OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 85, 3 (2023) 0305-9049
doi: 10.1111/obes.12539

Measuring Poverty Dynamics with Synthetic Panels Based on Repeated Cross Sections

HAI-ANH H. DANG†,‡,§,¶,# and PETER F. LANJOUW††,‡‡

†*Development Data Group, World Bank, Washington, DC, USA (e-mail: hdang@worldbank.org)*
‡*International School, Vietnam National University, Hanoi, Vietnam*
§*Paul H. O'Neill School of Public and Environmental Affairs, Indiana University, Bloomington, IN, USA*
¶*IZA, Bonn, Germany*
#*GLO, Essen, Germany*
††*School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands*
‡‡*Tinbergen Institute, Vrije Universiteit Amsterdam, Amsterdam, Netherlands*

Abstract

Panel data are rarely available for developing countries. Departing from traditional pseudo-panel methods that require multiple rounds of cross-sectional data to study poverty mobility at the cohort level, we develop a procedure that works with as few as two survey rounds and produces point estimates of transitions along the welfare distribution at the more disaggregated household level. Validation using Monte Carlo simulations and real cross-sectional and actual panel survey data – from several countries, spanning different income levels and geographical regions – perform well under various deviations from model assumptions. The method could also inform investigation of other welfare outcome dynamics.

I. Introduction

Following the steady progress of the past few decades in global poverty reduction, policymakers in both richer and poorer countries are devoting more attention to the nuanced

JEL Classification numbers: C53, D31, I32, O15.

*We would like to thank the editor Climent Quintana-Domeque, two anonymous reviewers, Francois Bourguignon, Fiona Burlig, Alan Dorfman, Chris Elbers, Francisco Ferreira, Gary Fields, Paul Glewwe, Bill Greene, Bo Honore, Stephen Jenkins, Dean Jolliffe, Aart Kraay, Christoph Lakner, Yue Man Lee, Michael Lokshin, Andy McKay, David McKenzie, David Newhouse, Reema Nayar, Franco Peracchi, Tuoc Van Phan, Menno Pradhan, Sergiy Radyakin, Carolina Sanchez-Paramo, Erik Thorbecke, Renos Vakis, Roy van der Weide, Nobuo Yoshida, and participants at meetings of the Econometric Society in Asia (Singapore) and Latin America (Medellin), International Association for Applied Econometrics (London), International Conference on Panel Data (London), North East Universities Development Consortium (MIT), and conferences and seminars at Cornell, IFPRI, Oxford, University of New South Wales, and World Bank for helpful discussions on earlier versions of this paper. We further thank Renos Vakis and Leonardo Lucchetti for their help with the Peruvian data. We also thank the UK Foreign Commonwealth and Development Office (FCDO) for funding assistance through its various programmes with the World Bank, including the Strategic Research Program (SRP), Knowledge for Change (KCP) Program, and the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Program.

dynamics underlying poverty and income mobility (e.g. Stiglitz, 2013; Piketty, 2014; World Bank, 2017). Measuring and tracking economic mobility, especially for the lower income groups, are increasingly regarded as essential for improving shared prosperity.[1] Indeed, a better understanding of the factors that help households escape poverty, or induce them to remain in (or fall into) poverty, would lead to a more effective and efficient fight against poverty. Panel data are traditionally employed to answer these questions. Collecting such data, however, can be very costly and can pose a number of logistical and capacity-related challenges. The scarcity of panel data has thus rendered the analysis of welfare dynamics difficult, if not impossible, in many developing country settings.

To overcome the non-availability of (actual) panel data, there have been a variety of efforts to develop pseudo-panels (or synthetic panels) out of multiple rounds of cross-sectional data (see, e.g. Deaton, 1985; Pencavel, 2007; Inoue, 2008; Juodis, 2018). Notably, since cross-section samples are typically refreshed each time that the surveys are fielded, these synthetic panels are possibly less exposed to the concerns surrounding attrition and measurement error that are often levelled at panel data.[2] Yet, because of their emphasis on cohorts rather than the household or individual, synthetic panel methods have not been widely applied to the study of poverty dynamics. Two notable exceptions are Bourguignon, Goh, and Kim (2004) and Guell and Hu (2006) who construct synthetic panels at the household level. However, these two approaches require certain assumptions that may not always be easily satisfied in available cross sections: the former requires at least three rounds of cross section data and assumes a first-order auto-regression (AR(1)) process through which past household or individual incomes (earnings) can affect present outcomes; the latter is exclusively restricted to duration analysis.

Building on a poverty imputation technique described in Elbers, Lanjouw, and Lanjouw (2003), a recent paper by Dang *et al.* (2014) constructs synthetic panels from as few as two rounds of household-level cross sectional data that can provide lower-bound and upper-bound estimates of poverty transitions. Drawing on validation data from Vietnam and Indonesia, this paper finds that the 'true' estimates of poverty mobility (as revealed by the actual panel data) are generally sandwiched between the upper bounds and lower bounds derived from the synthetic panels. However, this method's practical appeal is limited since it often yields rather wide bounds on estimated mobility, and these can be narrowed only if certain key statistical parameters can be imported from externally available panel data.[3]

[1]For example, Reeves (2020) calls for using mobility metrics as the 'measure of the nation' for the US. Poverty mobility also stands out in a December 2013 address by US President Obama to the Center for American Progress (https://www.whitehouse.gov/the-press-office/2013/12/04/remarks-president-economic-mobility). See also Baulch (2011) for a collection of studies on poverty dynamics for developing countries.

[2]See, for example, Glewwe and Jacoby (2000) and Kalton (2009), respectively, for overviews of the advantages and disadvantages of cross sections and panel data in developing and richer country contexts. See also Lee, Ridder, and Strauss (2017) for a recent study that investigates the impacts of measurement errors on poverty mobility using several rounds of panel data from South Korea.

[3]This method focuses on constructing the synthetic panels from two or more rounds of cross sections, each of which has consumption data. See also Gibson (2001) for a somewhat related study on how panel data on a subset of individuals can be used to infer chronic poverty for a larger sample. More broadly, this method is related to the literature on identifying the bounds on the joint distribution for outcomes in different samples (see, e.g. Cross and Manski, 2002) and the statistical literature on imputing missing data (see, e.g. Little and Rubin, 2020). See also Ridder and Moffitt (2007) and Dang, Jolliffe, and Carletto (2019), respectively, for reviews on the econometrics of data combination and poverty imputation.

We propose a significant refinement to the method introduced by Dang *et al.* (2014) to analyse mobility using *only* commonly available cross-sectional survey data. Our new method is predicated on some additional but fairly standard assumptions (based on asymptotic theory) that allow us to move beyond *bound* estimates to actual *point* estimates of poverty mobility. This offers greater accuracy, better interpretation, and potentially much wider application. In particular, we can more easily investigate multiple measures of poverty dynamics, such as the population shares in different poverty categories in both survey periods considered together (i.e. unconditional or joint probabilities) or the population shares in different poverty status categories in one period given their welfare status in the other period (i.e. conditional probabilities). We further provide new formulae for the standard errors on point estimates.

We also make additional contributions on both the time dimension and deeper treatment of income mobility. In particular, we extend the existing method to settings where more than two rounds of data are available to investigate richer inter-temporal profiles of movement into and out of poverty. Our framework also permits more general analysis of mobility among different income groups, rather than just the 2×2 poverty transition matrix. This expands analysis of mobility from merely focusing on the lower part of the income distribution to its entire range and offer relevant inputs for policy advice. For example, as living standards are rising globally and the global poverty rate has been decreasing, more attention is being focused on the vulnerable population groups that are currently not poor but have a high risk of falling into poverty (e.g. World Bank (2017)). As another example, it is common practice to present a 5×5 transition matrix to examine income mobility where this is permitted by available panel data (e.g. Fields (2001)).

On the empirical front, we first validate our estimates with Monte Carlo simulations for various data situations, including settings where variables are only partially observed to one where they are fully observed, as well as different sample sizes. We further implement a number of 'stress tests' of the estimators under deviations from the model assumptions. We subsequently validate our proposed methods with multiple rounds of cross sectional and panel survey data from several countries including Bosnia-Herzegovina, Lao PDR, Peru, the USA and Vietnam. These countries represent diverse settings ranging from developing to high-income countries in different geographical locations, covering both household income (the USA) and household consumption data (the remaining countries). We find that in many cases our synthetic panel estimates are close to those derived from panel data – often lying within the 95% confidence intervals (CIs) or even one SE of the latter.

Recent validations and applications of (earlier versions of) our synthetic panel methods by various researchers for different country contexts ranging from India to Africa, Latin America, and Europe have been yielding encouraging results (Ferreira *et al.*, 2012; Beegle *et al.*, 2016; UNDP, 2016; OECD, 2018; Dang *et al.*, 2019; Salvucci and Tarp, 2021). Even in those cases where our synthetic panel estimates fall outside the CIs surrounding the true panel estimates, the observed qualitative patterns of poverty mobility are generally quite similar between the panel and synthetic panel estimates. Herault and Jenkins (2019) and Garces-Urzainqui (2017) similarly document examples where strict statistical criteria

are not satisfied, but the qualitative conclusions needed for policy design remain fairly robust.[4]

This paper consists of six sections. We discuss the basic framework and theoretical results in the next section, and the Monte Carlo simulation exercise in Section III. Our data are described in Section IV and we report on the empirical validations using actual panel data in Section V. Section VI offers concluding remarks. We leave most of the technical details to Appendix S1, describe in more detail the Monte Carlo simulation in Appendix S2, offer more data description, robustness checks and additional estimation results in Appendix S3, and summarize the estimation procedures in Appendix S4.

