Program Evaluation
Using Multidimensional Poverty Measures:
Evidence from TUP[1]

T M Tonmoy Islam
Center for Poverty Research
University of Kentucky
tmisla2@uky.edu

James P Ziliak
Center for Poverty Research
University of Kentucky
jziliak@uky.edu

April 2010

Abstract:
Anti-poverty programs implemented in recent years in developing countries typically have multiple goals beyond improving income levels. Assessing the effectiveness of these interventions has gained increased currency in light of the UN's Millennium Development Goals. We introduce multidimensional poverty measures to the program evaluation literature in order to provide a comprehensive measure of program effectiveness in the presence of multiple outcomes. We combine robust multidimensional poverty measures with difference-in-difference matching estimators to evaluate the effectiveness of the Targeting the Ultra-Poor (TUP) program undertaken in Bangladesh that aimed to improve the well being of families living in extreme poverty across a number of dimensions. We show that the TUP program reduced multidimensional poverty among the treated group by 18-21 percentage points relative to the comparison group, or nearly 25 percent relative to the baseline poverty rate.

[1] We would like to thank Susan Davis of BRAC USA and Munshi Sulaiman of BRAC for providing us with the dataset to do this research work.

I. Introduction

Measuring the effectiveness of anti-poverty programs has gained increased urgency as governments of developing countries work to reduce poverty in order to meet the United Nation's Millennium Development Goals (MDGs). One such program, "Targeting the Ultra Poor" (TUP), was implemented in Bangladesh by a non-governmental organization called Bridging Resources Across Communities (BRAC). This was a random assignment experiment launched in 2002 that addressed a number of outcomes such as income, health, empowerment, food security, and other dimensions of well being among those living in chronic poverty. Typically the program evaluation literature in economics would assess each outcome in the TUP program separately and then use an ad hoc method to judge the overall success or failure of the program. The ad hoc approach, however, has no theoretical basis and may not offer robust policy evaluation. In this paper we use an axiomatically derived multidimensional measure of poverty to assess the overall effectiveness of the TUP program.

The problem with a unidimensional measure of poverty in evaluating a program such as TUP is that it does not give a full picture of a person living in poverty, and in some cases, may underestimate the number of people living in poverty or the severity of poverty. For example, suppose there are three dimensions of well being—income, consumption and health—and that a person earns above the poverty line. Using a headcount ratio or the income-gap ratio or any other poverty measure that uses income as a dimension to measure poverty, the person is said to be non-poor. However, the person may be suffering from a serious illness that requires substantial out-of-pocket medical expenditures for treatment and so, after medical expenses are met, the person's remaining income falls well below the poverty line. Therefore, this person's discretionary consumption level is the same as that of a person living in poverty, even though his

income is above the poverty line. It would be logical to consider this person to be in poverty, but the typical income poverty measure does not allow this person to be counted as poor.[2]

Taking this example a step further suggests that it may be preferable to measure poverty based on consumption levels. For example, let us define a person as being poor if they consume less than a predetermined level of calories a day. A person may be income non-poor, but due to dietary restrictions, may consume less than the threshold level each day. This person then would be considered to be in poverty, even though their income is above the poverty line. As Sen (1983) argues, commodity ownership should not mean a person is not poor, because it does not tell us what the person is doing with it. In other words, taken alone it does not tell us what benefits a person can and cannot get from that commodity. This suggests a preferred alternative is to adopt a measure of poverty that incorporates all dimensions at once.

Recently, Emran, et al. (2008) used the TUP data to evaluate the effectiveness of the program; however, the authors measure each outcome separately. Their analysis showed that participation in the program by the poorest members of the society (the treated group) significantly improved their net income and food security measures, but did not have any significant impact on health, women's empowerment, and ownership of land. From their analysis, the overall success of the program cannot be ascertained.

We extend the analysis of Emran, et al. (2008) by using some of the multidimensional measures of poverty recently illustrated in Bourguignon and Chakravarty (2003) to evaluate the TUP program. This measure extends the unidimensional poverty measure of Foster, Greer, and Thorbecke (1984) to situations with multiple outcomes. The measure satisfies key axioms for robust poverty measures such as monotonicity, transfer, and transfer sensitivity (Sen 1983),

[2] The National Academy of Sciences (NAS) panel came up with an income poverty measure for the United States where they recommend subtracting out out-of-pocket medical expenses when calculating income poverty (Blank and Greenberg, Improving the Measurement of Poverty. The Brookings Institution. December, 2008)

which are important for among other things determining whether the rich are getting richer at the expense of the poor.

With a multi-dimensional measure of poverty we then use a number of recent techniques from the econometrics of program evaluation, including difference-in-difference estimators and difference-in-difference with matching estimators (Heckman, et al. 1998; Blundell and Costa-Dias 2009), to provide a more complete portrait of the success or lack thereof of the TUP program. The TUP program was set up as a random assignment experiment. However, as noted by Emran et al. (2008), the treatment and the control groups created by BRAC may suffer from selection bias because the poor were determined by the villagers, and thus participation in the program may reflect "cream skimming" on the part of villagers. To control for this possible selection bias, Emran, et al. (2008) created their own treatment and comparison groups using the objective selection criteria provided by BRAC, and then applied a variety of nonexperimental program evaluation estimators. We follow a similar approach to recreate the treatment and control groups but instead evaluate the program using multidimensional poverty measures using difference in difference with and without matching. We show that the TUP program reduced poverty of the treatment group by 18-21 percentage points between the years 2002 and 2005, when compared to the control group.

II. Multidimensional Poverty

The primary metric of a society's well being since establishment of national income and product accounts has been income (per capita), and indeed one of the UN's main MDGs is halving the number of people living on less than a $1 a day (PPP) by 2015.[3] The advantage of this measure is that it is transparent and once proper adjustment is made for purchasing power parity it offers a straightforward metric for comparisons across countries and time. However,

[3] "MDG Monitor". http://www.mdgmonitor.org/browse_goal.cfm

governments of both developed and developing nations are also stressing the need to focus on multiple dimensions of well-being, rather than one single outcome such as income. Recently Nobel laureates Joseph Stiglitz and Amartya Sen chaired a commission established by French President Nicholas Sarkozy to design a new way to measure well-being in France beyond GDP.[4] England has adopted an Index of Multiple Deprivation that combines 37 indicators to measure material hardship at the local level. Likewise, the UNDP's Human Development Index is a multi-dimensional measure designed to gauge the well-being of people living in different countries of the world.

