DRAFT -- PLEASE DO NOT CITE

Testing Alternative Aggregation Methods Using Ordinal Data for a Census Asset-Based Wealth Index

Rodrigo Lovatón Dávila

December 15, 2015

Abstract

The construction of wealth indices based on housing characteristics and asset ownership has been widely used when other measures of socioeconomic status are not available. A popular approach has been to apply principal components analysis (PCA) on data recoded to binary indicators (Filmer and Pritchett, 2001). However, this procedure has been criticized since standard PCA methods are not designed to handle discrete data. In this paper, I compare alternative aggregation procedures that have been proposed to overcome this issue. The paper uses data from twelve developing countries. The evidence indicates that methods based on ordinal data have high agreement in rankings, but the PCA procedure on dichotomized data also has reasonable agreement with these measures. The alternative measures do not have striking differences in their relationship with a set of education, fertility, and mortality outcomes, both based on wealth index quintiles and on regression analysis. Finally, none of the asset-based indices outperformed the rest in terms of similarities of rankings with the logarithm of income per capita. In this sense, despite recommendations given by previous research (Howe et al., 2008; Kolenikov and Angeles, 2009), results suggest a relatively similar performance of the PCA procedure on dichotomized data with respect to methods based on ordinal data.

1. Introduction

The asset-based index approach to measure socioeconomic status has been widely used as an alternative measure of that status when income and expenditure data are not available. Principal components analysis (PCA) on data recoded to binary indicators (Filmer and Pritchett,

2001) is one of the most frequently used procedures to construct such an index. However, this approach has been subject to criticism, given that the standard PCA method does not consider that many asset variables are in fact categorical or ordinal. Furthermore, the variable dichotomization procedure not only generates spurious negative correlations (across binary indicators derived from the same categorical or ordinal variable) but also neglects the ordering of categories that may contribute additional information to define the index (Howe *et al*., 2008; Kolenikov and Angeles, 2009).

The use of ordinal data and polychoric correlations has been proposed as an alternative to overcome these criticisms of the commonly used approach that applies PCA to binary data (Kolenikov and Angeles, 2009). The performance of aggregation procedures based on asset ordinal data has not been extensively tested. Howe *et al*. (2008) found that the choice of categorical versus binary data had a strong influence on the agreement between alternative indices defined from living conditions variables. Kolenikov and Angeles (2009) compared PCA applied to binary indicators to other methods using ordinal variables. Their results show better performance of indices based on ordinal data according to different criteria, including the proportion of data variability explained by the index and its statistical significance in explaining women's fertility. Thus, they do not recommend working with binary indicators unless there is no information at all regarding the ordering of categories. Other research on this topic has examined the question on aggregation procedures for asset variables, but not through methods appropriate to deal with discrete asset data (Montgomery *et al*., 2000; Bollen *et al*., 2002; Filmer and Scott, 2012).

In this paper, I use census data to test alternative aggregation procedures to define a asset-based wealth index based on information of housing characteristics and assets.[1] The type and number of variables available vary widely in census microdata, in comparison to the more standard asset information typically included in household surveys (such as in the case of the Demographic and Household Surveys). This data variability provides an appropriate setting to test the relative performance of asset-based indices produced by alternative PCA methods, some of which are designed to handle ordinal variables. In particular, I explore whether these

[1] Throughout the paper, I will refer to indices constructed from information on housing characteristics and assets simply as asset-based indices or asset indices (which they are frequently called).

alternative methods generate similar household rankings and whether there are differences in their relation with selected education, fertility, and mortality outcomes. Results are also compared against the logarithm of income per capita for those datasets with this information available.

The paper is organized as follows. In the next section, I discuss previous research on methods to aggregate data on housing characteristics and asset ownership to define a proxy measure of socioeconomic status. In Section 3, I describe the data and the methods to construct the indices that are analyzed, including principal component analysis and the use of polychoric correlations. Next, in Section 4, I show the results of the study. Section 5 has a discussion of the main findings of the study. The appendix to this paper includes additional tables.

2. Literature Review

Filmer and Pritchett (2001) examined the use of housing characteristics and asset ownership to define an alternative measure of household socioeconomic status. This practical approach is motivated by the fact that income or expenditures are not always available in microdata. In their application, categorical variables are transformed into binary indicators (where each category is recoded as a separate variable) and principal components analysis (PCA) is used to assign weights to each indicator to construct an index. The authors found not only comparable rankings of households based on asset or expenditure data but also that these measures had similar predictive power to explain school enrollment using microdata from India, Indonesia, Nepal, and Pakistan. The method proposed by Filmer and Pritchett (2001), which applied PCA to dichotomized asset data, has been widely used as a control for household socioeconomic status in other studies that examine a variety of outcomes (see, for example, Bollen *et al*., 2002; Minujin and Bang, 2002; Houweling *et al*., 2003; Rutstein and Johnson, 2004; McKenzie, 2005; Lindelow, 2006; Bollen *et al*., 2007; Filmer and Scott, 2012; Wagstaff and Watanabe, 2003).

The use of information on housing characteristics and assets to define a proxy measure for household socioeconomic status leads to the question about the methods used to aggregate (i.e. produce weights for) the data. This question has been previously explored by several studies in this field. Montgomery *et al*. (2000) analyzed the use of individual living conditions variables against an index represented by the simple sum of these indicators. Their evidence indicates that either of these alternatives had limited explanatory power for consumption expenditures per adult, but they were useful proxies in regressions explaining fertility, child mortality, or children's schooling. Bollen *et al*. (2002) applied four different aggregation methods on information from consumer durable goods, including the number of assets, their current and median value, and PCA on binary indicators. The authors conclude that the number of durable assets and the binary PCA have stronger effects on children ever born than the current or median value of assets. Howe *et al*. (2008) worked with several methods to calculate weights, including PCA on categorical and dichotomized data. The study suggests that the choice of data (categorical versus dichotomized variables) had more influence on the agreement of indices than the different methods that were used to weight the data, while all the aggregation procedures had similar moderate agreement with consumption expenditures. Filmer and Scott (2012) compared a variety of approaches to measure welfare based on living conditions data, which include indices derived from an asset count, the traditional PCA on binary indicators, item response theory (IRT), and predicted per capita household expenditures. Their results show that household rankings are not identical and they depend on which measure is used, but differences in outcomes across these rankings are robust to this choice. Overall, conclusions regarding the relative performance of these methods do not strongly advocate for the use of one of them before the rest.

The Filmer and Pritchett (2001) approach to produce the index has been subject to criticism by more recent research, given that many asset variables are ordinal (such as dwelling ownership, type of water supply, or predominant walls material). Some specific issues have been identified (Howe *et al*., 2008; Kolenikov and Angeles, 2009). PCA relies largely on the calculation of the variance of the data --as it will be later discussed-- to produce the weights for the index. However, the methods frequently used to calculate the variance-covariance matrix for PCA neglect the fact the asset data are primarily discrete. In fact, PCA is based on the assumption that the data follow a multivariate (joint) normal distribution, which is clearly violated when working with asset data (Kolenikov and Angeles, 2009).[2] Furthermore, the dichotomization process generates spurious negative correlations between binary indicators derived from the same categorical variable, because by construction these come from categories that are mutually exclusive.[3] Therefore, the variance of the data used for PCA is based not only on the (positive) correlations between asset variables (hypothesized to depend on unobserved household wealth), but also on artificial negative correlations between indicators defined from the same categorical variable. The index could then reflect variability associated to these spurious correlations rather than that of unobserved household wealth.[4] Finally, the ordering of categories implied in ordinal variables contributes additional information, but this information is lost when these variables are transformed into binary indicators. PCA on binary indicators is considered assumption-free, both in terms of the order of categories (which may be incorrect) and the scale given to the distance between categories. However, ignoring the ordering may exclude useful information to rank households if the options clearly follow a certain order.

Kolenikov and Angeles (2009) examined methods to define a wealth index based on ordinal data, to address the issues of the popular approach proposed by Filmer and Pritchett (2001). In particular, they compared PCA on dichotomized variables to PCA using polychoric correlations (on ordinal data) and PCA on ordinal variables using the standard methods to calculate the variance of the data.[5] The authors implemented a simulation study based on artificial data, where PCA based on binary indicators is compared against the two methods based on ordinal variables. Results showed that transforming categorical variables into binary

[2] Kolenikov and Angeles (2009) argue that normality should be at least a reasonable distribution approximation, given that non-normal distributions of indicators entail that "some of the properties of the principal components no longer hold or need to be revised" (p. 161). Other authors suggest that normality is not always required but imply that it may be necessary for certain properties of PCA; for example, Jolliffe (2002), p. 19 or Timm (2002), p. 447.

[3] For example, having a dirt floor implies that a household cannot declare any other alternative as predominant construction material of the dwelling floors. Thus, the binary indicators will have some negative correlation just for being defined based on categories from the same variable.

[4] On this matter, Kolenikov and Angeles (2009, p. 138) argue that the "PCA method then needs to take into account both the fundamental (usually positive) correlations between observed variables and the spurious (negative) correlations between the dummy variables produced from a single factor," such that the "PCA procedure may not be able to recover the SES from the data, as the directions of greater variability may now correspond to those spurious correlations."

[5] Polychoric correlations are designed to handle discrete data unlike the standard correlations used in PCA. They yield a different estimate of the variance-covariance matrix when the latent variables of interest are continuous but we observe ordinal data. The details about the calculation of PCA with polychoric correlations are presented in Section 3 on the methodology.

indicators leads to lower performance, mostly in terms of the proportion of variance of the data explained by the index. In addition, even if the ordering of categories is incorrect, there is no evidence that PCA on binary indicators produces better results. The study also includes an empirical example to illustrate the differences between these procedures using data from the Bangladesh 2000 Demographic and Health Survey. This empirical application showed a relatively comparable ranking of households by wealth quintiles for the standard PCA on ordinal variables and the polychoric PCA, but higher disagreements of both indices with that derived from PCA on binary indicators. Moreover, when using each index as a control variable to explain fertility, results are very similar for the methods based on ordinal variables, while the binary PCA has lower significance and smaller coefficients. Based on this evidence, the authors to conclude that PCA on binary indicators is "not recommended unless there is absolutely no information about the ordering of categories" (Kolenikov and Angeles, 2009, p. 162). This recommendation is very strong and I examine it in the rest of this paper.

3. <u>Methodology</u>

3.1. Principal Components Analysis (PCA) and Polychoric Correlations

Before describing the data and aggregation procedures that are applied in this study, I present a brief summary of principal components analysis (PCA) and the calculation of polychoric correlations, as these concepts are used throughout the analysis. The objective of PCA is to find a subset of principal components that represents most of the variation in a set of x_j variables. Therefore, PCA is a data dimensionality-reduction technique. Given a set of x_j variables (j=1, ...,p), PCA produces a linear combination, denoted by PC_1, that maximizes the variance of a weighted sum of the x_j variables by using weights w_{ij}, as follows:

$$PC_1 = W_{\boldsymbol{1}}^{'}\boldsymbol{X} = w_{11} \cdot x_1 + w_{12} \cdot x_2 + \ldots + w_{1p} \cdot x_p \ldots (1)$$

A second linear combination, PC_2, can likewise be defined, that also maximizes the remaining variance of the x_j variables that was not captured by PC_1 and is uncorrelated with PC_1. More generally, PC_k can be defined as the *kth* linear combination or principal component that maximizes the remaining variance of the x_j variables and is uncorrelated with

$PC_1, PC_2, ..., PC_{k-1}$. The total number of uncorrelated linear combinations or principal components that can be constructed is equal to the number of x_j variables in the data.

