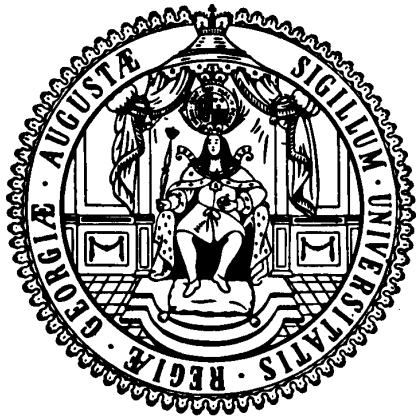
Ibero-Amerika Institut für Wirtschaftsforschung
Instituto Ibero-Americano de Investigaciones Económicas
Ibero-America Institute for Economic Research
(IAI)

Georg-August-Universität Göttingen
(founded in 1737)



Diskussionsbeiträge · Documentos de Trabajo · Discussion Papers

Nr. 231

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

Stephan Klasen, Simon Lange

March 2015

Platz der Göttinger Sieben 3 · 37073 Goettingen · Germany · Phone: +49-(0)551-398172 · Fax: +49-(0)551-398173
e-mail: uwia@gwdg.de · http://www.iai.wiwi.uni-goettingen.de

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

Stephan Klasen† Simon Lange‡

March 17, 2015

Abstract

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

Keywords: targeting; transfers; social assistance; proxy means tests; poverty; ROC-analysis; Latin America; Bolivia.

JEL Classification Numbers: C52; I38; O21.

*This work was supported by the German Federal Ministry for Economic Cooperation and Development in cooperation with the Kreditanstalt für Wiederaufbau (KfW). We are grateful to Julia Johannsen, Fernando Landa, Roland Pardo, Ramona Rischke, and Cecilia Vidal for helpful comments on earlier versions of this paper.

†Economics Department, University of Göttingen.

‡Corresponding author. Economics Department, University of Göttingen. Address: Platz der Göttinger Sieben 3, 37073 Göttingen, Germany, telephone: +49–551–39–10940 e-mail: *simon.lange@wiwi.uni-goettingen.de.*

1 Introduction

The idea of targeting social transfer programs to those in need is intuitively appealing—in the words of Amartya Sen: "the more accurate a subsidy in fact is in reaching the poor, the less the wastage, and the less it costs to achieve the desired objective. It is a matter of cost-effectiveness in securing a particular benefit. [...] it is one of maximizing the poverty-removal benefits accruing from a given burden of cost" (Sen, 1995).[1] If the argument in favor of targeting takes precedent over concerns,[2] the question is how to identify the *targets* in this exercise, i.e. how to identify households that have insufficient *means*. While direct means tests using verifiable data on incomes (e.g. from tax records) are usually viable in developed countries, identifying the needy is much more challenging in developing countries where a larger share of total income is generated in the urban informal sector and where people in rural areas tend to rely on subsistence agriculture for a living. Verifiable data on incomes or other direct measures of economic welfare are thus often lacking for the majority of the population.

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

A further alternative is to consider a number of readily observable and verifiable household characteristics to construct a proxy means test (PMT).[3] PMTs employ a limited number of variables in order to predict the welfare-level of households.[4] A statistical model, usually a linear model, is first calibrated with data from a representative household expenditure survey. A shortened questionnaire is then administered to a much larger set of households. In conjunction with the statistical model, data from the shortened questionnaire are finally used to derive scores

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

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

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

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

for the identification of beneficiaries. Whether this strategy is suitable for the identification of the deserving poor is largely an empirical question that depends on the underlying model, the quality of the available data, the joint distribution of the target measure and the proxies, and—as will be demonstrated in the present paper—the proportions of the population in poverty and targeted.

We analyze the potential for PMTs based on two waves of Bolivian household data, the *Encuesta de Hogares* 2008 and 2011. Based on *Receiver Operating Characteristics* (ROC)-analysis, a statistical tool to assess the accuracy of classifiers, we compare regression-based PMTs that differ in terms of complexity.[5] The paper thus contributes to the on-going debate about the usefulness of PMTs (e.g. Grosh and Baker, 1995; Kidd and Wylde, 2011; Alatas et al., 2012). However, in contrast to these studies, the analysis of ROC-curves allows us to illustrate targeting outcomes over the entire range of options that are available to policy-makers, i.e. the typical trade-off in terms of under-coverage of the poor on the one hand and leakage of benefits to the non-poor on the other.[6] We then take the analysis one step further by going from targeting measures such as error rates—often the only criterion to evaluate the accuracy of targeting mechanisms—to simulations of poverty effects of uniform transfers, the most common form of cash transfer in developing countries.[7,8]

Our findings suggest that PMTs are reasonably accurate in the Bolivian setting when a large fraction of the population is considered poor and an equally large share is targeted. For instance, our estimates suggest that only 52–60 percent of the poor are covered when ten percent are poor and ten percent are targeted compared to 70–78 percent when 50 percent are poor and an equal share is targeted. While this is very much in line with the literature,[9] we also find that the proportion of poor covered is responsive to an increase in the inclusion rate in the latter scenario: when 70 percent are targeted—despite only 50 percent being poor—around 90 percent of the poor are covered. In other words, leakage in this case is confined mostly to the poor close to the poverty line. We illustrate this point further based on non-parametric regression methods.