A major question confronting researchers is how to quantify poverty, whether across single or multiple dimensions (Bourguignon and Chakravarty, 2003). Sen (1976) mentions that there are two problems faced when trying to measure poverty - identifying who is poor from a population and using all the available information from those poor to produce a measure of poverty. Usually, countries measure poverty using a single, or unidimensional measure. A certain dimension, such as income or consumption is taken and a threshold level is calculated. In the case of income, if anyone earns below that threshold, then the person is said to be living in poverty, but if anyone earns above that threshold, then the person is said to be not poor. Foster, Greer and Thorbecke (1984) (FGT) proposed a generalized class of unidimensional poverty measures as:

$$P_a(y;z) = \frac{1}{n}\sum_{i=1}^{q}\left(1-\frac{y_i}{z}\right)^{\alpha} \qquad (1)$$

where z is a predetermined level of the poverty line, q is the number of people/households living in poverty (that is, earning less than z), y_i is the income of individual/household i who is living

[4] The *Commission on the Measurement of Economic Performance and Social Progress* released their report in September 2009, available at http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf.

in poverty, and *n* is the total number of people/households in the community. Here α, is a number which gives different measures of poverty, known as the "poverty aversion" index. When α=0, the measure $P_\alpha(y;z)$ becomes the headcount ratio, that is, it measures the ratio of the number of people living in poverty, compared to the whole population. When α=1, $P_\alpha(y;z)$ is the income gap ratio, which measures the average normalized shortfall of income of all the individuals/households living in poverty. When α=2, the measure is called the squared-poverty gap, and it measures the average normalized shortfall of income, like the $P_1(y;z)$ measure, but in this case, it puts more weight on poorer people in the community. As $\alpha \to \infty$, increasing weight is placed on the worse-off individuals.

The important theoretical value about $\alpha \geq 2$ poverty measure is that it satisfies the monotonicity and transfer axioms of Sen (1976). The monotonicity axiom states that, all else being equal, a fall in income of a poor household should increase $P_\alpha(y;z)$. The transfer axiom says that a transfer of income from a poor household to a richer household should increase $P_\alpha(y;z)$. However, it can be easily seen that the headcount ratio does not satisfy the transfer axiom or the monotonicity axiom. When α=1, the monotonicity axiom is satisfied, but the transfer axiom is not (Sen, 1976). Satisfying these axioms can be important in the measurement of poverty, so that the measure can change if the rich in the society are getting richer at the expense of the poor, even though the number of poor remains the same.

Although the FGT measure is quite robust in the class of unidimensional poverty measures, as with all such measures it fails to capture fully the poverty level of an individual. Instead of depending on one dimension to measure poverty, multi-dimensional measures of

poverty take into account differences in income, health, and other social characteristics of an individual/family to determine whether the person/family is in poverty. Bourguignon and Chakravarty (2003) extend the FGT measure to the multi-dimensional case as follows:

$$P_a(X;z) = \frac{1}{n}\sum_{j=1}^{m}\sum_{i\in S_j} a_j \cdot \left(1 - \frac{x_{ij}}{z_j}\right)^{\alpha} \qquad (2)$$

where, *m* is the number of dimensions being used to assess poverty of individuals/households, a_j is the weight given to each dimension, S_j is the indicator which is 1 if the dimension *j* is below the threshold level, and 0 otherwise. If $S_j = 0$, then the measure for that dimension of poverty is 0. So, this shows that the higher the value of $P_a(X;z)$ the higher is the poverty in that community. In our simple example above, we had three dimensions to measure poverty - income, health and consumption. The threshold for each of these three dimensions is then determined. A poverty measure, such as the gap ratio or the squared gap ratio is calculated for each dimension of the person and then a weighted average is taken to measure the depth and the severity of poverty. If a person falls well below the poverty threshold in all three dimensions, then this person is in extreme poverty. If the person is deprived in less than three dimensions, then he or she is in a different degree of poverty.

Multidimensional poverty measures as in equation (2) require the evaluator to assign weights to each dimension ex-ante and this can help in the proper measurement of the success of the program as the weights cannot be manipulated ex-post. Indeed, this measure differs from the univariate approach by virtue of the weighting scheme used in aggregating outcomes across multiple dimensions. This is the baseline measure we use to assess the effectiveness of the Targeting the Ultra Poor program.

Before proceeding, we note in passing that multidimensional poverty measurement does not need to use the same dimensions to measure poverty across communities. Poverty is an absolute notion when it comes to the set of capabilities (or what benefits a person can get from using that commodity), but it is a relative notion when it comes to the set of commodities (Sen, 1983). For example, a television can be a necessary item for a child in UK due to the educational programs shown on it, but it is a luxury item for a child in a developing country like Tanzania since it is not a necessary item there for educational purposes (Sen, 1983).

III. Program Evaluation with Multi-dimensional Poverty

Programs that do not involve random placement of individuals in the treatment and the control group confront the challenge of "re-creating" the experimental environment. One common way of evaluating non-experimental programs is by using the technique of matching. It involves comparing the outcome of a program participant (Y_1) with those of certain non-participants (Y_0) that have similar characteristics as the participants. Any difference in the outcomes between the participants and the non-participants can be said to be due to the impact of the program (Heckman, et al. 1998). Heckman, et al. (1998) show that the estimated gain of a participant from the program, after controlling for program participation (D=1) and characteristics X is:

$$E(Y_1\text{-}Y_0|D=1,X). \tag{3}$$

If a particular domain of characteristics, or region of common support X, is used then the equation for the evaluation of a program becomes:

$$\sum_{i\in I_1} w_{N_0,N_1}(i)\left(Y_{1i} - \sum_{j\in I_{01}} W_{N_0,N_1}(i,j)\cdot Y_{0j}\right) \tag{4}$$

where Y_{1i} is the outcome of a person in the treated sample, Y_{0j} is the outcome for matched persons j. I_{01} is the set of people in the control group. $W_{N_0,N_1}(i,j)$ is a positive weight function so that for all i, $\sum W_{N_0,N_1}(i,j)=1$, and N_0 and N_1 are the number of people in the treatment and the control group, respectively. $w_{N_0,N_1}(i)$ is the weight given to each participant, with I_1 being the set of participants, and therefore, $\sum w_{N_0,N_1}(i)=1$. If both treatment and control groups have a region of common support (similar X values), then the effect of treatment on the treated can be evaluated by matching (Heckman et al., 1998).

