.8 | 3.8 |
| 20-24 | 7 | 7.1 | 2.3 | 2.5 | 2.7 | 3.0 | 2.9 | 3.5 | 3.9 |
| 25-29 | 5.7 | 5.6 | 2.7 | 2.5 | 3.3 | 3.0 | 3.1 | 4.1 | 4.0 |
| 30-34 | -2.9 | -2.9 | 2.7 | 2.3 | 2.8 | 2.7 | 2.8 | 3.6 | 3.6 |
| 35-39 | -8.4 | -8.4 | 5.3 | 2.5 | 5.6 | 2.9 | 3.0 | 6.0 | 3.9 |
| 40-44 | 0.8 | 0.9 | 2.0 | 2.0 | 2.4 | 2.3 | 2.3 | 3.0 | 3.1 |
| 45-49 | 1.3 | 1.3 | 1.8 | 2.0 | 2.3 | 2.3 | 2.3 | 2.8 | 3.1 |
| 50-54 | 3.4 | 3.4 | 0.9 | 1.0 | 1.1 | 1.2 | 1.1 | 1.4 | 1.5 |
| 55-59 | 0.2 | 0.2 | 1.0 | 1.0 | 1.2 | 1.2 | 1.2 | 1.6 | 1.6 |
| 60-64 | -1.9 | -1.9 | 1.5 | 1.2 | 1.7 | 1.4 | 1.4 | 2.0 | 1.8 |
| 65-69 | -2.1 | -2.1 | 1.5 | 1.0 | 1.6 | 1.2 | 1.1 | 1.8 | 1.5 |
| 70-74 | 1.3 | 1.3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 75-79 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 80-84 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 85-89 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 90-94 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 95-100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Averages** | 0.32 | 0.33 | 1.49 | 1.18 | 1.68 | 1.41 | 1.41 | 2.03 | 1.85 |
| **Diff. in Avgs.** | | 0.01 | | -0.31 | | | -0.27 | | -0.18 |

* Both use same validation sample, 1,000 bootstrap samples, and $n$ = 16,384

† C.I. obtained from the empirical distribution

^ C.I. obtained from SAS command - uses average of the lower and upper bounds for all bootstrap samples.

Moving to the rest of the replicates, panels B, C, and D provide normal confidence intervals for the bias for each score at the 90, 95, and 99% confidence intervals, respectively. Although the replicates of bias (Diff.) are essentially identical, the replicates for the confidence intervals are slightly different and in general are narrower than the original PPI results. All the confidence intervals, with the exception of one in panel C that uses a SAS command, are obtained using the following formula from the Stata 11 Base Reference Manual (StataCorp, 2009, pp 216-217):

$$\widehat{se} = \left\{ \frac{1}{k-1} \sum_{i=1}^{1} \left(\hat{\theta}_i - \bar{\theta}\right)^2 \right\}^{1/2}$$

Where $\hat{\theta}$ is the observed value of the statistic (the difference between estimated and the true poverty likelihood) calculated with the original dataset (the validation sample), $i = 1, 2, \cdots, k$ denotes the respective bootstrap samples, $\hat{\theta}_i$ is the value of the statistic from the $i$th bootstrap sample, and where

$$\bar{\theta} = \frac{1}{k} \sum_{i=1}^{k} \hat{\theta}_i$$

The confidence intervals with nominal rates of coverage $(1 - \alpha)$ are given according to the normal-approximation method as follows:

$$\left[\hat{\theta} - z_{1-\alpha/2}\widehat{se}, \hat{\theta} + z_{1-\alpha/2}\widehat{se}\right]$$

Where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)th$ quantile of the standard normal distribution.

The replicate of the C.I. at the 95% level using the SAS command is much closer to the C.I.s obtained manually using the empirical distribution than the ones provided in Schreiner

(2009).[30] It will be informative to investigate why these replicates differ from the original results. Given that confidence intervals give us a sense on the level of precision for each statistic, this is an important point to investigate. In particular, the narrower the confidence intervals are the higher the level of precision, thus, our results indicate that the precision of the bias for each score replicated is better when applying this formula, above, than what is reported in Schreiner (2009).[31]

### *3.4.5 Discussion: Replication*

Section 3.4.1 described the process of replicating the construction of the poverty scorecard, the calibration, and the validation. The replication process did not uncover any major errors. While there were two 'errors' in the way two variables were created (see table 3.4.1.1), these are inconsequential and have no impact on the accuracy of the PPI. The results of this exercise show that the original results presented in Schreiner (2009) and our replicates are essentially identical with some very minor differences that may be due to rounding. This is an important finding in its own right as previous replication projects have shown that errors in empirical economic research in published articles may be quite common, while the success rate of pure replication studies is very low (Dewald et al., 1986; McCullough et al. 2008; Herndon et al. 2013).

[30] The C.I. using SAS is done for the poverty likelihood for each score rather than for the difference between the estimated and true poverty likelihoods. Yet, given that for each score in each bootstrap sample the true poverty likelihood is subtracted from the estimated poverty likelihoods, each statistic is being rescaled identically. Thus, the C.I. for the poverty rate for each score is identical to the difference between the estimated and true poverty likelihood. This is corroborated when we obtain the confidence intervals manually using the empirical distribution assuming asymptotic normality and after applying the formula noted above.

[31] Schreiner (2014, p. 9) explained that the most likely candidate for these differences is that this paper assumes that the distribution is asymptotically Normal, while he uses "the empirical confidence intervals (which in general would be wider than those found assuming Normality)."

In general, the replication process has been very beneficial for two key reasons. First, the success of the replication gives credibility to our study since our replicates are essentially identical to the results presented by Schreiner (2009), which means we are not introducing our own errors that could have potentially led to biases in our analysis. Second, this exercise gave us a deep and thorough understanding of how the PPI is created, which puts us in a position to make suggestions on how to, perhaps, improve the PPI's accuracy and relevance by making minor but important changes without altering the methodology much and by maintaining its appeal for being simple but accurate. While this replication exercise was successful in overcoming the obstacles typically faced by researchers who attempt to replicate previous results (Dewald et al. 1986; McCullough et al. 2006; McCullough, 2007; Anderson et al. 2005), much of this success is due to Schreiner's openness and willingness to share his data, as well as his generous guidance in helping us solve some issues we had with our initial results. Importantly, rather than sending the code to recreate the PPI, or to perform the validation with the bootstrap, the guidance we received was more conceptual,[32] which then helped with our own coding of the analysis. This, we argue, brings more credibility to the PPI's methodology and to our exercise in replicating the PPI's results since there was no exchange of any codes.

In terms of the construction of the poverty scorecard, there are three key features of the PPI methodology that are worth noting. First, we argue that the 10 questions capture different dimensions of poverty. Although the dependent variable is constructed from a measure of expenditures per capita, the questions included in the poverty scorecard capture the relationship between this measure and other dimensions of wellbeing (or poverty). For example, question 1 captures poverty embedded in demographics: having a higher number of younger kids is associated with higher levels of poverty, a higher dependency ratio, and/or a phase in the life-

[32] And on interpretation of key steps and variable creation.

cycle of young couples (Arriagada, 2002; Jelin and Díaz-Muñoz, 2003; St. Bernard, 2003). Likewise, question 2 captures poverty or wealth from the dimension of education (or human capital): lower levels of education are closely linked to higher levels of poverty. Importantly, question 2 also captures female-headedness, which is also associated with higher levels of poverty. Questions 3, 4, and 5 capture poverty (or wealth) from the point of view of physical capital (or standard of living). Similarly, questions 6 through 10 capture poverty from the dimension of assets. Finally, one can argue that questions 3 ("What is the main material of the floors?") and 6 ("What fuel does the household most frequently use for cooking?") can also be used as proxies for measuring poverty from the dimension of health, especially for women and for younger kids. For instance, Cattaneo, Galiani, Gertler, et al. (2009) find that replacing dirt floors with cement significantly improves the health of young children as well as improving cognitive development. Similarly, there is growing evidence that the use of solid fuels for cooking and heating, such as the use of wood, increase the likelihood of mortality and morbidity (WHO, 2013).

At this point, it might be valuable to compare the PPI with a more conventional measure of multi-dimensional poverty, namely, the Multidimensional Poverty Index developed by the Oxford Poverty & Human Development Initiative (OPHI). An MPI typically includes the following dimensions of poverty: education, health, and standard of living. The dimensions, indicators, and weights associated with the MPI, however, do not come from a statistical methodology that can estimate the contribution of each of these indicators to poverty or the index. For instance, the MPI for Peru developed by OPHI (2013) includes 10 indicators in these 3 dimensions that cover many of the variables included in the PPI for Peru 2007. For example, for the dimension of education, the MPI includes years of schooling, and school attendance. For

the dimension of health, it includes child mortality and nutrition. Finally, for the standard of living, the MPI includes cooking fuel, sanitation, water, electricity, floor, and asset ownership. Undoubtedly, while the MPI has great appeal for looking at poverty from various important dimensions, the choices of dimensions, indicators, and weights for these, are selected more from a philosophical point of view of what the researcher thinks is important for measuring poverty. There are some that argue that this level of flexibility is one of the appealing features of this methodology (Alkire and Foster 2011), since it can be applied widely to different contexts, but with the subjective judgment of the researcher. Yet, these choices do not come from a statistical analysis, which might be one of the major short-comings of this approach. Moreover, and in contrast to the PPI, not all of the indicators of the MPI are applicable to all households, such as child mortality.[33]

One can argue that the PPI also captures different dimensions of poverty—in fact, many of the indicators from the MPI are the same ones chosen by the PPI for Peru 2007; yet, these do come from an objective measure of poverty (expenditures) and the "dimensions", indicators, and weights come from the identification of the statistical relationship between these variables and poverty based on expenditures. Moreover, the indicators used in the PPI are simple to collect and verify, while the indicators apply to all households. Given this brief discussion, we argue that the PPI can also be considered to be a multi-dimensional poverty index.

Coming back from our digression, we want to highlight two additional features of the construction of the PPI that are appealing. First, the way the data has been structured, each variable can have different possible values (or categories or values), but for each additional value of the variable, one is able to capture nonlinearities in the model. This might be vital when scoring a household and might increase the level of accuracy. In fact, this approach has been

[33] This was noted by Schreiner (2014).

recommended when the relationship between the dependent and explanatory variable does not have a linear relationship across the distribution of the variable: one unit change in the right-hand side variable does not always have the same effect on the dependent variable at different points of the variable (Farrington and Loeber 2000). Second, the PPI uses a statistical approach, as well as judgment, to select the final set of variables included in the model, as well as the weights (points) that eventually go into the final poverty scorecard. Part of this judgment involves including variables that can potentially change over time in order to facilitate measuring changes (or proxies of this) of poverty over time. For example, changes in the number of kids under 17-years old may capture lower dependency ratios and different (and less demanding) phases in the life-cycle of households. Likewise, changes in the material of the floor—from earth to concrete, for example—may also capture improvements in the standard of living of the household (as well as improvements in health (Cattaneo, et al. 2009). Finally, as far as this paper is aware, when using a regression framework for poverty scoring or targeting, no other approach transforms the coefficients into non-negative numbers as it is done by the PPI, other than exponentiating the coefficients using the logit formula. This might play a role in the adoption of the PPI since the scorer does not have to deal with decimal places, positive and negative numbers, or exponentiating values.

In terms of the calibration process, there are three important features to highlight, although most of them are related to the distribution of households across scores. First, poverty likelihoods are a decreasing function of scores: the lower the score, the higher the probability of being poor. This is a very simple and intuitive result that can facilitate the adoption and use of the PPI. Second, there are very few households with scores of 0, while there are essentially no poor households with scores greater than the 60-64 score mark. Finally, it seems that most of the

poor households can be identified by looking at the scores of 50-54 or below. This means that when we look at bias, we should pay particular attention to the lower scores, which might be more relevant for the profile of clients that pro-poor organizations serve. For instance, although the bias reported by the PPI gives us a measure of how far the tool is from predicting poverty rates, these poverty rates are for the entire sample, which is representative of all of Peru; yet, the vast majority of households that are served by pro-poor organizations will undoubtedly lie in the lower end of the distribution of scores, while very few to none would be found in the upper tail of the distribution of scores.

Finally, the validation process revealed three important findings. First, the overall level of bias for the entire sample is very small at around a third of a percent, which is great for what the PPI is primarily intended to be used for: to accurately identify poor households. Yet, it is important to recall that the validation is done from data that is representative of the population of Peru for when the data was collected, while the construction of the scorecard comes from a different sub-sample that is also representative of Peru for the same time. Yet, this is a very important result that should not be minimized. On the other hand, when looking at the level of bias for scores at 35-39 or lower, the levels of bias are at times pretty significant. For instance, at scores of 10-14 and 35-39, the bias is -5.4 and -8.4, respectively, which represent an underestimate of poverty. On the other hand, at scores 15-19 and 20-24, the levels of bias are 6.9 and 7, respectively, representing an overestimate of poverty. This fact should be kept in mind when we look at the levels of bias, precisely for the reasons noted in the previous paragraph.

While we are able to replicate the variables used to create the scorecard, the scorecard and scores, the association between scorecards and poverty likelihoods, and the difference between the estimated and the true poverty likelihoods using the bootstrap, the only important

difference in the validation process between the original results and the replicates is in the confidence intervals that aim to measure the precision of the bias. More specifically, the replicates show, on average, narrower confidence intervals across all scores and across all confidence levels. This implies that the precision of the PPI for Peru is actually better than what is presented in Schreiner (2009). The more likely candidate for this difference is, perhaps, a difference in the formula used to obtain the confidence intervals. The one used here is the one presented in the Stata 11 documentation (StataCorp, 2009, pp 216-217), which relies on the empirical distribution from the 1,000 bootstrap samples and the assumption of asymptotic normality. Given that our results differ from those of Schreiner (2009), the remainder of this paper does not provide any other results on precision.

### 3.5 A more Nuanced Analysis of Scoring

Having successfully replicated the PPI, and having highlighted the most important features and findings from this exercise, section 3.5 suggests a more nuanced analysis of the PPI's measures of bias and targeting accuracy.

Figure 3.5.1, which is taken from the original PPI (Figure 5 in Schreiner 2009) provides an intuitive representation of the PPI's scoring approach, and it highlights important features that can facilitate a more focused and relevant way in which to gauge the accuracy and relevance of the PPI by concentrating on the lower end of the distribution of scores.

As noted, the poverty likelihoods are a decreasing function of the scores. Figure 3.5.1 also plots a log-fitted line of the distribution of poverty likelihoods across scores. The higher the fitted line in the lower scores and the flatter its slope overall, the better the tool is at identifying poor households. This can be another way in which differences in targeting accuracy across the PPI and possible alternatives of a modified PPI can be assessed visually.

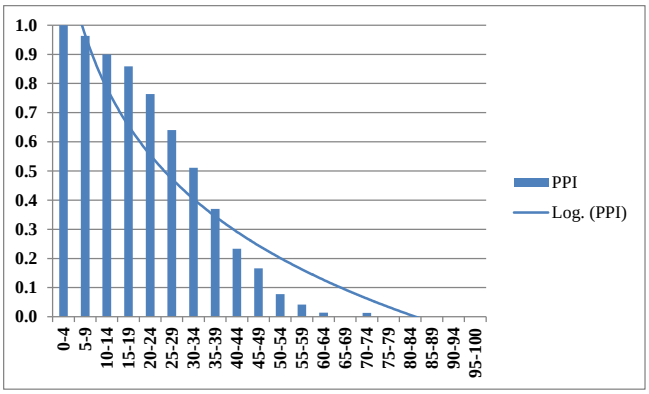


Figure 3.5.1: Distribution of Poverty likelihoods by Scores

For example, in section 3.6 when we modify the PPI scorecard (by adding rurality), we can plot the same distribution and fitted line to gauge which of the two scorecards yield a higher log-fitted line with a particular focus on the lower scores, where the poorest households should be scored. Indeed, given the way the PPI has been designed, one would expect the distribution of poor households to be skewed to the left. This is shown in figure 3.5.2. Importantly, figure 3.5.2 shows that, in percentage terms, very few poor households have scores over 54. Another helpful way to look at this distribution is by plotting the cumulative distribution of figure 3.5.2, as shown in figure 3.5.3.

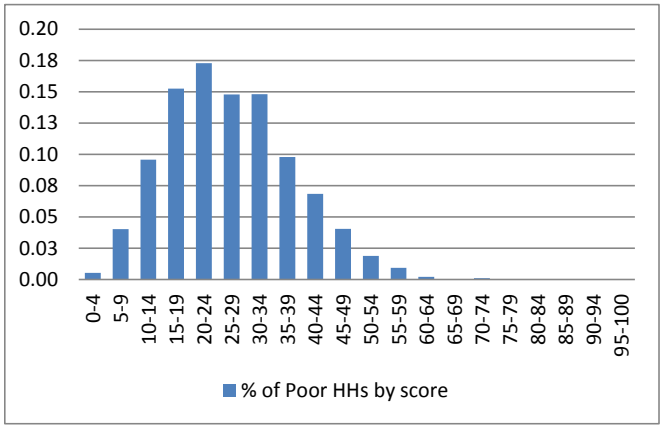


Figure 3.5.2: Distribution of % Poor HHs by Score Using Figure 5 in Schreiner (2009)

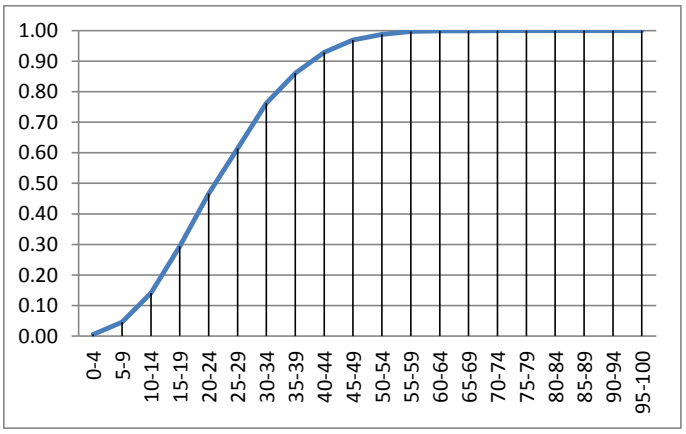


Figure 3.5.3: Cumulative Distribution of % of HHs by Score Using Figure 3.5.2

This simple exercise allows us to see that for the entire country, there is only a small percentage of poor households with the lowest possible scores, while there are essentially no poor households (in percentage terms) with scores greater than 50-54. Similarly, figure 3.5.3 shows that around 76% of poor households have a score at or below the 30-34 mark, while 86% of poor households have scores at or below the 35-39 mark. In fact, 99% of all poor households have a score of less than 55.

Looking at the cumulative distribution of the percentage of poor households that are identified by the PPI (figure 3.5.3) provides an additional important way in which to gauge the targeting accuracy and relevance of the PPI: the closer the slope to the northwest part of the graph the more accurate the tool and the faster it is at identifying poor households—we can call this the *speed of coverage* or SOC curve.

The *speed of coverage*, in fact, is simply a graphical representation of the percentage of poor households who are targeted at every score. Looking carefully at the PPI's documentation, one can see that figure 3.5.3 is the graphical representation of the first and third columns of Schreiner's (2009) Figure 14 for the National poverty line.[34] However, although the plotted

[34] The results presented in Schreiner (2009) are slightly different form this paper's, but this might be due to the rounding.

scores on the horizontal axis are ordinal, figure 3.5.3 does not tell us the distance—in terms of the percentile of scores in the entire sample for a given score—between scores. Thus, it is not possible to analyze the slope of the curve in an objective manner. Nevertheless, the *speed of coverage* gives us a sense of the percentage of households that are targeted at every score, which are essentially the inclusion rates for the poor.

One way to solve the issue of how to measure the slope of the *speed of coverage* curve is to use the "concentration curve", which is a variation of the speed of coverage and which yields the receiver operating characteristic (ROC) curve from where we can obtain the Area Under the Curve (AUC). [35] Recall that the ROC curve is a two-dimensional graphical illustration that plots the sensitivity on the Y-axis versus (1-specificity) on the X-axis for various values of a classification threshold (Baesens, Gestel, Viaene, et al., 2003), and it is the tool used in this paper to gauge the targeting power of the different models under evaluation.

In this sense the speed of coverage will complement the analysis using the ROC curve. The speed of coverage and the ROC curve, then, will be used to visually gauge how fast the PPI is identifying poor households: the former gives us a sense of how fast the tool is identifying poor households across the distribution of scores (in an ordinal scale), while the latter gives us a more objective measure of targeting power (in a cardinal scale). This will be useful in the subsequent sections when we compare some variations of the PPI that incorporate other key variables that may be helpful in making the tool more accurate and relevant. To be clear, the higher the ROC curve (the closer it is to the north-west corner) and the *speed of coverage,* the more accurate and relevant is the PPI in terms of targeting power.

[35] We thank Schreiner (2014) for making the suggestion to use the ROC curve in this paper.

Given these observations, three important comments are in order. First, although there have been some suggestions in the literature (Kidd and Wylde, 2011; IRIS 2005a; IRIS 2005b) that it might be beneficial to employ a two-step approach to better identify the poor, figures 3.5.2 and 3.5.3 show that this might be unnecessary: the tool, as is, is perfectly capable of identifying "all" the poor if one just concentrates on scores at or below the 50-54 mark. Second, given that most of the poor are identified in the lower tail of scores, one should expect that no poor households will be identified in the upper tail of the distribution of scores, as shown in table 3.4.3.1. This is an important point to keep in mind when assessing the accuracy of the tool by looking at the average of the bias (the difference between estimated poverty likelihoods with the true poverty likelihoods for the entire sample), as shown in table 3.4.4.1 of the previous section. Given that no poor households have scores greater than the 55-59 mark, one should expect the bias for the upper scores to be 0, as such, when averaging out the total bias (for all 20 scores), this will, by definition, yield a lower average of bias. This means that if we can look at bias just for households at, or below these scores, since we would be accounting for close to 100% of all the poor, then this focus might be warranted since looking primarily at all the possible poor households may be more relevant and informative for pro-poor organizations that primarily work with poor and very poor households. On the other hand, if one wanted to use the PPI to target a program for all of Peru, then looking at bias for the entire sample will be appropriate.

To illustrate more specifically, while the average bias is very small at around a third of a percent (0.3) for the entire validation sample, there are some big differences in bias, particularly for scores from 10 to 39, which contain the bulk of all poor households. Given this fact, the average of all the differences may mask the "true" level of bias or ***a more relevant level of bias*** if we only care about how accurate the tool is at identifying the typical profile of households that

are beneficiaries to pro-poor organizations. This is even more pertinent since the bulk of poor households have scores in this range (10 to 39) and since there are essentially no poor households (in terms of percentages) with scores greater than 59. As such, the average of the differences will be misleading. Of course, the bias is small, but only in the aggregate.

Figure 3.5.4 provides a graphical illustration of the bias taken from Schreiner (2009), (Figure 7 (National poverty line)). Figure 3.5.4 shows there are significant levels of bias in the lower scores, which is where the vast majority of poor households are scored. Although when we plot a log fitted line, this captures how the overestimates are cancelled out by the underestimates, yielding a fairly low and flat log-fitted line for the bias for the entire sample, which resembles the average of the overall bias at 0.32 shown in table 3.4.4.1.

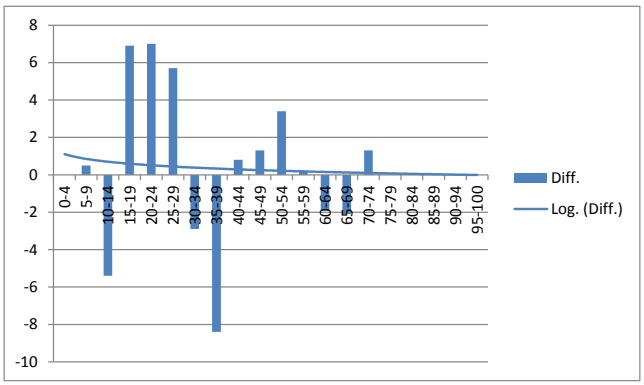


Figure 3.5.4: Bias (Diff. between Estimated Poverty Likelihoods and True Poverty Likelihoods)

We suggest two additional ways in which to also gauge the bias of the PPI, which might be more informative and relevant. The first is to provide the breakdown of the absolute bias,[36] along with the average of the bias, which is a simple modification of Table 3.4.4.1 (see Table 3.5.1a). The second way is to focus on the scores that cover 99% of all poor households by looking at the bias only at scores at or below the 50-54 mark (table 3.5.1b).

[36] Schreiner (2009) does not provide the absolute difference for the bias using the national poverty line, but he does note that "the average absolute difference is 0.8 percentage points across all eight lines" (p. 6).

These slight modifications to the way the results are presented may be more informative and relevant for using the PPI: we can see what is the level of absolute bias, which helps get a better sense on how far the tool is from its target, and we can focus our attention only on poor households, which is the group that we primarily care about. Yet, even with these modifications, in table 3.5.1.b one can see that the PPI, at least for the validation sample, is very accurate in terms of having a low level of average bias (0.81) still at under 1% for 99% of poor households, and a low level of absolute average bias at 3.85%.

Table 3.5.1a: Bias and Absolute Bias

| | PPI | |
|---|---|---|
| **Score *i*** | Diff. | \|Diff.\| |
| 0-4 | 0 | 0 |
| 5-9 | 0.5 | 0.5 |
| 10-14 | -5.4 | 5.4 |
| 15-19 | 6.9 | 6.9 |
| 20-24 | 7 | 7 |
| 25-29 | 5.7 | 5.7 |
| 30-34 | -2.9 | 2.9 |
| 35-39 | -8.4 | 8.4 |
| 40-44 | 0.8 | 0.8 |
| 45-49 | 1.3 | 1.3 |
| 50-54 | 3.4 | 3.4 |
| 55-59 | 0.2 | 0.2 |
| 60-64 | -1.9 | 1.9 |
| 65-69 | -2.1 | 2.1 |
| 70-74 | 1.3 | 1.3 |
| 75-79 | 0 | 0 |
| 80-84 | 0 | 0 |
| 85-89 | 0 | 0 |
| 90-94 | 0 | 0 |
| 95-100 | 0 | 0 |
| Avg. Diff. | 0.32 | 2.39 |

Table 3.5.1b: Bias and Absolute Bias for 99% of Poor Households

| | PPI | |
|---|---|---|
| **Score *i*** | Diff. | \|Diff.\| |
| 0-4 | 0 | 0 |
| 5-9 | 0.5 | 0.5 |
| 10-14 | -5.4 | 5.4 |
| 15-19 | 6.9 | 6.9 |
| 20-24 | 7 | 7 |
| 25-29 | 5.7 | 5.7 |
| 30-34 | -2.9 | 2.9 |
| 35-39 | -8.4 | 8.4 |
| 40-44 | 0.8 | 0.8 |
| 45-49 | 1.3 | 1.3 |
| 50-54 | 3.4 | 3.4 |
| Avg. Diff. | 0.81 | 3.85 |

This section provided a more nuanced analysis of how targeting accuracy can be measured, including a graphic representation of the distribution of the poverty likelihoods across scores (with a log-fitted line) (figure 3.5.2), a graphic representation of the *speed of coverage* which plots the cumulative distribution of the percentage of households identified by the scores (inclusion rates), which can be used for targeting purposes (figure 3.5.4). Similarly, this section suggested a complementary way to summarize the bias by also providing the absolute bias and

by focusing on the lower scores that identify 99% of poor households (tables 3.5.1a and 3.5.1b). These four additional ways to gauge the targeting accuracy and bias of the PPI will be helpful on the following sections that slightly modify the PPI in an attempt to improve its accuracy and relevance.

### 3.6 Application of the Original PPI to Rural and Urban Areas

The first set of analysis applies the original PPI to urban and rural areas by taking them as two independent samples. Recall the data is representative at these two levels. Thus, taking these two sub-samples as independent is justified. The main objective is to compare if the tool is equally as accurate in these two very different settings that are not nationally representative. There are three important reasons for undertaking this set of analysis. First, we would like to know if targeting accuracy and bias in these sub-groups falls by a lot or by very little. In the case that accuracy falls by a lot, it will be important to see if this can be remedied. Second, poverty is experienced very differently in rural and urban settings. Not only is the environment different, but so are the profiles of the poor who might have less education, be more likely to be indigenous, with different demographic characteristics, who might have very different sources for employment, have a different set of endowments, and whose dwellings and access to services might also be significantly different from their urban counterparts. Third, it might be likely that some microfinance institutions might primarily work in urban areas, while others might work primarily with rural beneficiaries. As such, if the PPI is (significantly) less accurate in either one of these settings, it might be desirable to consider creating two separate PPIs—one for rural areas and one for urban areas. Finally, the distribution of urban and rural areas may be quite different, as well as the distribution of the poor in these two settings. To put things in perspective, the ENAHO 2007 data shows that 37% of households live in rural areas, and while the poverty rate

for households is 34% for the nation, the vast majority of the poor live in rural areas: 57% of rural households are poor, compared to only 20% or urban households.[37]

Before proceeding with the results of the validation, it will be informative to look at the distribution of the 10 indicators used in the PPI scorecard across rural and urban areas to see if there are any major differences. These results are presented in table 3.6.1; column I presents the mean results for the nation (Pooled), columns II and III do this for rural and urban households, respectively, and column IV provides the difference between rural and urban areas. As expected, for some of these variables there are important differences between rural and urban areas (see column IV). More importantly, when looking at 7 of the final 10 indicators (questions 3, 4, 6, 7, 8, 9, and 10 [***the percentages have been bolded and italicized under*** column II]), we can see that households in rural areas seem to be fairly homogenous (at least based on these indicators)—either most of them have a particular characteristic or indicator or very few of them have it. For example, in rural areas: 86% have a dirt/earth floor (or similar) and only 1% have a parquet floor (or similar) (question 3); 72% have exterior walls made of adobe (or similar) (question 4); 64% use firewood (or similar) to cook, while only 3% use electricity (question 6); 94% do not own a refrigerator (question 7); 80% do not own a color television, while only 2% own two or more; 85% do not own a blender; and 86% do not own an iron. Since there is little variation in these variables across rural households, this means that when we use the logit to create the scores, this little variability will make it difficult for the model to discriminate between the poor and non-poor in rural areas.