II. Analytical framework for point estimates on poverty mobility

Basic framework

Let y_{ij} represent household consumption or income in survey round j for household i, where $i = 1, \ldots, N$, and $j = 1$ or 2. Let x_{ij} be a vector of time-invariant household characteristics that are observed in both survey rounds. Subject to data availability, these characteristics can include such variables as sex, ethnicity, religion, language, place of birth and parental education as well as variables that can be converted into time-invariant versions based, for example, on information about household heads' age and education. The vector x_{ij} can also include time-varying household characteristics if retrospective questions about the round-1 values of such characteristics are asked in the second round survey.

Consider the following projection of household consumption (or income) on household characteristics for survey round j

$$y_{ij} = \beta_j' x_{ij} + \varepsilon_{ij}. \tag{1}$$

We are interested in knowing such quantities of poverty dynamics as

$$P\left(y_{i1} \sim z_1 \text{ and } y_{i2} \sim z_2\right), \tag{2}$$

or

$$P\left(y_{i1} \sim z_1 \mid y_{i2} \sim z_2\right), \tag{3}$$

[4]Herault and Jenkins (2019) also suggest that their poverty mobility estimates based on household survey data from Australia and Great Britain are less accurate than those using data from lower-income countries in other studies. Yet, two notable features stand out from their validation study that may contribute (to some extent) to the lower accuracy in their study. One, their estimated R^2's using household survey data from Australia and Great Britain hover around 0.1–0.2 for regressions with more than 30 independent variables (regressors), which are generally lower than those shown in previous studies using much fewer regressors. For example, our estimated R^2's are predominantly between 0.2 and 0.5 for regressions using seven regressors only (Appendix S3, Table 3.1). Second, between two-thirds and three-fourths of the estimated coefficients on these regressors in Herault and Jenkins (2019) are statistically insignificant, which stand in contrast to the generally strongly statistically significant estimated coefficient in other validation studies. Indeed, adding more regressors in a misspecified model could result in less accurate estimates for both the correlation coefficients and the income model as a whole (Snijders and Bosker, 1994; Nakagawa, Johnson, and Schielzeth, 2017; Luca *et al.*, 2018). A deeper concern raised by Herault and Jenkins (2019), and also echoed in Garces-Urzainqui (2017) and Colgan (2022), relates to the potential sensitivity of results to cohort definition.

where the vector x_{ij} includes a vector of ones, z_j is the poverty line in period j, and the relation sign ($\sim$) indicates either the larger sign ($>$) or smaller or equal sign ($\leq$). For example, $P\left(y_{i1} \leq z_1 \; and \; y_{i2} > z_2\right)$ represents the probability that household i is poor in the first period but non-poor in the second period (considered together for two periods), and $P\left(y_{i2} > z_2 | y_{i1} \leq z_1\right)$ represents the conditional probability that household i, who are poor in the first period, escape poverty in the second period. These probabilities can also be interpreted as population quantities; for example, $P\left(y_{i2} > z_2 | y_{i1} \leq z_1\right)$ corresponds to the percentage of poor households in the first period that escape poverty in the second period. We also refer to those who are either poor or non-poor in both periods as the immobile, and those who escape or fall into poverty over time, respectively, as the upward and downward mobile. For convenience, we also refer to quantities (2) and (3), respectively, as unconditional mobility and conditional mobility.[5]

If panel data are available, we can easily estimate these quantities; otherwise, we can use synthetic panels for this purpose. To further operationalize the framework, we make the following two assumptions.

Assumption 1. The underlying population sampled is the same in survey round 1 and survey round 2.

Assumption 1 ensures that the distributions of the time-invariant household characteristics in the two survey rounds would be the same. As such, these time-invariant household characteristics can be employed as the connectors of household consumption between the two periods (i.e. $x_{i1} \equiv x_{i2}$). Coupled with equation (1), this assumption implies that households in period 2 with identical characteristics to those of households in period 1 would have achieved the same consumption levels in period 1 and vice versa (given the same error term). Assumption 1 will be violated if the underlying population changes due to major events as births, deaths, or migration; these events can be caused by natural disasters or economic crises or simply because the two survey rounds are too far apart. We can thus test this assumption by examining whether the observable time-invariant characteristics of the population of interest change significantly from one survey round to the next.

Assumption 2. ε_{i1} and ε_{i2} have a bivariate normal distribution with zero mean, (the partial) correlation coefficient ρ, and standard deviations σ_{ε_1} and σ_{ε_2}, respectively.

Assumption 2 is basic and fairly standard for most analysis of household consumption (income) data. Since we often convert household consumption y_{ij} to the logarithmic scale for better analysis, the normality assumption for log consumption is equivalent to the log-normality assumption of consumption. Furthermore, we can also employ other popular transformation methods (such as the Box-Cox technique) to make the dependent variable as close to normally distributed as possible.[6] But unlike Assumption 1, the assumption

[5] We restrict our discussion in this paper to a money-metric measure of poverty; for a multidimensional measure see Alkire and Foster (2011). Also see Calvo and Dercon (2009) and Foster (2009) for discussion on other definitions of chronic poverty.

[6] We return to more discussion with the empirical analysis in Section V. See, e.g. Weisberg (2014) for a textbook discussion on transformation methods.

of joint normality is widely used in practice but cannot be tested without panel data. We come back to relaxing this assumption in the empirical analysis.

The partial (conditional) correlation coefficient ρ is usually non-negative in most household surveys, for several reasons. First, since household poverty status tends to be strongly related over time, the joint probability that a household is poor in both survey rounds considered *together* is expected to be higher than the product of the probability that this household is poor in round one and poor in round two. Second, if shocks to consumption or income (for example, finding or losing a job) have some persistence, and consumption reacts to these income shocks, then consumption errors will also exhibit positive autocorrelation. Finally, although some households may experience negatively correlated incomes over time (e.g. reducing expenditure in one period in order to prepare for a wedding in the next), factors leading to such a correlation are unlikely to apply to the majority of households at the same time.

Assumption 2 is also simpler and less data-demanding than the assumptions typically employed in other pseudo panel models that analyse multiple rounds of repeated cross sections. Put differently, we assume that no cohort (or time) specific effects exist; neither do we explicitly assume individual level heterogeneity (as in, for example, Inoue, 2008). While we acknowledge that this rather simplistic departure from the literature could result in more restrictive analysis, it is motivated by the dearth of (even cross-sectional) survey data we typically face with in practice, particularly for poorer countries.[7] Notably, in situations where only two rounds of repeated cross sections exist, Assumption 2 is crucial for implementing our proposed model. But we return to relax this assumption and allow for the cohort fixed effects in the error terms in an alternative approach in Proposition 2. We further discuss potential heterogeneity and heteroscedasticity of the error terms with Monte Carlo simulation in Section III (and Appendix S2) and examine heterogeneity analysis in Section V.3.

If ρ is known, we can estimate quantity (2) by

$$P\left(y_{i1} \sim z_1 \text{ and } y_{i2} \sim z_2\right) = \Phi_2\left(d_1 \frac{z_1 - \beta_1{}' x_{ij}}{\sigma_{\varepsilon_1}}, d_2 \frac{z_2 - \beta_2{}' x_{ij}}{\sigma_{\varepsilon_2}}, \rho_d\right), \tag{4}$$

where $\Phi_2(.)$ stands for the standard bivariate normal cumulative distribution function (cdf), d_j is an indicator function that equals 1 if the household is poor and equals -1 if the household is non-poor in period j, and $\rho_d = d_1 d_2 \rho$.

We discuss next our point estimates method which addresses the limitations of, and significantly extends, the bounds method introduced in Dang *et al.* (2014).

Theoretical estimates for ρ

We offer the following proposition to obtain ρ, which helps provide the point estimate for poverty mobility.

[7] Serajuddin *et al.* (2015) find that, over the period 2002–11, more than one-third (57) of the 155 countries for which the World Bank monitors poverty data have only one poverty data point or no data at all. Even where countries collect data on poverty, these data may not be comparable over time. Indeed, Beegle *et al.* (2016) point out that around half of 48 Sub-Saharan African countries did not have two comparable household surveys for the period 1990–2012.

Proposition 1. Point estimate of ρ

Given equation (1) and Assumptions 1 and 2, and assuming that the simple (unconditional) correlation coefficient between household consumption in two survey rounds $\rho_{y_{i1}y_{i2}}$ is known, the partial correlation coefficient ρ is given by

$$\rho = \frac{\rho_{y_{i1}y_{i2}}\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})} - \beta_1{}'\text{var}(x_i)\,\beta_2}{\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}} \tag{5}$$

Proof. See Appendix S1. □

Central to the estimation of ρ in Proposition 1 is the value of $\rho_{y_{i1}y_{i2}}$, with the latter correlation coefficient being larger than the former correlation coefficient under most circumstances.[8] We propose next a simple way to approximate $\rho_{y_{i1}y_{i2}}$ based on cohort-level averages from the survey data.