One popular matching method is to use propensity scores (Rosenbaum and Rubin 1983). The basic idea is to predict the probability of program participation as a function of observables Xs, and then match members of the treatment and comparison groups based only on a scalar probability value rather than matching across the potentially large dimension of Xs. Although this approach controls for 'selection on observables', a potentially more robust approach that also controls for time-invariant unobservable heterogeneity is to combine matching with a difference-in-differences estimator (Smith and Todd 2005; Todd 2006):

$$\frac{1}{N_0}\sum_{i\in I_1}\left(Y_{it1}-\sum_{j\in I_{01}}W_{N_0,N_1}(i,j)\cdot Y_{jt0}-\left(Y_{it'1}-\sum_{j\in I_{01}}W'_{N_0,N_1}(i,j)\cdot Y_{jt'0}\right)\right) \quad , \qquad (5)$$

where t' and t are the time periods before and after the implementation of the program, and $W_{N_0,N_1}(i,j)$ and $W'_{N_0,N_1}(i,j)$ are the weights given to the members of the control group that have been matched with a member of the treatment group using propensity score matching at time t and t'. N_0 is the number of people in the treated group.

The typical evaluation focuses on a single outcome, Y, and when more than one outcome is addressed it is most common to evaluate each component sequentially. In the presence of

multidimensional poverty measures such as in equation (2) the difference-in-difference matching estimator in equation (5) is modified slightly by replacing Y with P_α as follows:

$$\frac{1}{N_0}\sum_{i\in I_1}\left(P_\alpha(y;z)_{it1}-\sum_{j\in I_{01}}W_{N_0,N_1}(i,j)\cdot P_\alpha(y;z)_{jt0}-\left(P_\alpha(y;z)_{it'1}-\sum_{j\in I_{01}}W'_{N_0,N_1}(i,j)\cdot P_\alpha(y;z)_{jt'0}\right)\right) \quad (6)$$

Since a decrease in the value of $P_\alpha(y;z)$ means that poverty level has decreased, we would expect the above equation to show a negative value if the program was successful and a positive value if the program was not successful. We estimate two versions of equation (6), a difference-in-difference without matching and a difference-in-difference with matching. In the case of difference in difference without matching, the weights are set as $W_{N_0,N_1}(i,j)=\frac{1}{N_1}$ and $W'_{N_0,N_1}(i,j)=\frac{1}{N_1}$.

The estimation of equation (6) is implemented following the algorithm of Dehejia and Wahba (2002):

1) A logit model for the probability of an individual participating is estimated using all the entries in the treatment group and the control group for the year before the beginning of the program and estimate the propensity score for each person;

2) Data is then sorted by the propensity score and then stratified according to some score range. In each score range, we note which one is a member of the treatment group and which ones are the members of the control group.

3) The difference in means of the different characteristic variables between the treatment group and control group in each score range are then tested. If they are not significantly different from 0, then the matching was a success, otherwise, we would have to make the range of sorting using propensity scores narrower to make the difference in means of the treatment and control group

similar to each other. However, if most of the t-values show that the means are significantly different between the treatment and the control group, then the logit needs to be respecified and the whole process repeated.

In this approach, the members of the treatment and the control groups can be matched using propensity score. After matching, the poverty value $P_\alpha(y;z)$ for each member is calculated. Then the $P_\alpha(y;z)$ measure of the control group of each score range is used to calculate the kernel density function for the year *t'*. Using the kernel density function, the weights given to each member of the control group $W'_{N_0,N_1}(i,j)$ in each score range is calculated. The process is repeated again to find the weights $W_{N_0,N_1}(i,j)$ in time *t*. This is done because it can be assumed that $P_\alpha(y;z)$ has changed over time, and so, its distribution has also changed. Therefore, the weights given to each member of the control group need to be updated. After doing the matching and calculating the weights, the values are then plugged into the equation above to measure the effectiveness of the program. The values of $P_\alpha(y;z)$ are calculated for $\alpha=\{1,2,3\}$ and then the difference-in-difference with matching estimate is calculated to test the overall impact of the program.

IV. Data

The data for the analysis has been obtained from *Bridging Resources Across Communities* (BRAC). BRAC is a non-governmental organization situated in Bangladesh that focuses on development work. It has offices in every district of Bangladesh, employing over 37,000 full-time workers and 53,000 community school teachers, with an annual expenditure of $250 million. It also has a number of development programs in Afghanistan, Sri Lanka, East Africa and the United Kingdom (Hulme and Moore, 2007). In 2002, BRAC launched the "Targeting the Ultra Poor" (TUP) program as an experiment to help people living in chronic

poverty. BRAC noted that its microfinance program did not really benefit the poorest women in Bangladesh, mainly due to self and social exclusion (Hulme and Moore, 2007).

BRAC commenced its TUP operations in three of the northern districts in Bangladesh - Rangpur, Kurigram and Nilphamari. BRAC then developed some exclusion and inclusion criteria to recruit people in the program. A household had to satisfy all three exclusion and two of the five inclusion criteria to be eligible for the TUP program (Matin, et al. 2004). The exclusion criteria were: 1) any member of a household was a member of an NGO; 2) any member of the household receiving government benefits; 3) the household has no physically able adult woman present. The inclusion criteria were: 1) the household owning less than 10 decimals of land[5]; 2) no adult man working in the household or disabled adult man present in the household; 3) presence of an adult woman in the household selling labor; 4) school aged children working in the household; and 5) no productive assets in the household.