Table 3.6.1. Distribution of 10 PPI Indicators: Pooled (Nation), Rural, Urban, and Difference

| | I | II | III | IV |
|---|---|---|---|---|

[37] These numbers are obtained from the construction sample. We should note that these poverty rates differ, slightly, from what is reported in INEI (2008) where poverty rates for rural areas are 64.6% and 25.7% for urban areas. However, INEI reports poverty rates based on population, whereas what we report here is based on households.

| Indicator | Value | Pooled | Rural | Urban | Diff. |
|---|---|---|---|---|---|
| 1. How many household members are 17-years old or younger? | A. Four or more | 10% | 17% | 6% | 11% |
| | B. Three | 12% | 14% | 11% | 3% |
| | C. Two | 21% | 19% | 23% | -3% |
| | D. One | 23% | 18% | 26% | -8% |
| | E. None | 34% | 32% | 35% | -3% |
| 2. What is the highest educational level that the female head/spouse completed? | A. None, pre-school, or kindergarten | 14% | 27% | 6% | 21% |
| | B. Grade school (incomplete) | 20% | 31% | 14% | 17% |
| | C. Grade school (complete) | 14% | 15% | 13% | 2% |
| | D. High School (incomplete) | 11% | 8% | 12% | -4% |
| | E. High School (complete), non-university superior (incomplete) or no female head | 29% | 17% | 37% | -19% |
| | F. Non-university superior (complete) or higher | 12% | 2% | 18% | -17% |
| 3. What is the main material of the floors? | A. Earth, wood planks, other, or no residence | 47% | ***86%*** | 24% | 61% |
| | B. Cement | 40% | ***14%*** | 56% | -42% |
| | D. Parquet, polished wood, linoleum, vinyl, tile, or similar | 13% | ***1%*** | 20% | -20% |
| 4. What is the main material of the exterior walls? | A. Adobe, mud or matting | 41% | ***72%*** | 23% | 49% |
| | B. Wattle and daub, wood, brick or cement blocks, stone blocks with lime, other, or no residence | 59% | 28% | 77% | -49% |
| 5. Excluding bathrooms, kitchen, hallways, and garage, how many rooms does the residence have? | A. One | 17% | 20% | 15% | 5% |
| | B. Two | 22% | 30% | 17% | 12% |
| | C. Three, four or five | 50% | 43% | 54% | -11% |
| | D. Six or more | 11% | 7% | 14% | -7% |
| 6. What fuel does the household most frequently use for cooking? | A. Other | 10% | 25% | 2% | 23% |
| | B. Firewood, charcoal, or kerosene | 37% | ***64%*** | 20% | 44% |
| | C. Gas (LPG or natural) | 48% | ***8%*** | 71% | -63% |
| | D. Electricity or does not cook | 5% | ***3%*** | 7% | -4% |
| 7. Does the household have a refrigerator/freezer? | A. No | 66% | ***94%*** | 50% | 44% |
| | B. Yes | 34% | ***6%*** | 50% | -44% |
| 8. How many color televisions does the household have? | A. None | 44% | ***80%*** | 23% | 56% |
| | B. One | 41% | 18% | 55% | -37% |
| | C. Two or more | 15% | ***2%*** | 22% | -20% |
| 9. Does the household have a blender? | A. No | 52% | ***85%*** | 33% | 52% |
| | B. Yes | 48% | 15% | 67% | -52% |
| 10. Does the household have an iron? | A. No | 49% | ***86%*** | 27% | 59% |
| | B. Yes | 51% | 14% | 73% | -59% |

On the other hand, the distribution of the variables for urban areas (III) shows greater variability, which might facilitate a better discrimination or identification of poor households.

Given the three points noted above, along with the results of table 3.6.1, it makes sense from a statistical, theoretical, and intuitive point of view not only to gauge if the PPI's accuracy differs across rural and urban settings, but to try to account for these differences in order to make the tool more accurate and relevant.

Let us turn to some key results using the validation sub-sample to look at poverty likelihoods and the speed of coverage for urban and rural areas.[38]

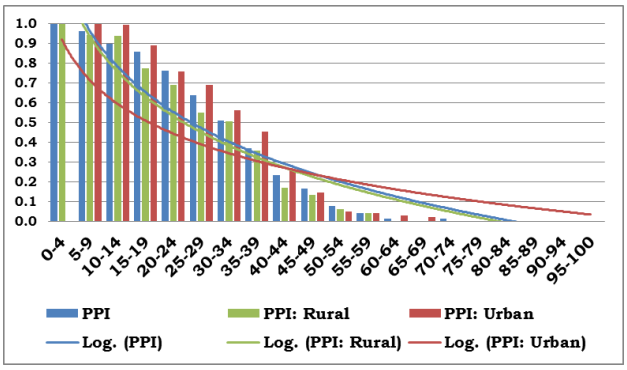



Figure 3.6.1a: Poverty Likelihoods: Nation (PPI), Urban & Rural Areas

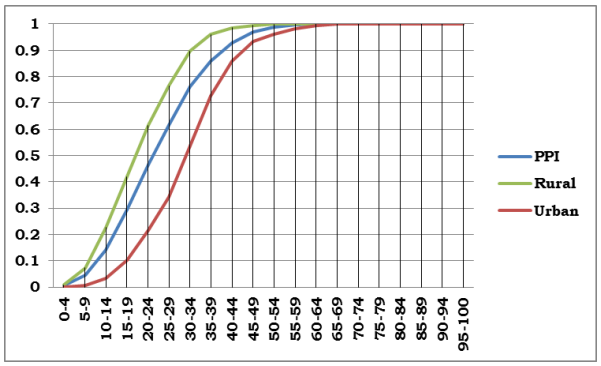



Figure 3.6.1b: Speed of Coverage: Nation (PPI), Urban & Rural Areas

The first panel of figure 3.6.1 provides a graphical comparison of the poverty likelihoods across scores for urban and rural areas with a point of reference using the poverty likelihoods for the original PPI (in blue). Recall from section 3.5 that a higher log-fitted line, especially in the lower scores, provides an indication of better targeting accuracy: i.e., at the lower scores, the tool is able to identify a higher percentage of households that are poor, as expected and as the tool was designed to do. While the log-fitted lines for the original PPI applied to the nation and the one applied only in rural areas are fairly close to each other, the one applied at the national level

[38] Following Schreiner's suggestions, the poverty likelihood and speed of coverage graphs are now presented using the validation sample. A previous version did this using the calibration sample. Overall, the results are fairly similar, with the exception of 3.6.1a.

seems more accurate at identifying poor households. However, the log-fitted line for urban areas is a bit deceiving: although it lies below the other two up the 40-44 mark, this is primarily due to no households in urban areas having a score of 0. Yet, urban areas have higher poverty likelihoods starting at the 45-49 mark.

Figure 3.6.1b shows the *speed of coverage* (SOC) for the three samples. Recall that the closer the curve to the north-west corner the faster the tool is identifying all poor households in the data. As with the poverty likelihoods, the speed of coverage is better for rural areas than for the urban areas. This means that a higher percentage of all the poor living in rural areas are scored in the lower distribution of scores (higher inclusion rates). In fact, nearly all poor households in rural areas have scores of less than 55. On the other hand, the distribution or the slope of the SOC clearly shows that the urban poor, on average, have higher scores; and in fact all the poor households in urban areas are only targeted after the 65-69 mark.

In terms of targeting accuracy, figure 3.6.1c provides the results of the ROC curve for the three areas. The original PPI applied to urban areas is significantly better than when applied to rural areas. Not surprisingly, the national ROC (PPI) falls in between the rural and urban ROC curves; however, we should note that it is almost identical to the urban ROC curve. This provides clear evidence that the PPI has much better targeting power in urban areas.

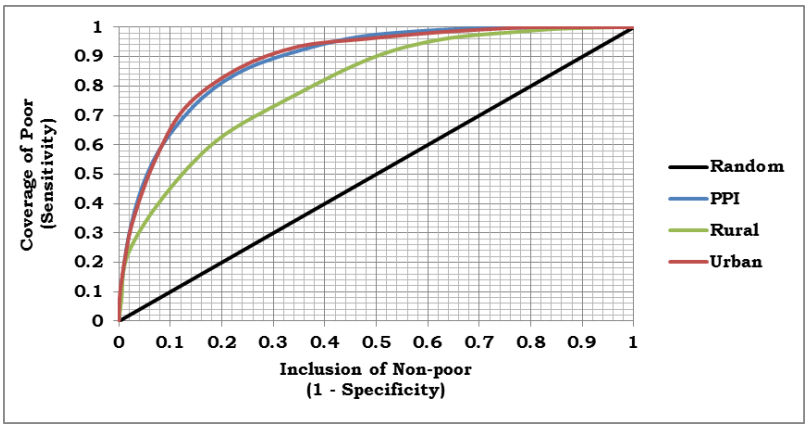



Figure 3.6.1c: ROC: Nation (PPI), Rural & Urban Areas

Yet, this seems to contradict the results of the SOC curve, which shows that the rural poor have lower scores and therefore a faster SOC. Thus, we need to reconcile these findings to get a clearer picture of what each of these two curves are telling us.

First, let us consider the speed in which the SOC curve identifies poor households in rural and urban settings. If we selected the score of 30-34 to be the cut-off, the PPI applied to the nation would target (or cover) 76.22% of the poor, it would target 53.86% of the poor in urban areas, but it would target an impressive 89.71% of poor households in rural areas. On the other hand, when the ROC curve is applied using the PPI at the national level (and using the same cut-off of 30-34) it targets 16.3% of the non-poor, it targets 6.76% of the non-poor in urban areas, while targeting a huge 49.33% of the non-poor in rural areas. The SOC curve, then, gives us a sense of inclusion along the distribution of scores but says nothing about leakage. On the other hand, the ROC curve plots the trade-off between inclusion (sensitivity) and leakage (1 – specificity) as a share of all the poor and all the non-poor, respectively, and thereby providing an objective visual illustration of the targeting power of the PPI applied to different settings.

Importantly, though, the fact that there are so many non-poor rural households that are scored at the lower end of the scores (leakage of close to 50%) provides some evidence that the PPI might improve its targeting accuracy if it used two different sets of indicators, one for rural areas and one for urban areas (although the current PPI already does very well in urban areas). It is plausible that there is a significant level of leakage in rural areas due the fact that most rural households are much more homogenous—in terms of the 10 indicators selected—than urban households, as shown in table 3.6.1. Recall that there is very little variation across 7 of the 10 indicators in rural areas. Thus, it seems that the PPI is not as successful at discriminating/identifying the poor from the non-poor in rural areas. Another way to say this is

that many non-poor households in rural areas do not have many of the 'endowments' used in the PPI to create the scorecard, such as electricity-based cooking appliances (16 points), or the highest possible education attained (16 points), or parquet floors (15 points), which would yield much lower scores. In other words, these items seem to be just much rarer in rural areas not only for the poor, but also for the non-poor, which might make it difficult to discriminate or identify the poor from the non-poor.

Table 3.6.2a and table 3.6.2b provide the results of the bias in the first column labeled, A: i.e., the difference between the estimated poverty likelihoods and the true poverty likelihoods from the bootstrap. The second column for each table, labeled B, provides the absolute bias for all samples. Table 3.6.2a provides the bias (Diff.) for the entire sample, while table 3.6.2b focuses on the scores that cover ~99% of all poor households who are scored at the 50-54 mark and below, which makes the analysis of bias more relevant since our main interest is on identifying the profile of households that pro-poor organizations might serve. Finally, the second to last row of these tables provides the average of the bias for each sample, while the last row gives the difference in the averages (sample x – Nation (original PPI)) in relation to the original PPI (Nation) to get a sense on the differences in the average bias when applying the original PPI.

Looking at table 3.6.2a, column A, we see that the average bias of the PPI applied in rural areas is greater than the one applied to the nation, indicating that in rural areas, on average, the PPI overestimates poverty by 2.19, which is more than the 0.32 percent for the nation, although still low. On the other hand, when applied to urban areas the PPI, on average, underestimates poverty by -1.41 percent. These results provide evidence that the PPI's accuracy in terms of bias differs across rural and urban settings, not only in terms of the magnitude of the average bias, but in terms of over or underestimating poverty. Moving to column B of table 3.6.2a, the results

show that the average absolute biases are fairly similar across samples, although slightly higher in rural and urban areas than when applied to the whole sample, while the absolute bias is higher for urban areas. These results are maintained in table 3.6.2b, which focuses on the scores that cover all poor households, although the magnitude of the bias and the differences in the bias are magnified, as expected. Looking at column A in table 3.6.2b, the PPI applied in rural areas shows that poverty is overestimated by 3.57 percent, compared to 0.81 percent when applied to the nation. On the other hand, when applied to urban areas the PPI underestimates poverty by -2.32 percent. In terms of the absolute bias shown in column B, in general, these biases are higher, especially for urban areas.

Table 3.6.2a: Bias and Absolute Bias: PPI Applied Separately to Rural & Urban Areas

| | A | | | B | | |
|---|---|---|---|---|---|---|
| | Diff. | | | \|Diff.\| | | |
| Score i | Nation | In Rural | In Urban | Nation | In Rural | In Urban |
| 0-4 | 0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 5-9 | 0.5 | 0.8 | -3.7 | 0.50 | 0.8 | 3.7 |
| 10-14 | -5.4 | -4.9 | -9.9 | 5.40 | 4.9 | 9.9 |
| 15-19 | 6.9 | 8.2 | 2.5 | 6.90 | 8.2 | 2.5 |
| 20-24 | 7 | 7.6 | 5.6 | 7.00 | 7.6 | 5.6 |
| 25-29 | 5.7 | 10.9 | -4.3 | 5.70 | 10.9 | 4.3 |
| 30-34 | -2.9 | 0.6 | -6.1 | 2.90 | 0.6 | 6.1 |
| 35-39 | -8.4 | 0.5 | -12.5 | 8.40 | 0.5 | 12.5 |
| 40-44 | 0.8 | 8.2 | -1.5 | 0.80 | 8.2 | 1.5 |
| 45-49 | 1.3 | 5.7 | 0.8 | 1.30 | 5.7 | 0.8 |
| 50-54 | 3.4 | 1.8 | 3.6 | 3.40 | 1.8 | 3.6 |
| 55-59 | 0.2 | 1.9 | 0.1 | 0.20 | 1.9 | 0.1 |
| 60-64 | -1.9 | 1.4 | -2.0 | 1.90 | 1.4 | 2.0 |
| 65-69 | -2.1 | 0.0 | -2.2 | 2.10 | 0.0 | 2.2 |
| 70-74 | 1.3 | 1.3 | 1.3 | 1.30 | 1.3 | 1.3 |
| 75-79 | 0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 80-84 | 0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 85-89 | 0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 90-94 | 0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 95-100 | 0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| **Averages** | 0.32 | 2.19 | -1.41 | 2.39 | 2.68 | 2.81 |
| **Diff. in Avgs.** | | 1.87 | -1.73 | | 0.29 | 0.13 |

Table 3.6.2b: Bias and Absolute Bias for 99% of Poor Households: PPI Applied Separately to Rural & Urban Areas

| | A | | | B | | |
|---|---|---|---|---|---|---|
| | Diff. | | | \|Diff.\| | | |
| Score i | Nation | In Rural | In Urban | Nation | In Rural | In Urban |
| 0-4 | 0 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 5-9 | 0.5 | 0.8 | -3.7 | 0.50 | 0.8 | 3.7 |
| 10-14 | -5.4 | -4.9 | -9.9 | 5.40 | 4.9 | 9.9 |
| 15-19 | 6.9 | 8.2 | 2.5 | 6.90 | 8.2 | 2.5 |
| 20-24 | 7 | 7.6 | 5.6 | 7.00 | 7.6 | 5.6 |
| 25-29 | 5.7 | 10.9 | -4.3 | 5.70 | 10.9 | 4.3 |
| 30-34 | -2.9 | 0.6 | -6.1 | 2.90 | 0.6 | 6.1 |
| 35-39 | -8.4 | 0.5 | -12.5 | 8.40 | 0.5 | 12.5 |
| 40-44 | 0.8 | 8.2 | -1.5 | 0.80 | 8.2 | 1.5 |
| 45-49 | 1.3 | 5.7 | 0.8 | 1.30 | 5.7 | 0.8 |
| 50-54 | 3.4 | 1.8 | 3.6 | 3.40 | 1.8 | 3.6 |
| **Averages** | 0.81 | 3.57 | -2.32 | 3.85 | 4.46 | 4.60 |
| **Diff. in Avgs.** | | 2.76 | -3.13 | | 0.61 | 0.14 |

In order to make these comparisons easier to see, figure 3.6.2 visually summarizes columns A and B of table 3.6.2a.

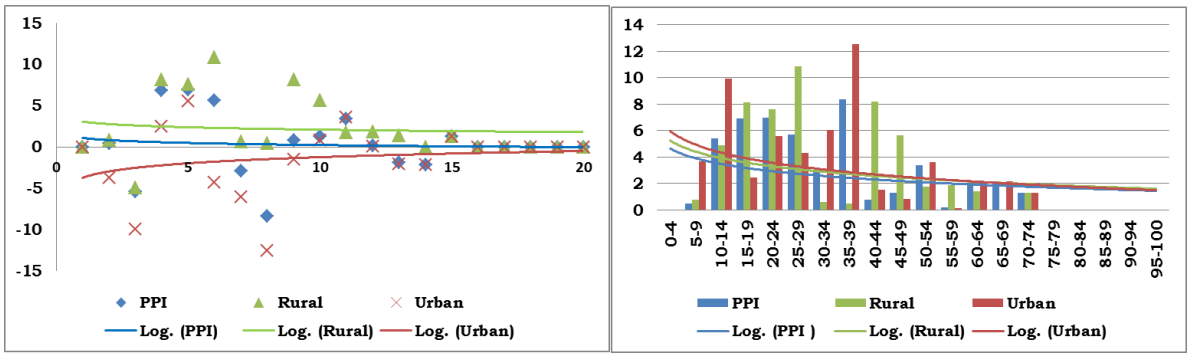



Figure 3.6.2: Graphical Representation of the Bias and Absolute Bias: Rural and Urban Areas
*Note*: National PPI = PPI

The first panel of figure 3.6.2 captures the performance of the PPI in terms of bias when applied to rural and urban areas, which is in the opposite direction of each other: while the PPI, on average, overestimates poverty in rural areas, it actually underestimates it in urban areas. On the other hand, the second panel of figure 3.6.2 shows that the absolute bias, on average, is smaller for the PPI when applied to the entire nation and with a greater absolute bias in urban areas; although in absolute terms, these differences are minimal as shown in the graph.

The results of this section show that the PPI has better poverty likelihoods, better *speed of coverage* (higher inclusion rates) in rural areas, but much higher leakage. If we were primarily concerned with inclusion, then the PPI performs much better in rural areas. However, the results of the ROC curve show that if we equally value inclusion and leakage, then the PPI has much better targeting accuracy in urban areas across all possible cut-offs, while having a pretty high level of leakage in rural areas. In terms of bias, the PPI has better bias and absolute bias measures in rural areas than in urban areas. Importantly, the bias is in the opposite direction for rural areas where poverty is overestimated, while the tool underestimates it in urban areas. Moreover, when looking at some of the lower scores shown in table 3.6.2, we see that the biases are at times quite substantial—even more than when just looking at the PPI applied to the entire nation. For example, the tool overestimates poverty in rural areas by 8.2 for scores at 15-19, and 40-44 and it yields an even greater overestimate of 10.9 for scores at 25-29. Similar high levels

of bias are found in urban areas: -9.9 for scores at 10-14, and -12.5 for scores at 35-39. This comes out clearer in the second panel of figure 3.6.2 when we look at the absolute bias.

Importantly, the results of the SOC and ROC curves provide evidence that it might be beneficial to create a separate scorecard for rural areas. This is because a significant share of non-poor households is also located in the lower end of the distribution of scores (SOC curve), which yields pretty high levels of leakage (ROC curve). We hypothesize that this might be due to the type of indicators used in the PPI, which might be either very common (little variation) or much rarer (little variation) in rural areas, which might make it difficult for the logit model to discriminate between the rural poor and non-poor. In fact, table 3.6.1 has shown that for 7 out of 10 indicators used in the PPI, there is very little variation in these indicators in rural areas

### 3.7 Slight modifications of the PPI to improve accuracy by adding regional variables?

This section uses the methodology developed by Schreiner and adds two important modifications to the original construction process by incorporating rurality and regional variables to the scorecard. This will lead to an additional question(s) in the scorecard, a modification to the points for each question and respective values, and to new poverty likelihoods or estimated poverty rates for each score. This set of analysis, then, involves replicating the construction, calibration, and validation of the original PPI but with these added variables. The results of the construction of the two new poverty scorecards are presented in table 3.7.1. Column A provides the points for the original PPI, while columns B and C provide two new scorecards with new points for models that add rurality, and rurality plus regional variables, respectively.

Table 3.7.1: Modified Replicate of Figure 1 in Schreiner (2009): A simple Poverty Scorecard for Peru Accounting for Rurality & Regional Variables

| | | A | B | C |
|---|---|---|---|---|
| | | PPI 2007 | PPI + Rural | PPI + Rural & Regional |
| Indicator | Value | Points | Points | Points |
| 1. How many household members are 17-years old or younger? | A. Four or more | 0 | 0 | 0 |
| | B. Three | 5 | 5 | 5 |
| | C. Two | 9 | 8 | 8 |
| | D. One | 16 | 15 | 14 |
| | E. None | 24 | **22** | 21 |
| 2. What is the highest educational level that the female head/spouse completed? | A. None, pre-school, or kindergarten | 0 | 0 | 0 |
| | B. Grade school (incomplete) | 5 | 5 | 4 |
| | C. Grade school (complete) | 7 | 7 | 6 |
| | D. High School (incomplete) | 9 | 8 | 8 |
| | E. High School (complete), non-university superior (incomplete) or no female head | 10 | 9 | 8 |
| | F. Non-university superior (complete) or higher | 16 | 14 | 14 |
| 3. What is the main material of the floors? | A. Earth, wood planks, other, or no residence | 0 | 0 | 0 |
| | B. Cement | 2 | 2 | 3 |
| | D. Parquet, polished wood, linoleum, vinyl, tile, or similar | 15 | 14 | 15 |
| 4. What is the main material of the exterior walls? | A. Adobe, mud or matting | 0 | 0 | 0 |
| | B. Wattle and daub, wood, brick or cement blocks, stone blocks with lime, other, or no residence | 2 | 3 | 2 |
| 5. Excluding bathrooms, kitchen, hallways, and garage, how many rooms does the residence have? | A. One | 0 | 0 | 0 |
| | B. Two | 1 | 1 | 1 |
| | C. Three, four or five | 5 | 5 | 4 |
| | D. Six or more | 10 | 9 | 8 |
| 6. What fuel does the household most frequently use for cooking? | A. Other | 0 | 0 | 0 |
| | B. Firewood, charcoal, or kerosene | 5 | 5 | 4 |
| | C. Gas (LPG or natural) | 9 | 10 | 10 |
| | D. Electricity or does not cook | 16 | 17 | 15 |
| 7. Does the household have a refrigerator/freezer? | A. No | 0 | 0 | 0 |
| | B. Yes | 5 | 5 | 5 |
| 8. How many color televisions does the household have? | A. None | 0 | 0 | 0 |
| | B. One | 3 | 3 | 3 |
| | C. Two or more | 7 | 6 | 6 |
| 9. Does the household have a blender? | A. No | 0 | 0 | 0 |
| | B. Yes | 3 | 3 | 3 |
| 10. Does the household have an iron? | A. No | 0 | 0 | 0 |
| | B. Yes | 2 | 3 | 2 |
| 11. Does household live in rural area? | A. No | N/A | 0 | 0 |
| | B. Yes | N/A | 4 | 5 |
| 12. Does the household live in region…? | A. 4: Sierra Norte | N/A | | 0 |
| | B. 8: Lima Metropolitana | N/A | | -4 |
| | C. 6: Sierra Sur | N/A | | -3 |
| | D. 5: Sierra Centro | N/A | | -2 |
| | E. 7: Selva | N/A | | 2 |
| | F. 1: Costa Norte | N/A | | 3 |
| | G. 3: Costa Sur | N/A | | 4 |
| | H. 2: Costa Centro | N/A | | 4 |
| | Max Score | 100 | 100 | 100 |
| | Min Score | 0 | 0 | -4 |

The results for the two new scorecards are very similar to the original PPI with some minor differences: in general, in the two new scorecards there is a slightly lesser role played by

demographics (Q.1), education of the female head/spouse (Q.2), the number of rooms in the dwelling (Q.5), and the number of TVs owned by the household (Q.8). Looking at the rurality variable in both new scorecards, a household has a higher probability of not being poor in rural areas. Given that there are more poor households in rural areas than in urban areas, this result is counterintuitive; however, this is simply the result of applying fixed effects in the logit regression, i.e., holding everything else constant, the rurality variable explains the within variability that might explain the likelihood of a household being poor, beyond what is explained by the other 10 indicators. The same interpretation applies to the model that adds rurality and regional variables.

As for the regional variables, it is worth mentioning that we set up the data in such a way to imitate the approach developed by Schreiner where the regional variable is 'ordered' in the sense that the base reference for all regions is the one most associated with poverty, i.e., the one with the highest poverty rate, which is the Sierra Norte, followed by the region with the second highest poverty rate, and continuing that way until reaching the final region, Costa Centro, which has the lowest poverty rate. Finally, while the scorecard in B (accounting for rurality) can also have a minimum value of 0 points and a maximum of 100 as the original PPI, the scorecard in C (accounting for rurality and region) can have a negative minimum value of -4, although still retaining a possible maximum value of 100. Although some might argue that having a negative number in the scorecard might make it cumbersome to apply the scorecard in the field, in practical terms, this should not be an issue. In other words, if an individual applying the scorecard can add 20 + 4, he or she should have no trouble subtracting 20 – 4.

We should note, however, that having negative and positive values by regions may create unwanted or unexpected complaints if this has political implications. For example, in a

collaborative effort between a team from IFPRI and the Egyptian Ministry of Trade and Supply to develop a proxy means test to target food subsidies in Egypt, Ahmed and Bouis (2002) explain that although the team found that all 26 regional fixed effects variables were significant, they opted not to include these in the final model since some were concerned this would lead to a disproportionate allocation of benefits by region implied by the results of the model. With this caveat in mind, let us look at the results.

Figure 3.7.1 summarizes the comparison in poverty likelihoods (panel a) and the SOC curve (panel b) for the original PPI applied to the nation, the original PPI accounting for rurality, and then accounting for rurality and regional variables using the validation sample.

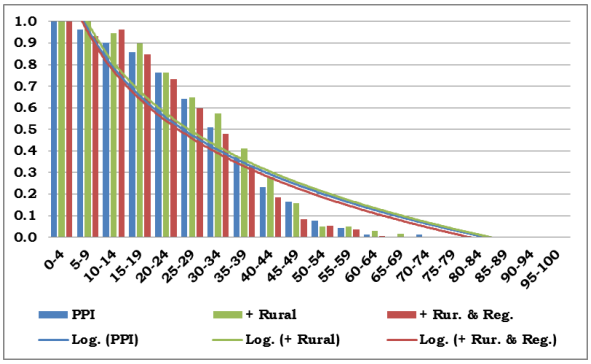



Figure 3.7.1a: Poverty Likelihoods: Accounting for Rurality & Regional Variables

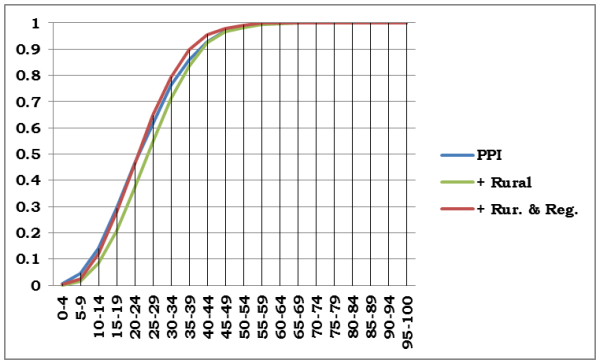



Figure 3.7.1b: Speed of Coverage: Accounting for Rurality & Regional Variables

As with the results of the previous section, while the log-fitted lines for the three models are almost identical, the log-fitted line for the model that includes rurality only is slightly higher than the other two. On the other hand, the second panel (b) of figure 3.7.1 shows that the SOC curve is slightly better for the third model that incorporates rurality and regional variables. Using the cut-off at the 30-34 mark, we can see that the original PPI applied to the nation targets 76.22% of all poor households, but when adding rurality it targets 71.21% of poor households. On the other hand, when adding rurality plus regional variables, the PPI targets 79.41% of poor households. This means that when adding rurality and regional variables, the modified PPI

would improve its inclusion rate by 3.19 percentage points. In terms of leakage and using the same cut-off, when applied to the nation, leakage is 16.32% versus 14.34% when adding the rural variable, and 19.27% when adding the rural and regional variables. This indicates that adding the rural variable has a small improvement in leakage, while adding the regional variables makes the leakage slightly worse.

Moving to the trade-off between inclusion (specificity) and leakage (1 – specificity), figure 3.7.1c provides the results of the ROC curve.

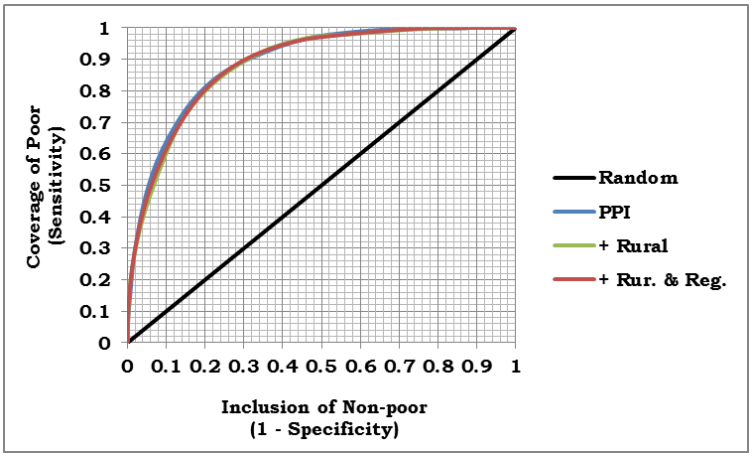



Figure 3.7.1c: ROC: Nation (PPI), + Rural, and + Rural & Regional Variables

Figure 3.7.1c for shows the targeting power of the PPI is essentially identical in all three models, indicating that adding rurality and regional variables essentially has no effect on targeting accuracy.