Lemma 1. Approximation of $\rho_{y_{i1}y_{i2}}$

Assume the following simple linear projection of household consumption between period 1 and period 2

$$y_{i2} = \delta y_{i1} + \eta_{i2}, \tag{6}$$

where δ is a scalar, η_{i2} is the random error term. Further assume there are no other control (x_{ij}) variables in equation (6) and x_{ij} have no cohort-specific first moment. Also assume that the sample size of each household survey round is large enough (or $N \to \infty$) and the number of cohorts (C) constructed from the survey data is fixed. The simple correlation coefficient $\rho_{y_{i1}y_{i2}}$ can then be approximated with the synthetic panel cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$, where c indexes the cohorts constructed from the household survey data.

See Appendix S1 for further discussion.

We can rely on the existing literature on pseudo-panel data to construct cohorts. For example, cohorts can be based on age (Deaton, 1985; Pencavel, 2007) or some combination of age and other characteristics such as education (e.g. Blundell, Duncan, and Meghir, 1988) or region (e.g. Propper, Rees, and Green, 2001). In the same spirit, other time-invariant characteristics such as gender or ethnicity may also qualify as candidates for cohort construction. The implicit assumption underlying traditional pseudo-panel analysis is that cohort dummy variables have a strong relationship with household consumption.[9] The assumption stated in Lemma 1 on a fixed number of cohorts is standard in the

[8] ρ typically has an inverse relationship with the R^2's obtained from equation (1). Put differently, ρ is smaller when the income equation has a better model fit; see Appendix S1 for further discussion.

[9] In addition, we can obtain good estimates of correlation at the cohort-level aggregated data if the individual data within a cohort show very similar values (or the intraclass correlation is close to 1 (Snijders and Bosker, 2011)). Furthermore, if these cohort dummy variables do not capture any variation in household consumption, the synthetic panel cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$ would simply be 0. In the extreme case, consumption (or poverty) mobility can happen entirely within cohorts, but this case would be easily detected with the cohort means remaining largely unchanged across the two survey rounds (or $\rho_{y_{c1}y_{c2}}$ changes very little over time in contexts with three or more cross sectional survey rounds). We return to more discussion in the next section on Monte Carlo simulation.

traditional pseudo-panel literature (Moffitt, 1993; Verbeek and Vella, 2005; Juodis, 2018) and helps preclude measurement errors with cohort means. It is also defined as the Type 1 asymptotics of pseudo panel data (Verbeek, 2008).

Notably, given the small number of cohorts in practice (where we may have only two rounds of repeated cross sections), we do not include other control variables in equation (6). Similar to Assumption 2 that is discussed earlier, equation (6) represents a simplification of the typical linear dynamic model employed in the pseudo-panel literature due to data constraints (see, e.g. Moffitt (1993)). The assumption that x_{ij} have no cohort-specific first moment helps ensure that δ is consistently estimable when it is linked to equation (1) (Inoue, 2008). Lemma 1 can be straightforwardly extended to multiple waves to obtain $\rho_{y_{cj}y_{ck}}$ for any pair of survey rounds j and k, but a longer time interval between survey rounds tends to decrease $\rho_{y_{cj}y_{ck}}$. For example, Kopczuk, Saez, and Song (2010) find that the (rank) correlation of earnings decreases over longer time intervals for panel data from the US Social Security Administration between 1937 and 2004.[10]

Alternatively, we can instead assume that the number of cohorts is large enough (instead of being fixed). This is the Type 2 asymptotics of pseudo panel data, which was proposed by Deaton (1985) and subsequently used in various studies including Verbeek and Nijman (1993) and Collado (1997). Using this different assumption allows us to employ a richer assumption for the error terms that includes the cohort fixed effects in the error term, which we refer to as Assumption 3 below.

Assumption 3. Further assume that the error terms ε_{i1} and ε_{i2} include a cohort fixed effect.

This offers another way of estimating ρ.

Proposition 2. - Alternative estimate of ρ

Given equation (1) and Assumptions 1, 2 and 3, and assume that the sample size of each household survey round is large enough (or $N \to \infty$) and the number of cohorts (C) constructed from the survey data is large enough (or $C \to \infty$). The partial correlation coefficient ρ can then be estimated from a modified version of equation (1) where all the variables are aggregated to the cohort level

$$y_{cj} = \beta_j' x_{cj} + \varepsilon_{cj}, \tag{7}$$

where the error term $\varepsilon_{\mathrm{cj}}$ includes a cohort fixed effect τ_{c} and the error ν_{cj}.

Proof. See Appendix S1. □

Notably, the different assumptions over whether the number of cohorts is fixed (Lemma 1) or goes to infinity (Proposition 2) result in two different ways to construct cohorts. Lemma 1 suggests that we can construct cohorts based on age (or age interacted with another variable), but Proposition 2 suggests that we can construct cohorts based on a combination of all the different values of the time-invariant variables in x_{ij}. The latter approach provides many more cohorts than the former, if there are enough time-invariant

[10]This result also holds for actual panel data from various other countries such as China, India, Peru, Vietnam, and the UK (Chaudhuri and Ravallion, 1994; Khor and Pencavel, 2006; Jenkins, 2011; our estimates).

variables. For instance, using the US's Panel Study of Income Dynamics (PSID) data in 2007–09 with a sample size of around 3,400 observations, the number of constructed cohorts is 31 with Lemma 1 (using age as the cohort variable, with a restriction of heads between age 25 and 55), but the corresponding figure using Proposition 2 (for a combination of age, gender, years of schooling, ethnicity, and urban residence) is 1,120. Given a typical sample size of 5,000 for most current household surveys, the number of cohorts (and cohort cell sizes) can be slightly larger.

While it appears reasonable to assume that N tends to infinity with most current household surveys, there is no consensus in the literature on how large cohort sizes should be. Monte Carlo simulations by Verbeek and Nijman (1992) suggest that cohort sizes of 100–200 are sufficient, while Devereux (2007) argues for larger cohort sizes in the thousands. Khan (2021) offers a new metric to calculate cell sizes; yet, it is a complex function that is sensitive to variations within and across cohorts, over time for cohorts, as well as autocorrelation and covariance of the control variables. Indeed, our validation results, shown in Section V, suggest that we can obtain reasonably good estimates for total sample sizes ranging from slightly more than 1,300 observations (Bosnia-Herzegovina) to 9,100 observations (Peru) and the results do not appear to strongly depend on the sample sizes.

We note the caveat that Lemma 1 provides approximates of $\rho_{y_{i1}y_{i2}}$, and both Proposition 1 and Proposition 2 are based on asymptotic theory (using a large number of cross sections). How well these estimates turn out to be in practice (using only two rounds of cross sections) is an empirical issue. A simple (but partial) diagnostic test for Proposition 1 to work is that the cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$ is statistically different from 0; the corresponding test for Proposition 2 is that the β_j's in equation (7) are jointly statistically different from 0. We offer a sample Stata command in Appendix S4 to estimate equation (7). Our preferred method for the empirical illustrations in this paper is Proposition 1 and Lemma 1, since this approach lays out more clearly the relationship between $\rho_{y_{i1}y_{i2}}$ and ρ. But we also use Proposition 2 for alternative estimates.[11]

Mobility for three (or more) periods or consumption groups

We next provide Proposition 3, which shows the asymptotics of the point estimates in equation (4).

Proposition 3. Asymptotic results for point estimates for two periods.

Assume that equation (1) and Assumptions 1 and 2 hold, and assume further that all the standard regularity conditions are satisfied for equations (1), (i.e. $X'\varepsilon/N \xrightarrow{p} 0$ and $X'X/N \xrightarrow{p} M$ finite and positive definite).[12] Let P be the population parameter of interest (e.g. $P = P\,(y_{i1} \leq z_1 \; and \; y_{i2} > z_2)$ for household i, i = 1, . . . , N), d_j an indicator function that equals 1 if the household is poor and equals −1 if the household is non-poor

[11] We further discuss theoretical bounds on ρ and another way to approximate it in Appendix S1.

[12] As is the usual practice, the vectors of time-invariant characteristics x_i's ($k \times 1$) are transposed into row vectors and stacked on top of each other to form the matrix X ($n \times k$), and the vectors of error terms ε's ($n \times 1$) are formed similarly from the scalars ε_i's.

in period j, j = 1, 2, $\rho_d = d_1 d_2 \rho$, and $\rho_{y_{i1}y_{i2},d} = d_1 d_2 \rho_{y_{i1}y_{i2}}$, and the (ˆ) sign represents the estimate. Our point estimates are distributed as

$$\sqrt{n}\left[P - \Phi_2\left(d_1 \frac{z_1 - \widehat{\beta}_1' x_{ij}}{\widehat{\sigma}_{\varepsilon_1}}, d_2 \frac{z_2 - \widehat{\beta}_2' x_{ij}}{\widehat{\sigma}_{\varepsilon_2}}, \widehat{\rho}_d\right)\right] \sim N(0, V), \tag{8}$$

where $\widehat{\Phi}_2(.) = \Phi_2\left(d_1 \frac{z_1 - \widehat{\beta}_1' x_{ij}}{\widehat{\sigma}_{\varepsilon_1}}, d_2 \frac{z_2 - \widehat{\beta}_2' x_{ij}}{\widehat{\sigma}_{\varepsilon_2}}, \widehat{\rho}_d\right)$ is the estimated quantities of poverty dynamics for household *i*.