Conditional on satisfying the criteria for participation, participants were randomly assigned to a treatment group and a control group. There were 2543 households in the treatment group (denoted as SUP) and 2524 households in the control group (denoted as NSUP) selected by BRAC to the program. The treatment and control households were asked a number of social and economic questions in 2002. The treatment group then participated in the TUP program, where the women from the treatment household were given training on income generation. They were also given repeated training on income generation, and then an asset was transferred to them for income generation. The program gave participants a monthly stipend, healthcare, technical support, follow-up to see whether they were improving in different economic and social indicators, and education on different social issues by BRAC field workers (Matin et al.,

[5] A decimal is a unit of measurement of area used in India and Bangladesh. It is approximately equal to 1/100 acre or about 40.46 m² (http://en.wikipedia.org/wiki/Decimal_(unit))

2004). At the end of the program in 2005, the treatment and the control households were interviewed again to measure the outcomes.

An important portion of the program was the transfer of knowledge to these ultra-poor households about health and well-being. For example, before the start of the program, 63 percent of the households did not own any soap, showing that they either did not have any information about how important soap was for hygiene, or they were too poor to purchase soap. Therefore, it was important to transfer some basic knowledge about health, hygiene and rights to these ultra-poor people. Besides income generation, the program educated households on the importance of using soap, sanitary bathroom, brushing teeth and consuming salt with iodine in it. It also educated people on basic laws concerning inheritance and divorce, importance of vitamins in a person's diet, and human rights. Some of these ordinal variables can be included in the poverty measure to show how the capabilities of households improved due to participation in TUP.

The dataset contains a number of important outcomes that can be used for multidimensional poverty analysis. However, as noted by Emran et al. (2008), the treatment and the control groups created by BRAC may suffer from selection bias because the poor were determined by the villagers, and there may have been some degree of favoritism or preference in the selection process. Therefore, there may be type 1 and type 2 errors. A type 1 error can occur when people who should have qualified to participate in the program were not included in the program. A type 2 error can occur when people who did not qualify to participate in the program were actually included to participate in it. Emran et al (2008) recreated their own treatment and control groups using the selection criteria provided by BRAC in order to conduct their analysis. Using the guidelines from the exclusion and inclusion criteria of BRAC and the methodology used by Emran et al. (2008), we created four different groups, namely TUP1 - households who

received treatment and they qualified to receive the treatment; TUP2 - households who received treatment but should not have received the treatment; TUP3 - households who did not receive the treatment but they qualified to receive the treatment; and TUP4 - households who did not receive the treatment and did not qualify to get the treatment. There are 1997 households in the TUP1 group, 513 households in TUP2, 1667 in TUP3 and 857 in TUP4 group. Thirty three households were unclassified. Therefore, our treatment group is TUP1 and comparison group is TUP3. The method used to create our treatment and control groups is illustrated in Appendix 1, and the description of each group is in Appendix Table 1. The main analysis will compare the outcomes between households in groups TUP1 and TUP3. We use a simple difference-in-difference estimator and also a difference-in-difference with matching estimator to measure the impact of the program on the treated. In the difference-in-difference estimate, we also look into the differences between TUP1 and TUP2 and between TUP2 and TUP4 to see how people were included due to a type 2 error fared in the program. For completeness we also present results using BRACs original treatment and control groups, SUP and NSUP.

Income per capita had to be calculated because Bangladesh does not define poverty level by income, but rather by consumption level. Bangladesh Bureau of Statistics considers a household that consumes less than 2122 kcal of food per person per day as a household living in poverty. If a household consumes less than 1805 kcal of food per person per day, then the household is said to be in 'hard core' poverty (Nizamuddin et al., 1999). The price of rice in 2002 was estimated to be 11.96 taka per kilogram (Dorosh and Shahabuddin, 2002). The USDA measure is that one kilogram of brown, medium-grain rice gave 1120 kcal when cooked (USDA, 2009). Using these numbers, the extreme poverty line for a given year was calculated to be 7035.31 taka. In this analysis, 7000 taka was considered to be the estimate for the extreme

poverty line, and so, anyone earning below that is considered to be extremely poor in the income dimension.

Based on data availability we use twenty-one variables as dimensions to measure poverty. The dimensions chosen and the corresponding threshold values for each dimension, or z, are shown in Table 1. The table shows that a number of cardinal and ordinal measures have been included in the analysis. Variables such as per-capita income, number of cows, goats, hens and land owned show how much wealth a household possesses. Other variables, such as whether the household is educated, uses soap, iodized salt and toothbrush show the capability and well-being of the household. Including these dimensions are important in assessing multidimensional poverty of the ultra-poor, as many do not have access to such basic commodities. When these variables are included, then any improvements in these dimensions are captured in the measure as a decline in the poverty measure in equation (2).

Also, measures such as self-reported health and food deficit rating are included in the analysis. The reason to include them is to see whether the perception of poverty have changed among the ultra-poor because of program participation. It is important for the poor to feel that their economic condition has improved. So, any improvement in these variables will be captured in the poverty measure.

Because there are no pre-specified thresholds of *z*, we set the thresholds based on the dimension being measured. For example, owning soap was recorded by BRAC as a dichotomous outcome, and thus if the household does not own any soap they are considered to be poor in that dimension. On the other hand, schooling was recorded in units from 0 to 10.5 and thus a household head with 5 or less is recorded as poor in the schooling dimension. The summary statistics of the variables of the treatment group and the control group in the year 2002 are

illustrated in Table 2. These values are the normalized shortfalls if their values are below the poverty threshold. A value closer to 1 shows a higher degree of deprivation in that dimension. The p-values show the difference in means between the treated (TUP1) and the control groups (TUP3) that have been constructed in this paper. The p-values show that most of the means are significantly different, showing that finer divisions between the treated and control groups need to be done in order to analyze the outcome of the program. As a result, later in the paper, we use propensity scores matching technique to match household of the treated group with those of the control group.

V. Results

In constructing $P_a(X;z)$ from equation (2) each of the 21 dimensions was given equal weight, which is consistent with Utilitarian social preferences, but the weights can vary for each dimension if the social planner has different preferences as we demonstrate in the sensitivity section. Given the appropriate weights, the poverty gap for each dimension is then calculated using $\alpha = 1, 2$, and 3. Sample means for these three measures are recorded for each pair of treatment and comparison groups in each year in Table 3. As indicated in the table, poverty fell between 2002 and 2005 for all poverty measures and groups, though it is necessary to compare how the treatment and comparison groups changed differentially over the period.