Table 3.7.2 provides the results for the bias (Diff.) and the absolute bias. Colum A of panel a shows that the bias for the model that incorporates rurality has an almost identical level of the overall average of bias (0.36) compared to the original PPI (0.32). On the other hand, the average bias the model that incorporates rurality and regional variables is slightly higher at 1.34 percent. However, when looking at the second panel that focuses on ~99% of poor households we see that the absolute bias is smaller for the two modified PPIs with the one accounting for

rurality having an average absolute bias of 2.43 compared to 3.85 for the original PPI, while the model that adds rurality and regional variables being close to the original at 3.8.

Table 3.7.2a: Bias and Absolute Bias: Modified PPI Accounting for Rurality & Region

| | A | | | B | | |
|---|---|---|---|---|---|---|
| | Diff. | | | \|Diff.\| | | |
| **Score *i*** | **PPI** | **Rural** | **Rur. & Reg.** | **PPI** | **Rural** | **Rur. & Reg.** |
| 0-4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5-9 | 0.5 | 0.0 | 11.6 | 0.5 | 0.0 | 11.6 |
| 10-14 | -5.4 | -1.2 | -3.4 | 5.4 | 1.2 | 3.4 |
| 15-19 | 6.9 | -0.9 | 2.6 | 6.9 | 0.9 | 2.6 |
| 20-24 | 7 | 9.8 | 9.4 | 7.0 | 9.8 | 9.4 |
| 25-29 | 5.7 | 3.7 | 4.8 | 5.7 | 3.7 | 4.8 |
| 30-34 | -2.9 | -0.6 | -2.5 | 2.9 | 0.6 | 2.5 |
| 35-39 | -8.4 | -5.6 | -1.2 | 8.4 | 5.6 | 1.2 |
| 40-44 | 0.8 | 0.6 | 3.5 | 0.8 | 0.6 | 3.5 |
| 45-49 | 1.3 | 0.0 | 2.6 | 1.3 | 0.0 | 2.6 |
| 50-54 | 3.4 | 4.4 | -0.4 | 3.4 | 4.4 | 0.4 |
| 55-59 | 0.2 | -0.1 | -1.0 | 0.2 | 0.1 | 1.0 |
| 60-64 | -1.9 | -2.3 | 0.9 | 1.9 | 2.3 | 0.9 |
| 65-69 | -2.1 | -1.8 | 0.0 | 2.1 | 1.8 | 0.0 |
| 70-74 | 1.3 | 1.2 | 0.0 | 1.3 | 1.2 | 0.0 |
| 75-79 | 0 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| 80-84 | 0 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| 85-89 | 0 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| 90-94 | 0 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| 95-100 | 0 | 0 | 0 | 0.0 | 0.0 | 0.0 |
| **Avg.** | 0.32 | 0.36 | 1.34 | 2.39 | 1.61 | 2.19 |
| **Diff. in Avgs.** | | 0.04 | 1.02 | | -0.78 | -0.20 |

Table 3.7.2b: Bias and Absolute Bias for 99% of Poor Households: Modified PPI Accounting for Rurality & Region

| | A | | | B | | |
|---|---|---|---|---|---|---|
| | Diff. | | | \|Diff.\| | | |
| **Score *i*** | **PPI** | **Rural** | **Rur. & Reg.** | **PPI** | **Rural** | **Rur. & Reg.** |
| 0-4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5-9 | 0.5 | 0.0 | 11.6 | 0.5 | 0.0 | 11.6 |
| 10-14 | -5.4 | -1.2 | -3.4 | 5.4 | 1.2 | 3.4 |
| 15-19 | 6.9 | -0.9 | 2.6 | 6.9 | 0.9 | 2.6 |
| 20-24 | 7 | 9.8 | 9.4 | 7.0 | 9.8 | 9.4 |
| 25-29 | 5.7 | 3.7 | 4.8 | 5.7 | 3.7 | 4.8 |
| 30-34 | -2.9 | -0.6 | -2.5 | 2.9 | 0.6 | 2.5 |
| 35-39 | -8.4 | -5.6 | -1.2 | 8.4 | 5.6 | 1.2 |
| 40-44 | 0.8 | 0.6 | 3.5 | 0.8 | 0.6 | 3.5 |
| 45-49 | 1.3 | 0.0 | 2.6 | 1.3 | 0.0 | 2.6 |
| 50-54 | 3.4 | 4.4 | -0.4 | 3.4 | 4.4 | 0.4 |
| **Avg.** | 0.81 | 0.93 | 2.46 | 3.85 | 2.43 | 3.80 |
| **Diff. in Avgs.** | | 0.12 | 1.65 | | -1.42 | -0.04 |

In order to provide a clearer visual comparison of bias and absolute bias across all scores, figure 3.7.2 summarizes these results. The first panel shows that the log-fitted line for the original PPI (applied to the nation) and the model that accounts for rurality are essentially identical; however, when looking at the absolute bias, the second panel shows that, on average, the absolute bias is substantially smaller for the model that incorporates rurality. This is an important finding since by adding the rural variable and even the regional variable, this significantly reduces the bias and absolute bias—especially in the lower scores, which is what we would like; although this comes out clearer when we add the rural variable only.

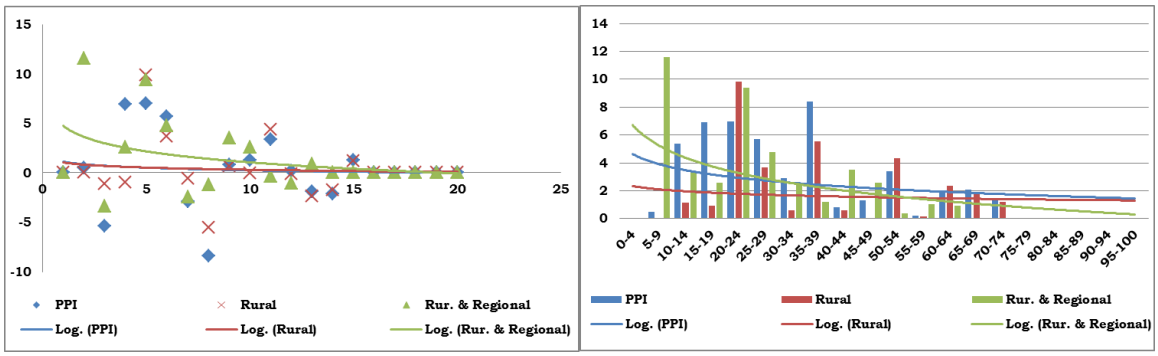


Figure 3.7.2: Graphical Representation of the Bias and Absolute Bias: Accounting for Rurality and Regional Variables

The results presented in this section provide mixed results. On the one hand, adding rurality and regional variables slightly improves poverty likelihoods, SOC curve (rurality & regional), higher inclusion rates (rurality & regional), lower leakage (rurality); however, these small modifications do not translate into any improvements in terms of absolute targeting accuracy as shown in the ROC plot. On the other hand, adding these variables yield small improvements in bias, although only for the absolute bias. Yet, these improvements are fairly small; although the improvements are greater in the lower end of the distribution of scores for rural areas.

Let us now see what happens when we create two separate sets of scorecards: one for rural areas and one for urban areas.

## 3.8 A Slightly Modified PPI: New Scorecards for Rural and Urban Areas

As noted above, there is reason to believe that poverty is experienced differently in rural and urban areas. Although there might be clear similarities between poor households in these two settings in terms of demographics, low levels of education, lacking in appropriate infrastructure for the dwelling, and having low levels of assets and endowments, it is plausible that the intensity in the association between these variables and poverty will differ across the two settings. For example, education may play a bigger role in the likelihood of exiting poverty in rural areas, or having assets, such as a refrigerator or a TV may be a much stronger signal for not

being poor. Moreover, it might be the case that some variables that go into the final poverty scorecard, such as having an electric stove for example, might be more relevant to identify poor households (or in this case to proxy wellbeing) in urban than in rural areas. Along these lines, it might be the case that having parquet floors (Q.3) may be irrelevant in rural areas since none or very few homes in fact might be able to afford this material or it might be hard to find (lack of supply). The same can be said for having an electric stove, a refrigerator, or a color TV, which might be more relevant for identifying the non-poor in urban areas, especially since these goods are very rare in rural areas as shown in table 3.6.1 in section 3.6.

All of this gives theoretical reasons to believe that it might be beneficial to create two separate scorecards—one for urban areas and one for rural areas—and perhaps even creating two sets of indicators. The previous two sections provided some evidence that the PPI performs differently (and in opposite directions) in rural and urban areas. Likewise, it has been shown that adding a variable to capture rurality, as well as regional variables, might improve the poverty likelihoods, the SOC and absolute bias, although in a fairly minor way. With this in mind, this section aims to create two new poverty scorecards, one for rural areas and one for urban areas. This involves splitting the data into urban and rural areas and replicating Schreiner's methodology from the construction, calibration, and validation. All of this is done twice—once for rural areas and once for urban areas, but always using the same validation sample as the one used by Schreiner.

For the construction of the scorecard, we begin by presenting the results of the logit in table 3.8.1. Column A gives the coefficients for the original logit, followed by columns B and C, which give the coefficients for urban and rural areas, respectively. In general, we see that the coefficients of the original PPI (A) fall exactly in between those of urban (B) and rural areas (C),

as expected; after all, we have split the data into two independent samples that together make up the original data set (A).

Table 3.8.1: Logit Results: Original, in Urban Areas Only, and in Rural Areas Only

| | | A | B | C |
|---|---|---|---|---|
| **Indicator** | **Value** | **Original Coeff.** | **Urban Only Coeff.** | **Rural Only Coeff.** |
| 1. How many household members are 17-years old or younger? | A. Four or more | | ***BASE*** | |
| | B. Three | -0.694*** | -0.676* | -0.917*** |
| | C. Two | -1.148*** | -1.222*** | -1.211*** |
| | D. One | -2.100*** | -2.317*** | -1.999*** |
| | E. None | -3.072*** | -3.411*** | -3.001*** |
| 2. What is the highest educational level that the female head/spouse completed? | A. None, pre-school, or kindergarten | | ***BASE*** | |
| | B. Grade school (incomplete) | -0.643*** | -0.352 | -0.700*** |
| | C. Grade school (complete) | -0.945*** | -0.771** | -0.931*** |
| | D. High School (incomplete) | -1.170*** | -0.983** | -1.337*** |
| | E. High School (complete), non-university superior (incomplete) or no female head | -1.260*** | -1.162*** | -1.253*** |
| | F. Non-university superior (complete) or higher | -2.041*** | -1.950*** | -2.225*** |
| 3. What is the main material of the floors? | A. Earth, wood planks, other, or no residence | | ***BASE*** | |
| | B. Cement | -0.242* | -0.268 | -0.564** |
| | D. Parquet, polished wood, linoleum, vinyl, tile, or similar | -1.952** | -2.027** | **N/A** |
| 4. What is the main material of the exterior walls? | A. Adobe, mud or matting | | ***BASE*** | |
| | B. Wattle and daub, wood, brick or cement blocks, stone blocks with lime, other, or no residence | -0.325*** | -0.370** | -0.521*** |
| 5. Excluding bathrooms, kitchen, hallways, and garage, how many rooms does the residence have? | A. One | | ***BASE*** | |
| | B. Two | -0.197 | 0.0237 | -0.435** |
| | C. Three, four or five | -0.680*** | -0.507** | -0.908*** |
| | D. Six or more | -1.319*** | -1.489*** | -1.136*** |
| 6. What fuel does the household most frequently use for cooking? | A. Other | | ***BASE*** | |
| | B. Firewood, charcoal, or kerosene | -0.668*** | -0.0951 | -0.769*** |
| | C. Gas (LPG or natural) | -1.205*** | -0.870* | -2.049*** |
| | D. Electricity or does not cook | -2.144*** | -1.857** | -2.018*** |
| 7. Does the household have a refrigerator/freezer? | A. No | | ***BASE*** | |
| | B. Yes | -0.657*** | -0.719*** | -1.000* |
| 8. How many color televisions does the household have? | A. None | | ***BASE*** | |
| | B. One | -0.334** | -0.463** | -0.11 |
| | C. Two or more | -0.913** | -1.023** | -1.455 |
| 9. Does the household have a blender? | A. No | | ***BASE*** | |
| | B. Yes | -0.438*** | -0.339* | -0.843*** |
| 10. Does the household have an iron? | A. No | | ***BASE*** | |
| | B. Yes | -0.265* | -0.274 | -0.716** |
| | _cons | 4.291*** | 4.005*** | 4.564*** |
| '* p<0.05; ** p<0.01 | N | 6274 | 3765 | 2495 |

The results of table 3.8.1 show that the relationship between variables, and respective values, and poverty is still the same across all models (all coefficients are negative); however, the magnitude of the association of each variable and of each possible value to the probability of being poor is different. Yet, it is still possible to see some of the important non-linearities noted

earlier. In order to see this more clearly, we turn to table 3.8.2 which has transformed the coefficients into points for the two new poverty scorecards.

Table 3.8.2 shows that demographics play a bigger role in the model in urban areas: households with no kids under 17-years old are associated with a lower likelihood of being poor (higher scores). On the other hand, education has a higher predictive ability in rural areas, which implies that higher education is rarer in rural areas.[39] However, the variable for the main material of the floor shows the biggest difference across urban and rural areas. The predictive ability in the model is very big for the third category (option C) in urban areas. On the other hand, that same third category does not exist in rural areas and so it is not relevant to the model. In fact, the final contribution to the model for Q.3 is 4% in rural areas, compared to a significantly different contribution of 15% for urban areas. This significant difference echoes Zeller's (2004) conclusion that some variables are irrelevant for predicting poverty since they may only be appropriate for one particular setting, but not in others.

Table 3.8.2 also illustrates the difference in the intensity of the relationship between variables and poverty and in relation to the base category used for reference. For instance, nonlinearities for cooking fuel are much more pronounced in rural areas: households that use a better type of fuel, such as gas or LGP or natural gas are much less likely of being poor. Similarly, in rural areas going from no TV to having one TV yields one point on the scorecard; however, going from one TV to two or more gives 11 points, compared to 8 points in urban areas. These nonlinearities are also seen in urban areas but they are not as pronounced as in rural areas. This might be because having a TV in rural areas is much less common than in urban areas. As such, having more than 2 TVs yields a greater likelihood of being non-poor in rural

[39] This interpretation was highlighted and made clearer by Schreiner on a previous version of this paper.

areas. The difference in the intensity of these non-linearities in urban and rural areas is also captured in Q. 7, 9, and 10 for owning a refrigerator, a blender, or an iron, respectively.

Table 3.8.2: Modified Replicate of Figure 1 in Schreiner (2009): A simple Poverty Scorecard for Peru Accounting for Rurality & Regional Variables

| | | A | B | C |
|---|---|---|---|---|
| | | Original PPI | Separate Scores | |
| | | PPI 2007 | Urban | Rural |
| Indicator | Value | Points | Points | Points |
| 1. How many household members are 17-years old or younger? | A. Four or more | 0 | 0 | 0 |
| | B. Three | 5 | 5 | 7 |
| | C. Two | 9 | 9 | 9 |
| | D. One | 16 | 17 | 15 |
| | E. None | 24 | 25 | 22 |
| 2. What is the highest educational level that the female head/spouse completed? | A. None, pre-school, or kindergarten | 0 | 0 | 0 |
| | B. Grade school (incomplete) | 5 | 3 | 5 |
| | C. Grade school (complete) | 7 | 6 | 7 |
| | D. High School (incomplete) | 9 | 7 | 10 |
| | E. High School (complete), non-university superior (incomplete) or no female head | 10 | 9 | 9 |
| | F. Non-university superior (complete) or higher | 16 | **14** | 17 |
| 3. What is the main material of the floors? | A. Earth, wood planks, other, or no residence | 0 | 0 | 0 |
| | B. Cement | 2 | 2 | 4 |
| | D. Parquet, polished wood, linoleum, vinyl, tile, or similar | 15 | 15 | N/A |
| 4. What is the main material of the exterior walls? | A. Adobe, mud or matting | 0 | 0 | 0 |
| | B. Wattle and daub, wood, brick or cement blocks, stone blocks with lime, other, or no residence | 2 | 3 | 4 |
| 5. Excluding bathrooms, kitchen, hallways, and garage, how many rooms does the residence have? | A. One | 0 | 0 | 0 |
| | B. Two | 1 | 0 | 3 |
| | C. Three, four or five | 5 | 4 | 7 |
| | D. Six or more | 10 | 11 | **9** |
| 6. What fuel does the household most frequently use for cooking? | A. Other | 0 | 0 | 0 |
| | B. Firewood, charcoal, or kerosene | 5 | 1 | 6 |
| | C. Gas (LPG or natural) | 9 | 6 | 15 |
| | D. Electricity or does not cook | 16 | 14 | 15 |
| 7. Does the household have a refrigerator/freezer? | A. No | 0 | 0 | 0 |
| | B. Yes | 5 | 5 | 7 |
| 8. How many color televisions does the household have? | A. None | 0 | 0 | 0 |
| | B. One | 3 | 3 | 1 |
| | C. Two or more | 7 | 8 | 11 |
| 9. Does the household have a blender? | A. No | 0 | 0 | 0 |
| | B. Yes | 3 | 3 | 6 |
| 10. Does the household have an iron? | A. No | 0 | 0 | 0 |
| | B. Yes | 2 | 2 | 5 |
| | Max Score | 100 | 100 | 100 |
| | Min Score | 0 | 0 | 0 |

**Note: Bold =** Figures rounded so that all scores add to 100.
Using decimals for all variables yields a max score of 100.

Based on these results, it seems that having scorecards for rural and urban areas might improve poverty likelihoods, the SOC curve, the ROC curve and the bias and absolute bias of the PPI. We explore this, below.

Figure 3.8.1a presents the poverty likelihoods across scores for the original PPI applied to the nation, the new scorecards for rural and urban areas, while figure 3.8.1b presents the SOC curve for these three samples.

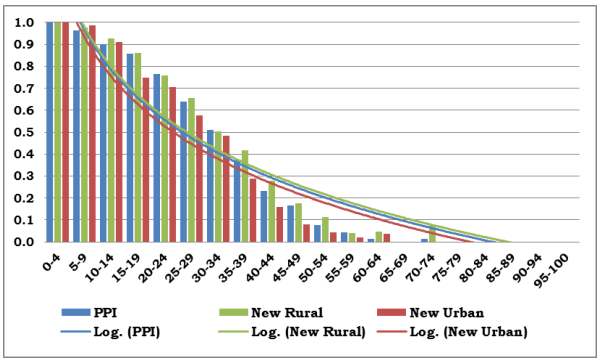



Figure 3.8.1a: Poverty Likelihoods: New Scorecard for Urban & Rural Areas

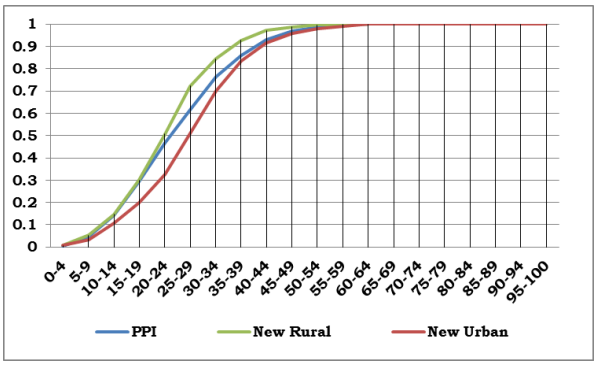



Figure 3.8.1b: Speed of Coverage: New Scorecard for Urban & Rural Areas

The log-fitted lines presented in the first panel of figure 3.8.1 show that the new scorecard for rural areas, on average, has higher poverty likelihoods than the original PPI applied to the nation and the new scorecard for urban areas; although the lines are very close to each other. Moreover, the second panel shows that the SOC curve for the new rural scorecard is fairly identical to the PPI applied to the nation up to the 20-24 mark, but it then gets steeper thereafter, indicating that with the new rural scorecard more poor households have lower scores. Indeed, the inclusion rates at the 30-34 cut-off is 84.43% for the new rural scorecard, while it is 76.22% for the nation, and 70.17% for the urban scorecard. Comparing these results to those in section 3.6 where we apply the original PPI to urban and rural areas separately, there is an important improvement for the SOC curve (or inclusion) in urban areas when we look at the same cut-off of 30-34—going from 53.86% using the original PPI in urban areas to 70.17% using the new urban scorecard. This is a 30.28% improvement, which is significant. On the other hand, the inclusion rate in rural areas goes from 89.43% using the original PPI in rural areas to 84.43% using the new rural scorecard, which although this represents a drop in accuracy, the inclusion

rate is still very high. In terms of leakage, the new rural scorecard lowered its leakage from 49.33% using the original PPI in rural areas, to 40% using the new rural scorecard; this is an 18.91% improvement. On the other hand, leakage deteriorated in urban areas, going from 6.76% using the original PPI in urban areas, to 11.54% with the new urban scorecard; although this is still fairly low.

In terms of the trade-off between coverage and leakage, the ROC curve in figure 3.8.1c shows that targeting accuracy is almost identical for the original PPI applied to the nation and for the new urban scorecard with perhaps a minor improvement for the new urban scorecard. On the other hand, the ROC curve for the new rural scorecard is still much lower than the other two.

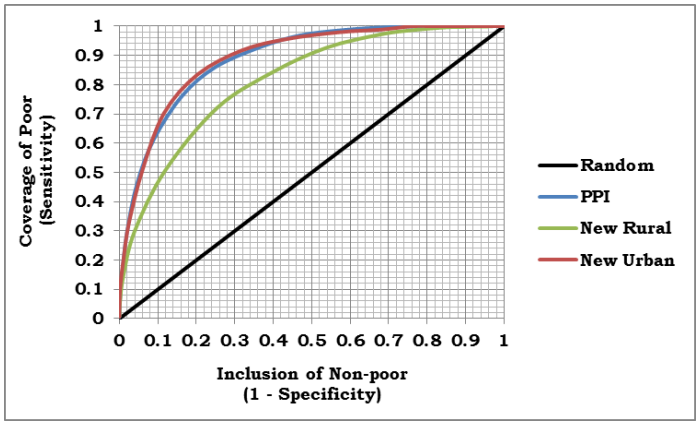


Figure 3.8.1c: ROC: Nation, New Rural & New Urban Scorecards

In order to see, however, if the new rural scorecard has better targeting accuracy overall than the original PPI applied only to rural areas, figure 3.8.1d plots these two ROC curves along with the national PPI ROC curve as a reference.

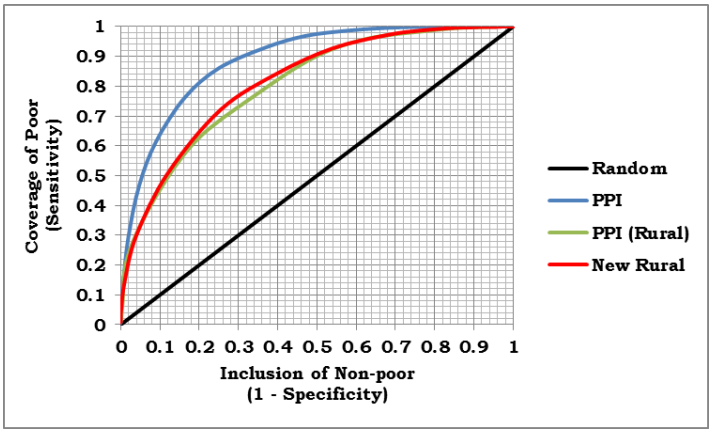


Figure 3.8.1d: ROC: Nation, Original in Rural & New Rural Scorecard

Figure 3.8.1d shows that the ROC curve for the new rural scorecard slightly improved targeting accuracy vis-à-vis the original PPI when applied only in rural areas. However, this improvement is minor, indicating that if we want to improve targeting accuracy in rural areas of Peru, which is where the majority of poor live, it might be necessary to find ways to ameliorate this other than obtaining different points for the questions already found in the PPI, as done here. It seems that the more appropriate choice would be to identify a different set of indicators better suited to predict poverty in rural areas. It might be very beneficial to replicate the entire process used to create the final PPI scorecard, but doing this separately for rural and urban areas. One would start by using the uncertainty coefficient to rank variables that are most strongly linked with higher poverty likelihoods in rural areas, and then one would move to construct the final PPI scorecard for rural areas. One would expect that this process would identify variables that are more relevant for rural areas, such as having agricultural land, livestock, distance to a road, or primary occupation of the head of household and of household members. In fact, the documentation for the new PPI for Peru has now incorporated some of these variables in their search for the 'best' poverty scorecard, and 5 out of the top 15 indicators that are most strongly linked to poverty using the uncertainty coefficient (See Figure 3 in Schreiner (2013)) are related to having agricultural land, or if the head of household or household members work in

agriculture as their main occupation. This provides evidence that there are variables that may be only pertinent to rural areas that are strongly linked to poverty likelihoods, and that can potentially be used to create a separate rural scorecard.

Moving to the measures of bias, table 3.8.3.a shows that the overall average of bias is now lower for urban and rural areas with the new scorecards: 0.32 for the original PPI, -0.12 for urban areas, and -0.09 for rural areas. Although the new scorecards are, on average, underestimating poverty, the average bias is smaller than the original PPI, although this is only by about two fifths of a percentage point. The same improvement holds true when looking at column B of table 8.14a for the absolute difference of the bias: the absolute bias is 2.39 for the original PPI, 2.29 for the new urban scorecard, and 2.04 for the new rural scorecard.

The results of this improvement in the bias and absolute bias are maintained when looking at table 3.8.3b for scores at the 50-54 mark or lower (see the Diff. in Avgs.).

Table 3.8.3a: Bias and Absolute Bias: New Scorecard Urban and Rural Areas

| | A | | | B | | |
|---|---|---|---|---|---|---|
| | **Diff.** | | | **\|Diff.\|** | | |
| **Score *i*** | **PPI** | **Urban** | **Rural** | **PPI** | **Urban** | **Rural** |
| 0-4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5-9 | 0.5 | -4.3 | 0.5 | 0.5 | 4.3 | 0.5 |
| 10-14 | -5.4 | 5.2 | 0.2 | 5.4 | 5.2 | 0.2 |
| 15-19 | 6.9 | 6.2 | -0.5 | 6.9 | 6.2 | 0.5 |
| 20-24 | 7 | -3.5 | 7.8 | 7 | 3.5 | 7.8 |
| 25-29 | 5.7 | -4.6 | 5.6 | 5.7 | 4.6 | 5.6 |
| 30-34 | -2.9 | -8.1 | 5.0 | 2.9 | 8.1 | 5.0 |
| 35-39 | -8.4 | 1.7 | -2.5 | 8.4 | 1.7 | 2.5 |
| 40-44 | 0.8 | 2.7 | -2.5 | 0.8 | 2.7 | 2.5 |
| 45-49 | 1.3 | 1.4 | -4.6 | 1.3 | 1.4 | 4.6 |
| 50-54 | 3.4 | 2.5 | -3.3 | 3.4 | 2.5 | 3.3 |
| 55-59 | 0.2 | 0.1 | 0.5 | 0.2 | 0.1 | 0.5 |
| 60-64 | -1.9 | -3.6 | -0.2 | 1.9 | 3.6 | 0.2 |
| 65-69 | -2.1 | 0.1 | 0.0 | 2.1 | 0.1 | 0.0 |
| 70-74 | 1.3 | 1.6 | -7.6 | 1.3 | 1.6 | 7.6 |
| 75-79 | 0 | 0 | 0 | 0 | 0 | 0 |
| 80-84 | 0 | 0 | 0 | 0 | 0 | 0 |
| 85-89 | 0 | 0 | 0 | 0 | 0 | 0 |
| 90-94 | 0 | 0 | 0 | 0 | 0 | 0 |
| 95-100 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Avg.** | 0.32 | -0.12 | -0.09 | 2.39 | 2.29 | 2.04 |
| **Diff. in Avgs.** | | -0.44 | -0.41 | | -0.10 | -0.35 |

Table 3.8.3.b: Bias and Absolute Bias for 99% of Poor Households: New Scorecard Urban and Rural Areas

| | A | | | B | | |
|---|---|---|---|---|---|---|
| | **Diff.** | | | **\|Diff.\|** | | |
| **Score *i*** | **PPI** | **Urban** | **Rural** | **PPI** | **Urban** | **Rural** |
| 0-4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5-9 | 0.5 | -4.3 | 0.5 | 0.5 | 4.3 | 0.5 |
| 10-14 | -5.4 | 5.2 | 0.2 | 5.4 | 5.2 | 0.2 |
| 15-19 | 6.9 | 6.2 | -0.5 | 6.9 | 6.2 | 0.5 |
| 20-24 | 7 | -3.5 | 7.8 | 7 | 3.5 | 7.8 |
| 25-29 | 5.7 | -4.6 | 5.6 | 5.7 | 4.6 | 5.6 |
| 30-34 | -2.9 | -8.1 | 5.0 | 2.9 | 8.1 | 5.0 |
| 35-39 | -8.4 | 1.7 | -2.5 | 8.4 | 1.7 | 2.5 |
| 40-44 | 0.8 | 2.7 | -2.5 | 0.8 | 2.7 | 2.5 |
| 45-49 | 1.3 | 1.4 | -4.6 | 1.3 | 1.4 | 4.6 |
| 50-54 | 3.4 | 2.5 | -3.3 | 3.4 | 2.5 | 3.3 |
| **Avg.** | 0.81 | -0.06 | 0.51 | 3.85 | 3.66 | 2.94 |
| **Diff. in Avgs.** | | -0.87 | -0.30 | | -0.18 | -0.90 |

Figure 3.8.2 provides a visual representation of table 3.8.3a. We see that in the first panel the log-fitted lines are very close to the 0 mark, indicating a very low level of bias. The second panel better illustrates that in most instances the bias is in opposite direction for urban and rural areas with the new scorecards: when the new urban scorecard underestimates poverty at a given score, the new rural scorecard overestimates it.