Our main finding, however, relates to the discrepancy between targeting accuracy and poverty impacts: our simulation study suggests that significant poverty effects are only attained with very narrow targeting. This is true for both the poverty gap and the squared poverty gap and for different initial poverty headcounts. There is thus an important trade-off between targeting

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

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

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

[8] Recent research suggests that the empirical relationship between poverty effects and targeting accuracy of transfer schemes is weak. See, in particular, Ravallion (2007).

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

accuracy, in particular, coverage of the poor, and poverty effects of transfer schemes that has received very little attention in the literature.

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

The remainder of the paper is organized as follows: in section 2, we review different targeting strategies and compare them to PMT-based targeting. Section 3 describes our datasets and our indicator of economic welfare. Section 4 introduces the proxy sets on which the remaining analysis is based and reports results from calibrating models. Section 5 introduces ROC-analysis and reports results from out-of-sample evaluations of PMTs in terms of targeting accuracy and section 6 reports results from poverty simulations. Final remarks are offered in section 7.

2 Targeting in developing countries

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

PMTs are a more sophisticated means testing device that relies on objective data rather

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

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

How do PMTs compare to universalistic transfers and alternative targeting strategies? Some targeting will usually result in a larger poverty effect as more resources will be available for those in need. See Akerlof (1978) for this argument and Grosh and Baker (1995) for evidence from simulations. In comparison with universalistic programs for which no data are required, PMTs are frequently found to bring about greater poverty effects (Grosh and Baker, 1995). This is also very much in line with what we find in our poverty simulations. More importantly, PMTs are often found to perform better in terms of targeting outcomes than alternative methods that are sometimes easier to administer (e.g. geographic targeting) (Coady et al., 2004). Finally, an additional advantage of PMTs as a targeting device is their cost-effectiveness. Administering detailed household surveys to all applicants will in most cases not be an option, especially when what is required are high-quality data on household expenditure rather than data on incomes.

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

There are at least two widely-used targeting strategies that try to circumvent the measurement problem. The first are community-based targeting schemes that explicitly aim to incorporate local knowledge by delegating the selection of beneficiaries to locals. While tapping into local knowledge may have advantages, this strategy is often criticized on the ground that it is prone to local capture (Conning and Kevane, 2002; Bardhan and Mookherjee, 2005). Alternatively, self-targeted programs such as public works programs circumvent the problem by adding a

[11] See neda (2005) for details on the SISBEN I-program.

requirement that results in disutility.[12] Presumably, they can be designed in such a way that only the truly needy will apply. See Besley and Coate (1992b) for the incentive argument and Drèze and Sen (1989) for evidence of the impact of such self-targeting in the case of India. However, Alatas et al. (2013) argue that such ordeal mechanisms may actually have ambiguous effects on targeting.

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

While our results confirm this broad pattern, our findings with regard to accuracy are somewhat more optimistic. For instance, we find that about 42–50 percent of all beneficiaries are erroneously included in the first scenario (ten percent poor and ten percent targeted). We also find that accuracy improves substantially when both the poverty headcount and the proportion targeted are higher. For instance, if the poverty headcount is 50 percent and 50 percent are targeted based on PMTs, less than 30 percent included are included erroneously.[13]

Our analysis is also broader in scope as we consider scenarios in which the proportion poor differs from the proportion targeted. In particular, we consider scenarios in which one deliberately allows for some leakage to the non-poor and investigate who benefits from leakage in section 5.3. Our findings suggest that if, for instance, at a headcount of 50 percent, 80 percent are targeted, we actually cover more than 90 percent of the poor. Hence, PMTs seem to be an attractive targeting device if the aim is not to identify the poorest of the poor, but to weed out the wealthiest households.

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

We certainly agree with Kidd and Wylde's observation that PMTs fail to accurately predict welfare of the poorest and the richest. Predictions from OLS-based regressions will always have a smaller variance than the outcome variable. However, this is irrelevant if the poverty rate is known and information on all potential beneficiaries is available. Note that the actual values of

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

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

scores are then no longer relevant; only the *ordering* among households induced by scores will matter.

3 Datasets and welfare indicator

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

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

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

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

4 Proxy Means Tests

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

We employ Ordinary Least Squares (OLS)-regressions of log real expenditure against proxies in order to evaluate the usefulness of PMTs in the present data. In order to mimic a real-life PMT exercise and to avoid 'overfitting' the data, all investigations are carried out using out-of-sample

predictions. In particular, we test the predictive power of models trained using (i) the 2008 data and (ii) a subsample of the 2011 data as *calibration samples*. The 2008 survey provides data on 3,937 households that have information on all relevant variables and household expenditure surveys of this size are frequent in Bolivia. We will therefore randomly set aside a subsample of comparable size—3,932 households—from the 2011 data, from which we retain 8,842 household records in total. The remaining 4,910 2011 records are then employed to analyze the performance of PMTs, that is, they serve as the *validation sample*.[14,15]

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

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

4.2 Proxy sets

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

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

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

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

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

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

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

4.3 Results

Table 1 reports results from the above regression models based on the 2008 and 2011 calibration samples, respectively. However, since individual coefficients are hardly decisive in this exercise, we report only indicators of the regression fit, namely the number of parameters, R^2-statistics from regressions based on calibration samples, and, for comparison, the squared Pearson correlation coefficient between predicted values (or *PMT scores*, in the terminology of this method) and actual values in the validation sample, i.e. the 2011 households not included in the calibration

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

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

Table 1: Regression results for log per adult expenditure based on 2011 and 2008 calibration samples (3,932 and 3,937 observations, respectively).