The baseline results of difference in difference without matching follow directly from Table 3, and are shown in Table 4. Using the treatment (denoted by SUP) and the control group (denoted by NSUP) of BRAC, it is seen that the program improved the condition of the treatment group by about 20 percentage points regardless of the poverty measure. Using our constructed treatment and control groups, TUP1 and TUP3, respectively, we find that the program improved the well being of ultra poor by about a little over 1 percentage points less than with the BRAC

sample, but still a very impressive 18.6 to 18.8 points. An interesting thing to notice is that when comparing TUP1 with TUP2, it is seen that TUP1 actually did slightly better than TUP2. So the poorer cohort in the treated group did better than the richer cohort, but only by a slight margin. Similarly, when comparing between TUP2 and TUP4, it is seen that TUP2 improved by around 18-20 points when using the poverty gap measure. This shows that households that were included due to type II error fared much better due to participation. BRAC's treatment and control group seems to overestimate the success of the program, which is consistent with a cream skimming story on the part of program administrators, though the differences are not large. A better way to test if cream skimming is going on is to do a difference-in-difference with matching.

A. Difference-in-Difference with Matching

We next report on the difference-in-difference with matching estimates. First, the logit regression results for predicting treatment (1 if the person was selected in the program, 0 otherwise) are shown in Table 5. The estimates show that increases in amount of land owned, income per capita, household size, and value of the home are significantly reduce the probability of participation in the program. With these estimates we constructed the predicted probability, or propensity scores, and then the treatment and the control group are matched across thirteen different groups with propensity score ranges from 0 - 0.1, 0.1 - 0.2 ..., 0.6 - 0.7. The two-tailed t-test for testing equal means is then run between the treatment and the control groups for each range, and the results are shown in Table 6. From Table 6, it is seen that most of the t-statistics are insignificant, showing that the logit used to calculate the propensity scores is a valid matching algorithm and the samples are well balanced.

The result of the difference-in-difference matching estimator is shown in Table 7, where the control group matching weights are from an Epanechnikov kernel estimator. The first row shows the difference between the average value between treatment and the weighted control group for the year 2002. The difference is very small, which shows that the matching between the treated and the control group was effective. The average difference-in-difference with matching estimate shows that the program helped to reduce poverty in the treated group by just above 21 percentage points when compared to the comparison group across all three poverty measures. These estimates are about 3 percentage points higher than those obtained in Table 5 without matching, or about 16 percent, suggesting that matching based on observables is important in this determining overall program effectiveness.

B. Sensitivity Analysis

In this subsection we report the results of several robustness checks to our baseline specifications. First, we examine the sensitivity of our results to the assumption of equal weights across each dimension of our poverty measure P_α from equation (2). Suppose, for example, that the social planner is particularly interested in reducing income, food, and health deficits compared to hardship along other dimensions. To admit unequal weights, a_j, we assigned a weight of 0.2 to each of per-capita income, food deficit rating, and self-reported health dimensions, while a weight of 0.0222 was given to the remaining 18 dimensions. The poverty measures were recalculated using these new weights and then the difference-in-difference with matching estimator was applied. The results shown in Table 8 indicate that the poverty reduction of the TUP program continues to be substantial, but the decline to about 15-16 percentage points from the original 21 points suggests that in this example too little weight is given to other non-

food, health, and income dimensions where great strides were made by the treatment group relative to the comparison group.

In the second specification check we alter the poverty threshold values, *z*, of the cardinal values used in constructing P_α from equation (2). The threshold values of all the cardinal values were doubled (e.g. the threshold for per-capital income was increased to 14000, both male and female clothing were increased to 4, winter blanket increased to 1, cow and goat increased to 1, hen increased to 2 and land increased to 20 decimals). The ordinal values were not changed. The difference-in-differences with matching estimates are shown in Table 9. The results show that poverty decreased by about 18.5-20 percentage points in the treated group. This result is quite close to what the other estimates showed. The table reveals that our baseline estimate of a 21 percentage point gain is quite robust to substantial changes in the poverty threshold values.

In a final check we examined whether changing the actual FGT poverty measure had any effect on the robustness of the results. Instead of using equation (2) as the poverty measure, which as noted previously is an aggregated version of the unidimensional FGT poverty measure, we adopt an alternative multidimensional poverty measure proposed by Bourgninon and Chakravarty (2003):

$$P_a(X;z) = \frac{1}{n}\sum_{j=1}^{m}\sum_{i\in S_j}\left[a_j \cdot \left(1-\frac{x_{ij}}{z_j}\right)^{\theta}\right]^{\frac{\alpha}{\theta}}. \qquad (7)$$

In equation (7), all parameters are the same as in equation (2) except for the addition of θ, which represents a substitutability parameter. For example, when θ=1 we assume that each dimension in the poverty measure is perfectly substitutable. However, as θ increases the substitutability between the dimensions decline. For example, if there are two dimensions – health measure and income, then a higher value of θ signifies that it is harder for a person to substitute better health

dimension for higher income dimension. However, the values of α/θ need to be higher than 1 in order to satisfy the transfer axioms. The transfer axiom specifies that if a measure such as income is transferred from a poor to a richer person, poverty measure should increase and vice versa (Foster et al., 1984). Using different values of θ and α, and equal weights to all the dimensions (a_j), the difference-in-differences results are presented Table 10. The table shows that using different BC measures, where the value of $\alpha/\theta>1$, shows that the program was able to reduce poverty by 18-19 percentage points among the treated group when comparing them to the control group.

VI. Conclusion

This paper shows how the multi-dimensional poverty measure can be used to assess the overall effectiveness of an intervention by using program evaluation estimators. Specifically we combine multidimensional poverty measures with difference-in-difference and difference-in-difference with matching estimators to evaluate the Targeting the Ultra Poor program implemented in Bangladesh between 2002 and 2005.