In order to determine the weights, consider the optimization problem where the objective function is the variance of the principal component $VAR\left(W_j^{'}X\right)$ subject to the constraint of $W_j^{'}W_j = 1$ (imposed to find a single maximum) (Jolliffe, 2002; Timm, 2002):

$$L = VAR\left(W_j^{'}X\right) - \eta\left(W_j^{'}W_j - 1\right) = W_j^{'}\Sigma W_j - \eta\left(W_j^{'}W_j - 1\right) \ldots \text{ (2)}$$

In equation (2), $\sum$ is the variance-covariance matrix of the x_j variables in the data. Differencing equation (2) with respect to the weights W_j gives the first order condition for this (constrained) maximization problem:

$$\frac{\partial L}{\partial W_j} = \Sigma W_j - \eta W_j = 0 \ldots \text{ (3a)}$$

$$(\Sigma - \eta I)W_j = 0 \ldots \text{ (3b)}$$

This optimization problem is equivalent to finding the eigenvalues of the matrix $\sum$, where λ_j is an eigenvalue and the weight W_j is its corresponding eigenvector (Jolliffe, 2002; Timm, 2002). From equation (3b), it is also possible to infer that the eigenvalues are equal to the variance of the corresponding principal component or $VAR\left(W_j^{'}X\right) = \lambda_j$ (Jolliffe, 2002; Kolenikov and Angeles, 2009). A common measure to assess the performance of PCA is the proportion of total variance explained by each principal component. Following the previous result, this proportion can be calculated as:

$$\frac{VAR\left(W_j^{'}X\right)}{\sum_{i=1}^{p} VAR\left(W_i^{'}X\right)} = \frac{\lambda_j}{\sum_{i=1}^{p}\lambda_i} \ldots \text{ (4)}$$

In the standard calculation of PCA, the covariance matrix $\sum$ is estimated using the sample variance and covariance formulas:

$$S = \frac{1}{n-1} X'X \ \ldots \text{ (5)}$$

In equation (5), X is a matrix with each element derived from the original x_j variables after subtracting their sample means: $x_{ij} = \left(x_{ij} - \bar{x}_j\right)$. Given that the x_j variables could have different scales or measurement units, a common procedure is to work with the standardized data (zero mean and unit variance), which leads to an optimization problem analogous to equation (2) but based on the correlation matrix of the data (Jolliffe, 2002; Kolenikov and Angeles, 2009).

Principal components analysis (PCA) was not originally designed to handle discrete variables, but continuous (and normally-distributed) data. Polychoric correlations have been proposed as an alternative to calculate the correlation matrix necessary to perform PCA on discrete asset data (Kolenikov and Angeles, 2009). Consider ordinal variables y$_k$ with d_k categories derived from an underlying continuous variable y$_k$* using a set of thresholds $\gamma_1^k, ..., \gamma_{d-1}^k$, such that (Drasgow, 2006; Kolenikov and Angeles, 2009):

$$y_k = y_{ki} \ \ if \ \ \gamma_{i-1}^k \leq y_k * < \gamma_i^k \ \ldots \ (6)$$

In equation (6), i=1, 2, …, d_k and the first and last thresholds are defined as $\gamma_0^k = -\infty$ and $\gamma_d^k = \infty$, respectively. Furthermore, assuming underlying continuous variables y$_1$* and y$_2$* that are distributed following a bivariate normal distribution, the probability of an observation (y_{1i}, y_{2j}) is given by (Olsson, 1979; Drasgow, 2006; Holgado-Tello *et al.*, 2010):

$$Pr\left(y_{1i}, y_{2j}\right) = \int_{\gamma_{i-1}^1}^{\gamma_i^1} \int_{\gamma_{j-1}^2}^{\gamma_j^2} \phi\left(y_1, y_2; \rho\right) dy_1 dy_2 \ \ldots \ (7)$$

In equation (7), $\phi\left(Y_1, Y_2; \rho\right)$ is the bivariate normal distribution function with correlation ρ. Based on this probability, the likelihood function for a sample with n_{ij} observations of values (y_{1i}, y_{2j}) can be defined as:

$$L = \prod_{i=1}^{d_1} \prod_{j=1}^{d_2} \Pr\left(y_{1i}, y_{2j}\right)^{n_{ij}} \ \ldots \ (8a)$$

$$\ln L = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} n_{ij} \ln \Pr\left(y_{1i}, y_{2j}\right) \ldots \ (8b)$$

In equations (8a) and (8b), d_1 and d_2 are the number of categories for y$_1$ and y$_2$. The likelihood function is then maximized with respect to ρ and the set of thresholds ($\gamma_1^1, ..., \gamma_{d-1}^1; \gamma_1^2, ..., \gamma_{d-1}^2$) to obtain maximum likelihood estimates of the model parameters (Drasgow, 2006). The solution for ρ from the resulting system of equations is then used to create each element of the correlation matrix for PCA. Therefore, polychoric correlations yield an estimate derived from the underlying (unobserved) continuous variables, which is argued to be the "true" measurement of the correlation structure. In practical terms, standard correlation methods (such as the sample variance and covariance) seem to produce smaller-sized estimates, when applied to discrete data, relative to polychoric correlations (Kolenikov and Angeles, 2009).

3.2. Data

The analysis in this study was performed using selected census samples from the IPUMS-International project. The data from IPUMS-International offers the main advantage that most of the variables necessary for the analysis have been previously harmonized, which produces more comparable results. The selected census samples are shown in Table 1. The main criterion for the selection of datasets was the availability of income data, given only a small proportion of censuses collect this information.[6] Further, I excluded census samples that did not have data for all the outcomes of interest for this study, particularly those without questions on children ever born and on children surviving. Three additional census datasets were included to have more variability in the total number of assets available, which include Cambodia 1998 (only 6 items), Colombia 2005 (35 items), and Peru 1993 (30 items).

The countries represented in Table 1 are mainly from Latin America (Brazil, Colombia, Dominican Republic, Mexico, Panama, and Peru), while two census samples correspond to other regions (Cambodia and South Africa). The datasets are 10 percent samples of the corresponding censuses, except for Mexico 1970 (1 percent), Brazil 1970 (5 percent), and Brazil 2000 (6 percent). Therefore, given the large proportion of the population included, all the data are nationally representative and are also representative of lower geographical units.

[6] In particular, only 36 of the 277 census samples available (at the time this paper was written) through the IPUMS-International project included income data. However, given that the asset-based approach is primarily relevant for developing countries, then Canada, Puerto Rico, and the United States were excluded from the possible datasets for the analysis.

Table 1: Asset Variables Available for Selected Census Samples

Census sample	Variables available	Type of variables available					Household income
		Housing characteristics	Assets	Binary	Ordinal	Count	
Brazil 2000	21	11	10	8	7	6	Yes
Brazil 2010	20	11	9	10	8	2	Yes
Cambodia 1998	6	6	0	1	4	1	No
Colombia 1973	13	13	0	2	9	2	Yes
Colombia 2005	35	18	17	18	10	7	No
Dominican Republic 2002	26	13	13	14	10	2	Yes
Mexico 1970	13	11	2	8	4	1	Yes
Mexico 2000	24	14	10	12	10	2	Yes
Panama 1980	19	13	6	7	10	2	Yes
Panama 2010	26	14	12	7	10	9	Yes
Peru 1993	30	11	19	20	9	1	No
South Africa 1996	10	9	1	1	8	1	Yes

Data source: Integrated Public Use Microdata Series (IPUMS) - International.

The set of selected samples include varied information on housing characteristics and asset ownership at the household level (Table 1). The detailed list of variables available in each census sample is shown in Table A1 in the Appendix to this paper. The datasets have a wide number of variables available, ranging from 6 for Cambodia 1998 to 35 for Colombia 2005. The vast majority of data are discrete, given about half of the variables are binary and a slightly smaller proportion are ordinal. The specific variables included depend on the dataset, but cover aspects such as the predominant construction materials of the roof, dwelling ownership, type of water source, number of rooms in the dwelling, ownership of diverse durable assets, among others.

Household income information was collected in nine of the selected census samples datasets. The availability of income data allows me to compare the indices based on housing characteristics and assets against a more traditional measure of household socioeconomic status. Income is not a perfect measure of socioeconomic status because of issues such as measurement error (given the complexity of data collection) and year-to-year variability (Montgomery *et al*., 2000; Bollen *et al*., 2002). However, acknowledging these possible limitations, the income data provide an additional criterion in this analysis, to examine the relative performance of the asset-based indices. Income is adjusted by the number of household members to obtain a per capita variable, and it is also log-transformed due to its highly skewed distribution.

3.3. How are the indices defined?

Data on housing characteristics and assets are summarized to define a proxy measure of socioeconomic status. In general, this household measure is defined as:

$$WI_i = a_1 \cdot x_{1i} + a_2 \cdot x_{2i} + \ldots + a_n \cdot x_{ni} \ \ldots (9)$$

In equation (9), WI_i is the index for household i, x_j are variables representing housing characteristics and assets, and a_j is the weight assigned to variable x_j. All the variables available in the data, which were described in Table 1, are used for the construction of the indices.

I tested four alternative aggregation methods that have been proposed in the literature (Filmer and Pritchett, 2001; Kolenikov and Angeles, 2009). Two methods work with the original unrecoded data (including binary, ordinal, and count variables), while the other two only use binary indicators:

(i) PCA based on the standard calculation of correlations and applied to the unrecoded (discrete) asset data, which is called "ordinal PCA" in the rest of the paper.

(ii) PCA based on polychoric correlations on the ordinal data or "polychoric PCA."

(iii) PCA on the recoded data, where all ordinal variables have been transformed into binary indicators, and using the standard calculation of the correlation matrix. This is the common method performed in previous studies and that is identified in this paper as "binary PCA."

(iv) A count index, as a simple aggregation procedure to compare against the more sophisticated PCA-based methods. For this last index, each household variable receives a weight equal to 1, such that we count the number of assets that the household "owns."[7] The count index was defined only using durable asset variables (i.e. it excludes housing characteristics), so it is not fully comparable with the other methods.

[7] Certain variables available in the data registered the number of assets of a certain type. For example, the Brazil 2000 census sample included not only a binary response (yes/no) for televisions, cars, and air conditioning units, but also the number of such assets owned by the household. In these cases, the count index considered only ownership, such that multiple assets of the same kind were not double counted. The PCA-based indices do include the number of assets and not only its ownership, if this information is available.

The inclusion of the three PCA-based aggregation methods allows examination of changes in household rankings due to the use of polychoric correlations (by comparing the polychoric PCA against the ordinal PCA) and to the use of ordinal versus binary data (by comparing the ordinal PCA against the binary PCA). For all the PCA-based methods, only the first component is retained. The first component represents by definition the maximum variance extracted from the asset data. Therefore, I will follow the usual assumption in the literature that the first component "represents" (or is a proxy for) household socioeconomic status based on the various asset indicators used to produce it (Filmer and Scott, 2012).[8] Other components may be related to other dimensions of "household socioeconomic status," but were not used in the analysis.