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

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

sample. All quantities are estimated using individual inflation factors as weights.

Several observations can be made from table 1: as can be seen from the R^2-statistics, the regression fit with the 2008 calibration sample is much better than the fit obtained with the 2011 calibration sample for M1, the most parsimonious model. The R^2 is 0.44 for the model fitted to 2008 data compared to 0.36 for the model fitted to 2011 data. The difference is less pronounced for more sophisticated models but still noticeable. This implies that households' geographic and demographic characteristics were better predictors in 2008 as compared to 2011, a somewhat surprising finding.

As one would expect, the squared correlation coefficient, denoted ρ^2 in table 1, is lower than the R^2 when the 2008 data are used to obtain PMT scores.[19] This is particularly true for PMT scores obtained from the models calibrated on 2008 data. There are virtually no differences in the case of scores obtained from models calibrated on 2011 data.

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

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

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

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

While one could likely find proxies for such shocks in the dataset at hand, it seems clear that these would be hard to verify. It therefore seems questionable whether the explanatory power of these models could be improved substantially by finding and adding more proxies if proxies are to be verifiable.

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

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

5 Assessing accuracy

5.1 ROC-analysis

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

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

Table 2: Classification matrix: taxonomy of targeting errors

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

Consider table 2, a *confusion matrix*: each cell contains information about the number of households in a specific group. For instance, n_1 is the number of poor that are classified as beneficiaries and n_3 is the number of non-poor classified as beneficiaries. Granted we have data as in table 2, we can define the *true positive rate* (TPR), the proportion of poor (or *positives*) identified as beneficiaries, and the *false positive rate* (FPR), the proportion of non-poor identified as beneficiaries as

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

respectively.

TPRs and FPRs depend on the proportion admitted to the program which may vary from zero to one. Applied to targeting of poverty-alleviation programs, ROC-curves are plots of the TPR against the FPR for variations in the cut-off used to identify the poor in the space of PMT scores. Two closely related concepts in the targeting literature are *under-coverage*, the proportion of poor not targeted, and *leakage*, the proportion of non-poor erroneously receiving benefits (Grosh and Baker, 1995).

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

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

While a classifier is strictly superior whenever it produces greater TPRs for all FPRs, some of the curves may intersect. In other words, one model may perform better at low levels of the FPR, but worse at higher levels. At the same time, both models may generate ROC-curves that yield (almost) identical (partial) AUCs. Wodon (1997) shows that one can still find a criterion

that allows selection of one model as the superior model as long as society's (or a policy-maker's) preferences can be represented by a strictly quasi-concave utility function that takes the TPR and the FPR as arguments. However, we find that PMTs that involve more data are superior with only minor exceptions.

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

5.2 Results

The ROC-curves generated by our PMTs are plotted in figure 1. As described above, ROC-curves are generated by varying the proportion of beneficiaries based on predicted real expenditure and comparing classification matrices. We define as poor successively ten, 25, and 50 percent of the population, where the latter two figures correspond roughly to the incidence of extreme and moderate poverty in Bolivia at the time. Recall that we employ individual-level frequency weights, defined as household expansion factors multiplied with the number of household members, in order to obtain results representative of the entire Bolivian population in 2011.

Part of the information conveyed in figure 1 is tabulated in table 3 where we report TPRs at different inclusion rates for each PMT as well as AUCs for poverty rates of ten (panel A), 25 (B), and 50 percent (C), respectively. For instance, the first entry in panel A suggests that if ten percent are poor and ten percent are targeted, 52.6 percent of those receiving transfers will actually be poor. Note that we do not have to report in addition FPRs as they are functions of TPRs, the poverty rate, and the inclusion rate.[20] Therefore, from a targeting accuracy-perspective and conditional on inclusion and poverty rates, PMTs with higher TPRs are strictly preferred.

Reading the table from left to right we can infer the effect of increasing the complexity of PMTs for a given poverty headcount and beneficiary proportion. For instance, 57.1 percent of those targeted are poor under M2 and close to 60 percent under M3 and M4. However, the differences are small, probably smaller that would be suggested by differences in R^2-statistics reported in section 4.3. Given the caveats of using more complex models such as higher costs in generating the data and, possibly, higher susceptibility to manipulation and deceit (e.g. in the case of assets), it seems likely that policy-makers would want to opt for parsimonious PMTs.