We find that the TUP program reduced overall poverty among the chronically poor by around 18-21 percentage points, or nearly 25 percent relative to the baseline poverty rate. The sensitivity analysis shows that changing the multidimensional poverty measure, weights or the poverty threshold values does not significantly affect the overall success or failure measure of the program. Multi-dimensional program evaluation is an attractive alternative to the standard unidimensional approach to guide policymakers in identifying the effectiveness of comprehensive anti-poverty programs.

References

Blundell, R., Monica Costa Dias, "Alternative Approaches to Evaluation in Empirical Microeconomics," Journal of Human Resources, Vol. 44. No. 3, 2009

Bourguignon, F., S. Chakravarty. "The Measurement of Multidimensional Poverty," Journal of Economic Inequality. 1. 2003: 25-49

Dehejia, R., S. Wahba. "Propensity-score Matching Methods for Non-experimental Causal Studies," The Review of Economics and Statistics. Vol 84 No. 1, February, 2002: 151-161.

Dorosh, P., and Q. Shahabuddin. "Rice Price Stabilization in Bangladesh: An Analysis of Policy Options," MSSD Discussion Paper No. 46. IFPRI. 2002. http://www.ifpri.org/sites/default/files/publications/mssdp46.pdf

Emran, M., Virginia Robano, and Stephen Smith. "Assessing the Frontiers of Ultra-Poverty Reduction: Evidence from CFPR/TUP, an Innovative Program in Bangladesh," August, 2008. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1354158

Foster, J., J. Greer, and E. Thorbecke. "A Class of Decomposable Poverty Measures," Econometrica. Vol. 52, No. 3, May, 1984: 761-766.

Heckman, J., H. Ichimura, and P. Todd. "Matching as an Econometric Evaluation Estimator," The Review of Economic Studies. Vol. 65, No. 2, April, 1998: 261-294.

Hulme, D., and K. Moore. "Assisting the poorest in Bangladesh: Learning from BRAC's 'Targeting the Ultra Poor' Programme." University of Manchester Brooks World Poverty Institute. 2007 http://ssrn.com/abstract=1160303

Matin, I., A. Hadi, and S. Masud Ahmed. "Introduction." in Towards a Profile of the Ultra Poor in Bangladesh: Findings from the CFPR/TUP Baseline Survey. (Dhaka: Research and Evaluation Division, BRAC and Aga Khan Foundation, Canada). 2004:1-28

Nizamuddin, S., M. Ravallion, S. Shah, and Q. Wodon, "Annex: Chapter 5," in Bangladesh: From Counting the Poor to Making the Poor Count (Washington DC: The World Bank). 1999: 52

"Nutrition Data Laboratory," US Department of Agriculture. 2009 http://www.nal.usda.gov/fnic/foodcomp/search/

Rosenbaum, R., and D. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika 70:1, 1983: 41–55.

Sen, A. "Poverty: An Ordinal Approach to Measurement." Econometrica. Vol. 44, No. 2., March, 1976: 219-231

Sen, A. "Poor, Relatively Speaking," Oxford Economic Papers. Vol 35, No. 2, July, 1983: 153-169.

Smith, J., and P. Todd. "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods?" Journal of Econometrics, 125:1-2, 2005: 305-353.

Todd, P. "Matching Estimators," October, 2006. http://athena.sas.upenn.edu/~petra/papers/mpalgrave2.pdf

Appendix 1: Constructing the treatment and the control groups.

A number of the guidelines chosen to determine the treatment and the control groups have been taken from Emran el al. (2008) and the criteria used by BRAC to select people in the treatment and control groups.

The first exclusion condition, whether the person was an NGO member was determined if the person had any NGO savings, or was an NGO member, or took NGO loans. All these variables were given in the dataset. The second exclusion condition, whether the person received any government benefits, was determined from the variables if any government benefits were received, if any government were received recently or in last six months, and if one of the sources of income was government benefits. The third exclusion condition was hard to determine, as no data was given in the dataset, so it was assumed that all households had at least one able women.

For the inclusion condition ownership of land, the variable land ownership was used, Any household that owned less than 10 decimals of land was considered as a potential candidate. For productive assets, there was a binary variable that said if the person had productive assets or not. For child labor, if the earning member ratio = 1 and there was a child present in the household, then the house was considered for inclusion. Also, if the number of child in the house was greater than the number of individuals not working, then the house was considered in that category. For the category if the household had a disabled person or no adult men working in the household, the following was done: If the number of not disabled in the house was less than the household size, then it was included in the category. If the house was headed by a female, then it was also included in this category.

To measure if any woman from the household was working outside of the household, the following was done: if the household was headed by a woman, it was considered to be included in this category. Also if one of the main sources of income was daylabor (agriculture), daylabor (non-agriculture), small-business/trading, begging, servant and professional services.

If a household satisfied all the exclusion criteria and two of the five inclusion criteria, then it was considered to be a potential member. We find that 3618 households should have been included in the program and 1449 should not have been included in the program in 2002. Matching them with the treated and the control group of BRAC gives us TUP1, TUP2, TUP3, TUP4.

The description of each group is summarized in appendix table 1.

Appendix Table 1: Description of the Treated and Control Groups

Group	Description
SUP	The treated group created by BRAC for the TUP program. This group participated in the program.
NSUP	The control group created by BRAC. This group did not participate in the program
TUP1	The households in SUP who meet all the criteria to be in the program. This is our constructed treatment group.
TUP2	The households in SUP who received treatment, but should not have received any. They were included in the program due to type I error made in the selection process
TUP3	The households in NSUP who meet all the criteria to be in SUP. They were not included in the program due to type II error made in the selection process. This is our constructed control group.
TUP4	The households in NSUP who did not qualify to receive treatment, and who did not receive any treatment from TUP.