The ordering of categories may be an important input to define the indices, as it conveys additional information to rank households. The asset data had an implicit ordering in most cases and it was just necessary to assign the lowest value of each variable to the "worst" option and the highest to the "best" option. An example is shown in Table 2 for the main construction material of the roof from the Dominican Republic 2002 census sample, including both the ordinal and binary versions of the data. Some basic data manipulation was implemented to obtain ordinal versions of categorical variables. The original ordering of categories was modified only in a few cases where certain categories were clearly misclassified in the scale; for example, if dirt was second (or penultimate) in the scale rather than in one of the ends in the flooring material categories. Furthermore, some categories were dropped from the analysis (i.e. transformed to missing) if their position in the scale was unclear, which most often happened for "other" (residual) responses. For instance, "other" was excluded from the roof main construction material for Dominican Republic 2002, as it was unclear how to rank the construction material represented by this option with respect to the rest of the alternatives. In addition, a few categories were combined if they seemingly represented similar positions in a scale. For example, "palm leaves" and "cane" were combined for the roof main construction material for Dominican Republic 2002, given they did not appear to be qualitatively different, as shown in Table 2. In order to have comparable information to define the indices using the alternative aggregation methods, any categories that were dropped or combined were treated the same way for the dichotomized version of the data.

[8] For some further discussion on the use of higher order components, see Filmer and Scott (2012).

Table 2: Data Recoding, Main Construction Material of Roof, Dominican Republic 2002

Original variable	Ordinal	Binary
0 = Unoccupied households	*(dropped)*	*(dropped)*
1 = Concrete	4 = Concrete	1 = Concrete, 0 = No
2 = Zinc	3 = Zinc	1 = Zinc, 0 = No
3 = Asbestos	2 = Asbestos	1 = Asbestos, 0 = No
4 = Palm leaves	1 = Palm leaves or cane	1 = Palm leaves or cane, 0 = No
5 = Cane		
6 = Other	*(dropped)*	*(dropped)*

Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Based on these aggregation methods, I examine in the next section whether they produce similar household rankings and whether there are differences in their relation to selected education, fertility, and mortality outcomes. Comparisons with income data are included throughout the analysis. Household rankings produced by each measure are compared through Spearman rank correlations and cross-classification by wealth quintiles. The relationships with the selected outcomes were observed both through differences across wealth index quintiles and in regression analysis using each index as a control variable. The objective is to test whether the direction and strength of the relation of each proxy for socioeconomic status are consistent across diverse outcomes that are typically examined in social and economic research. The specific outcomes analyzed include school enrollment, literacy, completion of primary school, completion of secondary school, having any children, and having experienced any child death. For higher comparability of results, the same set of controls (i.e. available across datasets) was used for regressions estimated for the same outcome. The implicit assumption behind the proposed analysis is that a stronger relation with the outcome is considered to show a better performance of the socioeconomic status measure (Bollen *et al.*, 2007; Kolenikov and Angeles, 2009). This assumption is also followed by other research in this area. The analysis does not intend to estimate a causal relationship between socioeconomic status and the selected outcomes, but rather verify that they are indeed correlated. In the study by Kolenikov and Angeles (2009) this is distinguished as a "weaker requirement of internal validity" from the causal effect that needs to be verified for external validity of the measure (p. 159).

4. Results

4.1. Calculation of weights using PCA

The data on housing characteristics and asset ownership were used to define the proposed indices, where three of them had weights calculated by applying PCA. For the PCA-based indices, a first step was to examine the sign and size of the weights assigned to each item. Kolenikov and Angeles (2009) refer to a "natural ordering" of categories that follow a monotone relation among them in the case of ordinal variables. For example, it is expected that dirt be the worst and concrete be the best floor materials and that they would be assigned the lowest and highest value in the correspondent ordinal variable, while other flooring materials would be assigned a number within that range. If this ordering is meaningful in terms of the relation between the asset variable and (unobserved) household socioeconomic status, then weights assigned by PCA are expected to follow the ordering, assigning larger positive values to the most "desirable" household characteristics and larger negative values to the least "desirable" ones. Nevertheless, it should also be noted that in the case of binary PCA, weights tend to be larger for assets more unequally distributed across households (McKenzie, 2005; Vyas and Kumaranayake, 2006). That is, assets that are owned by all or by very few households will receive relatively smaller weights, as they do not vary much across households (and PCA is defined from the variability of the data).

An example of weights obtained by the three PCA-based methods is shown in Table 3 for the main flooring materials in the Colombia 2005 census sample. Categories in the table are ordered from best to worst. In the example, the weights assigned by the polychoric PCA method follow the ordering of categories, where carpet flooring ("best" option) is assigned the largest positive weight, dirt or sand ("worst" option) are assigned the largest negative weight, and the rest of categories receive intermediate values as weights. The binary PCA weights do not satisfy this monotonicity condition, given these are not strictly increasing as we move from the "worst" to the "best" flooring material. In particular, the weight assigned to tile flooring (0.153) is larger than the corresponding to carpet (0.064), while a similar issue occurs for cement or gravel flooring (-0.070) with respect to rough wood (-0.051). The disagreement between the implicit ordering of the variable and the weights calculated by the binary PCA seems to be driven by the larger frequencies associated to tile and cement (reported by about 30-40 percent of households) with respect to carpet and rough wood (reported only by 5-7 percent of households). In general, weights assigned by binary PCA to other ordinal variables do not necessarily follow their order of categories, similarly to this example.

Table 3: Weights for Flooring Material by Alternative Aggregation Methods, Colombia 2005 Census Sample

Main flooring material (best to worst)	Household proportion (%)	PCA aggregation method		
		Binary	Ordinal	Polychoric
Carpet, marble, parquet, or polished wood	6.8	0.064	0.247	0.388
Tile, vinyl, clay tile, or brick	44.8	0.153		0.119
Cement or gravel	33.9	-0.070		-0.103
Rough wood, board, plywood, or other vegetable	4.4	-0.051		-0.235
Dirt or sand	10.2	-0.166		-0.352

Data source: Integrated Public Use Microdata Series (IPUMS) - International.

The PCA based methods can also be compared by examining the proportion of variance of the data explained by each of them. The proportion of variance explained by the first principal component is calculated as the ratio of the first eigenvalue to the sum of all eigenvalues from the variance-covariance matrix, similarly for any of the PCA-based measures. Results are included in Table 4. As we observe, this criterion shows that the indices based on the calculation of polychoric correlations explain a larger proportion of the overall data variability, followed by ordinal PCA, and lastly by the indices based on dichotomized variables. The differences in proportion of variance explained by each of the indices are consistent and large across all datasets analyzed. For example, while the asset-based wealth index using binary PCA only explains 16 percent of the data variability for Peru 1993, PCA on ordinal variables achieves 26 percent, and the polychoric PCA 46 percent of the overall variability. In the bottom row of the table, on average, about 18 percent of variance is explained by the first component of binary PCA, 33 percent for ordinal PCA, and 49 percent for polychoric PCA. This evidence is consistent with previous findings by Kolenikov and Angeles (2009).

Table 4: Proportion of Variance Explained (%) by the First Principal Component, for Alternative Aggregation Methods

Census sample	PCA aggregation method		
	Binary	Ordinal	Polychoric
Brazil 2000	17.83	29.05	44.76
Brazil 2010	12.93	23.36	42.18
Cambodia 1998	20.15	35.45	45.81
Colombia 1973	15.62	37.46	45.30
Colombia 2005	15.51	27.18	45.14
Dominican Republic 2002	15.12	25.07	46.69
Mexico 1970	29.70	42.64	57.10
Mexico 2000	17.32	32.86	51.34
Panama 1980	17.68	35.16	49.51
Panama 2010	13.59	29.64	49.61
Peru 1993	15.56	25.88	46.15
South Africa 1996	23.29	52.28	60.91
Simple average	***17.86***	***33.00***	***48.71***

Data source: Integrated Public Use Microdata Series (IPUMS) - International.

4.2. Household Rankings

The household classifications resulting from each of the wealth indices can be compared to assess the consistency of rankings. As a first step to compare rankings, I calculated Spearman rank correlations between indices, both with respect to the index based on polychoric correlations (Table 5a) and to the logarithm of income per capita (Table 5b). The correlations show very high correspondence of the polychoric PCA index with other asset-based wealth indices (correlations larger than 0.9 in most cases), being the largest for the ordinal PCA and binary PCA, followed by the asset count index. Similarly, we observe a high congruence of rankings (around 0.5 to 0.6) if we compare any of these alternative aggregation methods against rankings based on the logarithm of income per capita. This evidence suggests not only that household rankings are (reasonably) similar to rankings based on the (logarithm of) household income per capita, but also that these rankings are highly consistent across the alternative aggregation methods for the asset-based measures analyzed.

Table 5a: Correlation Coefficients between Wealth Indices and Polychoric PCA Index

	Asset count	Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	0.920	0.977	0.998	1.000	0.716
Brazil 2010	0.821	0.956	0.990	1.000	0.594
Cambodia 1998	NA	0.886	0.983	1.000	NA
Colombia 1973	NA	0.983	0.998	1.000	0.531
Colombia 2005	0.858	0.981	0.997	1.000	NA
Dominican Republic 2002	0.901	0.975	0.999	1.000	0.487
Mexico 1970	0.714	0.989	0.999	1.000	0.555
Mexico 2000	0.930	0.988	0.997	1.000	0.628
Panama 1980	0.830	0.980	0.997	1.000	0.659
Panama 2010	0.910	0.973	0.995	1.000	0.375
Peru 1993	0.900	0.987	0.998	1.000	NA
South Africa 1996	NA	0.989	0.997	1.000	0.615
Simple average	***0.865***	***0.972***	***0.996***	***1.000***	***0.573***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

Table 5b: Correlation Coefficients between Wealth Indices and Log of Income per Capita

	Asset count	Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	0.677	0.693	0.721	0.716	1.000
Brazil 2010	0.523	0.524	0.567	0.594	1.000
Cambodia 1998	NA	NA	NA	NA	NA
Colombia 1973	NA	0.521	0.530	0.531	1.000
Colombia 2005	NA	NA	NA	NA	NA
Dominican Republic 2002	0.412	0.485	0.485	0.487	1.000
Mexico 1970	0.420	0.553	0.554	0.555	1.000
Mexico 2000	0.583	0.635	0.621	0.628	1.000
Panama 1980	0.518	0.649	0.665	0.659	1.000
Panama 2010	0.323	0.364	0.371	0.375	1.000
Peru 1993	NA	NA	NA	NA	NA
South Africa 1996	NA	0.606	0.605	0.615	1.000
Simple average	***0.494***	***0.559***	***0.569***	***0.573***	***1.000***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

Households were classified into quintiles based on the wealth index measures to further examine consistency of household rankings. In order to compare classifications, I calculated the proportion of households that were classified in the same, in a higher, or in a lower quintile across pairs of indices. The detailed classifications by quintile are available, but are not presented here. In particular, this procedure was performed to compare the classifications based on the polychoric PCA index and the log of income per capita against other measures. Results for the proportion of households classified in the same quintile are shown below in Tables 6a and 6b, while those for lower and higher quintiles are included in the Appendix to this paper.

The discrepancies in classifications by quintiles reveal larger differences than the Spearman rank correlation coefficients, but still a sizable overlap. Household classification by quintiles is highly consistent between the two PCA methods that use ordinal variables (polychoric and ordinal), relative to other measures. In fact, more than 90 percent of total households are classified in the same wealth quintile if we use the original unrecoded data (ordinal variables) for any of the census samples examined, disregarding the method applied to calculate the correlations matrix for PCA. The household classification by quintiles for the polychoric PCA also has a relatively large overlap with binary PCA for most of the datasets analyzed (around 75-85 percent), except for Cambodia 1998, which may be explained by the limited number of variables available for this census sample. The correspondence of wealth quintiles based on the polychoric PCA index is smaller with the asset count index (55 percent of households were classified in the same quintile on average) and the logarithm of income per capita (40 percent of households were classified in the same quintile on average).