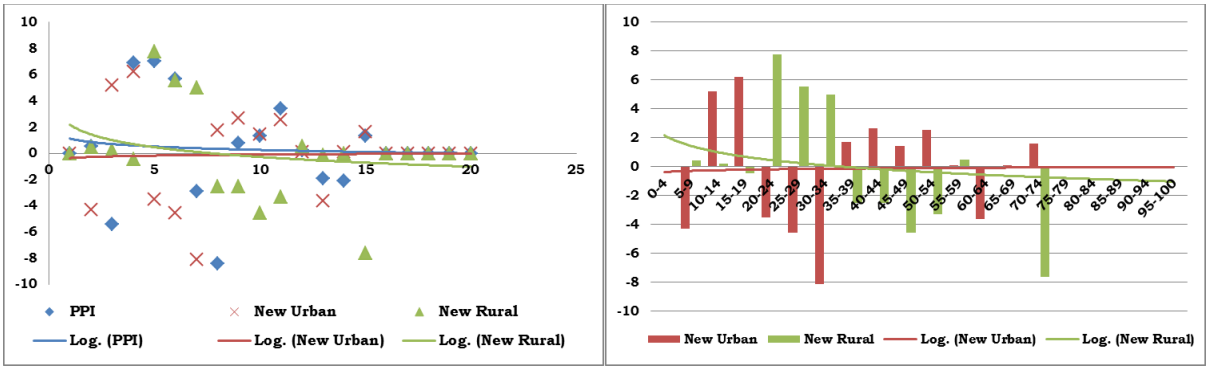



Figure 3.8.2: Bias (Diff.): New Scorecards for Urban and Rural Areas

Figure 3.8.3 visually summarizes the absolute bias presented in table 3.8.3a, column B. The log-fitted line for the new rural scorecard clearly shows that the bias for the new rural scorecard is much lower than for the original PPI and for the new urban scorecard. This is particularly more pronounced in the lower scores (0-4 to 20-24), which is ideally what we want.

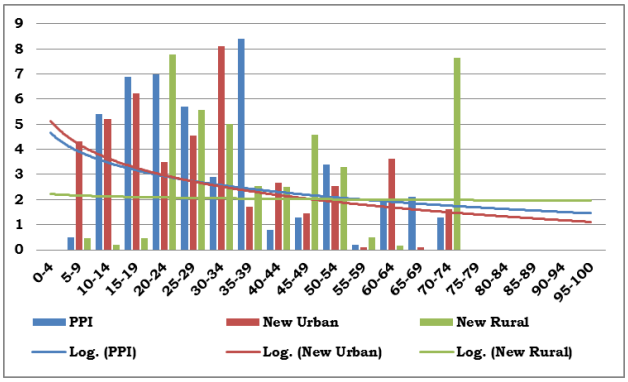



Figure 3.8.3: Absolute Bias (|Diff.|): New Scorecards for Urban and Rural Areas

The results of this section show that although the relationship between variables and poverty is the same across rural and urban settings, the intensity and relevance of this relationship differs across settings. For example, the contribution of each question to the model

differs in the two settings, which provide some indication of the need for using different points. Similarly, table 3.8.1 shows that there are some variables that are nor relevant in rural areas, such as having parquet floors (or similar). Recall that there is no coefficient for that variable, since only 1% of households in rural areas have parquet floors. This provides evidence that it might be necessary to select a set of variables that might be more appropriate for rural areas to predict poverty.

Moreover, the results in this section show that obtaining different points for the same indicators in rural and urban areas slightly improved targeting accuracy with slightly higher poverty likelihoods (new rural scorecard), slightly better SOC curve (new rural scorecard), much higher inclusion rates (at the 30-34 cut-off) for the new urban scorecard (compared to the original PPI applied to urban areas) and lower leakage in rural areas (compared to the original PPI applied in rural areas), and a slight improvement for the ROC curve for the new urban scores compared to the national ROC curve, and for the ROC curve for the new rural scorecard when compared to the PPI applied in rural areas. Yet, the results show that targeting accuracy is still much lower in rural areas across all models, which indicates the need to find a way to ameliorate this other than obtaining new scores for rural areas using the original 10 indicators in the PPI.

Finally, in terms of bias, the new scorecards showed an improvement across the board: less bias and absolute bias in rural and urban areas; however, when looking at the magnitude of these improvements, these do not seem very substantial overall. Yet, the greatest improvements came for the new rural scorecard (where most of the poor live) and for the lower distribution of scores (where most of the poor are located).

## 3.9 Conclusion and Recommendations

The objective of this paper was to evaluate the PPI in terms of its accuracy by looking at bias, precision, and targeting accuracy using data from the ENAHO 2007 for Peru as a case

study. The empirical approach used in this paper involved three steps. As a vital first step, this paper set out to do a pure replication of the results presented in Schreiner (2009). Second, this paper set out to do a more scientific replication (with a lower case "s") in order to answer three inter-related questions: 1) following previous research that looks at accuracy in non-nationally representative sub-groups we apply the original PPI to rural and urban areas to see if targeting accuracy and bias differ across these two settings; 2) this paper then tested if adding rural and regional variables might improve targeting accuracy and bias; and 3) we went a sept further and created two poverty scorecards with the same indicators for rural and urban areas but with different points to see if targeting accuracy and bias could be improved.

The paper successfully replicated the key results presented in Schreiner (2009). The results of this exercise show that the original results in Schreiner (2009) and our replicates are essentially identical with some very minor differences that may be due to rounding; in this sense, there were no errors found on the part of the original PPI. This is an important finding in its own right as previous replication projects have shown that errors in empirical economic research in published articles may be quite common, while the success rate of pure replication studies is very low (Dewald et al. 1986; McCullough et al 2008; Herndon et al., 2013).

Two additional points should be noted regarding the lessons learned from the pure replication. First, a more clear understanding of how the PPI is created and how accuracy is measured put us in a better position to make slight modifications introduced in sections 3.6 through.3.8. Second, we identified the following characteristics of the PPI that we argue are important: 1) the PPI can be considered a multi-dimensional measure of poverty (or wealth) that is superior than the ones used by OPHI (2013) since the 'dimensions', variables, and weights come from a statistical and objective approach; 2) from the way the variables are created, it is

possible to identify non-linearities that might be important for scoring purposes; 3) the process of creating the scorecard involves statistics, judgment on the part of the researcher (e.g. selecting variables that change over time, that are easy to verify, hard to falsify, and that apply to all households), field testing and user reviews, all of which should increase the accuracy of the scorecard, the relevance and clarity of questions, and the likelihood of adoption; 4) the final scores do not include negative numbers, decimals, or values that need to be exponentiated, all of which make the tool simple to use and more likely to be adopted; and 5) the vast majority of poor households are located in the lower end of the distribution of scores. The latter point means that when looking at bias, it might be more informative to concentrate in the lower end of the distribution of scores, which is where most of the potential pro-poor organizations' clients might be located.

The results of the "s"cientific replication showed mixed results. The key results are presented in table 3.9.1.

Table 3.9.1. Summary of "S"cientific Replication Results

| | **Original PPI Scorecard** | | | | | **New Scorecard** | |
|---|---|---|---|---|---|---|---|
| | | **I** | | **II** | | **III** | |
| **Indicator** | **Nation** | **In Rural** | **In Urban** | **Plus Rural** | **Plus Rural & Reg.** | **Rural** | **Urban** |
| Inclusion (SOC)^ | 76.20% | 89.34% | 53.86% | | 79.41% | 84.43% | 70.17% |
| Leakage^ | 16.30% | 49.33% | 6.76% | | | 40% | 11.54% |
| ROC | | ↓ | | | | ↓ | |
| Bias | | ↓ | ↓ | | ↓ | | |
| \|Bias\| | | | | | | | |
| Focused Bias | | ↓ | | | ↓ | | |
| \|Focused Bias\| | | | | ↑ | | | |

^ Cutoff of 30-34

↓ A decrease in accuracy of more than 1% for bias, or visually significant decrease of ROC; and the inverse interpretation for ↑.

For the first exercise for the "s"cientific replication presented under column I—applying the PPI to non-nationally representative samples—the results show that inclusion rates are much higher in rural areas but much lower in urban areas. On the other hand, leakage rates are extremely high in rural areas but fairly small in urban areas. This shows that the PPI has much lower discriminate ability in rural areas: while it is very accurate at scoring poor households in the lower end of the distribution of scores (as seen in the SOC curves), it is not as capable of discriminating between the poor and non-poor, as seen by a much lower ROC when the PPI is only applied to rural areas (see figure 3.6.1c, 3.8.1c and 3.8.1d). In terms of bias, when the PPI is applied separately to rural or urban areas, the bias increases by more than 1%, and this difference increases to more than 2% when looking at the focused bias (looking scores at 50-54 or below); although these levels of bias are still fairly small.

For the second "s"cientific replication where we add a variable for rurality and regional variables (column II in table 3.9.1), inclusion rates improve slightly but the accuracy in bias deteriorates by more than 1%; yet, the levels of bias are still fairly small.

Finally, looking at column III of table 3.9.1 for the results of the two separate and new scorecards, the results show that inclusion rates for rural areas are better when compared to those from the nation (column I), and are also better when compared to the model that adds rurality and regional variables (column II). Similarly, inclusion rates significantly improved for urban areas when compared to the original PPI applied to urban areas only (column I): going from 53.86% to 70.17%. Similarly, leakage rates improved for rural areas when compared to the original PPI applied to rural areas only (column I): going from 49.33% to 40%. However, leakage rates deteriorate for urban areas when compared to the original PPI applied to urban areas only (column I). In terms of the more powerful measure of targeting accuracy, the ROC

shows very similar results for rural areas with the new scorecard than when the original PPI is applied only to rural areas: in both instances, the ROC shows a much lower predictive ability than in any other model. As for bias, there are no significant changes in this regard; although all the measures of bias do improve across the board, as shown in tables 3.8.3a and 3.8.3.b even though there is really not much room for improvement to begin with.

Taking all these into account, there are five main result of this "s"cientific replication. First, when applied to the nation, the original PPI has excellent measures of accuracy in terms of bias, absolute bias (for the entire sample, and when the analysis concentrates on scores that cover ~99% of poor households), and targeting accuracy. Second, when the original PPI is applied to non-nationally representative samples, namely, to rural and urban areas, the PPI has very high leakage rates in rural areas and less than desirable inclusion rates in urban areas. Third, adding rurality and regional variables to the PPI does not make any significant improvement to the PPI's levels of accuracy (in terms of bias or targeting accuracy). Fourth, creating two separate scorecards using the same variables as the original PPI provides some important improvements, although there are still some tradeoffs: inclusion rates are better for rural and urban areas (84% and 70%, respectively) in comparison to these rates for the nation; overall, leakage rates are better in rural areas in comparison to the original PPI applied only to rural areas, and bias stays almost identical; but leakage is slightly greater with the urban scorecard than when the PPI is only applied to urban areas. Finally, and this may be the most important finding, the ROC, which plots the predictive ability of the tool, is always much lower in rural areas, independent on the model used. Although the speed of coverage curve (SOC) shows that any variation of the PPI is very good at scoring poor households in the lower end of the distribution of scores (high

inclusion rates), the same does not hold true when it comes down to discriminating non-poor households (high leakage).

These results indicate that if one is primarily interested in measuring poverty rates and the bias associated with this measure, then, the PPI does a great job in this regard, regardless of where the tool is applied (nationally, in urban areas, or in rural areas), while any modification to the PPI will not give us any major improvements in this regard. However, if we are also interested in being able to discriminate or differentiate between "truly" poor and non-poor households, the tool is not as successful when it is applied to rural areas—either when applying the original PPI or with the new scorecard for rural areas. This comes out clearly when looking at the ROC curve in both instances. On the other hand, the tool does a great job in terms of targeting accuracy when the PPI is applied at the national level or in urban areas, as shown by the ROC.

In the spirit of making objective suggestions to "improve" the PPI, the most important suggestion this paper can make based on the results of this paper, then, is for the developer of the PPI to create an additional set of indictors that are highly predictive of poverty status in rural areas so that two separate scorecards can be created. This is warranted for two reasons. First, table 3.6.1 shows that for 7 out of 10 indicators, there is little variation in these in rural areas—either most rural households do not have these indicators/characteristics, or very few of them do, such as parquet floors (or similar) from question 3, which dropped out of the logit regression for rural areas since only 1% of households have these type of floors. Second, and related to the first point, the ROC curve for rural areas clearly shows the PPI is not as successful at discriminating between poor and non-poor households. It will be important to see if a set of indicators that might be more relevant for rural areas yields a rural scorecard that can significantly improve the

targeting ability of the PPI in rural areas. This might be a worthwhile exercise. After all, most of the poor households in Peru live in rural areas. Some of these indicators might include having agricultural land, livestock, or having agriculture as the main occupation of household members. In fact, and as noted, some of the top indicators that are most strongly linked to higher poverty likelihoods presented in Figure 3 in Schreiner (2013) for the latest PPI for Peru already include some of these indicators; however, none of these make it to the final list of 10.

Of course, this only addresses the issue of whether the PPI can improve its targeting accuracy in rural areas—as all other measures of accuracy covered in this paper do not need much improvement. Assuming that the new rural scorecard improves targeting accuracy, it will also be important to consider whether having two scorecards—one for rural and one for urban areas—will be well accepted by the end-users. After all, if improving targeting accuracy comes at the expense of end-users being confused from having two tools, then this might be a high price to pay for improving accuracy, since it might lead to lower rates of adoption. Yet, if end-users of the PPI find that a rural and an urban scorecard can improve their targeting accuracy in each of these two settings, and if this is valuable for them, then it might be worth exploring the possibility of creating two scorecards.

# CHAPTER 4

# UNCONDITIONAL CASH TRANSFER PROGRAMS AND AGRICULTURAL PRODUCTION: THE CASE OF ZAMBIA

## 4.1 Introduction

Sub-Saharan Africa (SSA) is one of the world's regions with the highest levels of poverty: three-quarters of the rural population are considered poor (IFAD 2011). Agriculture remains one of the key economic activities for rural households in this part of the world. Addressing rural poverty in these settings will most likely require policy interventions that seek to improve agriculture production (Boone, et al. 2013). In fact, governments have recognized that the agricultural sector can be the backbone and potential source of growth where farming is the dominant livelihood activity for the majority of the poor (Devereux and Guenther 2007). Yet, increases in productivity in agriculture may be just one mechanism by which poor rural households can improve their wellbeing. Increasing their livelihood options through diversification in agricultural activities as well as non-agricultural activities may also facilitate poverty reduction (Ellis 2000).

A big share of the focus in poverty reduction efforts in the past two decades world-wide has been on the direct provision of conditional cash transfers (CCT) to the poor, starting with *PROGRESA* (later called *Oportunidades*) in Mexico and *Bolsa Familia* in Brazil, these programs have spread throughout Latin America and the world (Fiszbein et al. 2009). The ultimate goal of these CCTs is improving the short-term welfare of poor households by improving consumption levels, but also improving long-term welfare by investing in human capital formation for the young, particularly in the areas of education and health (Ibid.) While this is undoubtedly important, recent research indicates that focusing exclusively on investing in human capital formation for the young may miss opportunities to make these programs part of a broader

agricultural development strategy that can incorporate productive components which can also help alleviate poverty by increasing productivity (Handa and Davis 2006) and decreasing risk (Devereux and Guenther 2007).

A new line of research seeks to find the balance in channeling resources between these two competing approaches in a way that synergies can be obtained to make the most use of these transfers (Farrington, Harvey, and Slater 2005; de Janvry and Sadoulet 2009; Devereux and Guenther 2009; Devereux 2009; Sabates-Wheeler, Devereux, and Guenther 2009). For instance, the government of Ethiopia has shifted its social protection policies away from policies aimed to minimize risk (such as food aid programs) and towards policies aimed to promote growth by combining safety net programs with productive interventions aimed to improve market chains and agricultural exports, improve infrastructure with public works, and provide extension service packages for agricultural production (Devereux and Guenther 2007). Elsewhere in SSA, other policies to use the agricultural sector as the engine for growth have been considered, including increasing access to modern agricultural technologies, providing smart subsidies for fertilizers (Ricker-Gilbert et al. 2011), and investing in rural infrastructure and agricultural extension services (Johnson et al., 2003).

However, one key constraint to the strategy of promoting agriculture in SSA is related to the lack of access to credit. Adesina (2010) notes that less than 3% of commercial credit goes to agriculture in these regions. Cash transfers can serve as a vehicle to overcome credit constraints, particularly among smallholder and poorer farmers. Thus, cash transfers can be an important complement to a broader agricultural development agenda; they can serve not merely as a safety net but as a source of added support to promote farm level productivity gains (Boone et al. 2013, Devereux 2009; Gilligan, Hoddinott, and Taffesse 2009).

In the past decade, a growing number of African governments have launched cash transfer programs that are not conditional as part of their strategies of social protection. Davis et al., (2012) explain that most of these programs have focused on vulnerable populations in the context of HIV/AIDS leading to an emphasis on those that are ultra-poor, labor-constrained, and with prevalence of adverse health conditions, elderly and/or caring for orphans and vulnerable children. Thus, most of these programs focus on food security, health, nutritional and educational status, particularly of children; yet, most of the accompanying impact evaluations pay little attention to the analysis of livelihoods, or the current economic and productive activities of beneficiary households (Asfaw et al. 2012). Asfaw et al. (2012) note that there is good reason to believe that cash transfer programs will influence the productive dimension of beneficiary households, particularly, since a more clear exit path from poverty may be through self-employment generated by beneficiary households themselves, whether within or outside agriculture. This might be particularly more relevant for households that are not labor-constrained and that might face unemployment or underemployment.

At the same time, there is growing interest in investigating if the application of conditionalities to cash transfers is necessary for recipient households to invest in education and health. There are at least two reasons why economists might think it is disadvantageous to attach conditions to cash transfers, as noted by Fiszbein et al. (2009). First, some of the neediest households might find the conditions too costly to comply with if, for instance, the health clinics are too far away to attend, and may thus be deterred from taking up the benefit. Conditions might exclude some of the people the program aims to reach who might be the most in-need. This, in fact, has been a key finding by González-Flores, Heracleous and Winters (2012) for the case of the urban *Oportunidades*, in Mexico, and by Lund, Noble, Barnes, et al. (2009) in South

Africa for the case of the Child Support Grant program. Second, households that choose to comply with the conditions may incur a costly distortion to their own behavior to obtain a little extra cash in the short-run. For instance, if households know the precarious conditions of schools (or clinics), Fiszbein et al. (2009) note that perhaps it is wasteful for the children to spend time there rather than learning how to work in their farm. By "pushing poor households to do something that they would otherwise not be doing, CCTs might be imposing costly distractions on people who are trying to do the best thing for their families under conditions of severe scarcity" (Ibid., p. 46). Despite the lack of conditions, however, recent research in Africa shows that when targeting the very poor, unconditional cash transfers do have a positive impact on education and nutrition outcomes (Baird, McIntosh and Özler 2010; Robertson, Mushati, Eateon, et al. 2013; Agüero, Carter, and Woolard 2006). Given the major differences between Latin America and Africa in terms of the quality and quantity of service provision, the capacity to implement conditionalities, and cultural and political differences, Schubert and Slater (2006) conclude that the benefit/cost ratio of conditionalities may make the use of conditions in cash transfer programs in Africa inappropriate.

The main objective of this paper is to assess the potential role of cash transfers in promoting agricultural production among smallholder farmers. Towards this end, this paper will analyze the unconditional cash transfer Child Grant Program (CGP) implemented in 2010 by Zambia's Ministry of Community Development and Social Services. As with similar cash transfer programs the CGP aims to supplement household income, increase education and health outcomes and improve the overall nutrition of household members. While this is officially an unconditional cash transfer program, Winters (2013) notes that the CGP might be considered a program with "soft conditions" in the sense that households are told the program is aimed to

improve education, health, and nutrition, particularly for children. The CGP has six specific objectives: 1. To supplement and not replace household income; 2. To increase the number of children enrolled in and attending school; 3. To reduce the rate of mortality and morbidity among children under 5 years old; 4. To reduce stunting and wasting among children under 5 years old; 5. To increase the number of households owning assets such as livestock; and 6. To increase the number of households that have a second meal a day. In theory, 1 and 6 (and possibly 5) should improve numbers 3 and 4. Importantly, the CGP does not have any mechanisms, components, or conditions by which the health and education outcomes are to be achieved. The same is true for objective 5 related to the increase of livestock assets. In other words, the expectation is that even without conditions households will spend their cash transfers in such a way that they will eventually experience improvements in these indicators. Implied in this approach is the belief that households would spend in these items if they had access to financial means. This is very much in line with Hanlon, Barrientos, and Hulme (2010) who posit that if we want to help poor people, we need to just give money to the poor. The authors' main argument is that poor households know what is best for them and one of the key constraints they face access to funds.

Since the CGP targeted rural areas, the vast majority of program beneficiaries depend heavily on agriculture. This provides the opportunity to assess the influence of transfers on agricultural production, particularly objective 5, which aims to increase the number of households owning assets, such as livestock. Previous research in other sub-Saharan countries shows that cash transfers have an impact on agricultural and non-agricultural productive choices (Covarrubias et al. 2012). This paper aims to contribute to the literature in this regard by investigating if a cash transfer without conditions may lead to positive effects on productive investments that can potentially improve agricultural production. Nearly all evaluations of social

protection programs have concentrated on those in LA while the evidence in SSA is much more limited (Gilligan, Hoddinott and Taffesse 2009), although it is growing rapidly with several pilot cash transfer programs being implemented and evaluated in SSA (Davis, Gaarder, Handa, et al. 2012).

More specifically, the objective of this paper is two-fold: (1) to assess the impact of the CGP on agricultural production (value and size of harvest, and gross margins), and productive investments; and (2) to identify the technical efficiency gap for all households in the program and to assess the impact of the CGP in this regard. Most evaluations in the agricultural sector primarily concentrate on technological change (TC) indicators such as increases in yields (objective 1 of this paper); yet, little research has been done on the potential effect of programs on technical efficiency (TE) (Winters et al. 2010, 2011) (objective 2 of this paper). While a focus on TC indicators allows the researcher to identify impact on different components of production, it does not provide any information on whether farmers made the right use of the available inputs and technology at their disposal, i.e., managerial performance is ignored (González-Flores et al. 2014). Combining Stochastic Production Frontier Analysis (SPFA) with impact evaluation methodologies provides a useful alternative for measuring the productivity impact of agricultural projects, or in this case, the impact of the CGP. SPFA can help identify the levels of efficiency (or inefficiency) and therefore, this approach makes it possible to quantify the potential to increase agricultural output without the need for additional inputs or new technology (Coelli et al., 2005). This is attractive since policy makers are be able to target resources in the most appropriate manner (Solís et al. 2009). The narrowing of this inefficiency gap, in fact, has been put forth as an alternative to improve productivity levels in various African countries (Binam et al. 2008; Nyariki, 2011; Abdulai et al. 2013).

To this end, the remainder of the paper is structured as follows. Section 4.2 describes key aspects of the program. Section 4.3 provides the conceptual framework. Section 4.4 describes the data and Section 4.5 provides the econometric methods used in the analysis. Section 4.6 provides the results and Section 4.7 concludes.

**4.2 Description of the Child Grant Program**[40]

The Child Grant Program is one of seven unconditional cash transfer programs implemented in SSA under FAO's From Protection to Production (PtoP) project, which is a collaborative effort with the UNICEF Eastern and Southern Africa Regional Office and six countries in the region; it is supported by funding from the DFID Research and Evidence Division, as well as the FAO Multi-Partner Programme Support Mechanism (FMM). The project is part of a larger Transfer Project in partnership between FAO, UNICEF, Save the Children UK, and the University of North Carolina in supporting the design, implementation and impact evaluation of cash transfers in sub-Saharan Africa (FAO 2012).

The CGP was implemented by Zambia's Ministry of Community Development and Social Services in 2010 in three of the poorest districts of the country—Kalabo, Kaputa, and Shongombo—that have the highest rates of mortality, morbidity, stunting, and wasting among children under 5 years old. In addition to the geographic targeting, the CGP used categorical targeting where any household with a child under five years old was considered to be eligible. Beneficiary households receive 55,000 kwacha a month (equivalent to $11USD) independent of household size. To put this amount in perspective, it is deemed large enough to purchase one meal a day for everyone in the household for the entire month. The transfers are made every other month through a local pay-point manager. As with other cash transfer programs, the primary recipient of the transfer is the primary female care-giver of the household. In contrast to

[40] The description of the program and data come from Seidenfeld, et al. (2011).

some of the biggest cash transfer programs in the world, such as *Oportunidades*, the CGP does not impose any conditions attached to the cash transfer. However, it is thought to be a "soft conditional" cash transfer program in the sense that beneficiaries are told that the transfers are meant to be used to improve education, health, and overall nutrition.

The CGP was designed as a Randomized Controlled Trial (RTC) using randomized phase-in (Duflo, Glennerster, and Kremer 2008) that includes several levels of random selection. First, 90 out of 300 communities (Community Welfare Assistance Committees (CWACs)) in the three districts were randomly selected and ordered through a lottery to be considered in the program. The random selection was done in a transparent way that included the participation of CWACs members. In a second phase, CWAC members, and Ministry staff identified all eligible households with at least one child under the age of 5 living in these 90 randomly selected communities.

The baseline data collection began before CWACs were randomly assigned to treatment and control. Neither the households nor the enumerators knew who would benefit first and who would benefit later. The randomization of the communities, then, was done once baseline data had been collected. The randomization was done publicly with the participation of local officials, Ministry staff, and community members with the flip of a coin. Half of the selected communities were assigned to treatment and began receiving benefits in December of 2010. The second half of the communities serves as the control, and they were scheduled to receive the program at the end of 2013.

### 4.3 Conceptual Framework and Literature Review

The impact of cash transfers on households depends on the type of household or individual receiving the transfer, the manner in which transfers alter incentives, and in the differences in asset ownership and the constraints households face (Winters and Davis 2007).

For the case of the CGP, all households come from three of the poorest districts in Zambia, they have been identified to be very poor and to have at least one child less than 5 years old in the household. This reflects that targeting mechanism described in Section 4.2 and it suggests that households have low asset endowments and might face a series of significant constraints, including lack of access to credit, infrastructure, and input and output markets, among others, as it is typically the case in rural areas of Zambia (World Bank 2007).

Given this profile of households, one might expect that the transfers will have a strong income effect, even if there are no conditions attached to the transfer. For instance, by receiving cash transfers from the CGP, households may respond by demanding more goods and services. Similarly, the cash transfers may also induce households to make productive investments, especially if households are credit constrained (Fenwick and Lyne 1999; Scott 2000) and if investments are limited due to these constraints (Winters and Davis 2007).

A good starting point to understand the potential influence of cash transfers on agricultural production in the context of three of the poorest districts of rural Zambia, where the CGP was implemented, is to use the agricultural household model as presented by Singh, Squire and Strauss (1986). These household models are a staple for micro-level research on less-developed country rural economies; they form the basic building blocks for micro economy-wide models, and have been used to reflect imperfect-market environments that are typical in less-developed countries (Taylor and Adelman 2003). Building on the classic works by Singh et al. (1986), Asfaw, Daidone, Davis, et al. (2012) provide an analytical framework for evaluating the productive impacts of cash transfer programs on household behavior. This section relies primarily on this framework, unless other sources are cited.

Most rural households of the developing world earn at least part of their livelihoods through work in their own enterprises; particularly from agricultural endeavors. Households consume at least a portion of their own production, and household labor is often an important input into the production process of the household enterprise (Bardhan and Udry 1999). Singh et al. (1986) provide a relevant and useful approach to investigate household decision-making in these contexts by utilizing an agricultural household model where households are both utility-maximizing consumers of agricultural goods and are simultaneously profit-maximizing producers of said goods, while potentially facing market constraints (Singh et al., 1986). This hybrid of the economics of the firm and of the household is characteristic of the situation of most families in rural areas of the developing world and this provides the starting point for this analysis.

In the agricultural household model, when markets function perfectly, production and consumption decisions can be viewed as being "separable". All prices are assumed to be exogenous; households are price takers. Moreover, households do not face labor, credit, or other market constraints and are assumed capable to hire labor at the going real market wage and obtain credit at the going real interest rate. Households face no tradeoff between own consumption of agricultural goods or production for sale since there are no transaction costs. Given these assumptions, agricultural households in the "separable model" can solve the profit and utility maximization problem recursively in two steps: they first maximize profits by choosing the optimal level of labor. Thus, production decisions depend only on prices and the characteristics of the land and not on household's endowments or preferences (Bardhan and Udry 1999). In the second step, the profits from step 1 become part of the budget constraint and households then maximize their utility subject to this constraint. The solution yields standard

demand curves that depend on prices and income. Since income is partly determined by the household's production activities, it follows that changes in the factors of production will change income from profits, and hence, consumption behavior. Thus, consumption behavior is not independent of production behavior and this establishes the recursive property of the model (Ibid.).

An important characteristic of the "separable" agricultural household model is the fact that there is no subsistence constraint as noted by Taylor and Adelman (2003). Asfaw et al. (2012) further note that if this model reflects reality, cash transfers should have little effect on agricultural production and if markets are perfect, then, spending and investment in agriculture should be optimal and the effect of the transfer should be found only on consumption. While some studies find that the "separable" model is appropriate for some contexts, Carter and Yao (2002) note that market failures create non-separabilities that differentially constrain households depending on how integrated they are to the market. Given the level of extreme poverty and isolation in these three districts of Zambia, it seems that the "separable" model is not the appropriate one to use.

A more realistic model for these three districts in Zambia is found in the "non-separable" agricultural household model where one or more markets do not work. In contrast to the assumptions underlying the "separable" model, agricultural households in developing countries may face significant barriers in multiple markets, such as credit, insurance, labor, input and output. For instance, Key, Sadoulet and de Janvry (2000) find that high transaction costs in staple markets can often make self-sufficiency the optimal choice. When transportation costs, time, and information gathering add costs to selling food, this creates a price difference between the selling and buying price. In this sense, the more undeveloped the market is the higher these

costs. Depending on the level of variable and fixed transaction costs, it is possible to have farmers completely exit the market, while others may be more likely to be inserted in these markets. Similarly, labor transaction costs, such as monitoring work effort, can prevent households from purchasing labor in the market and may influence their preference to use family labor over hired labor making family and hired labor imperfect substitutes.