The covariance–variance matrix V can be decomposed into two components, one due to sampling errors (Σ_s) and the other due to model errors (Σ_m) assuming these two errors are uncorrelated such that $V = \Sigma_s + \Sigma_m$.

Proof. See Appendix S1. □

Several remarks are in order for this proposition. First, given a better fit for our regressions in equation (1), the model-based variances (i.e. synthetic panel estimates in our case) are usually smaller than the design-based variances (i.e. weighted estimates based on panel data) (Matloff, 1981; Binder and Roberts, 2009). Furthermore, a larger sample size would reduce the sampling variance; thus, this points to the advantages of cross sections over panel data when the former have larger sample sizes than the latter (see Appendix S4 for more discussion). While the reduction of variance can vary depending on the specific model or datasets under consideration (Binder and Roberts, 2009), our estimation results (Table 3) show that the model-based variances for the synthetic panels hover around 10%–50% of those for the design-based variances for different countries.[13]

Second, we can use data either from the first or the second survey round as the base year for Proposition 3, given the following identity

$$P(y_{i1} \leq z_1 \text{ and } y_{i2} > z_2) \equiv P(y_{i2} > z_2 \text{ and } y_{i1} \leq z_1). \tag{9}$$

We provide next Proposition 4 that further extends Proposition 3 to settings with more than two consumption groups.

Proposition 4. Asymptotic results for point estimates for mobility between different groups for two periods.

Given the same assumptions in Proposition 3, let P^{lm} represent household i's (i = 1, . . . , N) probability of moving from consumption group l in period 1 to consumption group m in period 2, that is $P^{lm} = P\left(z_1^{l-1} < y_{i1} \leq z_1^{l} \textit{ and } z_2^{m-1} < y_{i2} \leq z_2^{m}\right)$, where $l, m = 1, \ldots, k$, and z_j are the thresholds that separate the different consumption groups, with $z_j^0 = -\infty$ and $z_j^k = \infty$, for period j, j = 1, 2. Defining $F^{l,m}$ as $\Phi_2\left(\frac{z_1^l - \beta_1' x_{ij}}{\sigma_{\varepsilon_1}}, \frac{z_2^m - \beta_2' x_{ij}}{\sigma_{\varepsilon_2}}, \rho\right)$, and the (ˆ) sign represents the estimate, our point estimates are distributed as

$$\sqrt{n}\left[P^{lm} - \left(\widehat{F}^{l,m} - \widehat{F}^{l,(m-1)} - \widehat{F}^{(l-1),m} + \widehat{F}^{(l-1),(m-1)}\right)\right] \sim N(0, V). \tag{10}$$

[13]Our results are consistent with the findings in Binder and Roberts (2009), where the largest reduction in variances can depend on other factors and not just sample size differences. In particular, the reduction in variances for the synthetic panels are rather similar for Lao PDR and Peru, despite the ratio of the sample size for the cross sections over that of the actual panel is four times and 1.6 times for Peru and Lao PDR, respectively.

Proof. See Appendix S1. □

We provide in Appendix S1 several additional theoretical results. These include Corollary 3.1 (which provides the asymptotic results for conditional probabilities) and Proposition 5 (which extends Proposition 3 to the general setting where there are three or more survey rounds, i.e. $j \geq 3$).

III. Monte Carlo simulation

We first start in this section with assuming that both Assumptions 1 and 2 are satisfied before examining situations where these assumptions can be relaxed. Assume that household i's consumption can be generated for both periods using the following model

$$y_{i1} = \alpha_1 + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \beta_{13}x_{i3} + \beta_{14}x_{i4} + \beta_{15}x_{i5} + \beta_{16}x_{i6} + \beta_{17}x_{i7} + \beta_{18}x_{i8} + v_{i1}, \tag{11}$$

$$y_{i2} = \alpha_2 + \beta_{21}x_{i1} + \beta_{22}x_{i2} + \beta_{23}x_{i3} + \beta_{24}x_{i4} + \beta_{25}x_{i5} + \beta_{26}x_{i6} + \beta_{27}x_{i7} + \beta_{28}x_{i8} + v_{i2}, \tag{12}$$

where x_i's are household head's time-invariant characteristics, and v_i's the random error terms. We choose eight regressors for equations (11) and (12) to better mimic situations where we can employ up to seven time-invariant regressors when working with real household survey data (Appendix S3, Table 3.1).

Also assume the following parameter values

$$\alpha_1 = 1; \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = \beta_{17} = \beta_{18} = 1$$

$$\alpha_2 = 1.5; \beta_{21} = 1.2, \beta_{22} = 1.1, \beta_{23} = 1.05, \beta_{24} = 1.3, \beta_{25} = 0.9,$$

$$\beta_{26} = 1.15, \beta_{27} = 1.4, \beta_{28} = 0.6,$$

and

$$x_{i1} \sim N(0, 2.5), x_{i2} \sim N(0, 5), x_{i3} \sim N(0, 6), x_{i4} \sim N(0, 4), x_{i5} \sim N(0, 1), x_{i6} \sim N(0, 3),$$

$$x_{i7} \sim N(0, 2), x_{i8} \sim N(0, 1),$$

$$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6.5 & 1 \\ 1 & 6.5 \end{pmatrix}\right),$$

where $N(0, c)$ stands for the normal distribution with mean 0 and variance c; BVN(.,.) similarly represents the bivariate normal distribution with the vector of mean 0 and the given variance–covariance matrix. Without loss of generality, we assume a certain degree of correlation over time for the error terms v's ($\rho = 0.15$), which may be caused by time-varying factors such as unexpected shocks. Given these parameter values, we can calculate that $\text{var}(y_{i1}) = 27$, $\text{var}(y_{i2}) = 38.6$, $\rho_{y_{i1}y_{i2}} = 0.89$, as well as a range of values for R^2 for different equations that employ different numbers of time-invariant regressors (Appendix S2, Table 2.1).

The values for β's are motivated by the estimates for equation (1) using real household survey (Appendix S3, Table 3.1). For example, the ratios of the estimated coefficients between the two repeated cross sections range from 0.60 to 1.11 for Vietnam during 2006–08 and 0.73 to 1.39 for the USA during 2007–09.

We choose the value of 0.15 as a lower value of ρ (say, rather than 0) for two main reasons. First, in theory, it tends to be positive as earlier discussed with Assumption 2. Indeed, a zero correlation coefficient implies perfect income mobility between two periods (i.e. an average household's income in the second period has no relationship with its income in the first period), which rarely occurs perhaps except under extremely special circumstances such as overnight regime change. Second, empirical evidence using actual panel data from various countries suggest that ρ is often (much) larger than this value. For example, Khor and Pencavel (2006) estimate ρ to be 0.54 for China, and range from 0.62 to 0.78 for various richer countries such as Denmark, France, Germany, Italy, Sweden, the UK, and the USA in the late 1980s and early 1990s. Estimates by Dang *et al.* (2014) put ρ at 0.39 for Nepal (1995/96–2003/04) and 0.50 for Indonesia (1997–2000). Our estimation results (Table 1) suggest ρ ranges from 0.43 to 0.70 for countries with different income levels, such as Bosnia-Herzegovina, Lao PDR, the United States, Peru, and Vietnam during the 2000s. We discuss in more detail the Monte Carlo simulation procedures in Appendix S2.

We examine three main different data situations. These range from the most data-scarce situation where we only observe x_1 (i.e. $\rho = 0.88$ and is almost identical to $\rho_{y_{i1}y_{i2}}$) to a typical setting with a few such variables (i.e. $\rho = 0.58$), and to an unusual setting where we fully observe all the x's (i.e. $\rho = 0.15$). These data situations correspond to Models 1, 5 and 8 in Table 2.1, which also provides the values for ρ for additional data situations. We provide simulation results for these models at three different sample sizes $N = 1{,}000$ (small), 4,000 (medium), and 10,000 (large), with 1,000 simulations for each model run. We fix the poverty line in period 2 at the 30th percentile, and then graph in Figure 2 the true percentage of households that are poor in both periods (solid line), its 95% CIs (shaded bands), and the estimated percentage using simulated data (dashed line) against the whole spectrum of poverty rates in the first period.

Figure 1 shows that the estimated poverty rates generally track the true rates and fall within their 95% CIs. Unsurprisingly, more time-invariant variables result in better predictions. In other words, a lower ρ (resulting from a better model fit) helps improve accuracy.[14] Indeed, the dashed lines are almost indistinguishable from the solid line for the graphs where $\rho = 0.15$ (with the R^2's hovering around 0.8) or $\rho = 0.58$ (with the R^2's hovering around 0.5–0.6). When very limited information exists on the time-invariant variables ($\rho = 0.88$, or both the R^2's equal 0.09), the estimates (partially) fall outside the 95% CIs for the middle part of the distribution for mid-sized or unusually large sample sizes ($N = 4{,}000$ or 10,000), but still compares favourably well to the true poverty rates for small sample sizes ($N = 1{,}000$). Varying the model parameters or the poverty lines gives us similar results (not shown).