Table 1: Variables Used in Constructing the Multidimensional Measure

Name of Variable	Description of Variable	Measurement units	Poverty Threshold
Cash Saving	Does the household have any cash saving	0 – no 1 – yes	0 – in poverty
Female Clothing	How many units of female clothing (saree) does the household have	Cardinal values	2 or less – living in poverty
Male Clothing	How many units of male clothing (lungi) does the household have?	Cardinal values	2 or less – living in poverty
Winter Clothes	Does the household own any winter clothes?	0 – no 1 – yes	0 – in poverty
Food Deficit Rating	How does the house rate their food deficit?	0 – severe 1 - mild 2 – no deficit 3 – surplus	0 and 1 – in poverty
Soap	Does the household own soap?	0 – no/don't know 1 – yes	0 – in poverty
Toothbrush	Does the household own a toothbrush?	0 – no/don't know 1 – yes	0 – in poverty
Salt	Does the household use salt with iodine	0 – no/don't know 1 – yes	0 – in poverty
Health status	What is the self-assessed health status of the head of household?	0 – poor/bad 1 – fair 2 – good 3 – very good 4 – excellent	2 and below – in poverty
Eat twice a day	Is the household able to eat twice a day?	0 – no 1 – yes	0 – in poverty
House Condition	What is the condition of the house?	0 – bush roof 1 – low ceiling 2 – medium ceiling 3 – high ceiling	1 and below – in poverty

Latrine	What is the type of bathroom used by the household	0 – open 1- pit 2- sanitary/slab	0 – living in poverty
School	Years of schooling completed by the head of household	0 to 10.5	5 and below – living in poverty
Tubewell	Does the household own a water hand pump?	0 – no 1 – yes	0 – living in poverty
Wall	What material was used to construct the walls of the house?	0 – straw/soil/hemp /jute 1 – brick/bamboo /tin	0 – living in poverty
Roof	What material was used to construct the roof?	0 – straw/plastic /hemp 1 – tin	0 – living in poverty
Cow	How many cows do the household own?	Cardinal values	0 – living in poverty
Goat	How many goats do the household own?	Cardinal values	0 – living in poverty
Hen	How many hens do the household own?	Cardinal values	1 or less – living in poverty
Blanket	Does the house own any blankets?	Cardinal values	0 – living in poverty
Land	Amount of land owned by the household	Cardinal values	<10 decimals – living in poverty
Income	Income earned by the household in a given year	Cardinal values	<7000 – living in poverty

Table 2: Summary Statistics of Treatment and Control Groups in the TUP Program

	Treated Group (TUP-1)		Control Group (TUP-3)		
Variable	**Average**	**Standard Deviation**	**Average**	**Standard Deviation**	**P-values for Difference in Means**
Per Capita Income	0.6521	0.2058	0.6269	0.2146	0.000
Amount of Land Owned	0.8349	0.2454	0.7672	0.2880	0.000
Own Blanket	0.9860	0.1176	0.9712	0.1673	0.001
Own Hen	0.7629	0.3886	0.6995	0.4110	0.004
Own Goat	0.9504	0.2171	0.9268	0.2605	0.144
Own Cow	0.9770	0.1501	0.9190	0.2729	0.000
Type of Roof	0.4796	0.4269	0.4088	0.4305	0.002
Type of Wall	0.1669	0.1820	0.1982	0.2000	0.000
Own Tubewell	0.9900	0.0996	0.9700	0.1706	0.000
Average Schooling	0.9448	0.1026	0.9428	0.1032	0.566
Type of Bathroom	0.4775	0.1038	0.4646	0.1283	0.001
Eat Twice a Day	0.5193	0.4998	0.3437	0.4751	0.000
Self-reported Health	0.2651	0.2552	0.2681	0.2603	0.7350
Use Iodized Salt	0.9144	0.2799	0.9226	0.2673	0.981
Own Brush	0.9820	0.1331	0.9844	0.1239	0.924
Own Soap	0.6775	0.4675	0.6311	0.4827	0.002
Food Deficit Rating	0.5385	0.1763	0.4579	0.2053	0.000
Winter Clothes	0.9332	0.1702	0.9142	0.1886	0.001
Male Clothing	0.1107	0.2112	0.1110	0.2121	0.965
Female Clothing	0.1550	0.2361	0.1317	0.2216	0.000
Cash Saving	0.9304	0.2545	0.8908	0.3120	0.000
Number of Obs.	1997		1667		

Table 3: Poverty Gaps for Alternative Treatment and Comparison Groups in the TUP Program

Variable	Year	Poverty Gap	Squared Gap	Cubic Gap
SUP	2002	0.7682	0.7335	0.7135
		(0.0017)	(0.0017)	(0.0017)
SUP	2005	0.4805	0.4470	0.4296
		(0.0025)	(0.0025)	(0.0026)
NSUP	2002	0.7126	0.6743	0.6526
		(0.0020)	(0.0020)	(0.0020)
NSUP	2005	0.6221	0.5862	0.5672
		(0.0027)	(0.0028)	(0.0028)
TUP1	2002	0.7742	0.7397	0.7200
		(0.0017)	(0.0018)	(0.0018)
TUP1	2005	0.4933	0.4597	0.4425
		(0.0029)	(0.0030)	(0.0030)
TUP3	2002	0.7305	0.6921	0.6704
		(0.0022)	(0.0023)	(0.0023)
TUP3	2005	0.6360	0.6003	0.5813
		(0.0032)	(0.0032)	(0.0033)
TUP2	2002	0.7438	0.7082	0.6876
		(0.0045)	(0.0045)	(0.0045)
TUP2	2005	0.4719	0.4379	0.4203
		(0.0055)	(0.0056)	(0.0056)
TUP4	2002	0.6728	0.6346	0.6132
		(0.0037)	(0.0037)	(0.0037)
TUP4	2005	0.5913	0.5547	0.5355
		(0.0052)	(0.0052)	(0.0052)

Table 4: The Effect of TUP from Difference-in-Difference Estimation Results for Alternative Treatment and Comparison Groups