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

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

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

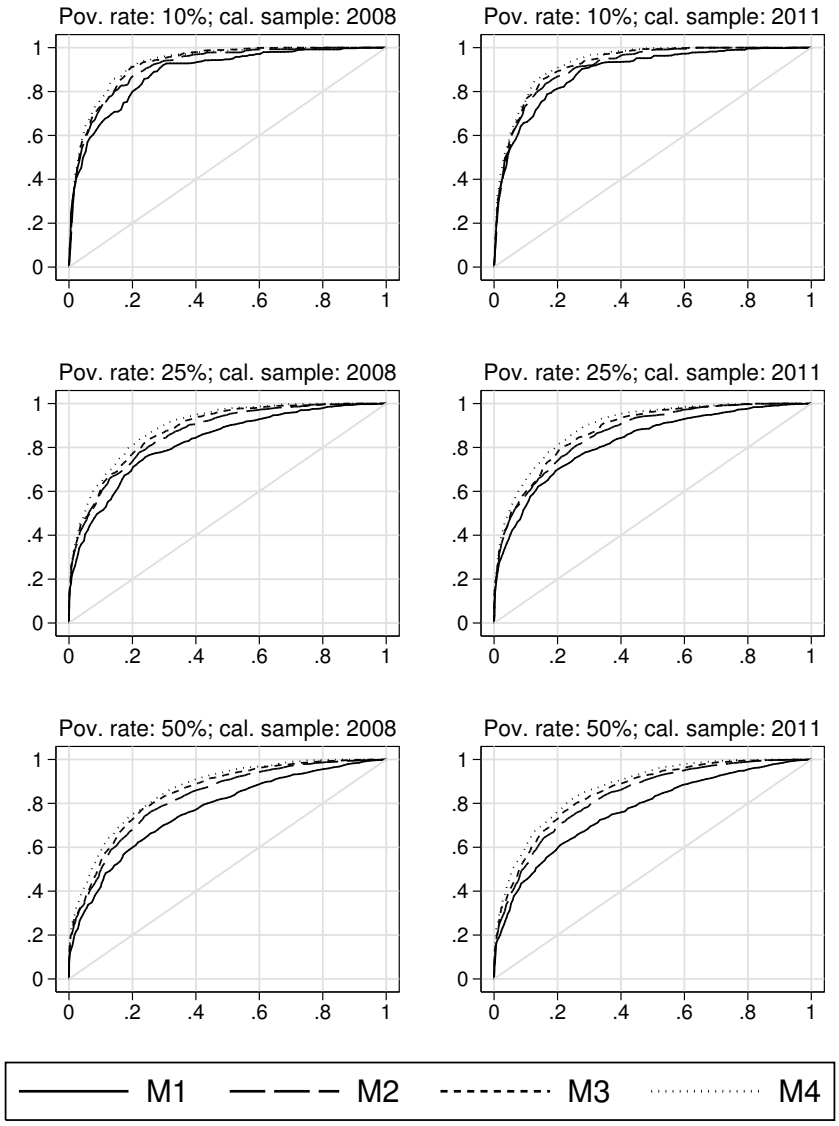


Figure 1: ROC-curves for PMTs M1–M4 based on 2008 and 2011 calibration sample.

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

We also find that if the proportion receiving benefits is equal to the proportion in poverty, PMTs produce the highest TPRs when half the population is poor. Under M1, for instance, the TPR is only 54.1 percent when ten percent are poor and the same number is marked for

Table 3: True positive rates at varying levels of inclusion rates and AUCs.

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

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

inclusion. However, when half the population is poor and half receive transfers, almost 70 percent of beneficiaries will be poor. We conclude that PMTs are less accurate in identifying the poor when only a small percentage of the population is poor and an equally small share is targeted. The appropriateness of PMTs when the goal is to reach the poorest with a very limited program is thus questionable.

This reaffirms findings from the literature, both from simulations and real world-experience. For instance, in their simulations based on data from Bangladesh, Rwanda, Sri Lanka, and Indonesia, Kidd and Wylde (2011) find that TPRs between 29 and 43 percent when ten percent are poor and ten percent are targeted. Our results suggest that targeting at this level would

be more accurate in Bolivia. For Mexico's *Opportunidades* and Brazil's *Bolsa Famĺia* programs which target around 25 and six percent of the population, respectively, Veras Soares et al. (2010) report TPRs of 20 and 41 percent, respectively.[21] Even though our results look somewhat better, it is not clear whether one would want to accept that only about half of the poorest ten percent are covered by the program.

PMTs perform much better when a larger portion of the population is considered poor and the program channels transfers to an equally larger share of beneficiaries. When half of the population is poor and half is targeted, TPRs range from 69 to 78 percent. Between 80 and 95 percent of the poor would be included if 80 percent of the population would be targeted. Thus, if leakage is not a major concern, broad targeting can be achieved with reasonably high TPRs with these models.

5.3 Assessing leakage

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

In our analysis, where will we find those erroneously identified as poor? Much can already be inferred from table 3 where we find that true positive rates increase when the poverty rate is increased together with the percentage targeted. This implies that leakage tends to benefit the 'poor among the non-poor.' This is further illustrated in figure 2, where we investigate the distribution of benefits when 25 and 50 are targeted. Again, this is done for our out-of-sample scores in order to mimic the real world situation where models are fitted with one dataset and the PMT is then applied to fresh data.

Figure 2 is generated by first defining dummy variables identifying beneficiaries for all PMTs, i.e. the poorest 25 and 50 percent based on PMT scores. These indicators are then regressed against percentiles of the log real expenditure distribution. We use *kernel-weighted local polynomial regression*, a non-parametric method that allows us to recover the non-linear relationship between the beneficiary dummy and real expenditure percentiles.[22] The resulting regression lines can be interpreted as the probability of receiving the transfer and, by the law of large numbers, as the proportion of beneficiaries *conditional on real expenditure*. The black lines depict the result from using the four PMTs with a targeting cut-off of 25 percent and the blue lines the

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

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

results for a cut-off of 50 percent.