The results remain reasonably robust where we relax Assumptions 1 and 2 in various ways. These include situations where the time-invariant household characteristics x's have

[14]This further highlights the importance of the explanatory of the income model as discussed earlier.

TABLE 1
Estimated ρ from actual panels and synthetic panels for different countries

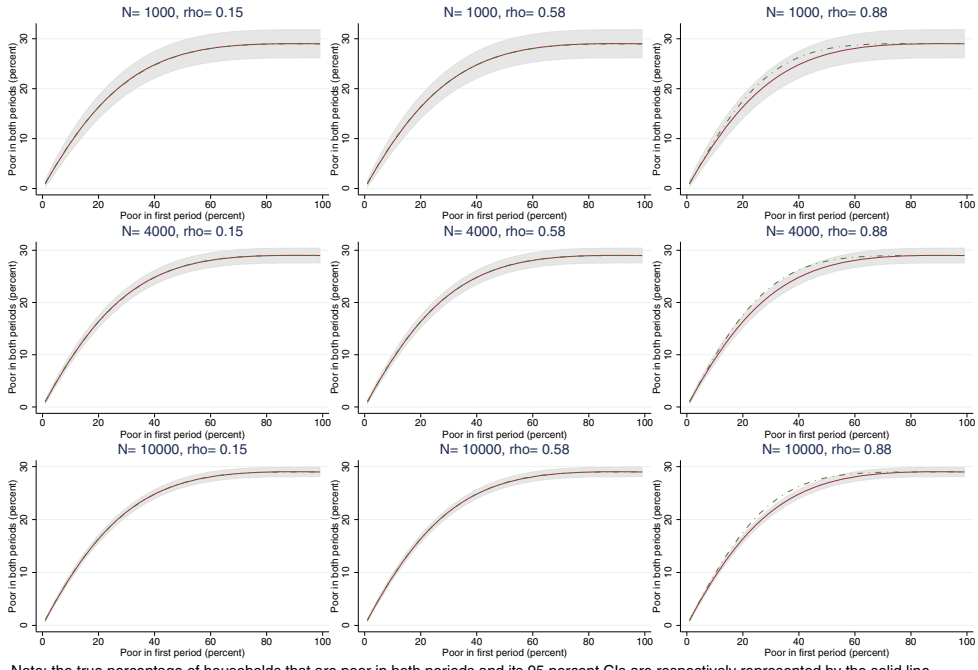
				Synthetic panels		
		Actual panels		*Method 1*		*Method 2*
Country	*Survey year*	$\rho_{y_{i1}y_{i2}}$	ρ	$\rho_{y_{i1}y_{i2}}$	ρ	ρ
Bosnia- Herzegovina	2001 2004	0.48	0.45	0.43	0.40	0.61
Lao PDR	2002–03 2007–08	0.51	0.43	0.56	0.46	N/A
Peru	2004 2005	0.82	0.64	0.82	0.69	0.67
	2005 2006	0.82	0.66	0.80	0.63	0.68
	2004 2006	0.79	0.63	0.73	0.51	0.68
Vietnam	2004 2006	0.81	0.66	0.85	0.73	0.61
	2006 2008	0.78	0.63	0.85	0.76	0.62
	2004 2008	0.75	0.58	0.84	0.74	0.47
USA	2005 2007	0.76	0.66	0.89	0.84	0.72
	2007 2009	0.82	0.70	0.86	0.79	0.74
	2005 2009	0.72	0.57	0.71	0.59	0.56

Note: The synthetic panel estimates are based on cross sectional data except for Bosnia-Herzegovina and the USA, where these estimates are based on two rounds of actual panel data. $\rho_{y_{i1}y_{i2}}$ is the simple correlation across two survey rounds for household consumption for all countries except for the USA, where it is the correlation for household income. ρ is the partial correlation, conditional on household head's gender, years of schooling, ethnicity, and residence areas. All estimates for $\rho_{y_{i1}y_{i2}}$ and ρ are significant at the 0.01 level. Household heads' ages are restricted to between 25 and 55 in the first survey round and adjusted accordingly for the second survey round.

different distributions or are correlated with each other, or where ρ may vary for different population groups, or the errors are heteroskedastic errors (Appendix S2).

IV. Data

To validate our method with real survey data, we analyse household panel survey data from Bosnia-Herzegovina (Bosnia-Herzegovina Living Standards Measurement Survey) in 2001–04, Lao PDR (Expenditure and Consumption Survey, LECS) in 2002/03–2007/08, the USA (PSID) in 2005, 2007 and 2009, Peru (Peruvian National Household Survey, ENAHO) in 2004, 2005 and 2006, and Vietnam (Vietnam Household Living Standards Survey, VHLSS) in 2004, 2006 and 2008. The number of households comprises 2,376 households for Bosnia-Herzegovina, 6,500 households for the LECS, 9,189 households for each round of the VHLSSs, more than 5,000 households for the PSIDs, and almost 20,000 households for the ENAHOs. These data are of high quality and are typically



Note: the true percentage of households that are poor in both periods and its 95 percent CIs are respectively represented by the solid line and shaded bands; the estimated percentage using simulated data is represented by the dashed line.

Figure 1. Predicted poverty rates vs. true poverty rates for two periods based on simulated data [Colour figure can be viewed at wileyonlinelibrary.com]

employed to produce estimates of poverty and income for these countries. We discuss further details of these datasets in Appendix S3.

Consistent with the literature on pseudo-panel data, we restrict the household heads' age range to 25−55 for the first survey round and adjust this appropriately for later survey rounds to ensure stable household formation (e.g. looking at the age cohort 27−57 if the next survey round is 2 years later). While this age range can be extended to include older people, it may be ill-advised to include those who are younger, at least since most household heads tend to be older than 25 in all the countries we look at. The time-invariant variables that we use include the household head's age, years of schooling, ethnicity (i.e. whether belonging to ethnic majority groups), and whether the household resides in urban areas.[15] We provide the estimated parameters for equation (1) in Appendix S3, Table 3.1.

V. Empirical validation for poverty dynamics

Our Monte Carlo simulation suggests that the proposed method works reasonably well even when our assumptions are not fully satisfied (Section III and Appendix S2). We examine in this section how well the method performs with real household survey data.

[15]In contexts where there is (much) migration, the urban residence dummy variable may not satisfy Assumption 1. We return to testing this assumption with real household survey data in Section V.1.

Testing assumptions and estimates for ρ

Regarding testing for Assumptions 1 and 2 using the real household survey data, since Assumption 1 is automatically satisfied for true panel data, we only test it for the cross-sectional components with Lao PDR, Peru and Vietnam. The *t*-tests for the null hypothesis that the distributions of the time-invariant variables are the same across survey rounds are not rejected at the 5% level for the latest period for Peru and Vietnam, but not for Lao PDR. While this suggests that Lao PDR may not offer the best data for validation purposes, we still show validation results for this country since these differences may not be practically very large (e.g. half a year of schooling between the two rounds).[16] Assumption 2 is not testable for the cross sections, but can be tested using the actual panels. Formal multivariate normality tests, including the Doornik and Hansen (2008) test, reject the hypothesis of univariate or bivariate normality distribution for all the countries. Nevertheless, plotting the estimated error terms (ε_{ij}) for both the cross sections and the panel data against the normal distribution (see Appendix S3, Figure 3.1 for Lao PDR, Peru and Vietnam only to save space), suggests that the former approximate the latter fairly closely in practice for each year.[17]

We next discuss the estimates for ρ. After obtaining an estimate for $\rho_{y_{i1}y_{i2}}$ from the synthetic panels based on age cohorts (using Lemma 1) – which are all highly statistically significant with p-values less than 0.01 – we provide the synthetic panels estimates for ρ (using Proposition 1) in Table 1, column Method 1. Estimates using the synthetic panels deviate from those using the actual panels from 0.02 (the USA during 2005–09) to 0.16 (Vietnam during 2004–08) in absolute terms, corresponding to a range of 4%–28% in relative terms. This is within the range of ±30%, where our Monte Carlo simulation (Appendix S2) indicates that estimates remain robust. Furthermore, estimates for ρ are smaller than those for $\rho_{y_{i1}y_{i2}}$, which is consistent with our earlier theoretical discussion.[18]

Alternatively, we also estimate ρ using Proposition 2. Assumption 3 for the cohort fixed effects is satisfied for all the countries, except for Lao PDR so Proposition 2 does not apply for this country. The estimates for ρ (column Method 2 for the synthetic panels) are somewhat better than the estimates using Method 1 for three countries: Peru, Vietnam and the USA, but are worse for Bosnia-Herzegovina. Note, however, that estimates for ρ are just an intermediate input in the estimation of poverty mobility, which is the focus of our analysis.