In terms of cross-classifications into lower or higher quintiles with respect to the polychoric PCA index (Tables A2a and A3a in the appendix to this paper), results show that a similar proportion of households move up or down for other measures, except for the asset count. For example, if we compare quintiles based on the polychoric PCA index for Brazil 2000, about 2.5 percent of households are classified in a higher or lower quintile for the ordinal PCA index, 9 percent for the binary PCA index, and 28 percent for the log of income per capita. However, 35.9 percent of households are relatively less "wealthy" using the asset count index but only 6.9 percent of households appear to be "wealthier" in this case. The larger discrepancies in classification into lower quintiles are mainly explained by the relatively lower variability of the count index, given this method only produces discrete values that range between zero and the total number of asset variables available in the dataset. This lower variability implies that a larger number of households are assigned the same number for the index, thus creating issues to calculate the cutoff points to define the quintiles.

Table 6a: Comparison of Classification by Quintiles

Households Classified in the Same Quintile with respect to Polychoric PCA Index (%)

	Asset count	Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	57.2	82.4	94.7	100.0	43.3
Brazil 2010	50.1	74.9	95.2	100.0	38.5
Cambodia 1998	NA	32.9	90.5	100.0	NA
Colombia 1973	NA	79.5	94.6	100.0	36.2
Colombia 2005	59.9	84.7	96.2	100.0	NA
Dominican Republic 2002	56.9	78.4	95.2	100.0	34.3
Mexico 1970	35.1	83.7	96.1	100.0	38.3
Mexico 2000	67.8	82.3	94.5	100.0	42.5
Panama 1980	53.4	79.2	93.1	100.0	41.7
Panama 2010	60.5	79.8	93.4	100.0	38.4
Peru 1993	49.0	83.3	96.0	100.0	NA
South Africa 1996	NA	85.0	92.4	100.0	38.4
Simple average	***54.4***	***77.2***	***94.3***	***100.0***	***39.1***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

The overlap in classifications by quintiles in Table 6b show that all indices based on housing characteristics and assets have a relatively similar performance when compared to quintiles based on the logarithm of household income per capita. In all cases, about 35 to 45 percent of households coincide in the same wealth quintile defined by income, even following the simple approach of counting household durable assets. Furthermore, analogous to previous results, the asset count tends to classify households more often into lower quintiles than higher quintiles when compared to income, as it can be observed in Tables A2b and A3b in the Appendix to this paper.

Overall, the consistency of household classification by quintiles is more moderate (but sizable) between income and all the indices based on housing characteristics. The existing discrepancies could be explained by the concept of household wealth that each measure is capturing: while housing characteristics and assets reflect accumulation of material well-being for a household (a stock), income reflects monetary gains from household members over some specified period of time (a flow). Additionally, it is possible that income is a noisy measure due to measurement error and year-to-year fluctuations, issues that are not expected to affect asset information. Nevertheless, we do not observe a relatively better or worse performance by any of the material wealth measures based on alternative aggregation procedures when compared to income per capita. All of them classify about the same proportion of households into the same wealth quintile as income.

Table 6b: Comparison of Classification by Quintiles

Households Classified in the Same Quintile with respect to Log of Income per Capita (%)

	Asset count	Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	40.4	42.8	43.9	43.3	100.0
Brazil 2010	35.7	36.8	38.4	38.5	100.0
Cambodia 1998	NA	NA	NA	NA	NA
Colombia 1973	NA	35.9	36.1	36.2	100.0
Colombia 2005	NA	NA	NA	NA	NA
Dominican Republic 2002	32.5	34.3	34.3	34.3	100.0
Mexico 1970	28.3	38.6	38.3	38.3	100.0
Mexico 2000	40.9	42.7	42.4	42.5	100.0
Panama 1980	35.4	41.0	42.3	41.7	100.0
Panama 2010	35.5	38.4	38.5	38.4	100.0
Peru 1993	NA	NA	NA	NA	NA
South Africa 1996	NA	39.4	38.7	38.4	100.0
Simple average	***35.5***	***38.9***	***39.2***	***39.1***	***100.0***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

Finally, graphical analysis of kernel density estimates for each of the asset-based indices was used to identify whether there are any differences along the resulting distribution of wealth (see Figure A1 in the Appendix to this paper). The purpose is to visually inspect the resulting distributions based on the alternative aggregation methods applied. In order to have comparable scales, all indices were standardized (i.e. they have zero mean and unit variance). Even though alternative aggregation procedures were implemented to calculate the variable weights, all the PCA-based indices appear to produce very similar distributions, either using the dichotomized or the original unrecoded data. This result was surprising given that the binary PCA works with transformed (dichotomized) data. For these three indices, the only noticeable discrepancy happens for Cambodia 1998, where there seems to be larger disagreement in left tail of the distribution, possibly explained by the small number of variables available for this census sample.

The most important differences correspond to the comparisons between the asset count indices against the PCA-based indices. For all the graphs, as it would be expected, the density mass is concentrated around a more limited set of values for the asset count (which produces a less smooth distribution). As previously discussed, this distribution can be explained by the definition of the index, which has discrete values that range from zero up to the maximum number of assets available in the data. For example, the Mexico 1970 sample has only two asset variables available to define the asset count; therefore, the resulting index could have only values of zero, one, or two. The three possible values can be clearly observed in the kernel density estimate for this dataset. However, the asset count does produce a comparable (but not smooth) distribution with respect to the PCA-based indices in the case of Panama 2010. The reason for these similarities is that PCA-based methods assigned weights of similar size to many items included in this dataset, such that the equal (unit) weights used in the count index resemble the weights produced by PCA. For instance, ordinal PCA assigned weights of 0.231 to the type of lighting, 0.224 to the main method of garbage disposal, 0.207 to the fuel used for cooking, 0.204 to ownership of a stove, and so forth.

4.3. How are outcomes changing across the indices?

Measures of socioeconomic status are often used in economic research to examine differences across outcomes of interest and as a control in regression analysis. In this sub-section, I examine the extent to which the proposed indices produce similar-sized differences across quintiles of household wealth and have comparable coefficients in regression analysis for a set of six selected education, fertility, and mortality outcomes. The selected outcomes are some of those frequently analyzed in social and economic research.

The differences for school enrollment (for children ages 7 to 14) by quintile of wealth for each of the alternative proxies of socioeconomic status are shown in Table 7a. As expected, across almost all datasets and measures, I obtained increasing proportions of children enrolled in school when moving from the bottom ("poorest") to the top ("richest") quintile. The detailed proportions of children enrolled in school are not shown, but are available upon request. In Table 7a, I present a summary figure: the average difference in school enrollment across quintiles, which was calculated as the difference between the top and bottom quintile divided by four.[9] This number represents the average change in the outcome of interest when we move along wealth quintiles. In Table 7a, the differences in school enrollment across quintiles are strikingly similar for all the PCA-based measures. For example, the change in school enrollment is about 2.8 percent across quintiles for Mexico 2000 and 3.6 percent for Panama 1980 for the binary PCA, ordinal PCA, or polychoric PCA. If these measures are compared to quintiles based on the logarithm of income per capita, the numbers for the average difference across quintiles are also similar but are consistently smaller than the rest. Across datasets and measures, the largest differences in school enrollment between quintiles correspond to the two census samples from the 1970s (Mexico 1970 and Colombia 1973), which also coincide with the two lowest average school enrollment rates.

Table 7a: Average Change in School Enrollment (children ages 7-14) across Wealth Quintiles, for Alternative Aggregation Methods [1/]

	Mean	Average difference across quintiles			
		Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	94.5	2.29	2.33	2.34	1.75
Brazil 2010	96.9	0.59	0.64	0.64	0.43
Cambodia 1998	64.5	5.66	4.53	4.17	NA
Colombia 1973	62.2	11.49	11.42	11.30	8.10
Colombia 2005	90.8	3.96	4.02	4.10	NA
Dominican Republic 2002	87.4	0.79	0.85	0.88	0.20
Mexico 1970	69.4	7.93	8.11	8.10	6.67
Mexico 2000	92.8	2.84	2.88	2.87	1.87
Panama 1980	87.7	3.56	3.54	3.59	3.06
Panama 2010	97.0	1.00	1.02	1.00	0.09
Peru 1993	87.0	3.36	3.43	3.42	NA
South Africa 1996	88.9	3.15	3.20	3.19	2.45
Simple average	***84.9***	***3.89***	***3.83***	***3.80***	***2.73***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. The average difference is calculated as the difference between top and bottom quintiles divided by four.

The differences in school enrollment by household socioeconomic status are also examined through regressions controlling for each of the measures based on the alternative

[9] The average difference presented is equivalent to calculating the difference between second and bottom, third and second, fourth and third, and top and fourth quintile, and averaging across these numbers.

aggregation methods. The regression results for school enrollment are shown in Table 7b. The table shows the odds-ratio for each census sample and socioeconomic status measure. As it is expected, almost all measures show a positive (odds-ratio larger than one) and a statistically significant coefficient for SES in explaining school enrollment. The only notable exception is observed for the log of income per capita for Panama 2010, which is marginally statistically significant and implies a negative effect on school enrollment.

The coefficients for the asset count and the three PCA-based methods are very similar in size, and all of them are larger than the effect of the logarithm of income per capita. If we further examine the (small) differences in the effects across samples, none of the measures is consistently larger than the rest. However, it is possible to identify a pattern based on the (small) differences in coefficients: the polychoric PCA tends to have the first or second largest coefficient (for ten out of twelve samples), followed by the ordinal PCA (for six out of twelve samples). But differences are small on average.

Table 7b: Logit model for School Enrollment (Children Ages 7-14)
Wealth Index Coefficient (Odds-ratio) 1/

	Asset count	Binary PCA	Polychoric PCA	Ordinal PCA	Log income per capita
Brazil 2000	**2.224**	**2.000**	**2.184**	**2.100**	**1.444**
	0.088	*0.059*	*0.074*	*0.060*	*0.034*
Brazil 2010	**1.405**	**1.378**	**1.434**	**1.393**	**1.088**
	0.050	*0.033*	*0.039*	*0.030*	*0.024*
Cambodia 1998	**NA**	**1.489**	**1.363**	**1.442**	**NA**
		0.040	*0.033*	*0.031*	
Colombia 1973	**NA**	**1.883**	**1.792**	**1.804**	**1.249**
		0.030	*0.026*	*0.026*	*0.025*
Colombia 2005	**2.078**	**1.927**	**2.059**	**1.963**	**NA**
	0.044	*0.096*	*0.097*	*0.089*	
Dominican Republic 2002	**1.131**	**1.126**	**1.142**	**1.145**	**1.012** #
	0.023	*0.025*	*0.028*	*0.029*	*0.018*
Mexico 1970	**1.314**	**1.543**	**1.553**	**1.558**	**1.321**
	0.037	*0.073*	*0.071*	*0.071*	*0.040*
Mexico 2000	**1.780**	**1.775**	**1.820**	**1.774**	**1.123**
	0.093	*0.105*	*0.099*	*0.088*	*0.034*
Panama 1980	**1.647**	**1.816**	**1.863**	**1.842**	**1.283**
	0.071	*0.094*	*0.090*	*0.096*	*0.055*
Panama 2010	**1.709**	**1.552**	**1.689**	**1.611**	**0.943** ##
	0.078	*0.075*	*0.091*	*0.079*	*0.027*
Peru 1993	**1.330**	**1.337**	**1.399**	**1.386**	**NA**
	0.074	*0.062*	*0.093*	*0.101*	
South Africa 1996	**NA**	**1.786**	**1.775**	**1.729**	**1.300**
		0.046	*0.046*	*0.046*	*0.019*

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. All estimates are statistically significant at the 1 percent level unless otherwise noted (### p>.01, ## p>.05, # p>.10). The table includes odds-ratio coefficients in bold and clustered standard errors in italic. Standard errors are clustered using mesoregions for Brazil, districts for Cambodia and Panama, municipalities for Colombia and Dominican Republic, states for Mexico, provinces for Peru, and magisterial districts for South Africa. The estimation sample is restricted to persons 7 to 14 years old that are not household heads.