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

We know from table 3 that when 50 percent of the population is targeted, more than 90 percent of the individuals in the bottom decile of real expenditure will be covered. In line with these high TPRs reported in the table, we find that even with the most parsimonious model, the probability of receiving the transfer is greater than 90 percent everywhere below a poverty line that renders the bottom ten percent poor. At the same time, leakage will be considerable as more than four out of five households receiving the transfer will not be poor. The question remains, however, where in the distribution of real expenditures these 'false positives' will be located. We find that the probability of receiving the transfer drops to about 50-55 percent at the median of the distribution and to about ten percent by the time we get to the richest ten percent with all PMTs except M1 for which it is still about 20 percent. With 25 percent of the population admitted to the program, we find that the probability of receiving the transfer is only between 70 and 80 percent for an individual at the tenth percentile. The probability drops to only 20 percent for an individual at the median of the distribution and to less than ten percent for an individual at the 90th percentile the top decile.

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

6 Poverty simulations

6.1 Set-up

This final section offers simple simulations that illustrate the potential poverty effects of PMT-based monetary transfers. Ravallion (2007) shows that programs can be well-targeted and still have little effect on poverty. It is therefore paramount to also simulate poverty effects of transfers based on the PMT instruments discussed above.

Our set-up is simple: we assume a fixed budget b for a transfer scheme is available. We consider a total budget that amounts to one percent of total household expenditure or 97bs. million per month. We use household inflation factors in order to arrive at total expenditure

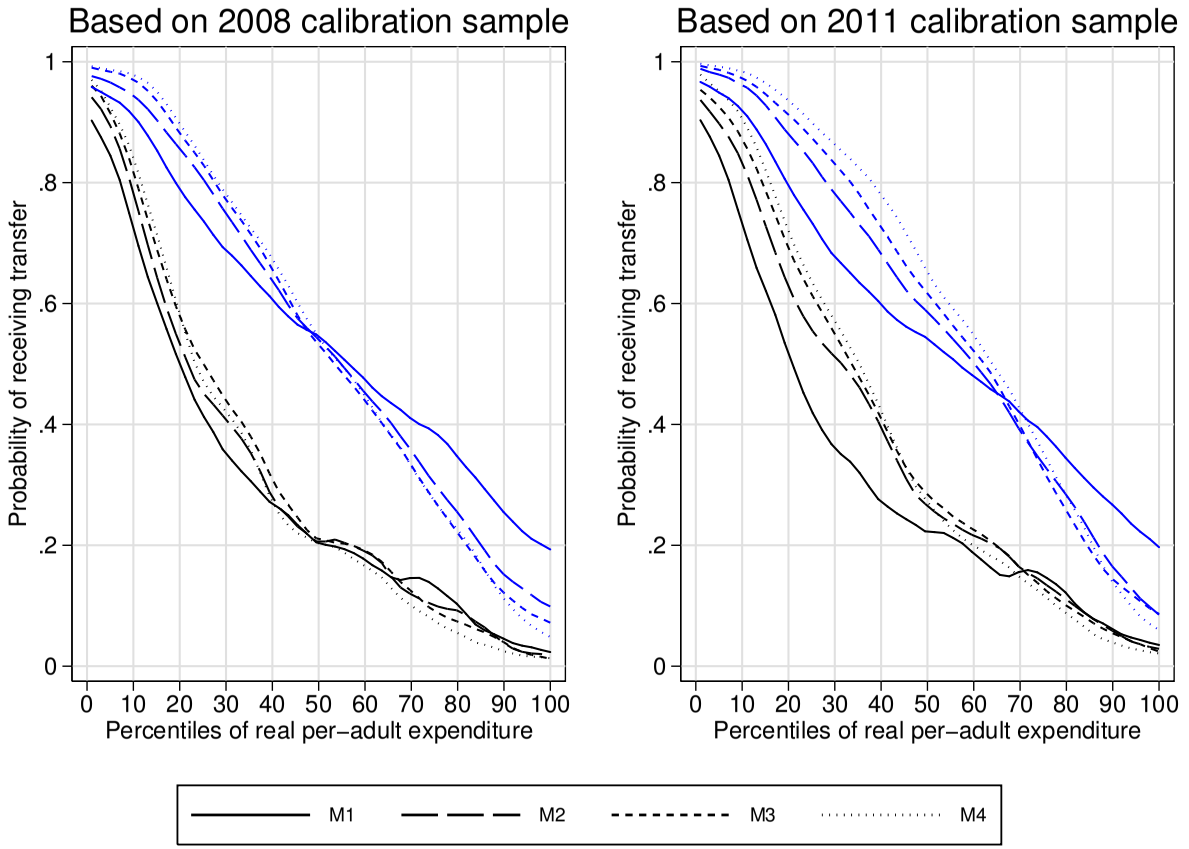


Figure 2: Local polynomial regression of beneficiary status against percentile of per-adult real expenditure. Based on PMTs trained on the 2008 (left panel) and 2011 calibration samples (right panel) and evaluated with the validation sample. The solid, long-dashed, short-dashed, and dotted regression lines indicate models M1–M4; the black and blue regression lines correspond to targeting the poorest 25 and 50 percent of the population, respectively, when an equal share is poor.