[16]Assumption 1 is also satisfied for Vietnam in 2004–06, and mostly satisfied for Peru in 2004–05 except for heads' years of schooling and urban residence. However, similar to Lao PDR, these differences appear not very large (e.g. a difference of 0.2 years of schooling between two rounds). Our earlier Monte Carlo simulation results suggest that this assumption can be violated to some extent.

[17]We convert all the countries' consumption (income) variables to logarithmic scale. For the PSID, we further apply Box-Cox transformation to remove skewness for log income. Still, as an alternative to making a parametric bivariate normal distribution as in Assumption 2, we also experiment with relaxing this assumption and employ a copula approach. Estimation results are rather similar and are further discussed in Appendix S3.

[18]Estimation results using a variant of Method 1 (Corollary 1.1 in Appendix S1) are very similar to those using Proposition 1, with the differences being at most 0.01.

Overall poverty mobility

It can be useful to briefly examine the performance of the bound estimates first. To save space, we show in Table 2 the bound estimates for unconditional poverty dynamics using the latest two survey rounds available for each country. While all the true poverty rates are reassuringly encompassed within the estimated bounds, the bound estimates are generally quite wide and can be hard to interpret. For example, the true upward and downward mobility rates for Vietnam are respectively 5.9% and 4.9%. Yet, the bound estimates for both upward mobility and downward mobility for this country are almost identical at [0.5, 9.8] and [0.6, 9.9].[19] This points further to the value of seeking improvements on the bound estimates.

We show the point estimates for the same countries and periods in Table 3, using data in the second survey round (x_{i2}) as the base year for predictions. To evaluate the goodness-of-fit for estimation results, we show comparison with the 95% CIs and one SE around estimates based on the panels. We also consider the efficiency of the synthetic panel estimates by looking at the proportion of the overlap between the 95% CIs of the synthetic panel estimates and the true estimates over the 95% CI of the synthetic panel estimates. The larger this overlap, the more efficient the synthetic panel estimates are; for instance, an overlap of 100% indicates that the 95% CIs of the synthetic panel estimates falls well within those of the true estimates. We show both the averaged proportions of the overlap (or mean coverage) for all the dynamics calculations and the number of times that the overlap reaches 100% (coverage of 100%).

Results appear very encouraging with the synthetic panel point estimates being close to the true point estimates and lying within the 95% CIs around the true estimates for all the cases (i.e. 20 out of 20). Furthermore, more than half of the synthetic panel point estimates fall within one standard error of the actual panel estimates (i.e. 11 out of 20). For the efficiency tests, the mean coverage ranges from 83% to 100% and there is 100% overlap for more than four fifths (i.e. 17 out of 20) of the cases.

In addition, for the USA and Bosnia-Herzegovina where the sample size is the same for both actual panel and synthetic panel estimates, the standard errors for the latter are smaller than those for the former, which is consistent with our earlier discussion. We discuss in Appendix S3 various robustness checks including using data in the first survey round as the base year, or data in earlier survey rounds (x_{i1}), or bootstrap SEs, or using a copula approach as an alternative to assuming a bivariate normal distribution for the error terms.[20]

Further extensions

We further extend the proposed method to provide estimates for population subgroups; it is important to do so for at least two reasons. First, policymakers are usually interested in focusing on smaller population groups rather than the whole population in designing social

[19]We provide the bound estimates for the conditional mobility rates in Appendix S3, Table 3.2. These bounds form even wider intervals than those shown in Table 2.

[20]We provide the point estimates for the conditional mobility rates in Appendix S3, Table 3.3. Estimation results are, unsurprisingly, slightly less accurate than those in Table 3 since both the numerators and denominators in the ratios in Corollary 3.1 are estimated.

TABLE 2

Estimated bounds on poverty dynamics based on synthetic data for two periods, joint probabilities (percentage)

	Bosnia- Herzegovina		*Lao PDR*		*Peru*		*USA*		*Vietnam*	
Poverty status	*2001–04*		*2002/03–2007/08*		*2005–06*		*2007–09*		*2006–08*	
First period and second period	*Actual panel*	*Synthetic panel*	*Actual panel*	*Synthetic panel*	*Actual Panel*	*Synthetic Panel*	*Actual Panel*	*Synthetic Panel*	*Actual Panel*	*Synthetic Panel*
Poor, Poor	10.3 (1.7)	[4.7, 18.0]	13.8 (1.2)	[8.5, 24.1]	29.9 (1.3)	[23.2, 40.7]	6.0 (0.4)	[2.5, 8.8]	9.9 (0.8)	[4.7, 14.0]
Poor, Non-poor	12.6 (1.2)	[2.8, 16.1]	14.3 (1.1)	[2.3, 17.9]	11.6 (0.9)	[2.5, 20.0]	3.8 (0.3)	[0.6, 6.9]	5.9 (0.5)	[0.5, 9.8]
Non-poor, Poor	10.5 (1.4)	[2.2, 15.6]	10.9 (1.0)	[0.5, 16.1]	8.9 (0.8)	[0.2, 17.7]	4.6 (0.4)	[1.4, 7.7]	4.9 (0.5)	[0.6, 9.9]
Non-poor, Non-poor	66.5 (2.2)	[63.7, 77.0]	61.0 (1.6)	[57.5, 73.1]	49.7 (1.6)	[39.1, 56.6]	85.7 (0.6)	[82.9, 89.2]	79.3 (1.0)	[75.5, 84.8]
N	1,342	1,342	1,989	3,215	2,250	9,084	3,368	3,368	2,723	3,701

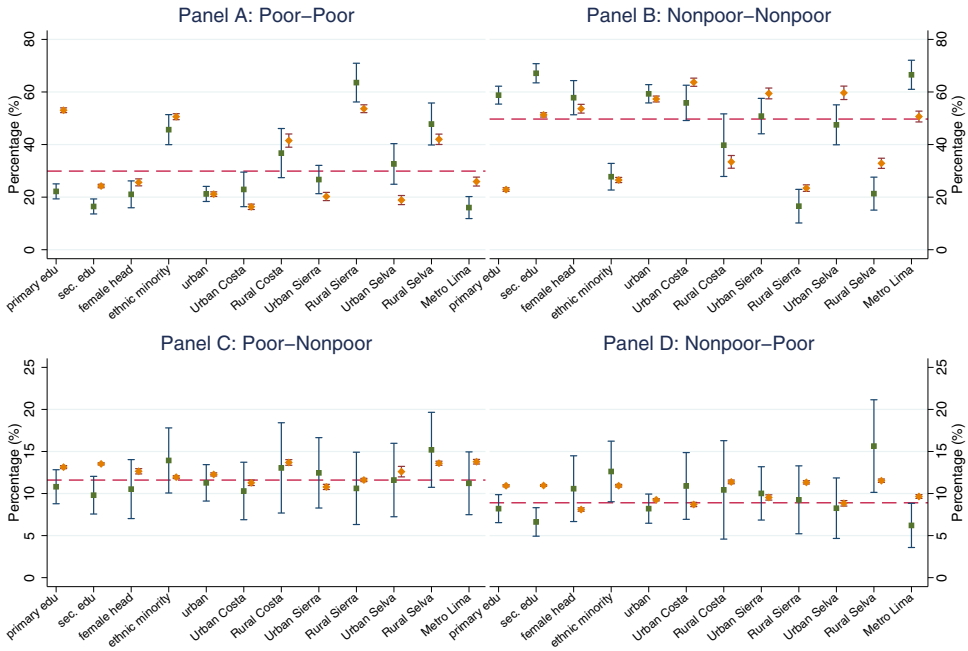
Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the USA. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. The estimated bounds are shown in brackets under the 'Synthetic Panel' for each country. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percentage. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round.

TABLE 3

Poverty dynamics based on synthetic panel data for two periods, joint probabilities (percentage)

	Bosnia- Herzegovina		*Lao PDR*		*Peru*		*USA*		*Vietnam*	
Poverty status	*2001–04*		*2002/03–2007/08*		*2005–06*		*2007–09*		*2006–08*	
First period and second period	*Actual panel*	*Synthetic panel*	*Actual panel*	*Synthetic panel*	*Actual Panel*	*Synthetic Panel*	*Actual Panel*	*Synthetic Panel*	*Actual Panel*	*Synthetic Panel*
Poor, Poor	10.3	8.2	13.8	13.2	29.9	30.9	6.0	6.2	9.9	9.6
	(1.7)	(0.2)	(1.2)	(0.4)	(1.3)	(0.4)	(0.4)	(0.2)	(0.8)	(0.3)
Poor, Non-poor	12.6	12.6	14.3	13.2	11.6	12.3	3.8	3.2	5.9	4.9
	(1.2)	(0.3)	(1.1)	(0.1)	(0.9)	(0.1)	(0.3)	(0.1)	(0.5)	(0.1)
Non-poor, Poor	10.5	12.1	10.9	11.4	8.9	10.0	4.6	4.0	4.9	5.0
	(1.4)	(0.2)	(1.0)	(0.2)	(0.8)	(0.1)	(0.4)	(0.1)	(0.5)	(0.1)
Non-poor, Non-poor	66.5	67.2	61.0	62.2	49.7	46.8	85.7	86.6	79.3	80.4
	(2.2)	(0.6)	(1.6)	(0.6)	(1.6)	(0.4)	(0.6)	(0.3)	(1.0)	(0.4)
Goodness-of-fit tests										
Within 95% CI	4/4		4/4		4/4		4/4		4/4	
Within 1 standard error	2/4		4/4		2/4		1/4		2/4	
Mean coverage (percent)	100		100		91.6		83.0		100	
Coverage of 100%	4/4		4/4		3/4		2/4		4/4	
N	1,342	1,342	1,989	3,215	2,250	9,084	3,368	3,368	2,723	3,701

Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the USA. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. SEs are obtained adjusting for complex survey design for all countries, except for the US PSID. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percentage. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. The 'Within 95% CI' row shows the number of times that the estimates based on the synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the 'Within 1 standard error' row shows a similar figure but using one standard error around the estimates based on the actual panels. The 'Mean coverage (percent)' row shows the mean proportion of the 95% CIs around the synthetic panel estimates that overlap with those based on the actual panels; the 'Coverage of 100%' row shows a similar figure for the number of times that the former fall completely inside the latter.