Control variables: sex, age, and age-squared of the child; sex, age, age-squared, and educational attainment of household head (dummies for primary, secondary, and university); urban/rural.

The differences across wealth quintiles were also measured for child mortality for women who ever gave birth, aged between 18 and 30 years old at the time of data collection. The information on child mortality was not directly available in the data, but it was approximated using children ever born and children surviving. Therefore, it does not refer to deaths of children within certain ages as it is typically reported (under one or under five years old), but to any child death implicitly declared in the data.

In almost all cases, I obtained decreasing proportions of child deaths when moving from the bottom ("poorest") to the top ("richest") quintile. That is, child mortality decreases with higher household socioeconomic status, as one would expect. The detailed results by quintile are not shown, but are also available upon request. In Table 8a, I present the same summary figure calculated before: the average change in child mortality across quintiles. The evidence is similar to results for school enrollment: the average change across quintiles is highly similar for all the PCA-based measures and it is larger (in absolute value) than the corresponding number for the logarithm of income per capita. For instance, the decrease in the child mortality rate across wealth quintiles using asset-based indices is about 4.6 percent for Colombia 1973 and 2.6 percent for South Africa 1996, while slightly smaller numbers (in absolute value) are found for quintiles based on the logarithm of income per capita.

Table 8a: Average Change in Child Mortality (Women Who Ever Gave Birth, Ages 18-30) across Wealth Quintiles, for Alternative Aggregation Methods [1/]

	Mean	Average difference across quintiles			
		Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	5.1	-1.67	-1.70	-1.72	-1.29
Brazil 2010	2.4	-0.60	-0.62	-0.62	-0.46
Cambodia 1998	16.5	-2.10	-1.87	-1.70	NA
Colombia 1973	17.5	-4.61	-4.64	-4.61	-4.11
Colombia 2005	2.8	-0.83	-0.83	-0.86	NA
Dominican Republic 2002	11.6	-0.78	-0.85	-0.84	-0.58
Mexico 1970	NA	NA	NA	NA	NA
Mexico 2000	6.6	-2.06	-2.04	-2.06	-1.64
Panama 1980	6.0	-1.50	-1.45	-1.44	-1.43
Panama 2010	3.3	-1.02	-1.02	-1.04	-0.34
Peru 1993	11.0	-4.46	-4.33	-4.36	NA
South Africa 1996	9.5	-2.57	-2.58	-2.58	-1.83
Simple average	***8.4***	***-2.02***	***-1.99***	***-1.98***	***-1.46***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. The average difference is calculated as the difference between top and bottom quintiles divided by four. Child mortality is derived from children ever born and children surviving, as declared by women who ever gave birth.

The regressions for child mortality are shown in Table 8b. The coefficients for household socioeconomic status as a control in a regression to explain child mortality appear to be statistically significant and consistently negative (except for only one statistically insignificant coefficient for the log of income per capita for Panama 2010). Similarly to school enrollment, the coefficients of all the asset-based indices are of similar size and tend to be stronger than the corresponding coefficient for the logarithm of income per capita. None of the measures is consistently larger than the rest if we examine the small differences across measures. The polychoric and ordinal PCA are again generally the first or second largest (negative) coefficients across samples, but differences are small on average.

Table 8b: Logit Model for Child Mortality (Women Who Ever Gave Birth, Ages 18-30) Wealth Index Coefficient (Odds-ratio) [1/]

	Asset count	Binary PCA	Polychoric PCA	Ordinal PCA	Log income per capita
Brazil 2000	**0.568**	**0.579**	**0.541**	**0.555**	**0.676**
	0.015	*0.011*	*0.012*	*0.012*	*0.012*
Brazil 2010	**0.739**	**0.777**	**0.731**	**0.762**	**0.816**
	0.010	*0.013*	*0.012*	*0.012*	*0.012*
Cambodia 1998	**NA**	**0.747**	**0.797**	**0.769**	**NA**
		0.013	*0.017*	*0.016*	
Colombia 1973	**NA**	**0.770**	**0.763**	**0.757**	**0.826**
		0.022	*0.018*	*0.018*	*0.019*
Colombia 2005	**0.754**	**0.716**	**0.680**	**0.706**	**NA**
	0.017	*0.016*	*0.016*	*0.016*	
Dominican Republic 2002	**0.906**	**0.921**	**0.900**	**0.903**	**0.919**
	0.022	*0.029*	*0.027*	*0.028*	*0.025*
Mexico 1970	**NA**	**NA**	**NA**	**NA**	**NA**
Mexico 2000	**0.718**	**0.693**	**0.696**	**0.708**	**0.864**
	0.010	*0.011*	*0.010*	*0.010*	*0.010*
Panama 1980	**0.884**	**0.780**	**0.780**	**0.785**	**0.863**
	0.039	*0.059*	*0.062*	*0.061*	*0.036*
Panama 2010	**0.606**	**0.574**	**0.548**	**0.571**	**1.005** #
	0.051	*0.042*	*0.048*	*0.571*	*0.030*
Peru 1993	**0.781**	**0.594**	**0.623**	**0.611**	**NA**
	0.031	*0.011*	*0.012*	*0.012*	
South Africa 1996	**NA**	**0.614**	**0.616**	**0.628**	**0.815**
		0.014	*0.015*	*0.015*	*0.014*

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. All estimates are statistically significant at the 1 percent level unless otherwise noted (### p>.01, ## p>.05, # p>.10). The table includes odds-ratio coefficients in bold and clustered standard errors in italic. Standard errors are clustered using mesoregions for Brazil, districts for Cambodia and Panama, municipalities for Colombia and Dominican Republic, states for Mexico, provinces for Peru, and magisterial districts for South Africa. The estimation sample is restricted to women who ever gave birth between 18 to 30 years old.

Control variables: age and age-squared, marital status, educational attainment (dummies for primary, secondary, and university), family size, and urban/rural.

The analysis was applied to four other outcomes: motherhood (having any children), primary school completion, secondary school completion, and literacy, all for persons between 18 and 30 years old. Results are shown in Tables A4a to A7b in the Appendix to this paper. The conclusions are qualitatively similar to the previous discussion for school enrollment and child mortality. The selected outcomes do change across the four measures of household socioeconomic status based on housing characteristics and assets, as it was expected. The estimated regression coefficients tend to be slightly larger for the asset-based measures than the logarithm of income per capita, but all are aligned in direction (positive or negative). The sign of the coefficients follows the hypothesized relation of the outcomes with household socioeconomic status (positive for the education ones and negative for having any children) and the largest coefficients are found for secondary school completion (the outcome with the largest variability in average rates across countries).

The (small) differences in coefficients between the four indices produced using the alternative aggregation methods also reveal patterns similar to those previously identified. The size of the coefficients is most comparable for the three PCA-based measures and the asset count. Even though none of the indices has consistently larger effects across samples and outcomes, the polychoric and the ordinal PCA are either the largest or second largest (or both) coefficients in almost all cases for any specific outcome. But differences in coefficients for the PCA-based measures are rather small in general.

5. **Discussion**

The construction of indices based on housing characteristics and asset ownership has been widely used when other measures of socioeconomic status are not available. This approach allows one to examine differences by socioeconomic status in outcomes of interest and to control for household socioeconomic status in regression analysis when the data do not include income or expenditures. Moreover, the use of asset-based indices also offers some advantages, such as smaller reporting errors and lower relative data collection cost than income or expenditures (Kolenikov and Angeles, 2009; Filmer and Scott, 2012). Principal components analysis (PCA) on data recoded to binary indicators (Filmer and Pritchett, 2001) is one of the most frequently used procedures to construct such an index. In this study, I compared the relative performance of the commonly used binary PCA to aggregation methods that are designed to handle discrete data.

The evidence presented in this paper indicates that methods based on ordinal asset data have higher agreement in rankings, but the common procedure of PCA on dichotomized data (and even the relatively simple method of counting assets) also has reasonable agreement with the other measures. More importantly, differences in variables of interest by wealth quintiles and the estimated coefficients on indicators of wealth in regression analysis are similar in size for all asset-based measures across a wide set of education, fertility, and mortality outcomes. Even though the asset-based indices show only moderate agreement with rankings based on the logarithm of income per capita, none of the asset-based indices outperformed the rest on this aspect. Furthermore, larger differences by wealth quintiles and in regression analysis are observed for the asset-based indices with respect to income, possibly due to noise in the income variable. Below, I discuss some specific aspects on the relative performance of each aggregation method to propose more specific recommendations.

The asset count is a relatively simple method to implement, given that it requires only adding the number of positive responses for a set of items. The coefficients for socioeconomic status measured through this proxy are of similar size but tend to be slightly smaller than the PCA-based methods. However, this method produces a more limited set of values that make more difficult to do a more fine classification of households, which created problems in the definition of wealth quintiles. Furthermore, in order to include other variables and not only durable assets, it would be necessary to recode them into binary indicators. For instance, the predominant type of flooring (an ordinal variable) cannot be added directly to an item count, but it could be recoded into "good flooring materials" (including cement and finished flooring types) and "bad flooring materials" (including dirt and similar quality materials). This recoding process may require to impose additional assumptions to determine which are "good" or "bad" categories in an ordinal variable. Therefore, the use of this measure is not recommended.

The PCA-based methods showed varying results regarding the different criteria applied to assess their performance. Similar to previous studies, weights assigned through binary PCA do not show the expected monotonicity property (from "worst" to "best" option in ordinal variables), in contrast to those derived from polychoric correlations. In addition, the proportion of variance explained had considerable differences across methods, being the largest for the polychoric PCA, followed by the ordinal PCA and then by the binary PCA. Nevertheless, all three PCA-based measures had reasonable (but not identical) agreement in terms of the household rankings, while results by quintiles or in regression analysis for the selected outcomes were strikingly similar. Therefore, if the objective is to examine differences by household socioeconomic status or the use of the measure as a control variable in a regression, any of the PCA-based measures appears to have a similar performance. In this sense, despite recommendations given by previous research (Howe *et al*., 2008; Kolenikov and Angeles, 2009), results suggest a relatively similar performance of the PCA procedure on dichotomized data with respect to methods based on ordinal data. However, given the discrepancies of rankings against income, the researcher should use the asset-based measures with more caution if the objective is different, such as identifying the poor within a certain population.