Two other important inter-related market failures add additional constraints to the agricultural household in the "non-separable" model, namely, credit and insurance. Poor households lacking assets or a legal document showing ownership of such asset have difficulties in accessing credit since they cannot use these as collateral to access credit (de Soto 2000; Bowles 2004, Ch. 9). Similarly, households may be credit rationed in the presence of adverse selection, asymmetric information, or due to government policies (Feder, Lau, Lin and Luo 1990). In the absence of credit, poor households often rely on assets, such as livestock, as a form of savings or insurance in the event of a shock. Asfaw et al., (2012) note that there are several issues with this type of risk coping strategy. First, many assets suffer from lumpiness, which increases the difficulties in saving in the first place. Second, when households in the same communities are faced with common risks rather than idiosyncratic ones, these assets usually suffer price drops as more households try to sell their assets as a coping mechanism. The interaction of market failures in the credit and insurance markets can significantly hamper investments in inputs, as well as on productivity-enhancing equipment or tools, and farm animals. Surely, liquidity and credit constraints are two of the main factors limiting poor agricultural households from investing optimally (Fenwick and Lyne 1999; Asfaw et al. 2012). For instance, fertilizers can greatly increase farm output, or the use of pesticides and fungicides can serve as damage abatement strategies to minimize output loss (Lichtenberg and Zilberman,

1986; Zhengfei et al., 2005; Cavatassi et al., 2011). Unfortunately, poor agricultural households often lack the resources to purchase these investments without credit (Dorward and Chirwa 2011). Feder et al., (1990) note that the types and quantities of inputs used by agricultural households differ from their optimal levels when faced with a binding constraint on liquidity, which has implications not only on production but also on technical efficiency. In fact, one of the key findings by Binam, Gockowski and Nkamleu (2008) for the case of cocoa farmers in West African countries is that having access to credit has the greatest effects on improving technical efficiency. Similarly, without access to credit or insurance markets, agricultural households adopt low-risk low-return strategies or sell more than the optimal amount of labor off farm in order to provide a variety of sources of income (Asfaw et al., 2012). This has implications for poverty not only at the household level but at the societal level as well (Bowles 2004, Ch. 9).

If household production and consumption decisions are "non-separable", then, cash transfers may be able to alleviate or overcome several of these issues. First, transfers provide a guaranteed steady source of income. This assurance, especially for agricultural households which are less likely to have regular sources of income, might allow households to adopt riskier strategies with a higher rate of return, such as increasing their use of productive seasonal inputs (Scott 2000). This guaranteed flow of income can help make up for failures in the insurance market, which might also mitigate stress sales and decumulation of assets. Second, the additional cash can be used for productive investments, such as investing in farm tools or animals, even if the cash is aimed to (unconditionally) improve education, nutrition, and health. Yet, even in the case where a cash transfer is attached to conditions, research in Latin America shows that CCT programs do alleviate liquidity constraints and facilitate productive investments. For instance, using data from *Oportunidades-Progresa*, in Mexico, Gertler, Martinez and Rubio-

Codina (2006) find that beneficiaries used part of their cash transfers (after consumption on goods and services) to invest in micro-enterprises, farm assets, and agricultural activities, which lead to increases in income, consumption, and overall better living standards in the mid- to long-term after the program ended. Also looking at *Oportunidades*, Todd, Winters, and Hertz (2010) find that the program increased land use, livestock ownership and spending on crop production, as well as the probability of consumption from own production, and the value and variety of consumption.

Similar results have been found with unconditional CT programs in Africa. For instance, using panel data from Kenya and Malawi, Zezza, de la Brière and Davis (2010) evaluate the impact of two social cash transfer programs on income-generation strategies and productive activities and find that beneficiary households are more likely to invest in durable goods in Kenya, and in small animals and tools in Malawi. Covarrubias, Davis, and Winters (2012) look at a social cash transfer program in Malawi and find that the program generates agricultural asset investments, reduces adult participation in low skilled labor, and limits child labor outside the home, while increasing child involvement in household farm activities. Finally, Asfaw, Davis, Dewbre, et al. (2012) using data from Kenya find that the Cash Transfer Program for Orphans and Vulnerable Children had a positive and significant impact on the accumulation of productive assets, especially on the ownership of small livestock such as sheep and goats.

These results show that the liquidity provided by cash transfers seems to move farmers closer to a better level of input use when credit markets have failed; they help diversify the sources of income, and facilitate investments in inputs, small farm animals, tools and assets. Such investments can be complementary by household labor and can lead to increased agricultural production by the household (Asfaw et al., 2012). Importantly, if farmers are not

acquainted with the use of inputs they might be using for the first time or with greater quantities and they are not properly trained in applying these inputs this might negatively affect their levels of efficiency (Schultz 1975). Finding the optimal way to apply these inputs might necessitate a process of learning by doing (Sipiläinen and Lansink 2005), learning from others (Conley and Udry 2010), or policies geared to improve technical efficiency (Binam et al. 2008; Nyariki, 2011), or more formal training such as extension services (Ali and Byerlee 1991; Abdulai, Nkegbe and Dongoh 2013; Binam, Gockowski, and Nkamleu 2008). This suggests that the use of new agricultural inputs might also have an impact on their levels of technical efficiency. Thus, this needs to be explored.

Given this discussion, the impact of the CGP can be manifested in various ways: the impact of cash transfers can be manifested through changes in household behavior, relaxing liquidity constraints, facilitating investments, and possibly in the levels of technical efficiency if new inputs are now being used. While we are ultimately interested in assessing whether the CGP led to increases in agricultural production or the value of harvest, and technical efficiency, it might be advisable to focus on more direct and intermediate impacts for two reasons (Asfaw et al., 2012): first, it will be informative to understand the mechanisms of impact—it is not enough to know whether cash transfers increase production or the value of harvest (income), but how they increased production (or income)—investment equipment, tools and farm animals, and/or different labor allocation, different use of inputs, shift in activities, etc. Second, given that income generating activities, including agricultural production are mediated by factors outside the control of the program and the household—such as prices, weather and shocks, and access to input and output markets—we may not see any impact on higher level outcomes, but it might be possible to identify an impact among the intermediate outcomes. Thus, while this paper looks at

increases in production and the value of harvest and technical efficiency, we should expect to find a more clear impact of the program on intermediate outcomes.

### 4.4 Data Collection and Description

In order to evaluate the impact of the CGP, this paper uses baseline and follow-up data encompassing four surveys—two are for baseline and two are from a follow-up two years after the initiation of the program. The data were collected by the American Institutes for Research (AIR). The baseline and follow-up rounds include household and community level data collected in 2010 and 2012, respectively.[41]

Surveys administered at the household level include information for household demographics, health, education, economic activity, income, assets, household amenities and conditions, access to facilities and programs, agricultural production, livestock and animal production, self-assessed poverty and food security, gender and reproduction, savings, and expenditures. The community level surveys include information on access to schools, migration patterns, agricultural prices, borrowing information, social capital, external shocks to the community, and wages and prices.

From this information it is possible to compare household level variables to verify if the randomization process yielded a reasonable control group at baseline. Given the nature of the data, it is possible to control for differences at baseline if they exist. Moreover, the detailed modules related to agricultural production have information on input use and family and hired labor, thus, it is possible to look at agricultural production and technical efficiency at the farm level. The baseline and follow-up data include information for 2,515 households corresponding to 14,565 individuals. Half of these households are in control communities and the other half are

[41] A third survey was implemented at follow-up to capture increases in own-business enterprises. This paper does no use that information.

in treated communities. AIR paid special attention "…to the process and timing of data collection, making sure that it was culturally appropriate, sensitive to Zambia's economic cycle, and consistently implemented" (AIR 2011, p. 9). For instance, the data at baseline and follow-up were conducted during the same time of the year between September and October to be sure the data was not picking up seasonal differences across years. In general, the planting season starts in December/January and harvest is in May/June (FAO 2014).

As with most RCTs in the social sciences, it is rare that all participants—treated and controls—remain in the study. For the CGP, follow-up data reveals that 9% of the sample was not reached or the surveys were not completed. If attrition was systematic, this would raise the issue of whether the RCT can still be considered an experimental design. To test this, we run a logit model with the same variables at baseline from table 4.1 (discussed below); the dependent variable takes the value of 1 for those that dropped out of the analysis (either because they could not be reached, they did not complete the survey, or because they relocated), and 0 for those that remained in the analysis. The results of the logit along with Odds Ratios are given in table B.1 in Appendix B with our main variable of interest being if the household was treated or not (Beneficiary = 1). The results show that there are no differences in the probability of having dropped out of the analysis between treated or control households. This confirms a preliminary analysis for attrition using the same data by Seidenfeld (2013) who finds no differential attrition issues. However, qualitative work undertaken the past summer reveals that a good number of households dropped out of the study from an area where cassava is one of the main foods households eat (Ibid.)[42]

Given that the results show no systematic differences in attrition, table 4.1 provides descriptive statistics at baseline for the treated and control households that remained in the

[42] This should be kept in mind when going over the results in section 4.6.

program, which together create a balanced panel with 2,263 households (1,153 in treated communities and 1,110 in controls). Although treatment and control status were assigned at the community level, it is still possible for the two groups to differ in some dimensions. This can be the case when randomization is done at a higher level rather than at the individual level which can pose difficulties in balancing the characteristics across groups (Miguel and Kramer 2004; Gertler, et al. 2011). Even in the case where randomization is done at the individual level, however, while a pure random assignment guarantees that the treatment and control groups will have identical characteristics, on average, in any random allocation the two groups will differ along some dimensions (Bruhn and McKenzie 2009), and it may therefore be necessary to do *ex-post* adjustments to correct for these imbalances.

Column I of table 4.1 provides the mean of the variables at baseline for the entire sample, column II does the same for treated households and column III does this for the controls; column IV provides the test of the difference in the means of the variables (accounting for clustering at the community level), for the two groups with a *t*-tests for continuous variables and $\chi^2$ for binary variables.[43] The results confirm that the randomization was successful: there is balance in all 31 variables at baseline between treated and control households with the exception of the variable for the distance to the water source.

[43] The analyses presented here use the characteristics of the head of household. The same set of analyses was done using the characteristics of the recipient of the transfer, but the results are very similar.

Table 4.1: Description of the Data: Baseline Means

| | Variables (units) | I Pooled | II Treated | III Controls | IV Test † |
|---|---|---|---|---|---|
| **Head of HH Chars.** | Head Age (yrs.) | 33.96 | 34.35 | 33.56 | |
| | Head Female (1,0) | 29% | 27% | 32% | |
| | Head Married/Co-habit (1,0) | 72% | 74% | 70% | |
| | Head Education (yrs.) | 5.26 | 5.48 | 5.03 | |
| **HH Characteristics** | Family Size | 5.70 | 5.76 | 5.63 | |
| | % Share Female | 55% | 55% | 54% | |
| | Dependency Ratio | 1.92 | 1.89 | 1.96 | |
| | # Orphans in HH | 0.34 | 0.33 | 0.35 | |
| | Max Education (yrs.) | 6.33 | 6.54 | 6.11 | |
| | Non-Ag. Business (1,0) | 0.72 | 0.75 | 0.69 | |
| | Remittances (Kwacha) | 0.17 | 0.17 | 0.17 | |
| **Wealth & Assets** | TLU Total^ | 0.37 | 0.45 | 0.28 | |
| | PPI Score | 17.0 | 17.3 | 16.8 | |
| | Land Operated (Has.) | 0.5 | 0.5 | 0.5 | |
| | Land Owned (Has.) | 1.7 | 1.8 | 1.5 | |
| **Access to Svcs.** | Water Src.: River/Lake/Spr./Dam/Rain (1,0) | 21% | 20% | 22% | |
| | Water Src.: Protected Well (1,0) | 7% | 6% | 8% | |
| | Water Src.: Unprotected Well (1,0) | 57% | 57% | 57% | |
| | Water Src.: Boreholel (1,0) | 11% | 12% | 11% | |
| | Water Src.: Tap or Other (1,0) | 4% | 5% | 2% | |
| | Distance to Water Source (Km.) | 0.86 | 1.03 | 0.67 | ** |
| | Garbage Disp.: Collected (1,0) | 1% | 0.6% | 1.2% | |
| | Garbage Disp.: Pit (1,0) | 37% | 36% | 38% | |
| | Garbage Disp.: Dumping (1,0) | 54% | 53% | 55% | |
| | Garbage Disp.: Burning or Other (1,0) | 8% | 10% | 6% | |
| **Shocks** | # Positive Shocks | 0.02 | 0.01 | 0.03 | |
| | # Negative Shocks | 0.26 | 0.24 | 0.28 | |
| **Institutional & Other Supp.** | Receives Gov't Support (1,0) | 13% | 14% | 11% | |
| | Receives NGO Support (1,0) | 2% | 2% | 1% | |
| | Receives Other Support (1,0) | 19% | 21% | 17% | |
| **Qual. Control** | Interviewer Speaks Diff. Lang. | 1% | 1% | 1% | |
| ***N*** | | 2,263 | 1,153 | 1,110 | |

† Tests are for differences in means with respect to treated households accounting for clustering.

^ Tropical Livestock Units provides a convenient method for quantifying a wide range of different livestock. We use the same conversion units used by a similar paper by Daidone, et al. (2013).

* $p<0.10$; ** $p<0.05$; *** $p<0.01$

The results of table 4.1 show that, on average, the head of household is around 34 years old, with a high share being women, and with low levels of education. On average, family size is over 5, women make a big share of household members, and households have a dependency ratio of close to 2, and a big share of households has orphans. In terms of wealth and assets, households have low endowments of animals (TLU Total), land (on average operated land is 0.5 hectares), and have high poverty rates (the PPI score of 17 indicates that 92.2% of households are below the national poverty line (Schreiner 2013). In terms of shocks[44] and support, households are more likely to have experienced a negative shock rather than a positive one, and over 10% of households receive government support or other kind of support (19%). The description of these variables at baseline indicate that CGP beneficiary households are, indeed, very poor with low levels of human capital and endowments.

Moving to the outcome variables for agriculture, table 4.2 provides baseline comparisons between treated and control households for the value of harvest and gross margins normalized by hectares.[45] It also provides information for total expenditures on crop production (and a disaggregation by expenditure type), harvest size, consumption from harvest, harvest sales (size, value, and share of households selling), and operated farmland (and a disaggregation by crop grown), farm tools and farm animals.

[44] Negative shocks include drought, change in food prices, flood, livestock disease, illness, crop disease or pest, change in sale of price of crops, and change in agricultural prices. Positive shocks include better pay or work, or a rise of profit from business.

[45] All variables for amounts of harvest-related variables and input use are normalized by hectares for an appropriate comparison.

Table 4.2: Household Production & Input Variables: Baseline Means

| | Variables (units) | I Pooled | II Treated | III Controls | Test † |
|---|---|---|---|---|---|
| **Value of Harvest** | Crop Production (ZMK/ha.) | 752,851 | 774,474 | 730,390 | |
| **Value of Gross Margins** | Crop Production (ZMK/ha.) | 715,461 | 732,088 | 698,189 | |
| **Crop Expenditures** | Crop Production Total (ZMK/ha.) | 37,390 | 42,386 | 32,201 | |
| | Seeds (ZMK/ha.) | 13,617 | 13,456 | 13,784 | |
| | Pesticides (ZMK/ha.) | 206 | 0 | 419 | |
| | Fertilizers (ZMK/ha.) | 1,934 | 1,849 | 2,022 | |
| | Hired Labor (ZMK/ha.) | 9,232 | 14,298 | 3,971 | * |
| | Other Expenses (ZMK/ha.) | 12,402 | 12,784 | 12,006 | |
| **Crop Expenditures** | Crop Production (1,0) | 23% | 24% | 22% | |
| | Seeds (1,0) | 13% | 13% | 13% | |
| | Pesticides (1,0) | 0.1% | 0.0% | 0.2% | |
| | Fertilizers (1,0) | 1% | 1% | 1% | |
| | Hired Labor (1,0) | 3% | 4% | 2% | * |
| | Other Expenses (1,0) | 11% | 11% | 11% | |
| **Harvest Size** | Maize (Kg./ha.) | 445 | 450 | 439 | |
| | Cassava (Kg./ha.) | 392 | 429 | 354 | |
| | Rice (Kg./ha.) | 62 | 21 | 104 | * |
| **Harvest Consumption** | Maize (Kg./ha.) | 274 | 283 | 264 | |
| | Cassava (Kg./ha.) | 183 | 177 | 189 | |
| | Rice (Kg./ha.) | 76 | 81 | 70 | |
| **Harvest Sales (Size)** | Maize (Kg./ha.) | 46 | 36 | 57 | |
| | Cassava (Kg./ha.) | 22 | 24 | 20 | |
| | Rice (Kg./ha.) | 53 | 41 | 65 | |
| **Harvest Sales (Value)** | Value of all Sales (ZMK/ha.) | 94,387 | 86,317 | 102,770 | |
| | Maize (ZMK/ha.) | 56,508 | 42,471 | 71,089 | * |
| | Cassava (ZMK/ha.) | 15,007 | 15,493 | 14,502 | |
| | Rice (ZMK/ha.) | 65,869 | 65,202 | 66,563 | |
| **Probability of Selling** | Any Sales from Harvest (1,0) | 22% | 20% | 24% | |
| | Maize (1,0) | 11% | 9% | 12% | |
| | Cassava (1,0) | 4% | 3% | 4% | |
| | Rice (1,0) | 8% | 8% | 8% | |
| **Farm Land** | Operated Land Total (Has.) | 0.5 | 0.5 | 0.5 | |
| | For Maize (Has.) | 0.24 | 0.26 | 0.22 | |
| | For Cassava (Has.) | 0.10 | 0.09 | 0.11 | |
| | For Rice (Has.) | 0.09 | 0.10 | 0.09 | |
| | For Millet (Has.) | 0.02 | 0.03 | 0.02 | |

*Continued on Next Page…*

Table 4.2 - *Continued from previous page.*

| | | | | | |
|---|---|---|---|---|---|
| | For Groundnut (Has.) | 0.013 | 0.012 | 0.013 | |
| | For Sweet Potatoes (Has.) | 0.009 | 0.007 | 0.011 | |
| | For Sorghum (Has.) | 0.011 | 0.008 | 0.015 | |
| | For Oth. Beans (Has.) | 0.004 | 0.003 | 0.004 | |
| | For Oth. Beans (Has.) | 0.002 | 0.003 | 0.002 | |
| **Farm Tools** | Axe (1,0) | 77% | 80% | 75% | |
| | Hoe (1,0) | 92% | 92% | 92% | |
| | Pick (1,0) | 3% | 3% | 3% | |
| | Shovel (1,0) | 6% | 7% | 5% | |
| | Hammer (1,0) | 5% | 5% | 5% | |
| | Plough (1,0) | 7% | 7% | 6% | |
| **Farm Animals** | Cows (#) | 0.21 | 0.29 | 0.13 | |
| | Cattle (#) | 0.45 | 0.52 | 0.38 | |
| | Goats (#) | 0.05 | 0.08 | 0.02 | ** |
| | Chickens (#) | 2.0 | 2.1 | 1.9 | |
| | Ducks (#) | 0.12 | 0.10 | 0.15 | |
| | TLU Total | 0.37 | 0.45 | 0.28 | |
| | ***N*** | 2,263 | 1,153 | 1,110 | |

† Tests are for differences in means with respect to treated households accounting for clustering.
* $p<0.10$; ** $p<0.05$; *** $p<0.01$

This comparison of 53 baseline variables shows that, on average, both treated and control households have fairly identical values with the exception of five variables: expenditures on hired labor and share of households hiring labor (both greater for treated households), amount of rice harvest (Kg./ha.) (greater for control households), value of maize sold (ZMK/ha.) (greater for control households), and the number of goats owned (greater for treated households). It should be noted, though, that 4 out of these 5 variables are significant only at the 10% level, while the magnitude of these difference is fairly small.

The results of table 4.2 show that these households have very low values of harvest and gross margins, they have low levels of expenditures on crops, and very low yields per hectare and only about a fifth of households sell part of their produce. Moreover, the total size of

operated land is very small (0.5 ha.), with around half of it devoted to maize (0.24 ha.), and the other half devoted to around 7 other crops.

In order to put these values in perspective, table 4.3 provides a comparison (of selected variables from table 4.2) between households in the CGP sample with results presented by Burke, Hichaambwa, Banda, et al. (2011) who use information form a nationally representative sample of smallholders in Zambia. The study is the first attempt to provide information on maize production costs for smallholders, defined as households with an area under crops of 4.99 hectares or less using the 2010 Crop Forecast Survey (CFS) conducted by the Central Statistics Office and Ministry of Agriculture and Cooperatives.[46] It should be noted that maize is grown by 80% of smallholders in Zambia (Feed the Future (FTF) 2011), and for the CGP sample, nearly 50% of the land is devoted to this crop.

Table 4.3 Comparison of Agricultural Variables: National Representative Sample of Smallholders in Zambia versus the CGP Sample for 2009-2010

| | **National Smallholders** | **CGP Sample** | **% of National Smallholders** |
|---|---|---|---|
| **Variable (units)** | **Maize** | **All Crops** | |
| Value of Harvest (ZMK/ha.) | 2,957,341 | 752,851 | 25% |
| Gross Margins (ZMK/ha.)^ | 2,315,526 | 715,461 | 31% |
| Crop Expenditures (ZMK/ha.)^ | 641,815 | 37,390 | 6% |
| | **Maize** | **Maize** | |
| Yield (Kg.ha) | 2,275 | 445 | 20% |

^ Excludes costs of land and family labor

*Source:* For National Smallholders, we use Burke et al. (2011), Table 5, who use the 2010 Crop Forecast Survey (CFS), conducted by the Central Statistics Office and Ministry of Agriculture and Cooperatives

[46] A document from the Ministry of Agriculture & Cooperatives and The Central Statistical Office (2011) define Small Scale Households as a household cultivating 4.99 hectares of crops or less. Of this, there is Category A: Area under crops of 2 hectares or less, and Category B: Area under crops 2.0- 4.99 ha. For Medium Scale Households, there is Category C: Area under crops 5.0 – 19.99 ha. And all households cultivating 20 hectares or more are classified as Large Scale farmers.

Table 4.3 shows that the CGP households obtain a value of harvest that is only about a quarter of what the nationally representative sample of smallholder maize growers obtained per hectare in 2009-2010. Likewise, in contrast to this nationally representative sample of smallholders, CGP households obtain about a third of gross margins and spend 6% of what the average smallholder spends. More specifically on maize yields, table 4.3 shows that the CGP sample obtains only about a fifth of the yields than those reported in Burke et al. (2011). We should highlight the fact that the average land operated by households in the CGP sample is about half a hectare, while the data presented by Burke et al. (2011) includes households that have an area under crop of 4.99 hectares or less.

Moreover, the results of table 4.2 show that CGP households have a high level of diversification: on average, operated land is devoted to around 9 different crops. This high level of diversification, however, also means a high level of fragmentation of already very small operated land. While this level of diversification may help minimize risks, this might also have implications on productivity. Table 4.2 also shows that households in the CGP sample have low participation rates in output markets (sales), and low endowments of farm tools and farm animals. For instance, only around 22% of households sell any of their produce; and while a big share of households own an ax (70%) or a hoe (92%), only 7% of households own a plough. Likewise, only a small percentage of households own large farm animals, like cows (6%), and cattle (10%), while very few households own medium-size farm animals like goats (2%), or small farm animals like ducks (3%); although around 41% of households own chickens.

Moving to the variables used in the SPFA, table 4.4 shows the variables used in the model. It should be noted that no information on family labor for crop production was collected at baseline, thus, it is not possible to use a SPFA using panel data. We restrict the SPFA using

data only at follow-up. Moreover, we restrict this part of the analysis to households that cultivated any crop and harvested.[47] The second panel of table 4.4 shows the variables used in the SPFA at follow-up, while the first panel shows the variables at baseline to get a sense on whether there were significant differences at baseline between treated and control households. This provides evidence on the assumption that both sets of households started off with the same levels for these variables and that any changes in the outcome variables after the program got underway is due to the CGP.

Looking at the first panel for baseline, we see that while there are no differences in the value of harvest at baseline, there are some differences in expenditures on crop production and the share of households who use hired labor, both higher among treated households; however, these differences are small, especially for hired labor. The results of the first panel provide evidence that treated and control households started off fairly equal at baseline.

However, the lower panel shows treated households have a greater value of harvest, spend significantly more in inputs, have a greater share of operated land, a greater share of households hire labor, while a lower share of households do not have any expenditures on crop production.

Tables 4.1, 4.2, and 4.4 show that the randomization into the program at the community level yielded a very good counterfactual for the treated households. However, the results also show that the CGP households are quite poor, with low levels of human capital, low land endowments with high fragmentation of land, and low levels of productivity, gross margins, and

[47] We run a logit regression to be sure there are no systematic differences between beneficiary and control households in terms of their probability of not having cultivated land nor harvested. The results are presented in Table B.3 in Appendix B. We find no systematic differences in this regard.

value of harvest, as corroborated by table 4.3. Let us turn to the econometric methods used in this paper.

Table.4.4: Variables Used in the Stochastic Production Frontier Analysis: Means at Baseline and Follow-up

| Variables (units) at Baseline | I Pooled | II Treated | III Controls | IV Test † |
|---|---|---|---|---|
| Value of Harvest (ZMK) | 416,146 | 429,990 | 401,110 | |
| Crop Expenditures (ZMK total) | 23,283 | 30,990 | 14,912 | ** |
| Land Operated (Has.) | 0.51 | 0.52 | 0.50 | |
| Family Labor (#days)^ | *Not Available at Baseline* | | | |
| Hired Labor (0,1) | 0.03 | 0.04 | 0.02 | ** |
| # Positive Shocks | 0.02 | 0.01 | 0.03 | |
| # Negative Shocks | 0.26 | 0.23 | 0.29 | |
| ***N*** | 1,936 | 1,008 | 928 | |
| **Variables (units) at Follow-up** | | | | |
| Value of Harvest (ZMK) | 691,613 | 762,897 | 614,183 | * |
| Crop Expenditures (ZMK total) | 53,180 | 74,571 | 29,945 | *** |
| Land Operated (Has.) | 0.84 | 0.95 | 0.73 | ** |
| Family Labor (#days) | 146 | 153 | 138 | |
| Hired Labor (0,1) | 0.09 | 0.12 | 0.05 | *** |
| No Family Labor (0,1)^^ | 0.01 | 0.01 | 0.01 | |
| No Crop Expenditures (0,1)^^ | 0.52 | 0.44 | 0.62 | *** |
| # Positive Shocks | 0.17 | 0.18 | 0.16 | |
| # Negative Shocks | 1.91 | 1.85 | 1.97 | |
| ***N*** | 1,936 | 1,008 | 928 | |

† Tests are for differences in means with respect to treated households accounting for clustering.

^ Data only available at Follow-up

^^ Battese (1997) correction for Log of 0

* $p<.10$; ** $p<.05$; *** $p<.01$

## 4.5 Econometric Approach

This paper aims to assess the impact of the program on: (1) agricultural production (value and size of harvest, and gross margins) and productive investments; and (2) technical efficiency. In order to evaluate this program along these dimensions, we need to know what would have happened to beneficiary households had they not received the program. This is the classical

problem of missing data within the evaluation literature where it is necessary to identify a proper counterfactual group that is similar in all respects to the treated group with the exception of not having received the program. The most direct way of ensuring a comparable group for an impact evaluation is with the use of an experimental design—a randomized control trial (RTC)—where households are randomly allocated between control and treated groups (Blundell and Costa Dias 2010; Gertler, et al. 2011). In principle, the RTC guarantees that the treatment status is uncorrelated with other observable and non-observable characteristics. As such, the potential outcomes are thought to be statistically independent of the treatment status. In this case, a simple mean comparison between treated and control groups can identify treatment impacts if the RCT was successful as follows:

$$E[\tau] = ATE = ATT = E[Y(1)] - E[Y(0)] \qquad (1)$$

Where $\tau$ is the change in the outcome measure, $Y(1)$ denotes the outcome of interest for the treated and $Y(0)$ is the outcome of interest for the controls. Under the experimental design, the average treatment effect (ATE) equals the average treatment on the treated (ATT).

However, even under an RTC, if the assignment to treatment or control is done at the village or community level, it is possible that while the two groups of households may be very similar there may still be some differences between the groups (Miguel and Kramer 2004; Gertler, et al. 2011), as is the case in this paper. Thus, in the estimation of impacts, it is important to control for these differences. This is done using a difference-in-differences (DD) estimator to evaluate the impact of the program for the first objective of this paper. The impact of the

program on the second objective will be assessed using a SPFA framework using only the second round of data, as explained earlier.[48] These two models are explained below.

*DD Estimator*

The impact of a program can be identified and in some cases improved by applying a DD estimator in the event that the randomization produces baseline differences between treated and control groups. One advantage of the DD estimator is that it helps control for unobserved heterogeneity, which may lead to selection bias. In this sense, the DD estimator helps control for baseline differences between the two groups, as well as time-invariant unobservable factors that cannot be accounted for otherwise (Asfaw et al. 2012). A key assumption is that differences between treated and control households remain constant overtime. Further, the DD estimator with conditioning variables helps minimize the standard errors as long as the effects are unrelated to the treatment and are constant overtime (Wooldridge 2002). The DD estimator is given in equation (2).

$$Y_{itv} = \beta_0 + +\beta_1 T_t + \beta_2 P_{it} + \beta_3 (T_t * P_{it}) + \sum \beta_i Z_{iv} + \mu_{itv} \qquad (2)$$

Where $T_t$ is a binary variable for time equal to 0 for baseline and equal to 1 at follow-up; $P_{it}$ is a binary variable equal to 1 if household $i$ received the program; $T_t * P_{it}$ is the interaction between time and program receipt; and $\mu_{itv}$ is the error term. In order to control for characteristics that may influence the outcome of interest beyond the effect of the program, $Z_{iv}$ is included in the regression, which is a vector of household and community characteristics at baseline, which could also affect our outcome of interest, $Y_{itv}$. However, we allow exogenous variables to change over time in order to also account for these changes from baseline. Under the DD model

[48] Daidone (2014) suggests using the size of the labor force in the household as a proxy for family labor so that a SPFA using panel data can be done. Perhaps this is something that might be done in a future version of this paper.

in (2) $\beta_1$ captures changes over time, $\beta_2$ captures the differences between treated and control households, and $\beta_3$ captures the impact of the program.

In general terms, we run two specifications: 1) the first uses the variables presented in table 4.1 to evaluate the impact of the program on input use and expenditures on inputs, use of farm tools, and ownership of farm animals; and 2) to estimate the impact of the program on all outcome variables related to harvest, we use the same control variables and we add all input variables. All the variables for harvest and input use are normalized to be on a per hectare basis.[49] For all models, we cluster standard errors at the village level. Moreover, we use information from the community survey to assess the change in prices from baseline to follow-up to take inflation into account. The final monetary values for baseline and follow up, then, reflect real prices using 2012 values. For outcome variables that are binary, we use the logit and report the average marginal effects following Bartus (2005).