Note: dashed lines represent the country average; green and orange represent the synthetic panel estimates and the actual values respectively; the whiskers stand for the 95 % CIs.

Figure 2. Profiles of poverty mobility, Peru 2005–06 [Colour figure can be viewed at wileyonlinelibrary.com]

safety net programs; and second, synthetic panels usually have larger sample sizes than panel data, which can help improve estimation accuracy for these population subgroups. We plot the estimated rates with their 95% CIs for the absolute measures of poverty dynamics against the true rates for the population categorized by ethnicity (i.e. ethnic minority groups), gender of household heads (i.e. female-headed households), education achievement (i.e. primary education or higher, lower secondary education or higher), and residence areas (i.e. urban households or regions the household live in) for Peru in Figure 2. Not surprisingly, the 95% CIs for synthetic panels estimates are much smaller than those for the true rates with the gaps between the standard errors amplified roughly twice (i.e. multiplied by 1.96). Our estimates appear to be reasonably good, and fall within the 95% CIs for the true rates around half of the times for the immobile; the corresponding figure is three-fourths or more for the mobile.

We next show in Table 4 the estimated consumption quintile transition matrix using data from Vietnam in 2006–08, where the actual and synthetic panel estimates are shown in panel A and panel B, respectively. Estimates are off with some of the row and column totals (which sum up to 20% by definition), but we focus on the inner transitions since the former do not offer as much insight into mobility as the latter.[21] Estimation results

[21]The row or column totals should sum up to 20% by definition and serve mostly as an indicator of prediction accuracy for these totals only. In addition, it may be useful to highlight the fact that our validation is predicated on the assumption that the true panel data for Vietnam have good quality. If the mobility in the true panel data is partly caused by spurious changes due to measurement errors (or attrition bias) in household consumption, our estimates based on the synthetic panel data would be more accurate since cross sections are free of such data issues.

TABLE 4
Consumption dynamics for two periods, Vietnam 2006–08 (percentage)

			2008					
			Poorest	*Quintile 2*	*Quintile 3*	*Quintile 4*	*Richest*	*Total*
Panel A: True panels	2006	Poorest	12.7	4.7	1.7	0.6	0.2	19.7
			(0.8)	(0.4)	(0.3)	(0.2)	(0.1)	(0.9)
		Quintile 2	4.8	7.5	4.6	2.0	0.6	19.6
			(0.4)	(0.6)	(0.5)	(0.3)	(0.1)	(0.9)
		Quintile 3	1.8	5.2	6.9	4.6	1.5	20.0
			(0.3)	(0.5)	(0.5)	(0.5)	(0.2)	(0.9)
		Quintile 4	0.6	2.0	5.0	7.8	4.8	20.2
			(0.2)	(0.3)	(0.5)	(0.6)	(0.5)	(0.9)
		Richest	0.1	0.6	1.8	4.9	12.9	20.5
			(0.1)	(0.2)	(0.3)	(0.5)	(0.7)	(0.8)
		Total	20.0	20.0	20.0	20.0	20.0	100
			(1.0)	(0.9)	(0.9)	(0.9)	(0.9)	
			2008					
			Poorest	*Quintile 2*	*Quintile 3*	*Quintile 4*	*Richest*	*Total*
Panel B: Synthetic panels	2006	Poorest	**13.7**	3.6	**1.6**	**0.4**	**0.0**	19.2
			(0.3)	(0.0)	(0.0)	(0.0)	(0.0)	(0.3)
		Quintile 2	**5.6**	5.4	**4.5**	**2.2**	0.3	17.8
			(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)
		Quintile 3	**2.3**	**4.5**	**6.4**	5.6	**1.5**	20.4
			(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)
		Quintile 4	**0.6**	**2.1**	**5.1**	**8.5**	**5.2**	21.4
			(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(0.1)
		Richest	**0.0**	0.3	**1.4**	**5.4**	**14.0**	21.1
			(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.2)
		Total	22.2	15.8	18.9	22.2	20.9	100
			(0.3)	(0.1)	(0.1)	(0.1)	(0.2)	

Note: Synthetic panels are constructed from cross sections for Vietnam. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. SEs are obtained adjusting for complex survey design. All numbers are weighted using population weights. Poverty rates are in percentage. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. Joint probabilities are shown. Estimates based on the synthetic panels that fall within the 95% CI of those based on the actual panels are shown in bold.

are, again, rather encouraging with the majority (i.e. four-fifths) of the inner transitions falling within the 95% CIs of the true estimates, which are presented in bold. These estimates also pass the 100% mark of the coverage test. Other useful statistics that can be calculated from Table 4, panel B, include the percentages of the population that have seen either an improvement or a decline or remained in the same quintile over time, which are respectively 24.7%, 27.3% and 48%. These estimates are within the 95% CIs around those based on the actual panels. Furthermore, some of the remaining estimates that fall just outside these 95% CIs around the true estimates appear practically close to the latter (e.g. the transition from quintile 3 to quintile 4 or from the richest quintile to quintile 2). We further discuss the poverty mobility estimates for three periods in Appendix S3.

VI. Discussion and conclusion

Panel data currently are still unavailable, or have low quality, in a large majority of developing countries, and this situation may persist for quite some time. In the absence of panel data, our proposed method offers a means to construct synthetic panels that allow study of poverty and welfare dynamics. While our estimates are not perfect, Monte Carlo simulations and analysis using real household survey data indicates that they perform reasonably well under various deviations from the model assumptions. Moreover, synthetic panels are constructed from cross sections, which are not affected by issues specific to actual panels such as attrition and measurement errors.

Our proposed method need not be restricted only to the analysis of poverty transition, and may be further applied to other analyses of dynamics, such as labour transitions or health consumption. In fact, there have recently been promising extensions of our method to other topics such as intergenerational mobility (see, e.g. Foster and Rothbaum, 2015), shared prosperity (Dang and Lanjouw, 2016), or more extensive analysis of welfare dynamics along the whole income distribution (see, e.g. Moreno, Bourguignon, and Dang, 2021).

However, it is worth re-stating several caveats about our proposed method. In particular, where possible one should attempt to check the underlying assumptions before constructing synthetic panels. Since our proposed estimates for the correlation coefficients are based on asymptotic theory, they may be biased in the smaller-sample surveys that are typically available for analysis. Although we can generally afford to tolerate a degree of imprecision in our estimates of ρ, and our validation tests suggest that these estimates are reasonably successful in a variety of very different empirical settings, care should be taken to validate estimation results wherever possible (say, by using older panel data for the same country) before producing new estimates. It will also be important to further investigate sensitivity to cohort definition, which can vary depending on the context. As a general rule, the explanatory power of the income model will play a decisive role in determining success of the synthetic panel approach.

Final Manuscript Received: March 2021

References

Alkire, S. and Foster, J. E. (2011). 'Counting and multidimensional poverty measurement', *Journal of Public Economics*, Vol. 95, pp. 476–487.

Baulch, B. (ed). (2011). *Why Poverty Persists: Poverty Dynamics in Asia and Africa*, Edward Elgar Publishing.

Beegle, K., Christiaensen, L., Dabalen, A. and Gaddis, I. (2016). *Poverty in a Rising Africa*, The World Bank, Washington, DC.

Binder, D. A. and Roberts, G. (2009). 'Design- and model-based inference for model parameters', in Pfeffermann D. and Rao C. R. (eds), *Handbook of Statistics* Sample Surveys: Inference and Analysis, Vol. 29B, Elsevier, North-Holland.

Blundell, R., Duncan, A. and Meghir, C. (1988). 'Estimating labor supply responses using tax reforms', *Econometrica*, Vol. 66, pp. 827–861.

Bourguignon, F., Goh, C.-C., and Kim, D. I. (2004). *Estimating Individual Vulnerability to Poverty with Pseudo-Panel Data*. World Bank Policy Research Working Paper No. 3375.