Polychoric correlations are designed to handle discrete data, in contrast to the standard correlation methods applied for PCA. However, there is one minor disadvantage of polychoric correlations. Leaving aside that they are computationally more intensive than standard correlation methods, the minor problem concerns their calculation for categories with small (or zero) frequencies. As discussed in the methodology section, the polychoric correlations are calculated from a likelihood function (equations 7, 8a, and 8b), which requires having non-zero frequencies for the combinations of values of the categorical variables. For example, the pairwise polychoric correlation may not be defined between variables such as electricity and ownership of an electric appliance, since at least one combination of categories may have zero frequencies. This is a minor disadvantage of polychoric correlations that does not recommend against their use, but that may require some data manipulation to work around the issue. For instance, a possible solution is to add a trivial small number to the frequencies associated to the different combinations of values of the categorical variables, so that none of these combinations has small (or zero) frequencies. In the few cases found in the data used in this study, alternatively, I combined categories with small frequencies with those that appear to represent similar positions in the scale. Based on the evidence shown in this paper on household rankings and the relation with selected outcomes, any of the PCA-based measures seem to produce similar results. However, given the possible difficulties in the calculation of polychoric correlations, a practical recommendation would be to implement either the binary PCA or the ordinal PCA. No striking differences in performance have been found between these two measures.

6. Appendix: Additional Tables and Figures

Table A1: Detailed Asset Variables Available for Selected Census Samples

Brazil 2000	Brazil 2010	Cambodia 1998
1 Dwelling type	1 Dwelling type	1 Ownership of dwelling
2 Rooms (number)	2 Dwelling ownership	2 Lighting
3 Sleeping rooms (number)	3 External walls material	3 Cooking fuel
4 Ownership of dwelling	4 Rooms (number)	4 Toilet
5 Ownership of land	5 Bedrooms (number)	5 Water supply
6 Water source	6 Toilet	6 Rooms (number)
7 Piped water	7 Sewage	
8 Bathrooms (number)	8 Water supply	
9 Toilet	9 Piped water	
10 Waste water	10 Garbage destination	
11 Trash	11 Electricity	
12 Radio	12 Radio	
13 Refrigerator	13 TV	
14 VCR	14 Washer	
15 Washing machine	15 Refrigerator	
16 Microwave	16 Cell phone	
17 Telephone	17 Phone	
18 Computer	18 Computer	
19 TV (number)	19 Motorcycle	
20 Private car (number)	20 Auto	
21 Air conditioner (number)		

Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Table A1 (continued): Detailed Asset Variables Available for Selected Census Samples

Colombia 1973	Colombia 2005	Dominican Republic 2002
1 Dwelling type	1 Dwelling type	1 Dwelling access
2 Predominant roof material	2 Wall material	2 Outer walls material
3 Outside wall material	3 Floor material	3 Roof material
4 Floor material	4 Trash removal	4 Floor material
5 Rooms (number)	5 Electricity	5 Rooms (number)
6 Bedrooms (number)	6 Sewage drains	6 Kitchen
7 Room for cooking	7 Running water	7 Tenancy
8 Water source	8 Natural gas	8 Bedrooms (number)
9 Toilet	9 Telephone	9 Cooking fuel
10 Use of toilet	10 Toilet type	10 Lighting
11 Location of toilet	11 Location of water service	11 Water source
12 Lighting	12 Bathrooms (number)	12 Toilet
13 Ownership of dwelling	13 Kitchen	13 Waste removal
	14 Ownership of dwelling	14 Refrigerator
	15 Rooms (number)	15 Stove
	16 Bedrooms (number)	16 Washing machine
	17 Source of water for cooking	17 Television
	18 Kitchen	18 Air conditioning
	19 Fuel for cooking	19 Radio/stereo
	20 Fridge	20 Car
	21 Washing machine	21 Cistern
	22 Stereo	22 Computer
	23 Water heater	23 Converter
	24 Shower	24 Generator
	25 Blender	25 Landline or cellphone
	26 Electric or gas oven	26 Internet
	27 Air conditioner	
	28 Fan	
	29 TV color	
	30 Computer	
	31 Microwave	
	32 Bike (number)	
	33 Motorcycle (number)	
	34 Ships, sailboats, boats (number)	
	35 Autos (number)	

Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Table A1 (continued): Detailed Asset Variables Available for Selected Census Samples

Mexico 1970	Mexico 2000	Panama 1980
1 Bath with running water	1 Dwelling type	1 Dwelling type
2 Kitchen	2 Walls	2 Rooms (number)
3 Rooms (number)	3 Roof	3 Bedrooms (number)
4 Ownership	4 Floors	4 Kitchen
5 Wall material	5 Kitchen	5 Ownership
6 Floor material	6 Bedrooms (number)	6 Exterior walls material
7 Roof material	7 Rooms (number)	7 Roof material
8 Piped water	8 Water	8 Floor material
9 Sewer connection	9 Toilet	9 Drinking water source
10 Fuel for cooking	10 Sewer	10 Sewer facilities
11 Electricity	11 Electricity	11 Bathroom
12 Radio	12 Fuel for cooking	12 Lighting
13 TV	13 Dwelling ownership	13 Fuel for cooking
	14 Radio	14 Television
	15 Television	15 Radio
	16 Videocassette player	16 Telephone
	17 Blender	17 Refrigerator
	18 Refrigerator	18 Washing machine
	19 Washing machine	19 Sewing machine
	20 Telephone	
	21 Hot water heater	
	22 Car, van, or light truck	
	23 Computer	
	24 Trash disposal	

Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Table A1 (continued): Detailed Asset Variables Available for Selected Census Samples

Panama 2010	Peru 1993	South Africa 1996
1 Dwelling type	1 Water source	1 Dwelling type
2 Dwelling ownership	2 Lighting (electricity)	2 Rooms (number)
3 Wall material	3 Rooms (number)	3 Dwelling ownership
4 Roof material	4 Walls	4 Fuel for cooking
5 Floor material	5 Floor	5 Fuel for heating
6 Rooms (number)	6 Sewage	6 Fuel for lighting
7 Bedrooms (number)	7 Roof	7 Water supply
8 Water supply	8 Dwelling ownership	8 Toilet
9 Toilet	9 Dwelling type	9 Refuse disposal
10 Lighting	10 Vacuum cleaner	10 Telephone
11 Garbage disposal	11 Car for private use	
12 Fuel for cooking	12 Car for work use	
13 Cook stove	13 Toilet	
14 Refrigerator	14 Bicycle	
15 Washing machine	15 Light truck for work	
16 Sewing machine	16 Room for cooking	
17 Residential phone	17 Computer	
18 Radio (number)	18 Washer	
19 Electric fan (number)	19 Floor polisher	
20 Air conditioner (number)	20 Sewing machine	
21 Cell phone (number)	21 Knitting machine	
22 Automobile (number)	22 Motorcycle	
23 TV (number)	23 Radio	
24 Cable TV	24 Refrigerator	
25 Computer (number)	25 Stereo	
26 Internet	26 Phone	
	27 Tricycle	
	28 TV black/white	
	29 TV color	
	30 Video camera	

Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Table A2a: Comparison of Classification by Quintiles
Households Classified in a Higher Quintile with respect to Polychoric PCA Index [1/]

	Asset count	Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	6.9	8.3	2.7	0.0	28.6
Brazil 2010	14.3	12.0	2.4	0.0	33.2
Cambodia 1998	NA	27.6	3.1	0.0	NA
Colombia 1973	NA	10.2	2.7	0.0	32.3
Colombia 2005	13.8	6.1	1.6	0.0	NA
Dominican Republic 2002	12.4	10.8	2.4	0.0	32.2
Mexico 1970	13.0	8.1	1.9	0.0	30.6
Mexico 2000	9.4	7.8	2.7	0.0	30.8
Panama 1980	12.1	10.4	3.5	0.0	32.4
Panama 2010	19.7	10.2	3.3	0.0	30.1
Peru 1993	7.7	8.3	2.0	0.0	NA
South Africa 1996	NA	7.2	3.0	0.0	29.8
Simple average	***12.1***	***10.6***	***2.6***	***0.0***	***31.1***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. Classified in a higher quintile by the alternative measure being compared to the polychoric PCA index.

Table A2b: Comparison of Classification by Quintiles
Households Classified in a Higher Quintile with respect to Log of Income per Capita [1/]

	Asset count	Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	19.1	28.4	27.9	28.1	0.0
Brazil 2010	23.2	29.1	28.5	28.3	0.0
Cambodia 1998	NA	NA	NA	NA	NA
Colombia 1973	NA	31.7	31.6	31.5	0.0
Colombia 2005	NA	NA	NA	NA	NA
Dominican Republic 2002	29.4	33.3	33.7	33.6	0.0
Mexico 1970	20.5	30.5	31.1	31.0	0.0
Mexico 2000	23.8	26.1	26.8	26.7	0.0
Panama 1980	19.1	26.0	25.5	25.9	0.0
Panama 2010	30.7	31.4	31.4	31.5	0.0
Peru 1993	NA	NA	NA	NA	NA
South Africa 1996	NA	31.4	31.3	31.8	0.0
Simple average	***23.7***	***29.8***	***29.7***	***29.8***	***0.0***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. Classified in a higher quintile by the alternative measure being compared to the logarithm of income per capita.

Table A3a: Comparison of Classification by Quintiles
Households Classified in a Lower Quintile with respect to Polychoric PCA Index [1/]

	Asset count	Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	35.9	9.3	2.5	0.0	28.1
Brazil 2010	35.6	13.1	2.4	0.0	28.3
Cambodia 1998	NA	39.6	6.4	0.0	NA
Colombia 1973	NA	10.3	2.7	0.0	31.5
Colombia 2005	26.3	9.2	2.2	0.0	NA
Dominican Republic 2002	30.6	10.8	2.4	0.0	33.6
Mexico 1970	51.9	8.2	2.0	0.0	31.0
Mexico 2000	22.8	9.9	2.7	0.0	26.7
Panama 1980	34.5	10.3	3.5	0.0	25.9
Panama 2010	19.8	10.1	3.3	0.0	31.5
Peru 1993	43.3	8.4	2.0	0.0	NA
South Africa 1996	NA	7.8	4.6	0.0	31.8
Simple average	***33.4***	***12.2***	***3.1***	***0.0***	***29.8***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. Classified in a lower quintile by the alternative measure being compared to the polychoric PCA index.