*SPFA*

This paper uses a SPFA framework to estimate the impact of the CGP on technical efficiency. The stochastic frontier model has been used in a large literature of studies of production, cost, revenue, and profit. The model was originally developed by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977). The canonical formulation that serves as the foundation for other variations is the Aigner et al. (1977) model as follows:

$$y = \boldsymbol{\beta}'\boldsymbol{x} + v - u, \qquad (3)$$

Where *y* is the observed outcome, $\boldsymbol{\beta}'\boldsymbol{x} + v$ is the optimal frontier goal, or the maximum production output, pursued by the farmer, $\boldsymbol{\beta}'\boldsymbol{x}$ is the deterministic part of the frontier and $v \sim N[0, \sigma_V^2]$ is the stochastic, two-sided and symmetrical part (can be positive, zero, or

[49] Except for the SPFA model, which uses land as a right-hand-variable.

negative). These two parts constitute the stochastic frontier. On the other hand, the amount by which the observed farmer fails to reach the optimum, or the frontier, is $u$, where $u = |U|$ and $U \sim N[0, \sigma_u^2]$. In this formulation, $u$ is one sided inefficiency term (which can be zero or positive) that is used to measure TE as follows

$$TE = \frac{y_i}{f(x_i; \beta)} \approx \exp(-u),$$

Thus, $y_i$ achieves its maximum feasible output iff $TE_i = 1$; otherwise, $TE_i < 1$ provides a measure of the shortfall of observed output from maximum feasible output (Kumbhakar and Lovell 2000).

We assume that the random part of the error structure ($v$) is heteroskedastic, with the variance function depending on a linear combination of variables that should have an effect on its behavior, namely, positive and negative shocks, such as having experienced a flood, drought, or pest disease. Similarly, we assume that the inefficiency part of the error term ($u$) is also heteroskedastic with the variance function depending on a linear combination of variables that theoretically affect efficiency: age and education of the head of household, access to credit, the PPI score that captures poverty likelihoods, use of hired labor, as well as the treatment variable to test if the CGP had an impact on TE.[50]

Under this framework, it is possible to disentangle "jumps" in the production function referred to as technological change (TC) stemming from the application of improved or new farming practices, as well as a relative measure of managerial ability for a given technology, i.e., technical efficiency (TE). These two effects are disentangled using the SPFA framework to determine if the CGP had an impact on the production function of beneficiary households, as well as technical efficiency. Unlike previous studies that need to be concerned with issues of

[50] We thank Daidone (2014) for suggesting adding the treatment variable in the ancillary equation.

self-selection using the SPFA (Bravo-Ureta, et al. 2012), we do not face this issue since the program is an RCT.

The key technology effect we are interested in identifying relates to the impact of participating in the CGP on the value of harvest. The main null hypothesis will be: Mean $TE_T$ = Mean $TE_C$ where the subscript *T* is for treated households and *C* is for control households. This will be captured in two ways: 1) including the treatment variable in the ancillary equation for inefficiency tests whether one group is more efficient than the other; and 2) following Fried et al. (2008) we infer TE by ranking the TE scores derived from a "best practice" production frontier. Failure to reject this hypothesis will indicate that the program had no effect on technical efficiency. To accomplish this, we use a conventional SPFA model using the combined sample of treatment and control farmers (pooled data) with the addition of the treatment variable to test whether treated farmers display a different set of technology than the control group. If no technological difference is found, then a single frontier combining all farmers (treated and control) is the more desirable option (Bravo-Ureta et al. 2012). We also evaluate alternative functional forms for the SPFA. We compare the Cobb-Douglas (CD) against the Translog (TL), which are the two most commonly used in efficiency studies (Bravo-Ureta et al., 2007). The results of maximum likelihood ratio tests and a test of joint significance on the second order and multiplicative part of the TL model (Greene, 2005, p. 103) will inform our final choice of the model. Given that a good number of households do not spend on crop expenditures, we correct for these zero logged values following Battese (1997). Within this overall framework, the predictor of TE can be obtained as the expectation of $u_i$ conditional on the composed error term following Jondrow et al. (1982). A full description of the estimation of the TE scores can be found in Greene (2010).

## 4.6 Results

This section provides the results of the paper. It begins with the results of the DD model for the impact of the CGP on agricultural production and productive investments, followed by the results of the SPFA model for the impact of the CGP on technical efficiency.

*Results: Difference-in-Difference*

Table 4.5 provides the results of the DD model: the first panel provides the results for the value of harvest, gross margins, crop expenditures and the share of households that spend on crop expenditures; the second panel provides the results for yields (size of harvest in Kg./ha.), amount of harvest consumed, amount of harvest sold, value of sales, and the share of households selling part of their produce; the third panel provides the results for land devoted to crop production, along with a disaggregation of land by crop produced; the fourth panel provides results for investments in farm tools; and the last panel provides results for investments in farm animals. In order to provide a clearer and more complete picture of the impact of the CGP on these indicators, column I provides the changes on these indicators over time (Time = $\beta_1$in [eq. 2]), column II provides the difference between treated and control households (Program = $\beta_2$ in [eq. 2]), and column III captures the impact of the program (DD = $\beta_3$ in [eq. 2]).

The first panel of table 4.5 shows that the CGP had no impact on the value of harvest, gross margins or total expenditures on crop production. However, column I shows that the trend overtime was for households to spend less on total crop production, and in all disaggregated types of expenditures of crop production (seeds, pesticides, fertilizers, hired labor, and other). Yet, column III shows that the program reversed this trend for treated households: all coefficients have a positive sign with the exception of hired labor. However, only the coefficient for expenditures on seeds is positive and significant.

Table 4.5: DD: Impact of CGP on Agricultural Production, Consumption, Sales, Input Use, Farm land, Tools, and Animals

| | | I | | II | | III | |
|---|---|---|---|---|---|---|---|
| | **Variables (units)** | **Time** | **Sig.** | **Program** | **Sig.** | **DD** | **Sig.** |
| **Value of Harvest** | Crop Production (ZMK/ha.) | -797 | | 56,105 | | -74,560 | |
| **Value of Gross Margins** | Crop Production (ZMK/ha.) | -8,658 | | 42,741 | | -56,024 | |
| **Crop Expenditures** | Crop Production Total (ZMK/ha.) | -26,149 | | 7,418 | | 27,631 | |
| | Seeds (ZMK/ha.) | -7,288 | | -3,189 | | 15,147 | ** |
| | Pesticides (ZMK/ha.) | -1,303 | | -460 | | 777 | |
| | Fertilizers (ZMK/ha.) | -453 | | -424 | | 4,521 | |
| | Hired Labor (ZMK/ha.) | -9,295 | | 10,838 | * | -2,901 | |
| | Other Expenses (ZMK/ha.) | -7,810 | | 653 | | 10,087 | |
| **Crop Expenditures** | Crop Production (1,0) | 0.04 | | 0.01 | | 0.14 | ** |
| | Seeds (1,0) | -0.01 | | -0.01 | | 0.05 | |
| | Pesticides (1,0) | *Model did not converge* | | | | | |
| | Fertilizers (1,0) | 0.013 | | -0.002 | | 0.010 | |
| | Hired Labor (1,0) | 0.03 | * | 0.03 | * | 0.003 | |
| | Other Expenses (1,0) | 0.06 | * | -0.003 | | 0.15 | ** |
| **Harvest Size** | Maize (Kg./ha.) | -56.29 | | 1.319 | | 30.86 | |
| | Cassava (Kg./ha.) | 36.88 | | 67.52 | | -233.6 | * |
| | Rice (Kg./ha.) | -39.91 | | -73.45 | * | 27.77 | |
| **Harvest Consumption** | Maize (Kg./ha.) | 19.17 | | 10.91 | | -18.37 | |
| | Cassava (Kg./ha.) | 8.99 | | -12.46 | | -42.02 | |
| | Rice (Kg./ha.) | -41.44 | ** | 14.89 | | 27.5 | |
| **Harvest Sales (Size)** | Maize (Kg./ha.) | -3.643 | | -19.8 | * | 66.35 | ** |
| | Cassava (Kg./ha.) | -7.388 | | 2.611 | | -5.551 | |
| | Rice (Kg./ha.) | -22.46 | | -23.92 | | 51.4 | ** |
| **Harvest Sales (Value)** | Value of all Sales (ZMK/ha.) | 4,764 | | -17,038 | | 91,115 | ** |
| | Maize (ZMK/ha.) | -19,954 | | -30,461 | * | 87,697 | ** |
| | Cassava (ZMK/ha.) | 4,384 | | -593 | | -2,123 | |
| | Rice (ZMK/ha.) | -6,391 | | -2,090 | | 52,390 | |
| **Probability of Selling** | Any Sales from Harvest (1,0) | -0.002 | | -0.05 | | 0.13 | *** |
| | Maize (1,0) | 0.01 | | -0.03 | | 0.08 | * |
| | Cassava (1,0) | 0.004 | | -0.011 | | 0.016 | |
| | Rice (1,0) | -0.02 | | -0.004 | | 0.07 | ** |
| **Farm Land** | Operated Land Total (Has.) | 0.05 | | 0.01 | | 0.23 | ** |
| | For Maize (Has.) | 0.14 | ** | 0.03 | | 0.14 | * |
| | For Cassava (Has.) | 0.01 | | -0.02 | | 0.001 | |
| | For Rice (Has.) | -0.07 | ** | 0.01 | | 0.06 | ** |

*Continued on Next Page…*

Table 4.4 - *Continued from previous page.*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | For Millet (Has.) | -0.01 | | 0.01 | | -0.001 | |
| | For Groundnut (Has.) | -0.001 | | -0.002 | | 0.004 | |
| | For Sweet Potatoes (Has.) | -0.01 | ** | 0.00 | | 0.01 | |
| | For Sorghum (Has.) | -0.01 | | -0.01 | | 0.02 | ** |
| | For Oth. Beans (Has.) | 0.000 | | -0.001 | | 0.01 | * |
| | For Oth. Crops (Has.) | | | | | | |
| **Farm Tools** | Axe (1,0) | 0.00 | | 0.04 | * | -0.01 | |
| | Hoe (1,0) | 0.03 | | 0.00 | | 0.02 | |
| | Pick (1,0) | 0.01 | | -0.01 | | 0.01 | |
| | Shovel (1,0) | -0.02 | | 0.01 | | 0.02 | |
| | Hammer (1,0) | -0.01 | | -0.01 | | 0.04 | * |
| | Plough (1,0) | -0.05 | *** | -0.004 | | 0.05 | |
| **Farm Animals** | Cows (#) | -0.03 | | 0.09 | | -0.24 | |
| | Cattle (#) | -0.10 | | -0.01 | | 0.44 | ** |
| | Goats (#) | 0.04 | ** | 0.06 | ** | 0.13 | ** |
| | Chickens (#) | -0.47 | * | 0.005 | | 1.11 | ** |
| | Ducks (#) | -0.12 | | -0.04 | | 0.20 | ** |
| | TLU Total | -0.17 | *** | 0.11 | | -0.09 | |
| | ***N*** | 4,525 | | 4,525 | | 4,525 | |

* $p<0.10$; ** $p<0.05$; *** $p<0.01$

This means that although the CGP led to increases on crop expenditures overall, the program only had a clear impact on expenditures on seeds. However, when we look at the impact of the CGP on the share of households that spend on crop production, we see that the program increased this share by 14% (baseline mean for treated was 22%). Similarly, the program increased the share of households that spend on other miscellaneous crop expenditures by 15% (baseline mean for treated was 11%).

For the second panel of table 4.5, column I shows a negative trend in the size of harvest for maize and rice. However, column III shows that the program also reversed this negative trend: the coefficient on harvest for maize and rice is positive, although not significant. On the other hand, while the size of harvest for cassava increased over time (column I), the program seems to have had a negative impact on this indicator. However, and as noted earlier, qualitative

work undertaken the past summer reveals that a good number of households dropped out of the study from an area where cassava is one of the main foods households eat. This result, then, is a problem attributed to the study and not to the CGP (Seidenfeld 2013). For consumption from harvest, the trend overtime was for households to increase consumption for maize and cassava (not statistically significant), while the trend shows that households significantly decreased their rice consumption (statistically significant). Yet, it is interesting to see that the sign of the coefficients for the impact of the program (column III) show the opposite sign for each of these indicators, which suggests that the program reversed these trends: when households were generally consuming more maize and cassava from own harvest over time, treated households were eating less of these crops; on the other hand, when households were generally eating less rice over time, treated households actually increased their consumption of rice from own harvest.

For sales, column I shows that households over time sold less maize, cassava and rice from harvest; however, the program reversed this trend and it actually increased the amount of harvest sold for maize and rice, as well as on the value of all sales, primarily driven by the value of maize sales. Similarly, when we look at the share of households that sold part of their produce, column I shows that a lower share of households sold their produce, especially rice; yet, once again, column III shows that the program reversed this trend and had a positive and significant impact in this regard: the CGP increased the share of households that sell part of their produce by 13% (baseline mean of 24% for treated households); it increased the share of households that sell maize by 8% (baseline mean of 12%); and it increased the share of households that sell rice by 7% (baseline mean of 8%). Taking all this together, one way to interpret these results is as follows: while the trend overtime was for households to spend less on crop production, for yields and sales of maize and rice to decrease, the program led to increases in expenditures on all inputs

(although only significant for seeds), except for hired labor, which helped reverse the decreases in yields over time for maize and rice (although not statistically significant), and it also facilitated an increase in the sales of maize and rice (positive and significant for both). This slight increase in sales is corroborated by the greater share of treated households that sold maize and rice (both positive and significant).

Now that panels 1 and 2 have shown that the program led to a slight increase in the use of inputs and sales of maize and rice (amount and value), let us see if the program had any impact on the land devoted to crop production. Recall that all the variables related to harvest are normalized by hectare; thus, it will be informative to see if farm land was expanded. Panel 3 shows these results. Column I shows that for 5 out of 8 crops cultivated, the trend over time shows that households generally reduced the amount of land devoted to rice (statistically significant), millet, groundnut, and sweet potatoes (statistically significant). Yet, column III shows that the program reversed this negative trend for rice (statistically significant), groundnut, sweet potatoes and sorghum (statistically significant). Moreover, while the trend over time was for total operated land to increase (not statistically significant), the CGP led to an increase of 0.23 ha. in operated land (baseline mean of 0.5 ha. for treated households), with the biggest increase for land devoted to maize (0.14 ha. compared to a baseline mean of 0.22 ha. for treated households, and beyond the 0.14 ha. increase over time shown in column I), followed by land devoted to rice (0.06 ha. compared to a baseline mean of 0.09 ha.), sorghum (0.02 ha. compared to a baseline mean of 0.015), and beans (0.01 ha. compared to a baseline mean of 0.004 ha). Contrasting these impacts to the baseline values, these increases represent an increase of 46% for total operated land, an increase of 63% of land devoted to maize, an increase of 66% of land devoted to rice, an increase of 133% of land devoted to sorghum, and an increase of 250% of

land devoted to beans. Yet, despite these significant percentage increases in land used, we should also note that the absolute size of these lands is small.

For investments in tools, the fourth panel of table 4.5 shows that although the sign of the coefficients have the expected positive sign for all the tools (except for axes), the program only had a clear impact on the share of households that own a hammer (4% compared to baseline mean of 5%).

On the other hand, looking at the last panel of table 4.5 for investments in farm animals, we see that while the trend over time was for households to decrease the number of animals they own (except for goats), the CGP reversed this trend and it actually had a positive impact on the number of animals owned for cattle (0.44 with a baseline of 0.38), goats (0.13 with a baseline of 0.02), chickens (1.11 with a baseline of 1.9) and ducks (0.2 with a baseline of 0.15).

Overall, these results show that the program had a positive impact on crop expenditures, primarily driven by expenditures in seeds (Kg./ha, and share of households that spend on seeds), and other miscellaneous expenditures (share of households), and it also had an impact on the sale of maize and rice (for Kg./ha., ZMK/ha., and the share that sell). This shift in increases in agricultural activities is also captured by the change in the size of operated land, which increased by around a quarter of a hectare, and was primarily driven by increases in land devoted to maize, rice, sorghum, and beans. However, this increase in agricultural activities did not translate into greater yields per hectare, higher value of harvest, greater gross margins, or own consumption from own crop production. In terms of investments in farm tools and animals, the CGP had a small impact in the share of households that own a hammer, while the program had a clearer impact on the number of animals owned by the household, with the greatest increase for small animals, like chickens.

Given these results, it will be beneficial to investigate if the program had a different impact across households that are labor-constrained versus households that have more members that can potentially work in the farm. Theoretically, one would expect that households with a greater number of working-age members to benefit more from the cash transfers if part of these transfers is invested in the farm which might induce their surplus labor to be used in the farm. This might be more plausible in there are high levels of unemployment or underemployment. To do this, we identify three types of households: (1) with one or no household members of working age (17 or older, but younger than 65) (20% of the sample); (2) with two household members of working age (80% of the sample); and (3) with three or more household members of working age (20% of the sample). Figure 4.1 provides the distribution of household members of working age for treated and control households. We do a test of means at baseline for these three groups (accounting for clustering) and find no differences in the composition of these groups between the treated and control households.[51] We should note that, not surprisingly, 90% of households in group 1 are single-headed households (never married, widowed, separated or divorced), 90% of households in group 2 are married or co-habiting, and 79% of households in group 3 are married and have other household members of working age. For the sake of organization and clarity, we will refer to these groups as follows: group 1 will be called single-headed households (SH-HH); group 2 will be called nuclear households (N-HH); and group 3 will be called surplus-labor households (SL-HH).

[51] This also corroborates the success of the randomization process in obtaining a very good counterfactual for the treated households even though this was done at the community level.

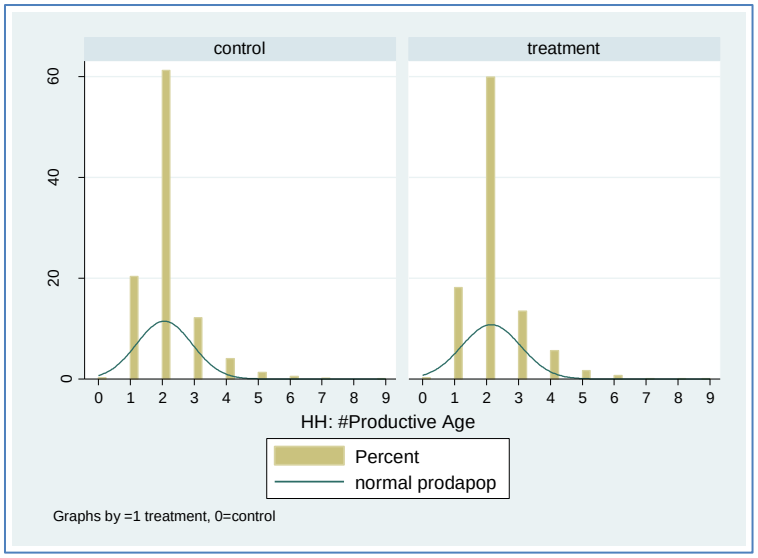



Figure 4.1: Number of Household Members of Working Age by Treatment Status

The results of the heterogeneous impacts by household type based on the size of the household labor-force using the DD model are presented in table 4.6. Column I provides the results of the impact of the program on single-headed households (S-HH), column II does the same for nuclear households (N-HH), and column III does this for surplus-labor households (SL-HH). For the sake of brevity, the table only provides the impact of the program ($\beta_3$, in [eq.] 2); we do not include the trends of over time, or the difference between the groups ($\beta_1$, and $\beta_2$, respectively, in [eq. 2]) as done in table 4.5.

The results of the first panel confirm that the program had no impact on the value of harvest or gross margins; however, recall that the sign of the coefficient for the impact of the program for both indicators was negative for the entire sample in table 4.5, while the sign of the coefficients are positive for single-headed (column I) and labor-surplus households (column III). For expenditures on crop production, the program did have an impact on this indicator, but only for nuclear households (column II); although the coefficient for surplus-labor households (column III) is also positive and it is almost identical in magnitude as the coefficient for nuclear households.

Table 4.6: DD: Heterogeneous Impacts by Household Labor Size: Single-headed Households (SH-HH), Nuclear Households (N-HH), and Labor-Surplus Households (LS-HH).

| | | I | | II | | III | |
|---|---|---|---|---|---|---|---|
| | | SH-HH | | N-HH | | SL-HH | |
| | Variables (units) | DD | Sig. | DD | Sig. | DD | Sig. |
| **Value of Harvest** | Crop Production (ZMK/ha.) | 28,862 | | -200,732 | | 191,054 | |
| **Value of Gross Margins** | Crop Production (ZMK/ha.) | 92,046 | | -168,373 | | 215,303 | |
| **Crop Expenditures** | Crop Production Total (ZMK/ha.) | -56,597 | | 47,046 | ** | 44,135 | |
| | Seeds (ZMK/ha.) | 13,214 | | 13,802 | ** | 22,314 | * |
| | Pesticides (ZMK/ha.) | *NC* | | 1,251 | | -86 | |
| | Fertilizers (ZMK/ha.) | -7,343 | | 6,467 | | 7,420 | |
| | Hired Labor (ZMK/ha.) | -55,402 | | 9,595 | * | 6,409 | |
| | Other Expenses (ZMK/ha.) | -7,066 | | 15,932 | ** | 8,077 | |
| **Crop Expenditures** | Crop Production (1,0) | 0.17 | | 0.15 | ** | 0.12 | |
| | Seeds (1,0) | 0.02 | | 0.05 | | 0.05 | |
| | Pesticides (1,0) | | | *NC* | | | |
| | Fertilizers (1,0) | *NC* | | 0.00 | | 0.01 | |
| | Hired Labor (1,0) | 0.00 | | 0.002 | | 0.031 | |
| | Other Expenses (1,0) | 0.21 | ** | 0.15 | ** | 0.12 | |
| **Harvest Size** | Maize (Kg./ha.) | 103.4 | | -90.4 | | 255 | * |
| | Cassava (Kg./ha.) | -139.2 | | -199.1 | | -343 | ** |
| | Rice (Kg./ha.) | 15.92 | | 37.25 | | 34 | |
| **Harvest Consumption** | Maize (Kg./ha.) | 73.65 | | -84.3 | | 58 | |
| | Cassava (Kg./ha.) | -44.05 | | -54.67 | | 19 | |
| | Rice (Kg./ha.) | -15.81 | | 34.24 | | 53 | |
| **Harvest Sales (Size)** | Maize (Kg./ha.) | 53.3 | * | 49.92 | ** | 117 | ** |
| | Cassava (Kg./ha.) | 24.14 | | -2.528 | | -19 | |
| | Rice (Kg./ha.) | 103.6 | | 55.6 | ** | -8 | |
| **Harvest Sales (Value)** | Value of all Sales (ZMK/ha.) | 116,583 | ** | 74,606 | ** | 115,152 | * |
| | Maize (ZMK/ha.) | 71,623 | ** | 82,006 | ** | 118,658 | * |
| | Cassava (ZMK/ha.) | 31,574 | | -3,199 | | -9,432 | |
| | Rice (ZMK/ha.) | 25,149 | | 68,404 | * | 18,874 | |
| **Probability of Selling** | Any Sales from Harvest (1,0) | 0.23 | ** | 0.13 | ** | 0.09 | |
| | Maize (1,0) | 0.09 | | 0.07 | | 0.08 | |
| | Cassava (1,0) | *NC* | | 0.007 | | *NC* | |
| | Rice (1,0) | *NC* | | 0.03 | ** | 0.02 | |
| **Farm Land** | Operated Land Total (Has.) | 0.12 | | 0.29 | ** | 0.16 | |
| | For Maize (Has.) | 0.08 | | 0.17 | ** | 0.06 | |
| | For Cassava (Has.) | -0.01 | | 0.013 | | 0.000 | |
| | For Rice (Has.) | -0.01 | | 0.08 | ** | 0.09 | * |

*Continued on Next Page…*

Table 4.6 - *Continued from previous page.*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | For Millet (Has.) | 0.02 | | -0.009 | | 0.012 | |
| | For Groundnut (Has.) | 0.005 | | 0.013 | ** | -0.022 | |
| | For Sweet Potatoes (Has.) | 0.01 | * | 0.004 | | 0.006 | |
| | For Sorghum (Has.) | 0.02 | ** | 0.01 | ** | 0.03 | |
| | For Oth. Beans (Has.) | 0.007 | ** | 0.01 | ** | 0.00 | |
| **Farm Tools** | Axe (1,0) | -0.04 | | 0.03 | | -0.2 | |
| | Hoe (1,0) | 0.01 | ** | 0.010 | | 0.0 | |
| | Pick (1,0) | 0.002 | | 0.00 | | 0.02 | |
| | Shovel (1,0) | 0.00 | | 0.01 | | 0.03 | |
| | Hammer (1,0) | *NC* | | 0.04 | * | 0.02 | |
| | Plough (1,0) | *NC* | | 0.04 | | 0.04 | |
| **Farm Animals** | Cows (#) | 0.02 | | 0.02 | | -0.81 | |
| | Cattle (#) | 0.38 | * | 0.22 | | 1.09 | ** |
| | Goats (#) | 0.13 | | 0.08 | ** | 0.26 | ** |
| | Chickens (#) | 1.1 | ** | 1.4 | ** | 0.5 | |
| | Ducks (#) | 0.04 | | 0.16 | * | 0.39 | ** |
| | TLU Total | 0.21 | | -0.05 | | -0.18 | |
| | ***N*** | 884 | | 2,739 | | 902 | |

*NC* = Model did not converge

* $p<0.10$; ** $p<0.05$; *** $p<0.01$

When we disaggregate crop expenditures, the results show that the program had the greatest impact for nuclear households (column II): the program had in impact on expenditures on seeds, hired labor, and other miscellaneous expenditures. Moreover, the program had an impact on expenditures on seeds for surplus-labor households (column III), although the magnitude of this impact is greater for surplus-labor households than that for nuclear-households. Note also that single-headed households (column I) followed the negative trend over time shown in column I of table 4.5 for crop expenditures: for 4 out of 5 expenditure indicators, the coefficient is negative, which gives an indication of a sort of retrenchment in agricultural production (although none of the coefficients are statistically significant) beyond the decreasing trend in spending on crop production. The contrasting results for hired labor across these three groups of households is

informative and it makes intuitive sense: single-headed households are labor constrained and since they seemed to have decreased their expenditures on crop production (beyond the decreasing trend), it makes sense that the sign of the coefficient on hired labor is also negative; on the other hand, nuclear households who showed the greatest positive impacts on crop expenditure (amount spent, and the share that spent) seem to have experienced a greater demand for hired labor as a result of this increase in input use, which is reflected on this indicator. Along these lines, the lower part of panel 1 also shows that the program had a greater impact on nuclear households for the share of households that spend on crop production: a greater share of nuclear households spends on crop production (15%) and on miscellaneous expenditures (15%); although the share of single-headed households that spend on crop production also increased substantially (21%) in comparison to their control group. On the other hand, even though the sign of the coefficients for all expenditures (lower part of panel 1) that capture the share of surplus-labor households that spend on these are positive, none of these are significant.

The second panel of table 4.6 shows the impact of the program on harvest, consumption, and sales. Recall that for the entire sample, table 4.5 showed that the program had no impact on yields. Yet, the disaggregation by household type based on household labor shows that the program did have an impact on maize yields for surplus-labor households (column III): the CGP increased maize yields for this group by 255 Kg./ha. Given the low levels of maize yields found at baseline, as shown in table 4.2, this increase is quite significant: baseline maize yields for treated households was 450 Kg./ha. This, then, represents a 56% increase in maize yields for surplus-labor households in relation to the baseline means for all treated households.[52] In terms of consumption, the results of table 4.6 corroborates that the program did not have an impact on

[52] Baseline means of maize yields for this group of treated households was 566 Kg./ha, which means that the 255 Kg./ha increase represents a 45% increase.

consumption from own production. Yet, we should note that the sign of the coefficients on consumption of maize, cassava, and rice are all positive for labor-surplus households. For sales, the program had a positive and significant impact on the amount of harvest sold for maize across all three groups; however, note that the magnitude of the impact is greatest for surplus-labor households (117 Kg./ha for surplus-labor households versus 53.3 Kg./ha for single-headed households and 49.92 Kg./ha for nuclear households). We should note that when comparing these increases in sales with the baseline means for all treated households for maize (36 Kg./ha), these are significant percentage increases: maize sales in Kg./ha increased by 325% for surplus-labor households; they increased by 234% for nuclear households; and they increased by 220% for single-headed households. Moreover, the program also had an impact on the amount of harvest sold for rice but only for nuclear households (an increase of 135% compared to the baseline mean for all treated). These positive results for the size of harvest sold for all three groups of households are also reflected in the value of sales, as one might expect: the CGP had a positive and significant impact in the value of harvest sold, the value of maize sold, and the value of rice sold (only for nuclear households). One interesting thing to note from these results is that the value of maize sold (per hectare) is greatest for surplus-labor households. One can, perhaps, speculate that this might be due to this type of household having more family labor, which might facilitate selling their maize in other markets where they might pay a higher price than the farm gate price. For the share of households that sell their harvest, the CGP had a positive impact for single-headed households (23% more households sell any of their harvest), and for nuclear households (13% more households sell any of their harvest and 3% more households sell rice); on the other hand, although the sign of all three coefficients for the share of households that sell their harvest is positive, none of these are significant for surplus-labor households.

Moving to the third panel for operated land, the CGP had the greatest impact on land expansion for nuclear households (column II), followed by single-headed households (column I), and with a modest impact on surplus-labor households (column III). For nuclear households (column II), the CGP had an impact on the total amount of operated land (an increase of 0.29 ha.), and when disaggregated this increase came from land devoted to maize (0.17 ha.), rice (0.08 ha.), groundnut (0.013), sorghum (0.01 ha.), and other beans (0.01 ha.). For single-headed households (column I), the CGP had an impact on the increase of land devoted to sweet potatoes (0.01 ha.), sorghum (0.02 ha.), and other beans (0.007 ha.). For surplus-labor households (column III), although most of the coefficients on increases of land are positive, the program only had an impact on land devoted to rice (an increase of 0.09 ha.). We should note that while the results for the increase in total operated land for nuclear households (0.29 ha.) is significant (and this is what is driving the results for the entire sample shown in table 4.5) in relation to the baseline means (0.5 ha.), which represents an increase of 53% of operated land, most of these increases in land expansion are fairly small in absolute terms. In fact, most of these increases might be considered to be similar to adding a vegetable garden or a garden plot; except, perhaps for the increases in land devoted to maize (increase of 0.17 ha.) and rice (0.08 ha.).