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

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

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

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

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

6.2 Results

The results of this simulation for an assumed pre-transfer headcount of ten, 25 and 50 percent (from left to right) are depicted in figure 3, where we plot relative changes in post-transfer poverty measures (headcount, poverty gap, and squared poverty gap from top to bottom) against the share of the population targeted. The dashed gray lines depict outcomes under perfect targeting, i.e. targeting based on actual real expenditure. In order to avoid over-populating graphs, we plot only results for PMTs trained on the 2011 calibration sample. As one would expect given our findings above, corresponding plots for PMTs trained on 2008 data are very similar. Part of the information conveyed in figure 3 is also tabulated in table 4, where we focus on the maximum poverty effect attained under ideal targeting as well as PMTs M1 and M4 and the inclusion rate at which this maximum occurs.[24]

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

The first key result from this exercise is that all four PMTs perform very similar in terms of the poverty impacts they bring about. This becomes clear from the data in table 4: the difference in poverty impacts between the two PMTs is barely greater than one percentage point. Only when ten percent of the population are poor and the focus is on the distribution-sensitive FGT2 measure is the difference more pronounced. Remember that we saw some differences in terms of targeting performance between these two PMTs and that they differ considerably in the number of proxies they take into account. Despite these differences, very similar poverty impacts can be had.

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

The results for the poverty headcount are more complicated. It is worth noting first that the headcount ratio is less appropriate for poverty comparisons in this exercise as it violates Sen's (1976) *monotonicity axiom*: all else equal, a reduction in welfare of a person below the poverty line must decrease poverty. For pre-transfer headcounts of 25 and 50 percent, we find that the poverty headcount is initially reduced substantially in the case of perfect targeting but

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

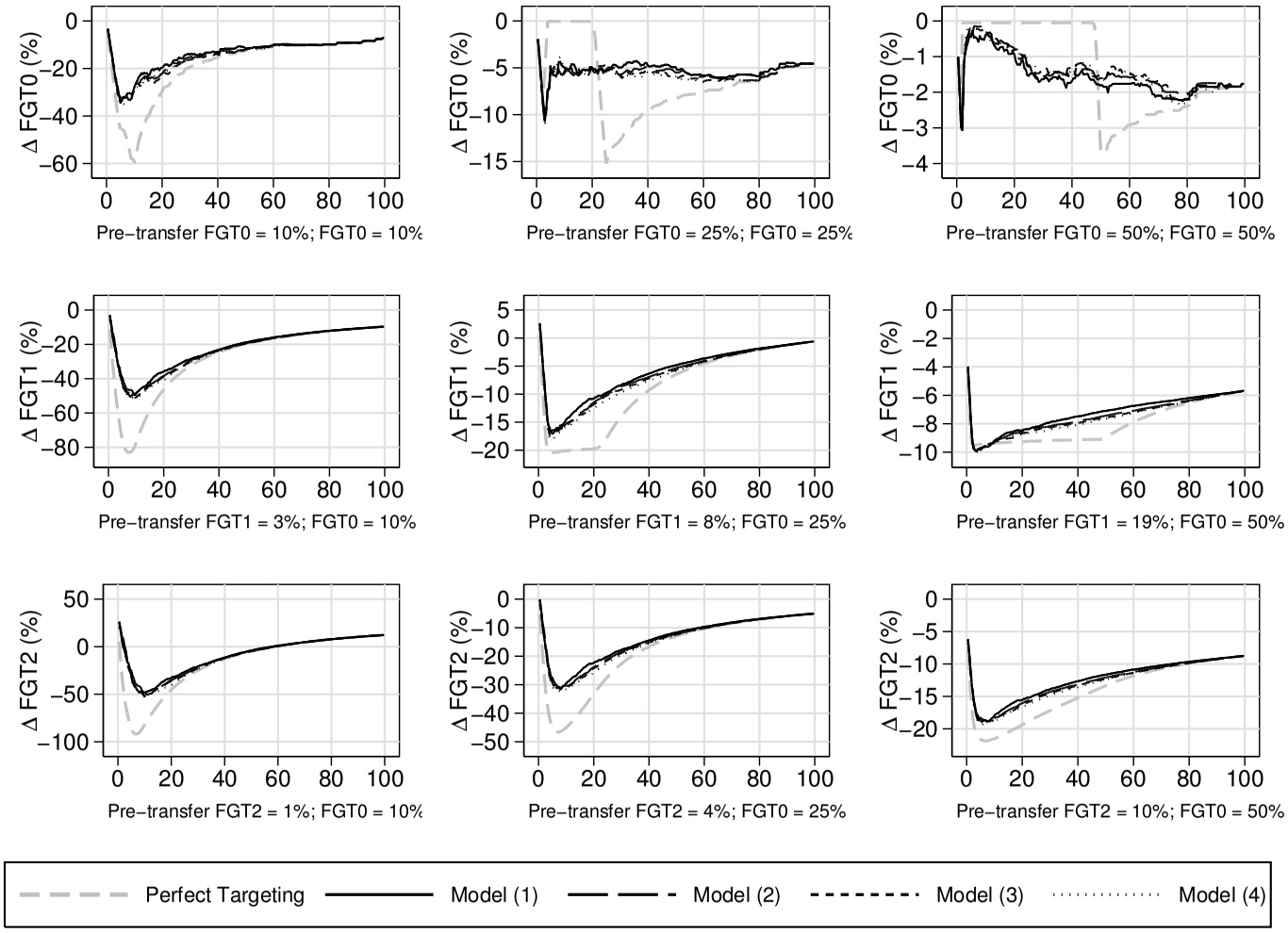


Figure 3: Simulated poverty impacts based on the FGT-class of poverty measures.