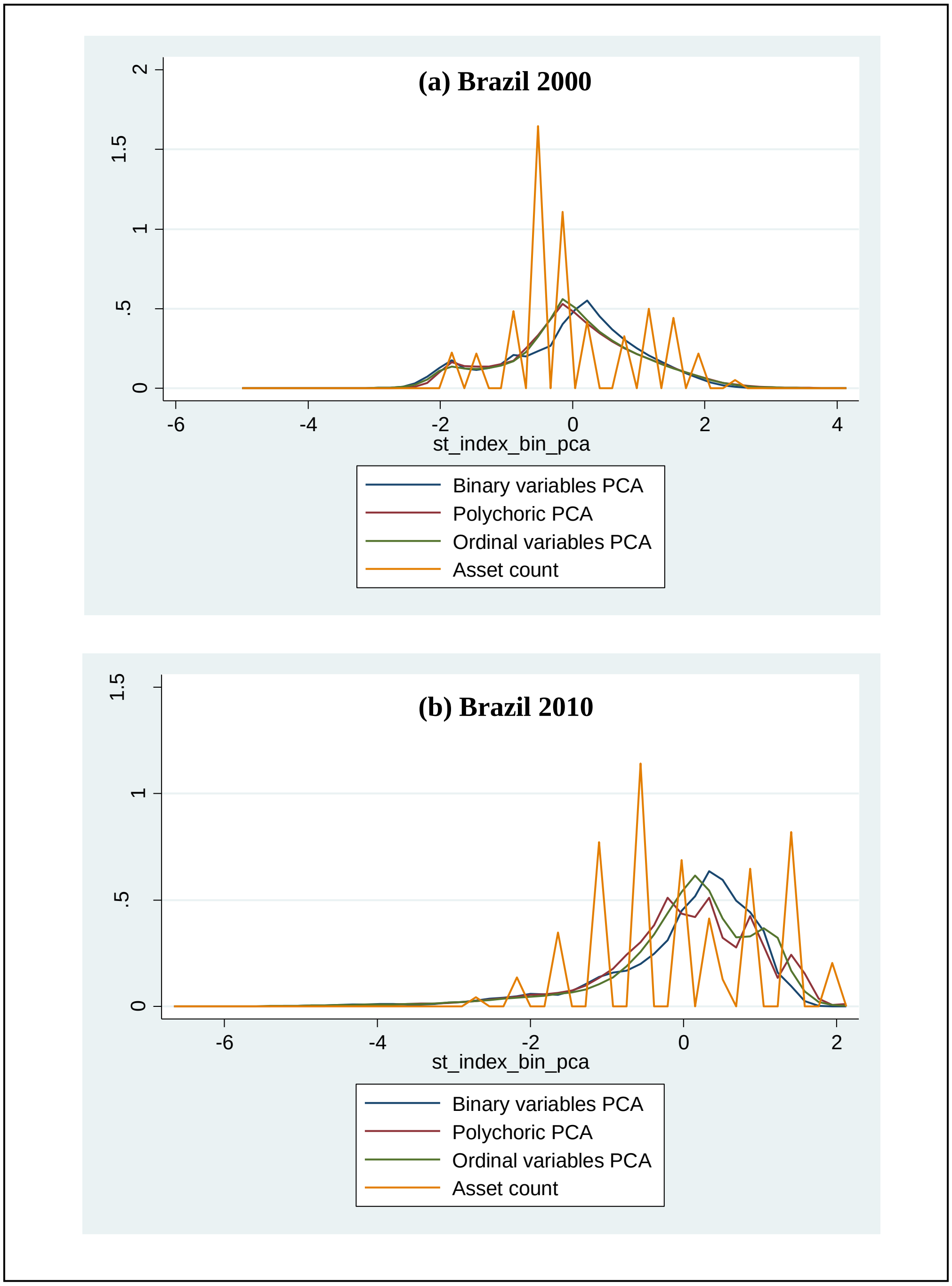
Table A3b: Comparison of Classification by Quintiles
Households Classified in a Lower Quintile with respect to Log of Income per Capita [1/]

	Asset count	Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	40.5	28.8	28.2	28.6	0.0
Brazil 2010	41.2	34.1	33.1	33.2	0.0
Cambodia 1998	NA	NA	NA	NA	NA
Colombia 1973	NA	32.3	32.3	32.3	0.0
Colombia 2005	NA	NA	NA	NA	NA
Dominican Republic 2002	38.1	32.3	32.1	32.2	0.0
Mexico 1970	51.2	30.9	30.6	30.6	0.0
Mexico 2000	35.3	31.2	30.8	30.8	0.0
Panama 1980	45.5	33.1	32.2	32.4	0.0
Panama 2010	33.8	30.2	30.1	30.1	0.0
Peru 1993	NA	NA	NA	NA	NA
South Africa 1996	NA	29.2	30.0	29.8	0.0
Simple average	***40.8***	***31.3***	***31.1***	***31.1***	***0.0***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

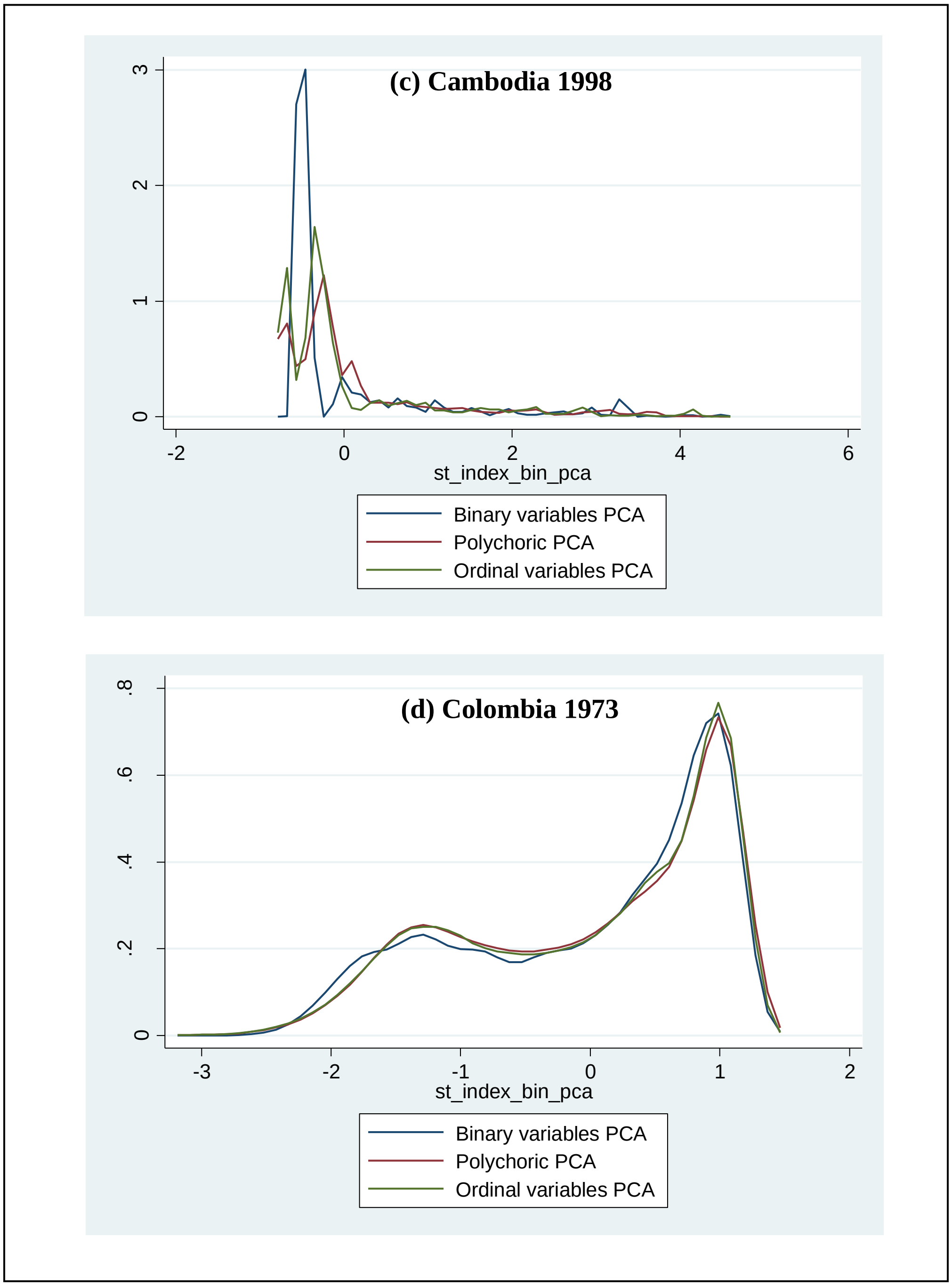
1. Classified in a lower quintile by the alternative measure being compared to the logarithm of income per capita.

Figure A1: Kernel Density Estimates for Wealth Indices Based on Alternative Aggregation Methods



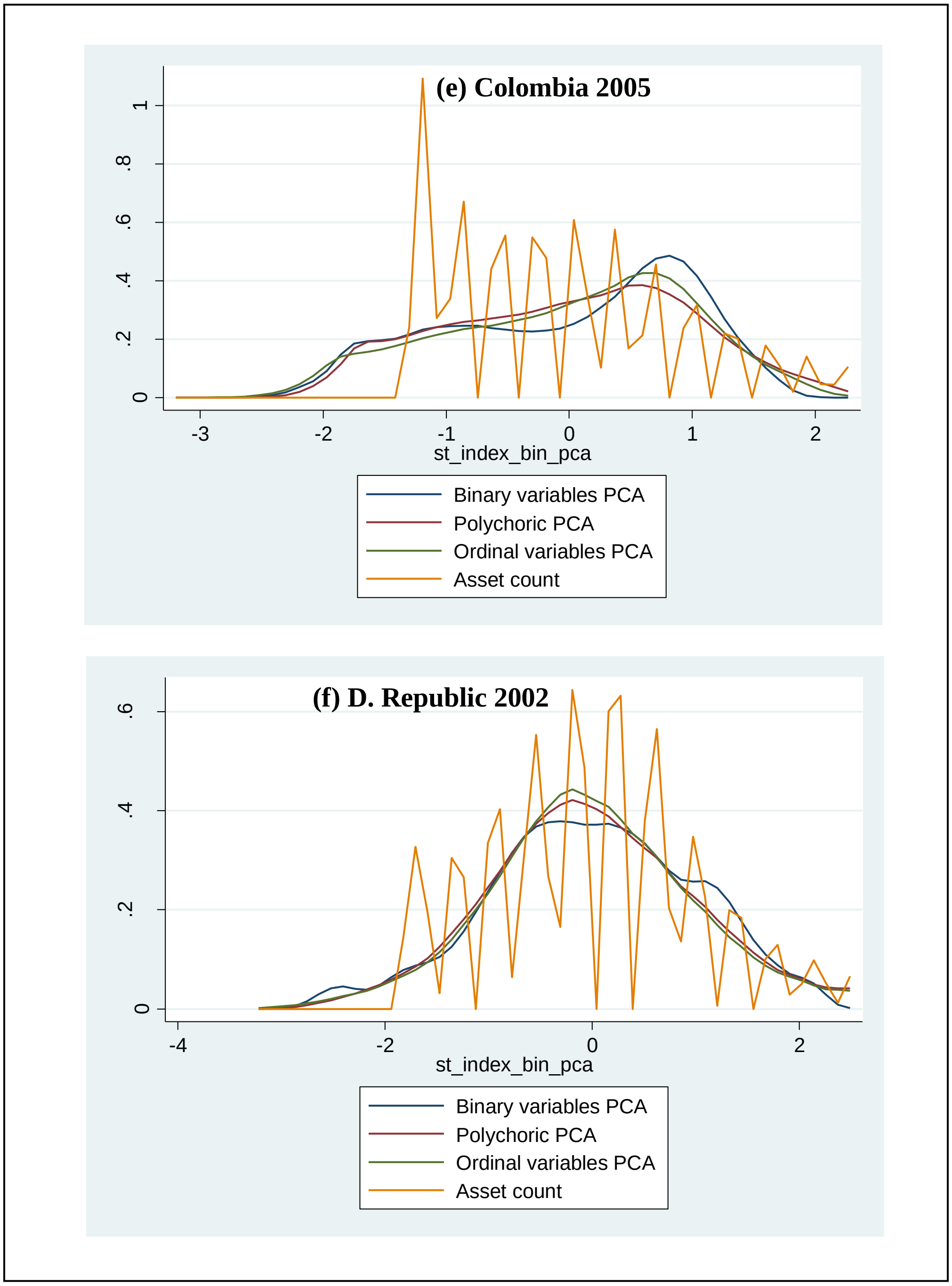
Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Figure A1 (continued): Kernel Density Estimates for Wealth Indices Based on Alternative Aggregation Methods