In contrast to the various impacts on harvest and sales, the impact of the CGP on investments in tools is very small: the program increased the share of households that own a hoe (1% for single-headed households, with a baseline of 92% for treated households), or a hammer (4% for nuclear households, with a baseline of 5% for treated households).

On the other hand, the program had a greater and clearer impact on investments in farm animals (particularly for labor-surplus households): the program increased the number of cattle owned (by 1.09 for surplus-labor households, and by 0.38 for single-headed households with a

baseline of 0.52 for treated households); it increased the number of goats owned (by 0.08 for nuclear households, and by 0.26 for labor-surplus households, with a baseline of 0.08 for treated households); it increased the number of chickens owned (by 1.1 for single-headed households, and by 1.4 for nuclear households, with a baseline of 2.1 for treated households); and it increased the number of ducks owned (by 0.16 for single-headed households, and by 0.39 for surplus-labor households, with a baseline of 0.10 for treated households).

Taken altogether, the heterogeneous impacts presented in table 4.6 paint a more nuanced and more complete picture of the impact of the program. These impacts can be summarized into three general impacts. First, the program had the greatest impact for nuclear households in terms of the number of indicators that were positively impacted: 23 indicators showed a positive and significant impact for nuclear households, followed by 11 indicators for single-headed households, and 9 indicators for labor-surplus households. In fact, when contrasting the results for the entire sample (table 4.5) with the heterogeneous impacts (table 4.6), we can see that the impacts are primarily driven by the impact of the program on nuclear households (for crop expenditures, rice sales, land expansion, and for investments in small farm animals). On the other hand, in terms of magnitude, the program had the greatest impact on surplus-labor households: they spent the most on seeds per hectare; the program had an impact on maize yields only for these households (and the increase was significant in relation to the baseline mean); the impact on harvest sales (Kg/ha.) was twice as big as the impact experienced by the other two types of households; the impact on the value of maize sales (ZMK/ha.) was greatest for these households; and the impact of the program on investments in large farm animals was greatest for this type of households (1.09 head of cattle versus 0.38 for single-headed households), and the impact on medium-sized animals was greatest for these households as well (an increase of goats by 0.26 in

contrast to an increase of 0.08 for nuclear households). Third, for labor-constrained households (single-headed households), the program seems to have retrenched these households from spending on crop inputs beyond the decreasing trends shown in column I of table 4.5 (except for the share of households that spend on miscellaneous expenditures), and when contrasting these results with those for investments in farm animals, it seems that this type of household preferred to invest in farm animals, rather than on crop expenditures. Yet, the program did help these households increase their value of harvest sales, especially for maize, and it increased the share of households that sell part of their produce. While these labor-constrained households did increase their share of land devoted to sweet potatoes, sorghum, and beans, these increases are about the size of a vegetable garden or garden plots, which gives an indication that these households might be more focused on subsistence production, which makes intuitive sense since they are labor-constrained.

Let us now turn to the results of the stochastic production frontier analysis to see if the program had an impact on technical efficiency.

*Results: Stochastic Production Frontier Analysis*

Table 4.7 provides the results of the SPFA model using maximum-likelihood. The results of the LR-test and the joint-test of significance (Table B.2 in Appendix B) show that the Translog fit the data better. Given that the variables were normalized to their geometric mean (GM) following common practice, the first-order-coefficients can be interpreted as partial production elasticities (Brato-Ureta, Greene, and Solis 2012). With the exception of expenditures on inputs, the model presents positive partial production elasticities for total family labor and total land operated in agricultural production. The results on input expenditures are contrary to those typically found in the literature; however, they are consistent with studies in

African countries that focus on smallholder farmers (Chirwa 2007; Nisrane, Berhane, Asraft et el. 2011). Yet, when looking at the second-order-coefficient for expenditures as well as its interaction with land, we can see that the coefficients are positive and significant. This provides evidence of the potential role of increasing input use. In fact, this result is consistent with Kalirajan (1991) who argues that budget restrictions are one of the main production constrains for small scale farmers in developing societies other than land. Moreover, the sum of all partial production elasticities is less than 1 revealing decreasing returns to scale; a result that is also consistent with previous research on small scale farmers in less favorable areas (e.g., González and López 2007; Solís et al. 2009; Chavas et al. 2005, as cited in Bravo-Ureta, Greene, and Solis 2012).

Moving to the other variables in the SPFA model, the first panel of table 4.7 shows that hired labor has a negative effect on production, which is consistent from the point of view that in the non-separable agricultural household model, family labor and hired labor are not perfect substitutes. The results also show that positive shocks increase the total value of harvest, while negative shocks have the inverse effect, as expected. Finally, the variable for participation in the program (Beneficiary) does not have any impact on the value of harvest, although the sign of the coefficient is positive. One way to interpret this result is that both sets of farmers essentially use the same set of technologies (González-Flores et al., 2014).

On the other hand we see that the variance of the inefficiency part of the error structure is a function of the education of the head of household, having access to credit, and the PPI score, which proxies wealth, while the beneficiary variable gives an indication that the program had no impact on technical efficiency, although the sign of the coefficient is positive. This means that having more education, having access to credit, and being less poor can significantly improve

technical efficiency (or decrease inefficiency). These results are also consistent with those for other African countries (Binam et al. 2004; Binam, Gockowski, and Nkamleu 2008; and Nisrane, Berhane, Asraft et el. 2011).

Table 4.7: SPFA for Total Value of Harvest, and Determinants of Random Shocks & Inefficiency

| Variable † | Coeff. | Sig. |
|---|---|---|
| Exp. On Inputs ($ Total) (x1) | -0.165 | |
| Tot. Fam. Labor (# days) (x2) | 0.201 | *** |
| Tot. Land Used (Has.) (x3) | 0.450 | *** |
| x1^2 | 0.072 | *** |
| x2^2 | 0.086 | ** |
| x3^2 | -0.190 | *** |
| x1*x2 | 0.001 | |
| x1*x3 | -0.001 | |
| x2*x3 | 0.058 | ** |
| Hired Labor (0,1) | -0.198 | * |
| No Fam. Labor (0,1) ^ | 0.127 | |
| No Exp. On Inputs (0,1) ^ | -1.715 | |
| # Positive Shocks | 0.144 | ** |
| # Negative Shocks | -0.068 | *** |
| Beneficiary (0,1) | 0.029 | |
| _cons | 13.647 | *** |
| RTS | 0.486 | |
| lnsig2v | | |
| # Positive Shocks | 0.491 | *** |
| # Negative Shocks | 0.138 | *** |
| _cons | -1.632 | *** |
| lnsig2u | | |
| Head Age | -0.006 | |
| Head Edu | -0.034 | ** |
| Access to credit (0.1) | -0.362 | *** |
| PPI Score | -0.011 | ** |
| Beneficiary (0,1) | 0.055 | |
| Hired Labor (0,1) | -0.020 | |
| _cons | 0.463 | ** |
| *N* | 1,936 | |

† Variables Transformed to Geometric Mean

^ = Battese (1997) Correction for Log of 0

* $p<.10$; ** $p<.05$; *** $p<.01$

Moving to the second and third panels, the results show that the variance of the idiosyncratic error term is a function of the positive and negative shocks, which is not surprising.

Finally, following Fried et al. (2008) we infer managerial ability by ranking the TE scores derived from a "best practice" production frontier obtained from the pooled model. These results are presented in table 4.8, where we see that TE is quite low for all households in the sample. These low levels of TE are in contrast to previous studies in African countries that look at mono-crop agriculture where TE levels are greater than 0.7 (Abdulai et al. (2013) and Abtania, Hailu and Mugera (2012) for the case of Ghana; Binam et al. (2004) for the case of Cameroon, and Binam and Gockowsky and Nkamleu (2008) for the case of Nigeria).

Table 4.8: Means of Technical Efficiency Scores

| | I | II | III | IV |
|---|---|---|---|---|
| | **Pooled** | **Treated** | **Control** | **Test** |
| | 0.556 | 0.554 | 0.558 | |
| ***N*** | 1,936 | 1,008 | 928 | |

* $p<.10$; ** $p<.05$; *** $p<.01$

Yet, these results are higher than other studies in African countries that focus on smallholder farmers (Binam, Sylla and Diarra et al. (2002) for the case of Ivory Coast, Chirwa (2007) for the case of Malawi and Nisrane, Berhane, Asraft et el. (2011) for the case of Ethiopia) where TE levels range from 0.415 in Ivory Coast, to 0.462 in Malawi. Nevertheless, table 4.8 shows that there are no differences in TE between treated and control households. Thus, we conclude that the CGP had no impact in this regard.

## 4.7 Conclusion & Discussion

This paper uses data collected between 2010 and 2012 from the Child Grant Program to assess whether the CGP had an impact on agricultural production (value and size of harvest, and gross margins), productive investments, and technical efficiency. The CGP was implemented

using an RCT phased-in approach where half of 90 communities were assigned to receive the treatment in 2010, while the other half were to receive the program starting in 2013. A combination of geographic and categorical targeting was used to identify households with at least one child under the age of 5. Beneficiary households receive 55,000 kwacha a month (equivalent to $11USD) independent of household size. In contrast to some of the biggest cash transfer programs in the world, such as *Oportunidades* in Mexico and *Bolsa Familia* in Brazil, the CGP does not impose any conditions attached to the cash transfer.

Using a difference-in-difference model for the entire sample, we find that the program did not have an impact on the value of harvest or gross margins. However, while there was a trend over time for households to spend less on crop production, the CGP seems to have reversed this negative trend (all coefficients related to crop expenditures are positive), and it had a positive and significant impact on expenditures in seeds, and the share of households that spend on crop production, including miscellaneous expenditures. Moreover, the program seems to also have reversed the negative trend over time for decreasing yields for maize and rice: the sign of the coefficients for the impact of the program on these two indicators is positive, although not significant. Furthermore, the increase in crop expenditures seems to be reflected in the amount of harvest sold for maize and rice (both positive and statistically significant), in the total value of sales, primarily driven by maize sales (both positive and significant), and the share of households that sell part of their harvest, primarily driven by sales of maize and rice (all positive and significant). We should note that the program had no impact on own consumption from harvest. However, if households are selling more of their harvest, it is plausible that the program had no impact in own consumption if the extra money earned from these sales is used to purchase other food items that the household does not produce. Thus, the result of no impact on consumption

makes intuitive sense. Given that the harvest-related variables were normalized by hectare, we also investigate if the program had an impact on the size of operated farm land. We find that the CGP had a positive and significant impact on total operated land (an increase of 0.23 ha.), primarily driven by increases in land devoted to maize (0.14 ha.), rice (0.06 ha.), sorghum (0.02), and beans (0.01)

The impact of the program on farm tools was very modest, although the impact on farm animals was greater: the CGP only increased the share of households that own a hammer (4%); while the program had a positive and significant impact on the number of cattle owned (increase of 0.44), goats (0.13), chickens (1.11), and ducks (0.2).

In order to provide a more nuanced and clearer picture of the impact of program, we run the same set of analyses across three household types based on differences in size of household labor, defined as household members 17 years old or older and younger than 65: (1) single-headed households; (2) nuclear households; (3) and surplus-labor households. The program impacted each of these types of households differently. For single-headed households, the program helped increase the share of households that spend on miscellaneous crop expenditures, it helped them increase the amount of maize sold, the total value of sales, primarily driven by maize, and it increased the share of households that sell part of their harvest. Moreover, the program increased their land devoted to sweet potatoes, sorghum, and other beans; although the sizes of these increases are equivalent to garden plots or vegetable gardens. Finally, the program helped single-headed households increase the number of cattle owned (by 0.28) and the number of chickens (by 1.1). These results suggest that for labor constrained households, the CGP is facilitating a diversification of income (more sales), and a diversification of crop production for subsistence (as signaled by the very small increases in land for three crops). On the other hand,

the CGP had the greatest impact on nuclear households for the total number of indicators that were positively impacted (23), and in fact, the positive impacts of the program presented for the entire sample in table 4.5 are driven primarily by the impact of the program on these nuclear households: for crop expenditures, rice sales, land expansion, and for investments in small farm animals. We should note that nuclear households also increased their expenditures on hired labor, which means that the CGP is also having positive externalities through these households that are providing employment opportunities for others. We should mention, however, that the results on the negative coefficients for gross margins and the null impact on yields per hectare for these households provide some evidence that increases in crop expenditures may not always lead to improvements in yields or gross margins, especially if households are not trained on the optimal ways in which these inputs should be applied. Thus, complementary programs, like extension programs might be necessary.

For surplus-labor households, the CGP had the greatest impact in terms of the magnitude of the benefits. For example, while the program had a positive impact on seed expenditures for nuclear and surplus-labor households, the magnitude of the increase was greatest for the latter. Moreover, surplus-labor households were the only ones able to translate increases in expenditures into greater yields, and in fact, in relation to the baseline means for treated households, the increase was quite substantial at 56%. Similarly, the impact of the program on maize sales—for Kg./ha and the value of maize sold—was greatest for surplus-labor households. Finally, the impact of the CGP on investments in large- and medium-size farm animals was also greatest for surplus-labor households: an increase of 1.09 for cattle (baseline of 0.38), and an increase of 0.26 for goats (baseline of 0.13). If it is easier for surplus-labor households to bring income from other sources other than the farm—since household labor is greater—then, it is

likely that the cash transfers provided by the CGP can create synergies with this additional income, which might facilitate greater expenditures on inputs, and greater investments in large- and medium-size animals, as the results of this paper have shown. Along these lines, having more household labor might facilitate finding markets that pay more than the farm gate price for agricultural produce, or it might facilitate selling in several markets, which might explain why surplus-labor households fetch a greater price on maize per hectare, as shown in this paper.

The results show that the program had no impact on technical efficiency; but we should note that technical efficiency scores are very low for all households in the sample. The results of the ancillary equation for the inefficiency part of the error structure shows that education and access to credit can significantly improve TE levels. These results are consistent with similar studies in African countries (Binam et al. 2004; Binam, Gockowski, and Nkamlew 2008; Nyariki 2011; Chirwa 2007) and indicate that the CGP might significantly benefit from complementary programs, such as extension programs, that can help reduce this large inefficiency gap.

In summary, this study provides evidence that the CGP improved agricultural production, but only for labor-surplus households, it increased expenditures on crop production, primarily driven by seeds and miscellaneous expenditures, it increased the sales of maize (amount and value) and the likelihood of selling, it increased the land devoted to agricultural production, and it increased investments in farm animals. One possible channel in which these unconditioned cash transfers lead to an increase in productive activities and investments is by relaxing liquidity constraints, as suggested by the agricultural household model (Boone et al. 2013). In this sense the CGP has made important headway in helping poor households improve their overall welfare and asset stock, as well as triggering a greater level of agricultural activity. These results are consistent with previous studies that evaluate the productive impact of similar programs in

various African countries (Zezza, de la Briere and Davis, 2010; Covarrubias, Davis, and Winters, 2012; Asfaw, Davis, Dewbre, et al. 2012). Importantly, the program had no impact on technical efficiency. Though, the technical efficiency scores found in this paper are very low. Thus, it seems that other complementary programs are also needed; particularly programs that can teach farmers the appropriate way to use and apply farm inputs. This might help improve gross margins and should improve productivity growth by decreasing the significant efficiency gap identified in this paper.

As a final comment, we should note that researchers and policy makers who are interested in finding synergies and complementarities between social and productive programs to promote agriculture as a tool for development and growth, should be clear that social programs such as the CGP target very poor households. As such, the potential for these households to make a significant contribution to the agricultural sector to make it a credible engine for growth may be unrealistic. After all, the total land owned by these households is very small (1.7 ha. at baseline), they barely use a third of it (0.5 ha. at baseline and 0.73 ha. at follow-up), and what land is used seems to be highly fragmented. Nevertheless, the CGP may have kick started some form of structural transformation: single-headed households who are labor-constrained seem to have retrenched from agricultural activities, while nuclear and surplus-labor households have intensified and expanded their agricultural endeavors.

# APPENDIX A

## PUNCTUAL TEST OF MEANS AND SAMPLE SELECTION EQUATION FOR CHAPTER 2

Table A.1: Punctual Test of Means

| | Mean | | % Reduction | |
|---|---|---|---|---|
| | **Treated** | **Control** | **of Bias** | **P>\|t\|** |
| ***Land*** | | | | |
| Land Owned (has.) | 2.75 | 2.72 | 9.5 | 0.944 |
| Owned Plots (#) | 3.37 | 3.25 | 75.5 | 0.602 |
| Black Soil (%) | 0.77 | 0.78 | 80.2 | 0.753 |
| Flat Land (%) | 0.38 | 0.38 | -1.6 | 0.890 |
| Irrigated Land (%) | 0.57 | 0.60 | 46.9 | 0.567 |
| ***Socio-Demographic*** | | | | |
| Family Size | 4.85 | 4.85 | 97.0 | 0.977 |
| Max Educ. In HH | 8.44 | 8.37 | 92.5 | 0.867 |
| % of Labor Force Male | 0.47 | 0.46 | 39.6 | 0.704 |
| Dependency Share | 0.29 | 0.29 | -40.0 | 0.818 |
| Credit Constrained (1,0) | 0.16 | 0.19 | 26.8 | 0.590 |
| Indigenous Head | 0.56 | 0.55 | 76.9 | 0.793 |
| Female Head (1,0) | 0.12 | 0.12 | 71.4 | 0.903 |
| Age of Head | 42.02 | 42.22 | 59.5 | 0.892 |
| ***Welfare*** | | | | |
| House (1,0) | 0.83 | 0.84 | 75.1 | 0.720 |
| Concrete/brick House (1,0) | 0.82 | 0.83 | 92.2 | 0.885 |
| Refrigerator (1,0) | 0.12 | 0.14 | 68.4 | 0.669 |
| Access to Water System (1,0) | 0.93 | 0.93 | 90.3 | 0.952 |
| Sewage (1,0) | 0.05 | 0.06 | -78.7 | 0.835 |
| Big Farm Animals (#) | 5.43 | 5.73 | 17.2 | 0.614 |
| ***Social Capital*** | | | | |
| Ag. Ass. Membership 5 yrs.+ (1,0) | 0.06 | 0.06 | 86.7 | 0.912 |
| Non-Ag. Ass. Membership 5 yrs.+ (1,0) | 0.63 | 0.62 | 96.7 | 0.972 |
| ***Community variables*** | | | | |
| Bus in Community (1,0) | 0.43 | 0.46 | 42.4 | 0.677 |
| Elementary School (1,0) | 0.88 | 0.89 | -349.5 | 0.631 |
| Distance to Closest City (km) | 27.36 | 27.29 | 98.2 | 0.966 |
| Chimborazo (1,0) | 0.49 | 0.50 | 82.5 | 0.835 |
| ***N*** | 164 | 324 | | |

Table A.2: Selection Equation: Treated Communities Only

| | **Unmatched** | | **Matched** | |
|---|---|---|---|---|
| | **Treated** | **Controls** | **Treated** | **Controls** |
| | **Coeff.** | **Coeff.** | **Coeff.** | **Coeff.** |
| Age of Head | .08031** | -.08031** | .0744** | -.07444** |
| Age of Head^2 | -.00079** | .00079** | -.00073** | .00073** |
| Education of Head | 0.03135 | -0.03135 | 0.03976 | -0.03976 |
| Education of Head^2 | -0.00102 | 0.00102 | -0.0013 | 0.00133 |
| Family Size | 0.01074 | -0.01074 | -0.00225 | 0.00225 |
| Land Owned (ha) | 0.01803 | -0.01803 | 0.01228 | -0.0123 |
| Constant | -2.03667*** | 2.03667*** | -1.89953** | 1.89953** |
| ***N*** | 340 | 340 | 327 | 327 |

* $p<0.10$; ** $p<0.05$; *** $p<0.01$

APPENDIX B

ATTRITION, LIKELIHOOD OF NO HARVEST, AND LR & JOINT TEST OF SIGNIFICANCE FOR CHAPTER 4

Table B.1: Logit on Attrition with Odds Ratio

| | Variables (units) | I Odds Ratio | II Sig |
|---|---|---|---|
| **Status** | Beneficary (0,1) | 0.91 | |
| **Head of HH Chars.** | Head Age (yrs.) | 1.06 | |
| | Head Age^2 | 1.00 | |
| | Head Female | 1.63 | |
| | Head Married/Co-habit | 1.12 | |
| | Head Never Married | Base | |
| | Head Divorced, Separated, Widowed | 1.17 | |
| | Head Education (yrs.) | 0.98 | |
| | Head Education^2 | 1.00 | |
| **HH Characteristics** | Family Size | 1.01 | |
| | % Share Female | 0.94 | |
| | Dependency Ratio | 0.94 | |
| | # Orphans in HH | 0.96 | |
| | HH member in wage labor | 0.81 | |
| | Non-Ag. Business | 0.87 | |
| | Remittances | 0.78 | |
| **Wealth & Assets** | TLU Total ^ | 0.81 | |
| | PPI Score | 1.02 | |
| | Land Operated | 1.10 | |
| **Access to Svcs.** | Water Src.: River/Lake/Spring/Dam/Rain (1,0) | 1.01 | |
| | Water Src.: Protected Well (1,0) | 1.49 | |
| | Water Src.: Unprotected Well (1,0) | Base | |
| | Water Src.: Boreholel (1,0) | 0.67 | |
| | Water Src.: Tap or Other (1,0) | 0.56 | |
| | Distance to Water Source (Km.) | 1.03 | ** |
| | Garbage Disp.: Collected (1,0) | Base | |
| | Garbage Disp.: Pit (1,0) | 0.84 | |
| | Garbage Disp.: Dumping (1,0) | 1.06 | |
| | Garbage Disp.: Burning or Other (1,0) | 1.26 | |
| **Shocks** | # Positive Shocks | 1.65 | * |
| | # Negative Shocks | 1.01 | |
| **Instit. & Other Supp.** | Receives Gov't Support (1,0) | 1.49 | * |
| | Receives NGO Support (1,0) | 0.65 | |

| | | |
|---|---|---|
| | Receives Other Support (1,0) | 0.94 |
| **Qual. Control** | Interviewr Speaks Diff. Lang. | 0.97 |
| **Regional FE** | YES | |
| ***N*** | | 2,515 |

^ Tropical Livestock Units provides a convenient method for quantifying a wide range of different livestock.

We use the same conversion units Daidone, et al. (2013).
* $p<0.10$;  ** $p<0.05$;  *** $p<0.01$

Table B.2: LR-Test and Joint Test of Significance

CD (model 1; nested): Log likelihood = -2238.9662

TL (model 2): Log likelihood = -2215.8541

LR-test = 2*(model 2 – model 1) = chi2 (6) = 46.2242; Prob > chi2 = 0.0000

Joint-test of Significance: CD versus TL

```
 test  $TL
 ( 1)  [ly]GMLX11 = 0
 ( 2)  [ly]GMLX22 = 0
 ( 3)  [ly]GMLX33 = 0
 ( 4)  [ly]GMLX12 = 0
 ( 5)  [ly]GMLX13 = 0
 ( 6)  [ly]GMLX23 = 0

         chi2(  6) =   44.42
       Prob > chi2 =    0.0000
```

Table B.3: Logit for Probability of Not Harvesting

| | Variables (units) | I Odds Ratio | II Sig |
|---|---|---|---|
| **Status** | Beneficiary (0,1) | 1.02 | |
| **Head of HH Chars.** | Head Age (yrs.) | 0.95 | |
| | Head Age^2 | 1.00 | |
| | Head Female | 1.04 | |
| | Head Married/Co-habit | 1.04 | |
| | Head Never Married | Base | |
| | Head Divorced, Separated, Widowed | 1.09 | |
| | Head Education (yrs.) | 0.92 | ** |
| | Head Education^2 | 1.00 | |
| **HH Characteristics** | Family Size | 0.92 | * |
| | % Share Female | 1.73 | |
| | Dependency Ratio | 1.04 | |
| | # Orphans in HH | 1.13 | ** |
| | HH member in wage labor | 1.20 | |
| | Non-Ag. Business | 1.02 | |
| | Remittances | 1.09 | |
| **Wealth & Assets** | TLU Total ^ | 0.84 | |
| | PPI Score | 1.00 | |
| | Land Operated | 1.10 | |
| **Access to Svcs.** | Water Src.: River/Lake/Spring/Dam/Rain (1,0) | 1.31 | |
| | Water Src.: Protected Well (1,0) | 1.83 | ** |
| | Water Src.: Unprotected Well (1,0) | Base | |
| | Water Src.: Boreholel (1,0) | 1.40 | |
| | Water Src.: Tap or Other (1,0) | 3.92 | *** |
| | Distance to Water Source (Km.) | 1.02 | * |
| | Garbage Disp.: Collected (1,0) | Base | |
| | Garbage Disp.: Pit (1,0) | 1,011,399 | *** |
| | Garbage Disp.: Dumping (1,0) | 1,111,429 | *** |
| | Garbage Disp.: Burning or Other (1,0) | 1,146,401 | *** |
| **Shocks** | # Positive Shocks | 1.51 | |
| | # Negative Shocks | 0.82 | |
| **Instit. & Other Supp.** | Receives Gov't Support (1,0) | 2.41 | *** |
| | Receives NGO Support (1,0) | 1.58 | |
| | Receives Other Support (1,0) | 0.96 | |
| **Qual. Control** | Interviewr Speaks Diff. Lang. | 1.42 | |
| **Regional FE** | YES | | |
| ***N*** | | 2,625 | |

^ Tropical Livestock Units provides a convenient method for quantifying a wide range of different livestock.

We use the same conversion units Daidone, et al. (2013).
* $p<0.10$;  ** $p<0.05$;  *** $p<0.01$

REFERENCES


Abadie, Alberto, and Guido W. Imbens. 2002. Simple and bias-corrected matching estimators for average treatment effects. Technical Working Paper 283. National Bureau of Economic Research.

Abatania, Luke N., Atakelty Hailu, and Amin W. Mugera. 2012. Analysis of farm household technical efficiency in Northern Ghana using bootstrap DEA. Paper presented at the 56th annual conference of the Australian Agricultural and Resource Economics Society. The Esplanade Hotel, Fremantle WA, 7-10 February 2012.

Abdulai, Shamsudeen, Paul K. Nkegbe, and Samuel A. Donkoh1. 2013. Technical efficiency of maize production in Northern Ghana. *African Journal of Agricultural Research*, Vol. 8(43), pp. 5251-5259, 7 November, 2013

Adesina, A., 2010. Conditioning trends shaping the agricultural and rural landscape in Africa. *Agricultural Economics* 41(S1), 73–82.

Agüero, Jorge M., Michael R. Carter and Ingrid Woolard. 2006. The Impact of Unconditional Cash Transfers on Nutrition: The South African Child Support Grant, Southern Africa Labour and Development Research Unit Working Paper Number 06/08. Cape Town: SALDRU, University of Cape Town.

Ahmed, A., & H Bouis. 2002. Weighing what's practical: Proxy means tests for targeting food subsidies in Egypt. *Food Policy*, 27(5–6), 519–540.

Ahmed, Akhter U., and Howarth E. Bouis. 2002. Weighing what's practical: proxy means tests for targeting food subsidies in Egypt. *Food Policy* 27.5 (2002): 519-540.

Aigner, D., K. Lovell, and P. Schmidt. 1977. Formulation and Estimation of Stochastic Frontier Production Function Models, *Journal of Econometrics*, 6, pp. 21-37.

Alderman, H. 1987. Allocation of Goods Through Non-Price Mechanisms: Evidence on Distribution by Willingness to Wait. *Journal of Development Economics* 25, 105-124.

Ali, Mubarak, and Derek Byerlee. 1991. Economic Efficiency of Small Farmers in a Changing World: A Survey of Recent Evidence. *Journal of International Development*, Vol. 3, No. 1, 1-27.

Alkire, Sabina, and James Foster. 2011. Understandings and misunderstandings of ultidimensional poverty measurement. *Journal of Economic Inequality*, 9 (2011), pp. 289–314

Anderson, Richard G., William H. Greene, B. D. McCullough, and H. D. Vinod. 2005. The

Roleof Data and Program Code Archives in the Future of Economic Research. Federal Reserve Bank of St. Louis. Working Paper 2005-014C.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics*. Princeton University Press, New Jersey.

Arriagada, Irma. 2002. Changes and inequality in Latin American families. *CEPAL Review*, 77. p. 135.

Asfaw, S., Daidone, S., Davis, B., Dewbre, J., Romeo, A., Djebbari, H., Winters, P., and Covarrubias, K. 2012. Analytical Framework for Evaluating the Productive Impact of Cash Transfer Programmes on Household Behaviour--Methodological Guidelines for the From Protection To Production Project. International Policy Centre for Inclusive Growth, United Nations Development Programme, Working Paper number 101 December, 2012.

Austin, Peter C, and Ewout W Steyerberg. 2012. *BMC Medical Research Methodology* 2012, 12:82

Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen. 2003. *The Journal of the Operational Research Society*, Vol. 54, No. 6 (Jun., 2003), pp. 627-635.

Baird, Sarah, Craig McIntosh, and Berk Özler. 2010. Cash or Condition? Evidence from a Randomized Cash Transfer Program. Policy Research Working Paper 5259, Impact Evaluation Series No. 45, The World Bank, Development Research Group, Poverty and Inequality Team.

Baker, Judy L. 2000. *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners.* Washington D.C.: LCSPR/PRMPO, The World Bank.

Bardhan, P. and Udry, C. 1999. *Development Microeconomics*, Oxford University Press: Oxford.

Barrett, Christopher B., and Michael R. Carter. 2010. The power and pitfalls of experiments in development economics: some non-random reflections. *Applied Economic Perspectives and Policy* 32, 515–548.

Barrett, Christopher B., M. Bezuneh, and A. Aboud. 2001. Income diversification, poverty traps and policy shocks in Cote d'Ivoire and Kenya, *Food Policy*, 26(4): 367–384.

Bartus, Tomás. 2005. Estimation of marginal effects using margeff. 2005. *The Stata Journal* (2005) 5, Number 3, pp. 309–329.