Table 4: Selected results from poverty simulations: ideal targeting, PMTs M1 and M2 based on 2011 calibration sample.

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

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

subsequently increases as targeting becomes less narrow. The initial transfer is large enough to lift the poorest above the poverty line. However, as the number of targeted individuals increases, the resources available *per individual* decrease rapidly and quickly become insufficient to lift the poorest households out of poverty. Despite the fact that the poorest households are better of, the headcount remains unaltered over a considerable range. It only drops sharply as the population proportion targeted approaches the headcount. This is where the maximum poverty effect occurs (see table 4). As a result, PMTs bring about a greater effect on poverty over a considerable range in the case of headcounts of 25 and 50 percent. This does not occur with higher-order FGT measures.[25]

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

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

transfers are 'wasted' on non-poor beneficiaries.

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

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

7 Conclusion

Based on Bolivian household survey-data, this study stipulates and evaluates a series of regression-based PMTs in terms of targeting accuracy and potential poverty effects. In line with similar studies, we find that PMT-based targeting results in fairly large margins of errors when the poverty headcount is small and only a small share of the population is targeted. On the other hand, targeting accuracy improves considerably when both the headcount and the targeting share increases. Naturally, leakage also increases but much of the non-poor receiving the transfer will still be close to the poverty line. Finally, a large share of the poor are covered even at high levels of poverty when only the rich are not targeted. Our findings further suggest that targeting using a lengthy list of proxies often does not increase accuracy over more parsimonious models, particularly when the entire population is considered. Interestingly, we find that PMTs are fairly robust over the time period of three years, i.e. PMTs trained based on somewhat dated datasets perform very similar in terms of accuracy and poverty effects. While one would conclude from this that PMTs should rather be used for broad targeting, our poverty simulations suggest that poverty effects are maximized when a small fraction—about five to ten percent—of the population is targeted. PMT-based targeting achieves a poverty effect of about three-fourths of what a perfectly targeted program would achieve at this level of targeting. Again, differences across PMTs are small.

While our study shows that the effect of PMT-based schemes on poverty should be evaluated in addition to traditional targeting measures such as total error rates and true and false positive rates, results from our poverty simulations should be interpreted cautiously. It is important to keep in mind that we do not consider any behavioral responses and that the budget available for the poverty-alleviation program likely depends on the population share targeted. Political-

economy considerations would suggest that schemes that target a larger fraction of the population command more political support and thus dispose of larger budgets. If these considerations were taken into account, it would seem reasonable to assume that the trade-off between targeting accuracy and poverty impact would be somewhat less severe and that the maximum poverty effect is reached at higher levels of targeting.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Drèze, J. and A. Sen (1989). *Hunger and Public Action*. Oxford: Oxford University Press.

Drèze, J. and P. Srinivasan (1997). Widowhood and Poverty in Rural India: Some Inferences from Households Survey Data. *Journal of Development Economics 54*, 217–234.

Dutta, P., R. Murgai, M. Ravallion, and D. van de Walle (2012). Does India's Employment Guarantee Scheme Guarantee Employment? *World Bank Policy Research Working Paper 6003*.

ECLAC (2006). *Compendium on Best Practices in Poverty Measurement*. Expert Group on Poverty Statistics (Rio Group), United Nations Economic Commission for Latin America and the Caribbean (ECLAC).

ECLAC (2010, September). Economic Survey of Latin America and the Caribbean 2009-2010. Technical report, Economic Commission for Latin America and the Caribbean.

Foster, J., J. Greer, and E. Thorbecke (1984). A Class of Decomposable Poverty Measures. *Econometrica 52*(3), 761–766.

Galasso, E. and M. Ravallion (2005). Decentralized Targeting of an Antipoverty Program. *Journal of Public Econhomics 89*, 705–727.

Gelbach, J. and L. Pritchett (2000). Indicator Targeting in a Political Economy: Leakier can be Better. *Journal of Policy Reform 85*, 705–727.

Gelbach, J. and L. Pritchett (2002). Is More for the Poor Less for the Poor? The Politics of Means-Tested Targeting. *The B.E. Journal of Economic Analysis & Policy 2*(1).

Grosh, M. E. and J. L. Baker (1995). Proxy Means Tests for Targeting Social Programs: Simulations and Speculation. *LSMS Working Paper No. 118*.

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

Johannsen, J. (2008). *Operational Assessment of Monetary Poverty by Proxy Means Tests: The Example of Peru*. Number 65 in Development Economics and Policy. Peter Lang Verlag.

Kidd, S. and E. Wylde (2011). Targeting the Poorest: An Assessment of the Proxy Means Test Methodology. Published by the Australian Agency for International Development (AUSAid).

Landau, K., S. Klasen, and W. Zucchini (2012). Measuring Vulnerability to Poverty Using Long-Term Panel Data. *Courant Research Centre Discussion Paper No. 18, Göttingen University*.

Lindert, K., E. Skoufias, and J. Shapiro (2006). Measuring Vulnerability to Poverty Using Long-Term Panel Data. *World Bank Social Protection Discussian Paper No. 0605*.