Calvo, C. and Dercon, S. (2009). 'Chronic poverty and all that: the measurement of poverty over time', in Addison T., Hulme D., and Kanbur R. (eds), *Poverty Dynamics: Interdisciplinary Perspectives*, Oxford University Press, New York.
Chaudhuri, S. and Ravallion, M. (1994). 'How well do static indicators identify the chronically poor?', *Journal of Public Economics*, Vol. 53, pp. 367–394.
Colgan, B. (2022). 'EU-SILC and the potential for synthetic panel estimates', *Empirical Economics*. https://doi.org/10.1007/s00181-022-02277-7.
Collado, M. D. (1997). 'Estimating dynamic models from time series of independent cross-sections', *Journal of Econometrics*, Vol. 82, pp. 37–62.
Cross, P. J. and Manski, C. F. (2002). 'Regressions, short and long', *Econometrica*, Vol. 70, pp. 357–368.
Dang, H.-A. and Lanjouw, P. (2013). *Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections*. World Bank Policy Research Working Paper No. 6504.
Dang, H.-A. and Lanjouw, P. (2016). 'Toward a new definition of shared prosperity: a dynamic perspective from three countries', in Basu K. and Stiglitz J. (eds), *Inequality and Growth: Patterns and Policy*. New York: Palgrave Macmillan.
Dang, H.-A., Lanjouw, P., Luoto, J. and McKenzie, D. (2014). 'Using repeated cross-sections to explore movements in and out of poverty', *Journal of Development Economics*, Vol. 107, pp. 112–128.
Dang, H.-A., Jolliffe, D. and Carletto, C. (2019). 'Data gaps, data incomparability, and data imputation: a review of poverty measurement methods for data-scarce environments', *Journal of Economic Surveys*, Vol. 33, pp. 757–797.
Deaton, A. (1985). 'Panel data from time series of cross-sections', *Journal of Econometrics*, Vol. 30, pp. 109–126.
Devereux, P. J. (2007). 'Small-sample bias in synthetic cohort models of labor supply', *Journal of Applied Econometrics*, Vol. 22, pp. 839–848.
Doornik, J. A. and Hansen, H. (2008). 'An omnibus test for univariate and multivariate normality', *Oxford Bulletin of Economics and Statistics*, Vol. 70, pp. 927–939.
Elbers, C., Lanjouw, J. O. and Lanjouw, P. (2003). 'Micro-level estimation of poverty and inequality', *Econometrica*, Vol. 71, pp. 355–364.
Ferreira, F. H. G., Messina, J., Rigolini, J., López-Calva, L.-F., López-Calva, L. F. and Vakis, R. (2012). *Economic Mobility and the Rise of the Latin American Middle Class*, World Bank, Washington DC.
Fields, G. S. (2001). *Distribution and Development: A New Look at the Developing World*, MIT Press, Cambridge.
Foster, J. E. (2009). 'A class of chronic poverty measures', in Addison T., Hulme D., and Kanbur R. (eds), *Poverty Dynamics: Interdisciplinary Perspectives*, Oxford University Press, New York.
Foster, J. E. and Rothbaum, J. (2015). *Using Synthetic Panels to Estimate Intergenerational Mobility*. Working Paper No. 013/2015. Espinosa Yglesias Research Centre.
Garces-Urzainqui, D. (2017). *Poverty Transitions Without Panel Data? An Appraisal of Synthetic Panel Methods*. Paper presented at the 7th Meeting of the Society for the Study of Economic Inequality, New York City.
Gibson, J. (2001). 'Measuring Chronic Poverty without a Panel', *Journal of Development Economics*, Vol. 65, pp. 243–266.
Glewwe, P. and Jacoby, H. (2000). 'Recommendations for collecting panel data', in Grosh M. and Glewwe P. (eds), *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*, The World Bank, Washington DC.
Guell, M. and Hu, L. (2006). 'Estimating the probability of leaving unemployment using uncompleted spells from repeated cross-section data', *Journal of Econometrics*, Vol. 133, pp. 307–341.
Herault, N. and Jenkins, S. (2019). 'How valid are synthetic panel estimates of poverty dynamics?', *Journal of Economic Inequality*, Vol. 17, pp. 51–76.
Inoue, A. (2008). 'Efficient estimation and inference in linear pseudo-panel data models', *Journal of Econometrics*, Vol. 142, pp. 449–466.

Jenkins, S. P. (2011). *Changing Fortunes: Income Mobility and Poverty Dynamics in Britain*, Oxford University Press, Oxford.
Juodis, A. (2018). 'Pseudo panel data models with cohort interactive effects', *Journal of Business and Economic Statistics*, Vol. 36, pp. 47–61.
Kalton, G. (2009). 'Designs for surveys over time', in Pfeffermann D. and Rao C. R. (eds), *Handbook of Statistics Sample Surveys: Design*, *Methods and Applications*Vol. 29A, Elsevier, North-Holland.
Khan, R. (2021). 'Assessing sampling error in pseudo-panel models', *Oxford Bulletin of Economics and Statistics*, Vol. 83, pp. 742–769.
Khor, N. and Pencavel, J. (2006). 'Income Mobility of Individuals in China and the United States', *The Economics of Transition*, Vol. 14, pp. 417–458.
Kopczuk, W., Saez, E. and Song, J. (2010). 'Earnings inequality and mobility in the United States: evidence from social security data since 1937', *Quarterly Journal of Economics*, Vol. 125, pp. 91–128.
Lee, N., Ridder, G. and Strauss, J. (2017). 'Estimation of poverty transition matrices with noisy data', *Journal of Applied Econometrics*, Vol. 32, pp. 37–55.
Little, R. J. A. and Rubin, D. B. (2020). *Statistical Analysis with Missing Data* 3rd ed., Wiley, Hoboken.
De Luca, G., Magnus, J. R. and Peracchi, F. (2018). 'Balanced variable addition in linear models', *Journal of Economic Surveys*, Vol. 32, pp. 1183–1200.
Matloff, N. S. (1981). 'Use of regression functions for improved estimation of means', *Biometrika*, Vol. 68, pp. 685–689.
Moffitt, R. (1993). 'Identification and estimation of dynamic models with a time series of repeated cross-sections', *Journal of Econometrics*, Vol. 59, pp. 99–123.
Moreno, H., Bourguignon, F., and Dang, H.-A. (2021). *On Synthetic Income Panels*. IZA Discussion Paper No. 14236.
Nakagawa, S., Johnson, P. C. D. and Schielzeth, H. (2017). 'The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded', *Journal of the Royal Society Interface*, Vol. 14, p. 20170213.
OECD. (2018). *A Broken Social Elevator? How to Promote Social Mobility*, OECD Publishing, Paris.
Pencavel, J. (2007). 'A Life Cycle Perspective on Changes in Earnings Inequality among Married Men and Women', *Review of Economics and Statistics*, Vol. 88, pp. 232–242.
Piketty, T. (2014). *Capital in the Twenty-First Century*, Belknap Press, Cambridge.
Propper, C., Rees, H. and Green, K. (2001). 'The demand for private medical insurance in the UK: a cohort analysis', *Economic Journal*, Vol. 111, pp. C180–C200.
Reeves, R. (2020). *Biden Should Restore the Office of Economic Opportunity Abolished by Reagan*. https://www.brookings.edu/opinions/biden-should-restore-the-office-of-economic-opportunity-abolished-by-reagan/
Ridder, G. and Moffitt, R. (2007). 'The econometrics of data combination', in Heckman J. and Leamer E. E. (eds), *Handbook of Econometrics* Vol. 6B, Amsterdam, Elsevier.
Salvucci, V. and Tarp, F. (2021). 'Poverty and vulnerability in Mozambique: an analysis of dynamics and correlates in light of the Covid-19 crisis using synthetic panels', *Review of Development Economics*, Vol. 25, pp. 1895–1918.
Serajuddin, U., Uematsu, H., Wieser, C., Yoshida, N., and Dabalen, A. (2015). *Data Deprivation: Another Deprivation To End*. World Bank Policy Research Paper No. 7252.
Snijders, T. A. B. and Bosker, R. J. (1994). 'Modeled variance in two-level models', *Sociological Methods & Research*, Vol. 22, pp. 342–363.
Snijders, T. A. B. and Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, Sage Publications, London.
Stiglitz, J. E. (2013). *The Price of Inequality- How Today's Divied Society Endangers Our Future*, W. W. Norton & Company, New York.
United Nations Development Programme (UNDP). (2016). *Multidimensional Progress: Well-being beyond Income*, United Nations Development Programme, New York.

Verbeek, M. (2008). 'Synthetic panels and repeated cross-sections', in Matyas L. and Sevestre P. (eds), *The Econometrics of Panel Data*, Springer-Verlag, Berlin, pp. 369–383.
Verbeek, M. and Nijman, T. (1992). 'Can cohort data be treated as genuine panel data?', *Empirical Economics*, Vol. 17, pp. 9–23.
Verbeek, M. and Nijman, T. (1993). 'Minimum MSE estimation of a regression model with fixed effects from a series of cross-sections', *Journal of Econometrics*, Vol. 59, pp. 125–113.
Verbeek, M. and Vella, F. (2005). 'Estimating dynamic models from repeated cross-sections', *Journal of Econometrics*, Vol. 127, pp. 83–102.
Weisberg, S. (2014). *Applied Linear Regression* 4th ed., John Wiley, Hoboken.
World Bank. (2017). *Monitoring Global Poverty: Report of the Commission on Global Poverty*, The World Bank, Washington, DC.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Supplementary Information