Difference	Year	Poverty Gap	Squared Gap	Cubic Gap
SUP-NSUP	2002	0.0556	0.0592	0.0609
		(0.0026)	(0.0026)	(0.0027)
SUP-NSUP	2005	-0.1416	-0.1392	-0.1375
		(0.0037)	(0.0038)	(0.0038)
D-in-D		**-0.1972**	**-0.1984**	**-0.1984**
		(0.0046)	**(0.0046)**	**(0.0047)**
TUP1-TUP3	2002	0.0437	0.0475	0.0495
		(0.0028)	(0.0028)	(0.0029)
TUP1-TUP3	2005	-0.1426	-0.1403	-0.1388
		(0.0044)	(0.0044)	(0.0045)
D-in-D		**-0.1863**	**-0.1878**	**-0.1883**
		(0.0052)	**(0.0053)**	**(0.0054)**
TUP1-TUP2	2002	0.0304	0.0315	0.0323
		(0.0059)	(0.0058)	(0.0059)
TUP1-TUP2	2005	0.0213	0.0220	0.0222
		(0.0076)	(0.0077)	(0.0077)
D-in-D		**-0.0091**	**-0.0095**	**-0.0101**
		(0.0074)	**(0.0096)**	**(0.0096)**
TUP2-TUP4	2002	0.0710	0.0735	0.0744
		(0.0059)	(0.0058)	(0.0058)
TUP2-TUP4	2005	-0.1193	-0.1168	-0.1152
		(0.0076)	(0.0077)	(0.0077)
D-in-D		**-0.1903**	**-0.1903**	**-0.1896**
		(0.0059)	**(0.0096)**	**(0.0096)**

Standard errors are in parenthesis

Table 5: Logistic Regression for Predicting Program Participation in 2002

Variable	Coefficient	Standard Error
Amount of Land Owned	-0.038*	0.0077
Per-capita Income	-0.00006*	0.00002
Household Size	-0.0463*	0.0220
School of Head	0.0158	0.0577
Age of Head	0.0031	0.0026
Present Value of House	-0.0002*	0.00002
Constant	0.68841*	0.1657

Dependent variable is a dummy, where 0 indicates that the individual was eligible, but was not selected, and 1 indicates that the eligible individual participated in the program. Standard errors are in parentheses. Asterisks indicate significance at the 5% level.

Table 6: Two-tailed Test of Equal Means Between the Treatment and the Control Groups.

Propensity Score	Per Capita Income	Age	Household Size	Land Owned	Schooling	Present Value of House
0-0.1	0.7118	-0.653	-0.2884	-0.0979	1.1494	0.5673
0.1-0.2	0.68	0.1775	-0.6339	-0.2113	1.508	-0.1451
0.2-0.3	-1.2516	-0.643	0.6596	1.0783	0.9662	-0.0976
0.3-0.4	-0.6269	-1.7882	2.1938*	0.3049	-0.8705	-0.2228
0.4-0.5	0.2201	-1.1192	0.4165	-0.9359	1.6949	1.874
0.5-0.6	0.8322	-0.5466	-1.3902	2.5265*	-0.1589	1.4087
0.6-0.7	-1.1961	2.5068*	0.8341	3.0915*	-0.0952	0.5744

The treatment and the control within a specified range of propensity score were put together into each group. Asterisks indicate that the t-value was significant at 5% level

Table 7: The Effect of TUP Program from Difference-in-Difference with Matching Estimator

Variable	Year	Poverty Gap	Squared Gap	Cubic Gap
TUP1 – wTUP3	2002	0.0224 (0.0017)	0.0239 (0.0017)	0.0251 (0.0017)
TUP1 – wTUP3	2005	-0.1934 (0.0030)	-0.1910 (0.0030)	-0.1887 (0.0030)
Difference-in-Difference		-0.2158 (0.0035)	-0.2149 (0.0035)	-0.2138 (0.0035)

Standard errors in parenthesis. TUP1 is the treatment group and TUP3 is the control group. w is the weight assigned to each member of the control group. The numbers represent the average difference between the treatment and control groups. The negative value indicates that the poverty measure has decreased. Standard errors are in parenthesis.

Table 8: Difference-in-Difference with Matching Estimator Result, When Weights Given to Each Dimension are Different

Variable	Year	Poverty Gap	Squared Poverty Gap	Cubic Poverty Gap
TUP1-wTUP3	2002	0.0198	0.0331	0.0421
		(0.0024)	(0.0033)	(0.0032)
TUP1-wTUP3	2005	-0.1499	-0.1287	-0.1138
		(0.0033)	(0.0034)	(0.0035)
Difference-in-Difference		-0.1698	-0.1618	-0.1559
		(0.0041)	(0.0045)	(0.0047)

Standard errors in parenthesis. TUP1 is the treatment group and TUP3 is the control group. w is the weight assigned to each member of the control group. The weights on each dimension are not equal anymore. The weights on per-capita income, food deficit rating and self-reported health were 0.2 each, while the rest were given a weight of 0.0222 each. The numbers represent the average difference between the treatment and control groups. The negative value indicates that the poverty measure has decreased.

Table 9: Difference-in-Difference with Matching Estimator Result, When Poverty Thresholds are Changed

Variable	Year	Poverty Gap	Squared Poverty Gap	Cubic Poverty Gap
TUP1-wTUP3	2002	0.0182	0.0215	0.0233
		(0.0015)	(0.0016)	(0.0017)
TUP1-wTUP3	2005	-0.1668	-0.1749	-0.1789
		(0.0027)	(0.0029)	(0.0030)
Difference-in-Difference		-0.1850	-0.1964	-0.2022
		(0.0031)	(0.0033)	(0.0034)

Standard errors in parenthesis. TUP1 is the treatment group and TUP3 is the control group. w is the weight assigned to each member of the control group. The threshold for per-capital income was increased to 14000, both male and female clothing were increased to 4, winter blanket increased to 1, cow and goat increased to 1, hen increased to 2 and land increased to 20 decimals. The numbers represent the average difference between the treatment and control groups. The negative value indicates that the poverty measure has decreased.

Table 10: Difference-in-Difference with Matching Using the Bourguignon-Chakravarty Multidimensional Poverty Measure

Variable	Year	Θ=2, α=4	Θ=2, α=3	Θ=3, α=5
TUP1-wTUP3	2002	0.0446	0.0350	0.0420
		(0.0024)	(0.0022)	(0.0023)
TUP1-wTUP3	2005	-0.1418	-0.1563	-0.1487
		(0.0030)	(0.0031)	(0.0031)
Difference-in-Difference		-0.1864	-0.1913	-0.1908
		(0.0039)	(0.0038)	(0.0038)

Standard errors in parenthesis. TUP1 is the treatment group and TUP3 is the control group. w is the weight assigned to each member of the control group. Equation (7) is used to measure poverty, instead of equation (2). Different values were calculated by changing the values of Θ and α.