Moene, K. O. and M. Wallerstein (2001). Targeting and the Political Support for Welfare Spending. *Economics of Governance 2*(1), 3–24.

Moffitt, R. (1983). An Economic Model of Welfare Stigma. *American Economic Review 73*(5), 1023–1035.

Moratti, M. (2010). Consumption Poverty and Pro-Poor Growth in Bolivia (1999–2007). *University of Sussex Economics Department Working Paper Series No. 13-2010*.

Morris, S. S., P. Olinto, R. Flores, E. Nils, and A. C. Figueiro (2004). Conditional Cash Transfers are Associated with a Small Reduction in the Rate of Weight Gain of Preschool Children in Northeast Brazil. *Journal of Nutrition 134*(9), 2336–2341.

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

Ravallion, M. (1998). Poverty Lines in Theory and Practice. *LSMS Working Paper No. 133*.

Ravallion, M. (2007). How Relevant is Targeting to the Success of an Antipoverty Program? *World Bank Policy Research Paper No. 4385*.

Ravallion, M. (2008). Miss-targeted or Miss-measured? *Economics Letters 100*(1), 9–12.

Schady, N. R. (2002). Picking the Poor: Indicators for Geographic Targeting in Peru. *Review of Income and Wealth 48*(3), 417–433.

Sen, A. (1976). Poverty: An Ordinal Approach to Measurement. *Econometrica 844*(2), 219–231.

Sen, A. (1995). The Political Economy of Targeting. In D. van de Walle and K. Nead (Eds.), *Public Spending and the Poor: Theory and Evidence*, pp. 11–24. John Hopkins University Press.

Skoufias, E., B. Davis, and S. de la Vega (2001). Targeting the Poor in Mexico: An Evaluation of The Selection of Household for PROGRESA. *IFPRI Food Consumption and Nutrition Division Discussion Paper No. 103*.

Thompson, M. L. and W. Zucchini (1989). On the Statistical Analysis of ROC Curves. *Statistics in Medicine 8*(10), 1277–1290.

Veras Soares, F., R. Perez Ribas, and R. Guerreiro Osório (2010). Evaluating the Impact of Brazil's Bolsa Família: Cash Transfer Programmes in Comparative Perspective. *Latin American Research Review 45*(2), 173–190.

Wodon, Q. (1997). Targeting the Poor Using ROC Curves. *World Development 25*(12), 2083–2092.

A Calculation of the welfare indicator

A.1 General considerations

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

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

Rents are actual rents payed or, in case the household owns the dwelling, hypothetical rents reported by the household. This may be problematic as home-owners may have no knowledge of the rental market. While potentially self-defeating, we have no choice but to rely on inputing values based on hedonic price regressions when information on both actual and hypothetical rents are missing. Rents, however, only account for a small share of total expenditure (see table 5).

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

Economic shocks such as unexpected hospital fees are a potential source of measurement error: such unexpected outlays will increase total expenditure even though necessary one-off spending

on hospitalization may have displaced essential expenditure on, say, food. We therefore exclude reported expenditure that is related to some adverse shock, in particular, hospitalization fees. Moreover, including repairs and enhancement of durables (e.g. repairs of a car or furniture) may lead to double counting when the increase in value resulting from the enhancement is already captured in the current value reported by households. We therefore exclude expenditure reported on car repairs. Furniture, another category that would be available, is excluded on the ground that it is likely included in the durables listings.

A.2 Equivalence scale

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

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

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

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

Deaton (1997, p. 205) notes that "the state of knowledge and agreement in the area is not such as to allow incontrovertible conclusions or recommendations." Theory-based methods to estimate equivalence scales are not yet fully convincing. The recommendation is thus to rely on arbitrary but reasonable equivalence scales. On the other hand, Drèze and Srinivasan (1997) show that if household expenditure can be divided between purely private and purely public components so as to maximize average utility among identical members, θ is equal to the share of private goods in total household expenditure. Expenditure shares are reported in table 5 below. Food is usually considered a purely private good and its expenditure share is about 60

Table 5: Distribution of expenditure and expenditure patterns.

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

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

percent. This can be considered a lower bound for θ since other goods and services are mostly neither fully private nor fully public. Our choice of $\theta = 0.8$ is thus sensible.

A.3 Differences in prices

Using consumption expenditure properly adjusted for household size and composition as an indicator of economic welfare will be misleading if prices for relevant goods and services vary across households. Rather, the welfare indicator should reflect the command over commodities of a given individual. A way forward is to use a *true cost-of-living index* that reflects the price of a reference bundle of commodities as a deflator (Ravallion, 1998). Here, we use the official series of consumer price indices (CPIs) for each of the nine departments. The data were provided by UDAPE-staff. We proceed by first normalizing CPIs to equal unity in the department of La Paz in 2011. We then deflate expenditure per adult equivalent by the normalized CPIs.

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

as

$$y_{rti} = \frac{x_i}{s_i p^*_{tr}}, \tag{A.1}$$

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

A.4 Expenditure patterns

Table 5 reports means of per capita expenditure, real per capita expenditure, and real per adult expenditure per month in Bolivia by quintiles of the later and by location for 2008 and 2011. For 2011 (2008) we find that average real expenditure *per adult* is 1,258bs (1,197bs) while real *per capita* expenditure is 897bs (852bs). This is roughly comparable to results reported by Moratti (2010).

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