Battese, George E. 1997. A Note on the estimation of Cobb-Douglas Production Functions when Some Explanatory Variables Have Zero Values. *Journal of Agricultural Economics*, 48(2) (1997) 250-252.

Baulch, Bob. 2002. Poverty monitoring and targeting using ROC curves: examples from

Vietnam. IDS Working Paper 161.

Besley, T.J. 1989. Means Testing Versus Universal Provision in Poverty Alleviation Programmes. *Economica*, 57(1): 119–129.

Besley, T.J, & R. Kanbur. 1993. Principles of targeting. In M. Lipton, & J. van der Gaag (Eds.), *Including the poor*. Washington, DC: World Bank.

Besley, T.J. and S. Coate. 1988. Workfare vs. Welfare: Incentive Arguments for Work Requirements in Poverty Alleviatiop Programs. Woodrow Wilson School Discussion Papers in Economics, Researh Program in DevelopmentS tudies, No. 142.

Binam, Joachim Nyemeck, Jean Tonye`, Njankoua wandji, Gwendoline Nyambi, and Mireille Akoa. 2004. Factors affecting the technical efficiency among smallholder farmers in the slash and burn agriculture zone of Cameroon. *Food Policy* 29 (2004) 531–545

Binam, Joachim Nyemeck, Jim Gockowski, and Guy Blaise Nkamleu. 2008. Technical Efficiency and Productivity potential of cocoa farmers in West African countries. *The Developing Economies*, XLVI-3 (September 2008): 242–63

Binam, Joachim Nyemeck, Kalilou Sylla, Ibrahim Diarra and Gwendoline Nyambi. 2003. Factors Affecting Technical Efficiency among Coffee Farmers in Coˆ te d'Ivoire: Evidence from the Centre West Region. *R&D Management* 15, 1, 2003.

Blanco, Antonio, Rafael Pino-Mejías, Juan Lara, and Salvador Rayo. 2013. Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications* 40 (2013) 356–364

Blundell, Richard, and Monica Costa Dias. 2008. Alternative Approaches to Evaluation in Empirical Microeconomics. CeMMAP Working Paper 26/08, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, London.

Boone, Ryan, Katia Covarrubias, Benjamin Davis, and Paul Winters. 2013. Cash Transfer Programs and Agricultural Production: The Case of Malawi. *Agricultural Economics* 44 (3): 365–78

Boucher, Stephen, Reuben Summerlin, and Meritxell Martinez. 2010. Poverty Targeting and Measurement Tools in Microfinance, Progress out of Poverty Index and the Poverty Assessment Tool. Report to the Social Performance Task Force.

Bowles, S. 2004. *Microeconomics: Behavior, Institutions, and Evolution*. New Jersey: Princeton University Press.

Bravo-Ureta, Boris E., Alexandre Nunes Almeida, Daniel Solís D., and Aarón Inestroza. 2011. The impact of MARENA's investments on sustainable agricultural systems in Honduras. *Journal of Agricultural Economics* 62, 429–448.

Bravo-Ureta, Boris E., Daniel Solís, Victor Moreira, José Maripani, Abdourahmane Thiam, and Teodoro Rivas. 2007. Technical efficiency in farming: a meta-regression analysis. *Journal of Productivity Analysis* 27, 57-72.

Bravo-Ureta, Boris E., William Greene, and Daniel Solís, D. 2012. Technical efficiency analysis correcting for biases from observed and unobserved variables: an application to a natural resource management project. *Empirical Economics* 43, 55-72.

Bruhn, M., and, McKenzie, D. 2009. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, Vol. 1, No. 4 (October 2009), pp. 200-232.

Burke, William J., Munguzwe Hichaambwa, Dingiswayo Banda, and T. S. Jayne. 2011. The Cost of Maize Production by Smallholder Farmers in Zambia. Working Paper No. 50, Food Security Research Project, Lusaka, Zambia

Burman, Leonard E., W. Robert Reed, and James Alm. 2010. *Public Finance Review* 38(6) 787-793.

Caliendo, Marco, and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22, 31–72.

Carter, M. R., and Yang Yao. 2002. Local versus Global Separability in Agricultural Household Models: The Factor Price Equalization Effect of Land Transfer Rights. *Amer. J. Agr. Econ*. 84(3) (August 2002): 702-715.

Cassell, David. 2007. Don't Be Loopy: Re-Sampling and Simulation the SAS® Way. Statistics and Data Analysis, SAS Global Forum. Paper 183-2007.

Cattaneo, Matias D., Sebastian Galiani, Paul J. Gertler, Sebastian Martinez, and Rocio Titiunik. 2009. Housing, Health, and Happiness. *American Economic Journal: Economic Policy* 2009, 1:1, 75–105.

Cavatassi, Romina, Mario González , Paul. C. Winters, Jorge Andrade-Piedra, Patricio Espinosa, and Graham Thiele. 2009. Linking smallholders to the new agricultural economy: an evaluation of the Plataformas program in Ecuador. ESA Working Paper No. 09-06, Food and Agricultural Organization.

Cavatassi, Romina, Mario González-Flores, Paul C. Winters, Jorge Andrade-Piedra, Patricio Espinosa, and Graham Thiele 2011a. Linking smallholders to the new agricultural economy: the case of the *Plataformas de Concertación* in Ecuador. *Journal of Development Studies* 47, 1545-1573.

Cavatassi, Romina, Mario Gonzalez-Flores, Lina Salazar, and Paul C. Winters. 2011b. Do

Agricultural Projects Alter Crop Production Technologies? Evidence from Ecuador. *Journal of Agricultural Economics* 62, 403-428.

Cerdán-Infantes, Pedro, Alessandro Maffioli, and Diego Ubfal. 2008. The impact of agricultural extension services: The case of grape production in Argentina. OVE Working Papers 2008/6, Inter-American Development Bank, Office of Evaluation and Oversight.

Chavas J, Petrie R, Roth M. 2005. Farm household production inefficiency in the Gambia: resource constraints and market failures. *Am J Agric Econ* 87:160–179

Chirwa, Ephraim W. 2007. Sources of Technical Efficiency among Smallholder Maize Farmers in Southern Malawi. AERC Research Paper 172. African Economic Research Consortium, Nairobi, November 2007.

Chromy, James R. and Savitri Abeyasekera. 2001. Chapter 19: Statistical analysis of survey data. In eds. *Household Surveys in Developing and Transition Countries: Design, Implementation and Analysis.*

Coelli, Timothy J., D.S. Prasada Rao, Christopher J. O'Donnell, and George E. Battese. 2005. *An Introduction to Efficiency and Productivity Analysis*. 2nd edition, Springer, New York.

Conley, Timothy G., and Christopher R. Udry. 2010. Learning about a New Technology: Pineapple in Ghana. *American Economic Review*, 100(1): 35-69.

Cornia, G., & Stewart, F. 1995. Chapter 13: Two errors of targeting. In D. Van de Walle, & K. Nead (Eds.), *Public spending and the poor: Theory and evidence*. Baltimore: Johns Hopkins University Press.

Covarrubias, K., Davis, B., Winters, P., 2012. From Protection to Production: Productive Impacts of the Malawi Social Cash Transfer', *Journal of Development Effectiveness* 4(1), 50-77.

Daidone, Silvio. 2014. Personal communication.

Dargatz, D.A., and G.W. Hill. 1996. Analysis of survey data. *Preventive Veterinary Medicine* 28(1996) 225-237.

Davis, B., Gaarder, M., Handa S. and Yablonski, J. 2012. Evaluating the impact of cash transfer programs in Sub Saharan Africa: an introduction to the special issue, *Journal of Development Effectiveness*, 4(1): 1–8, March.

de Janvry A. and Sadoulet, E., 2009. Agricultural Growth and Poverty Reduction: Additional Evidence. *World Bank Research Observer* 25(1), 1-20.

De Soto, H. 2000. *The Mystery of Capital*. New York: Basic Books. Harpercollins

Del Carpio, Ximena V., and Mywish Maredia. 2009. Measuring the Impacts of Agricultural Projects: A Meta-Analysis of the Evidence. Working Paper, Independent Evaluation Group, World Bank. Washington, D.C.: World Bank.

Devaux, André, Douglas Horton, Claudio Velasco, Graham Thiele, Gastón Lopez, Thomas Bernet, Iván Reinoso, and Miguel Ordinola. 2009. Collective action for market chain innovation in the Andes. *Food Policy* 34, 31–38.

Devereux, Stephen. 2009. Social Protection for Agricultural Growth in Africa. The Future of Agricultures, Working Paper No. SP06.

Devereux, Stephen, and Bruce Guenther. 2007. Social Protection and Agriculture in Ethiopia, Country case study paper prepared for a review commissioned by the FAO on 'Social Protection and Support to Small Farmer Development'.

Devereux, Stephen, and Bruce Guenther. 2009. Agriculture and Social Protection in Ethiopia. The Future of Agricultures, Working Paper 008.

Dewald, William G., Jerry G. Thursby and Richard G. Anderson. 1986. Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, Vol. 76, No. 4 (Sep., 1986), pp. 587-603.

Dorward, A. and Chirwa, E., 2011. The Malawi agricultural input subsidy programme: 2005/06 to 2008/09. *International Journal of Agricultural Sustainability* 9(1), 232-247.

Dreze, J. 1986. Famine Prevention in India. J.Drezea nd A.K. Sen (eds) *Hunger: Economics* and Policy, OxfordU niversity Press.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. Using randomization in development economics research: a toolkit. In *Handbook of development economics*, eds. T. Schultz, and J. Strauss J. Elsevier, Amsterdam, pp 3895–3962.

ENAHO (Encuesta Nacional de Hogares). 2007. Encuesta Nacional De Hogares Sobre Condiciones De Vida Y Pobreza –ENAHO 2007. Documento: Ficha Tecnica.

ENAHO. 2007b. Encuesta Nacional De Hogares 2007, Manual de Encuestador. Instituto Nacional de Estadistica e Informatica. Direccion Nacional de Censos y Encuestas, Doc. ENAHO. 08.01. Lima, Peru.

Efron, B. 1979. Boostrapt Methods: Another Look at the Jacknife, *Annals of Statistics*, 7: 1-26.

Efron, B. and R. Tibshirani. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, Vol. 1, No. 1 (Feb., 1986), pp. 54-75.

Elbers, Chris; Lanjouw, Peter; and Phillippe George Leite. 2008. Brazil within Brazil: Testin

the Poverty-Map Methodology in Minas Gerais. World Bank Policy Research Working Paper No. 4513, Washington, D.C., elibrary.worldbank.org/ doi/book/10.1596/1813-9450-4513, retrieved 19th, January 2014.

Ellis, F. 2000. *Rural Livelihoods and Diversity in Developing Countries*. Oxford: Oxford University Press

FAO. 2012. Protetcion to Production Website. URL: http://www.fao.org/economic/PtoP/en/ Accessed 02/02/13.

FAO. 2014. Giews Country Briefs, Zambia. URL: http://www.fao.org/giews/countrybrief/country.jsp?code=ZMB Accessed on March 19, 2014

Farrington, David P., and Rolf Loeber. 2000.Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour and Mental Health*, 10, 100–122.

Farrington, John, Paul Harvey, and Rachel Slater. 2005. Cash Transfers in the Context of Pro-Poor Growth. Discussion paper for OECD/DAC Povnet Risk and Vulnerability Task Group. Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) GmbH, and ODI.

Feder, G., Lau, L., Lin, J. and Luo, X. 1990. The Relationship between Credit and Productivity in Chinese Agriculture: A Microeconomic Model of Disequilibrium, *American Journal of Agricultural Economics*, 72(5): 1151–1157.

Feed the Future (FTF). 2011. Zambia, FY 2011-2015 Multi-Year Strategy. URL: http://feedthefuture.gov/sites/default/files/country/strategies/files/ZambiaFTFMulti-YearStrategy.pdf (accessed April 14, 2014).

Feldstein, Martin. 1974. Social Security, Induced Retirement, and Aggregate Capital Accumulation, *Journal of Political Economy*, September/October 1974, 82, 905-26.

Fenwick, L. and Lyne, M. 1999. The relative importance of liquidity and other constraints inhibiting the growth of small-scale farming in KwaZulu-Natal, *Development Southern Africa*, 16(1): 141–155.

Fiszbein et al. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty*. The World Bank, Washington, D.C.

Fried, Harold O., C.A. Knox Lovell, Shelton S. Schmidt. 2008. *The measurement of productive efficiency and productivity growth*. Oxford University Press, New York.

Gertler et al. 2011. *Impact Evaluation in Practice*. Washington, D.C.: The World Bank

Gertler, Paul, Sebastian Martinez and Marta Rubio-Codina. 2006. Investing Cash Transfers to

Raise Long-Term Living Standards. IMPACT EVALUATION SERIES NO. 6, World Bank Policy Research Working Paper 3994.

Gilligan, Daniel O., John Hoddinott & Alemayehu Seyoum Taffesse. 2009. The Impact of Ethiopia's Productive Safety Net Programme and its Linkages, *The Journal of Development Studies*, 45:10, 1684-1706, DOI: 10.1080/00220380902935907

González M, López L. 2007. Political violence and farm household efficiency in Colombia. *Econ Dev Cult Change* 55:367–392.

González-Flores, M., Bravo-Ureta, B.E., Solis, D., and Paul Winters. 2014. The impact of high value markets on smallholder productivity in the Ecuadorean Sierra: A Stochastic Production Frontier approach correcting for selectivity bias, *Food Policy* 44 (2014) 237–247.

González-Flores, M., M. Heracleous, and P. Winters. 2012. Leaving the Safety Net: An Analysis of Dropouts in an Urban Conditional Cash Transfer Program. *World Development* Vol. 40, No. 12, pp. 2505–2521.

Gönen, Mithat. 2003. Receiver Operating Characteristic (ROC) Curves. Statistics and Data Analysis. SAS SUGI, Paper 210-31.

Greene, William. 2005. *Econometric analysis*, 6th edn. Prentice Hall, New Jersey

Greene, William. 2010. A stochastic frontier model with correction for sample selection. *Journal of Productivity Analysis* 34, 15–24.

Grosh, Margaret, and Judy L. Baker. 1995. Proxy Means Tests for Targeting Social Programs: Simulations and Speculation. LSMS Working Paper No. 118, Washington, D.C.: World Bank.

Guan, Weihua. 2003. From the help desk: Bootstrapped standard errors. *The Stata Journal* 3, Number 1, pp. 71–80.

Guo, Shenyang Y., and Mark W. Fraser. 2010. *Propensity Score Analysis, Statistical Methods and Applications. Advanced Quantitative Techniques in the Social Sciences*, No. 12. SAGE Publications, California.

Hamermesh, Daniel S. 2007. Viewpoint: Replication in economics. *Canadian Journal of Economics / Revue canadienne d'Economique*, Vol. 40, No. 3

Handa, S. and Davis, B., 2006. The Experience of Conditional Cash Transfers in Latin America and the Caribbean. *Development Policy Review* 24(5), 513-536.

Hanlon Joseph, Armando Barrientos, and David Hulme, 2010. *Just Give Money to the Poor: The Development Revolution from the Global South*. CT: Kumarian Press.

Harris, S. D. 2007. State of the Microcredit Summit Campaign Report 2007, Washington, DC: Microcredit Summit Campaign.

Heckman, James J., 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–161.

Hermansen, Sigurd W. 2008. Data Mining and Predictive Modeling. SAS Global Forum 2008, Paper 143-2008.

Herndon, Thomas, Michael Ash, and Robert Pollin. 2013. Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogo. Political Economy Research Institute, University of Massachusetts Amherst, Working Paper Series, Number 322.

Horton Douglas, André Devaux, and Miguel Ordinola. 2011. Highlights of the Papa Andina Experience. In *Innovation for Development: The Papa Andina Experience*, eds. André Devaux, Miguel Ordinola, and Douglas Horton. International Potato Center, Lima, Peru, pp 1-36.

IDB (Inter-American Development Bank). 2010. *Development Effectiveness Overview*, *Special Topic, Assessing the Effectiveness of Agricultural Interventions*. Inter-American Development Bank, Washington, D.C.

IFAD. 2011. Rural Poverty Report 2011: New realities, new challenges: new opportunities for tomorrow's generation. International Fund for Agricultural Development (IFAD), Rome.

Imp-Act. 2003. Microfinance and Poverty: Developing systems for monitoring depth of poverty outreach and Impact, Report based on a one-day seminar held at Polokwane, South Africa, May, 2003.

INEI (Instituto Nacional de Estadística e Informática). 2008. Perú: Perfil de la Pobreza por departamentos, 2005-2007. Dirección Técnica de Demografía e Indicadores Sociales

IRIS. 2005a. Accuracy Results for 12 Poverty Assessment Tool Countries. http://www.povertytools.org/other_documents/Accuracy%20Results%2012.pdf

_____. 2005b. Note on Assessment and Improvement of Tool Accuracy. http://www.povertytools.org/training_documents/Introduction%20to%20PA/Accuracy_Note.pdf

_____. 2011. Accuracy Results for 31 Poverty Assessment Tools. Developing Poverty Assessment Tools Project.

Jelin, Elizabeth, and Ana Rita Díaz-Muñoz. 2003. Major trends affecting families: South America in perspective. Report prepared for United Nations Department of Economic and Social Affairs Division for Social Policy and Development Programme on the Family.

Johannsen J. 2006. Operational poverty targeting in Peru – Proxy Means Testing with non-

ncome indicators Working paper No 30. International Poverty Center/UNDP.

Johnson, M., Hazell, P., and Gulati, A., 2003. The Role of Intermediate Factor Markets in Asia's Green Revolution: Lessons for Africa? *American Journal of Agricultural Economics* 85(5), 1211-1216.

Jondrow James, C.A. Knox Lovel, Ivan S. Materov, Peter Schmidt. 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19:233–238.

Kalirajan, K.P. 1991. The importance of efficiency use in the adoption of technology: a micro panel data analysis. *Journal of Productivity Analysis* 2, 113–126

Key, N., Sadoulet, E. and de Janvry, A. 2000. Transaction Costs and Agricultural Household Supply Response, *American Journal of Agricultural Economics*, 82(2): 245–59.

Kidd Stephen and Emily Wylde. 2011. Targeting the Poorest: An assessment of the proxy means test methodology. Australian Government, AusAID.

Kumbhakar, Subal C., C.A. Knox Lovell. 2000. *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge.

Legovini, Arianna. 1999. Targeting Methods for Social Programs. Poverty & Inequality Technical Note 1. Washington D.C.: Inter-American Development Bank.

Leuven, Edwin, and Barbara Sianesi. 2003. PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing. http://ideas.repec.org/c/boc/bocode/s432001.html.

Lewis, W. Arthur. 1954. Economic Development with Unlimited Supplies of Labour, *The Manchester School*, 22: 139-191.

Lichtenberg, E. and Zilberman, D. The econometrics of damage control: Why specification Matters. 1986. *American Journal of Agricultural Economics*, Vol. 68, (1986) pp. 261–73.

Lund, Francie  Michael Noble, Helen Barnes, Gemma Wright. 2009. Is there a rationale for conditional cash transfers for children in South Africa? *Transformation: Critical Perspectives on Southern Africa*, Number 70, pp. 70-91. DOI: 10.1353/trn.0.0038

Maes, J. 2006. Microfinance Services for Very Poor People: Promising Approaches from the Field. Poverty Outreach Working Group, SEEP Network.

Maes, J., and Vekaria, K. 2008. Moving the World's Poorest Families Out of Poverty. Poverty Outreach Progress Brief, USAID, and The SEEP Network, October 2008.

McCrary, Justin. 2001. Do Electoral Cycles in Police Hiring Really Help Us Estimate the Effect of Police on Crime? Center for Labor Economics University Of California, Berkeley, Working Paper No. 5

McCullough, B.D. 2007. Got Replicability? The Journal of Money, Credit and Banking Archive. *Econ Journal Watch*, Volume 4, Number 3, September 2007, pp 326-337.

McCullough, B.D., Kerry Anne McGeary & Teresa D. Harrison. 2008. Do economics journal archives promote replicable research? *Canadian Journal of Economics*, Canadian Economics Association, vol. 41(4), pages 1406-1420, November

Meeusen, W. and J. van den Broeck. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error, *International Economic Review*, 1977, 8, 435–444

Miguel, Edward, and Michael Kremer. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72 (1): 159–217.

Ministry of Agriculture & Cooperatives and The Central Statistical Office, Zambia. 2011. 2010/2011 Crop Forecast Survey Report.

Murphy, Kevin M., and Robert H. Topel. 2002. Estimation and inference in two stem econometric models. *Journal of Business & Economic* Statistics 20:88–97

Narayan, Ambar; and Nobuo Yoshida. 2005. Proxy Means Tests for Targeting Welfare Benefits in Sri Lanka. Report No. SASPR ‐ 7, Washington, D.C.: World Bank.

Narayan, Ambar; and Nobuo Yoshida. 2005. Proxy Means Tests for Targeting Welfare Benefits in Sri Lanka, Report No. SASPR–7, Washington, D.C.: World Bank.

Nisrane, Fantu, Guush Berhane, Sina fikeh Asrat, Gera work Getachew, Alemayehu Seyoum Taffesse, and John Hoddinott. 2011. Sources of Inefficiency and Grow th in Agricultural Output in Subsistence Agriculture: A Stochastic Frontier Analysis. Development Strategy and Governance Division, International Food Policy Research Institute – Ethiopia Strategy Support Program II, Ethiopia. ESSP II Working Paper 019.

Nyariki, Dickson M. 2011. Farm Size, Modern Technology Adoption, and Efficiency of Small Holdings in Developing Countries: Evidence from Kenya. *The Journal of Developing Areas* 01/2011; 45(1):35-52. DOI:10.1353/jda.2011.0016

Oxford Poverty and Human Development Initiative (OPHI). 2013. Country Briefing: Multidimensional Poverty Index (MPI) At a Glance

Park, A., Wang, S., and Wu, G. 2002. Regional poverty targeting in China. *Journal of Public Economics*, 86, 123–153.

Pico, H.A. 2006. La cadena agroalimentaria de la papa a través de la metodología de plataformas

de concertación y proyectos compartidos. Paper presented at the 1st national Ecuadorian potato congress, INIAP, Ecuador, 17–19 May 2006. Accessed at http://www.quito.cipotato.org/presentambato/TEMATICAS%20DEL%20CONGRESO/COMERCIALIZACION/HPICO.doc.

Progress Out of Poverty (POP). 2011a. Uganda PPI: Field Test Overview.

POP. 2011b. Uganda PPI: User Review Feedback.

Rahman, Sanzidur, Aree Wiboonpongse, Songsak Sriboonchitta, and Yaovarate Chaovanapoonphol. 2009. Production efficiency of Jasmine rice producers in northern and north-eastern Thailand. *Journal of Agricultural Economics* 60, 419–435.

Ricker-Gilbert, et al. 2011. Subsidies and Crowding Out: a double-hurdle model of fertilizer demad in Malawi. *Amer. J. Agr. Econ*. 93(1): 26–42; doi: 10.1093/ajae/aaq122

Robertson, Laura, Phyllis Mushati, Jeff rey W Eaton, Lovemore Dumba, Gideon Mavise, Jeremiah Makoni, Christina Schumacher, Tom Crea, Roeland Monasch, Lorraine Sherr, Geoffrey P Garnett, Constance Nyamukapa, Simon Gregson. 2013. Effects of unconditional and conditional cash transfers on child health and development in Zimbabwe: a cluster-randomised trial, *Lancet* 2013; 381: 1283–92.

Rogers-Farmer, Antoinette Y., and Diane Davis. 2001. Analyzing complex survey data. *Social Work Research*; Sept 2001; 25, 3. ProQuest pg. 185.

Sabates-Wheeler, Rachel, Stephen Devereux and Bruce Guenther. 2009. Building synergies between social protection and smallholder agricultural policies, FAC Working Paper No. SP01.

Schreiner, Mark. 2005. Un Índice de Pobreza para México, memo for Grameen Foundation U.S.A.

_____. 2006. A Simple Poverty Scorecard for Bangladesh. Microfinance Risk Management, L.L.C.

_____. 2009. Progress out of Poverty Index, A Simple Poverty Scorecard for Peru.

_____. 2012. Progress out of Poverty Index, A Simple Poverty Scorecard for Peru.

_____. 2013. Progress out of Poverty Index (PPI), A Simple Poverty Scorecard for Zambia.

_____. 2014. Comments on draft Chapter 2 of thesis.

Schreiner, Mark, Matul, Michal, Pawlak, Ewa, and Sean Kline. 2004. Poverty Scorecards:

Lessons from a Microlender in Bosnia-Herzegovina. Microfinance Risk Management, L.L.C.

Schubert, Bernd and Rachel Slater. 2006. Social Cash Transfers in Low-Income African Countries: Conditional or Unconditional? *Development Policy Review*, 2006, 24 (5): 571-578.

Schultz, Theodore W. 1975. The Value of the ability to deal with disequilibria. *Journal of Economic Literature* 13, 827-846.

Scott, Kinnon. 2000. Chapter 21: Credit, in (eds.) Margaret Grosh and Paul Glewwe's *Designing Household Survey Questionnaires for Developing Countries, Lessons from 15 years of the Living Standards Measurement Study*. The World Bank, Volume 2.

SEEP (Small Enterprise Education and Promotion Network). 2008. Social Performance Map. Washington, D.C.: The SEEP Network.

Seidenfeld, David. 2013. Personal communication. Email exchange : Mon, Jun 3, 2013 4:54 pm

Seidenfeld et al. 2011. Zambia's Child Grant Program: Baseline Report. American Institutes for Research, Washington, D.C.

Singh, I., Squire L. and Strauss, J. (eds) 1986. *Agricultural household models: Extension, application and policy*. Baltimore, MD, Johns Hopkins University Press.

Sipiläinen, Timo, and Alfon G.J.M. Lansink. 2005. Learning in organic farming–an application on Finnish dairy farms. Paper prepared for presentation at the XIth Congress of the European Association of Agricultural Economists, Copenhagen, Denmark, August 24-27.

Solís, Daniel, Boris E. Bravo-Ureta, and Ricardo E. Quiroga. 2009. Soil conservation and technical efficiency among hillside farmers in Central America: a switching regression model. *Australian Journal of Agricultural and Resource Economics* 51, 491-510.

St. Bernard, Godfrey. 2003. Major Trends Affecting Families in Central America and the aribbean. Paper prepared for United Nations Division of Social Policy and Development Department of Economic and Social Affairs Program on the Family

StataCorp. 2009. Stata 11 Base Reference Manual.

Stewart, F., Laderchi, C. R., and Saith, R. 2007. Introduction: Four Approaches to Defining and Measuging Poverty. In F. Stewart, R. Saith, and B. Harris-White (Eds.) *Defining Poverty in the Developing World*.

Steyerberg, Ewout W., Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen,

Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattane. 2010. Assessing the Performance of Prediction Models A Framework for Traditional and Novel Measures. *Epidemiology*, Volume 21, Number 1, January 2010

Stine, Robert. An Introduction to Bootstrap Methods, Examples and Ideas. 1989. *Sociological Methods and Research*, Vol. 18, Nos. 2 & 3, November 1989/Februray 1990, 243-291.

Tarozzi, Alesandro; and Angus Deaton. 2007. Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas, *Review of Economics and Statistics*, Vol. 91, No. 4, pp. 773–792.

Taylor, J. E., & Adelman, I. 2003. Agricultural household models: Genesis, evolution, and extensions. *Review of Economics of the Household*, 1(1-2), 33-58.

Todd, Jessica Erin, Paul C. Winters, & Tom Hertz. 2010. Conditional Cash Transfers and Agricultural Production: Lessons from the Oportunidades Experience in Mexico. *The Journal of Development Studies*, 46:1, 39-67, DOI: 10.1080/00220380903197945

Vinod, Hrishikesh D. 2005. Evaluation of Archived Code with Perturbation Checks and Alternatives. Mimeo.

Winters, Paul C. 2013. Feedback during Dissertation Proposal Defense. Washington, DC, American University, April 19th, 2013.

Winters, Paul C., Alessandro Maffioli, and Lina Salazar. 2011. Evaluating the Impact of Agricultural Projects in Developing Countries: Introduction to the Special Feature. *Journal of Agricultural Economics* 62, 393-402.

Winters, Paul C., and Benjamin Davis. 2007. Designing a new PROCAMPO program: Lessons from Oportunidades. Report presented to the Inter-American Development Bank for the project Mexico: Estudios sobre politicas y gastos publico federal para el sector rural. Final Version.

Winters, Paul C., Lina Salazar, and Alessandro Maffioli. 2010. Designing impact evaluations for agricultural projects, SPD Working Papers 1007, Inter-American Development Bank, Office of Strategic Planning and Development Effectiveness.

Wollni, Meike, and Bernhard Brümmer. 2012. Productive efficiency of specialty and conventional coffee farmers in Costa Rica: Accounting for technological heterogeneity and self-selection. *Food Policy* 37: 67–76.

Wooldridge, J.M. 2002. *Econometric analysis of cross-section and panel data*. Cambridge, MA, MIT Press.

World Bank, 2006. Impact evaluation: the experience of the Independent Evaluation Group of the World Bank. Independent Evaluation Group, World Bank.

World Bank. 2007. Zambia Smallholder Agricultural Commercialization Strategy. Sustainable Development, AFTS1 Country Department 2, Zambia, Africa Region. Report No. 36573-ZM. Washington DC: WB.

World Health Organization (WHO). 2013. World Health Statistics 2013.

Yamauchi, C. 2010. Community-based targeting and initial local conditions: Evidence from Indonesia's IDT Program. *Economic Development and Cultural Change*, 59(1), 95–147.

Zeller et. al. 2004. Review of Poverty Assessment Tools. USAID: Accelerated Microenterprise Advancement Project.

Zeller, M., Sharma, M., Henry, C., and Lapenu, C. 2006. An Operational Method for Assessing the Poverty Outreach Performance of Development Policies and Projects: Results of Case Studies in Africa, Asia, and Latin American. *World Development* Vol. 34, No. 3. Pp 446—464.

Zezza, Alberto, Bénédicte de la Brière, and Benjamin Davis. 2010. The impact of social cash transfers on household economic decision making and development in Eastern and Southern Africa. *Draft October 28, 2010. Not for citation without permission*.

Zhengfei, G., Lansink, A. O., Wossink, A. and Huirne, R. 2005. Damage control inputs: A comparison of conventional and organic farming systems, *European Review of Agricultural Economics*, Vol. 32, (2005) pp. 167–189.