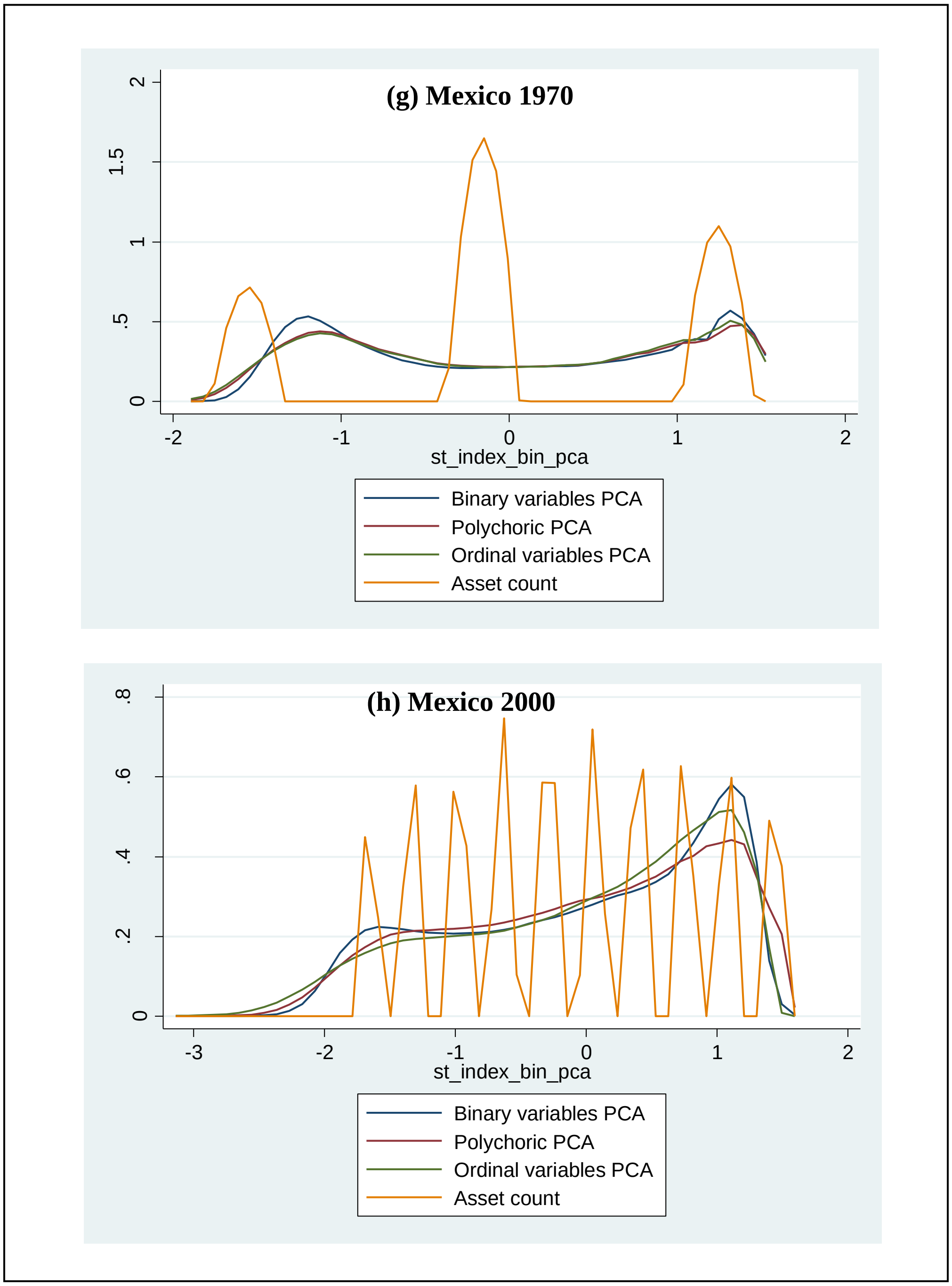
Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Figure A1 (continued): Kernel Density Estimates for Wealth Indices Based on Alternative Aggregation Methods



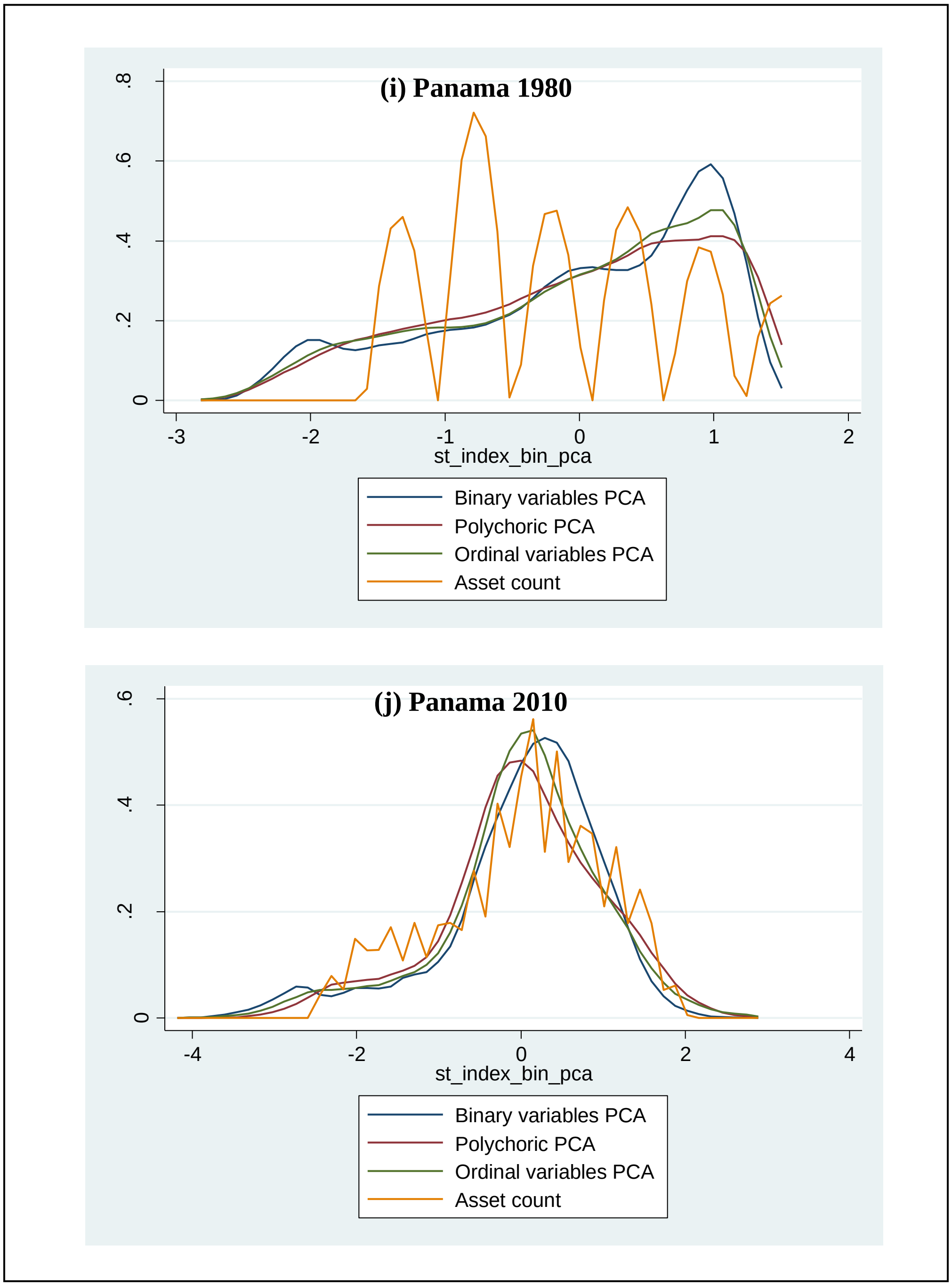
Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Figure A1 (continued): Kernel Density Estimates for Wealth Indices Based on Alternative Aggregation Methods



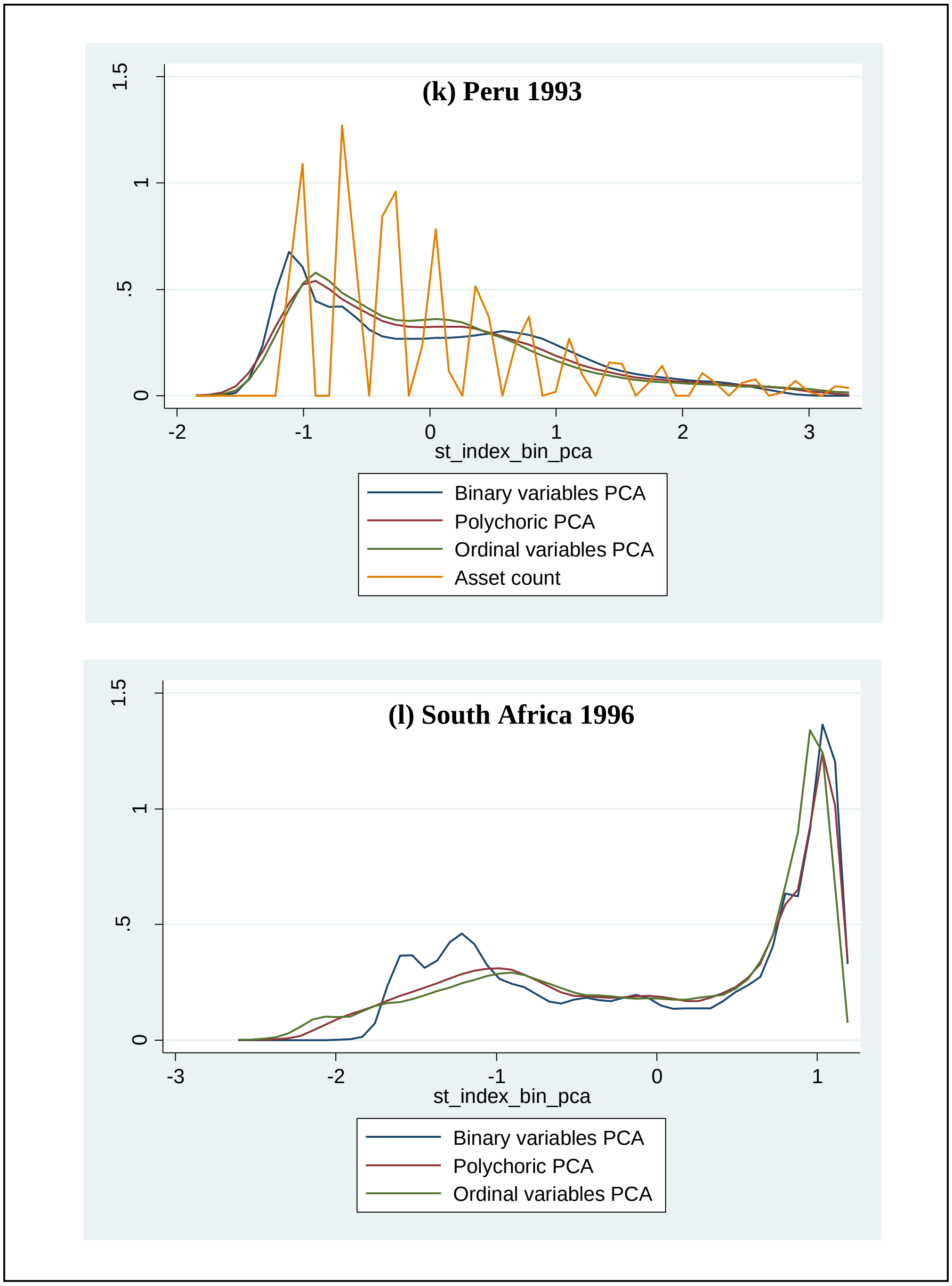
Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Figure A1 (continued): Kernel Density Estimates for Wealth Indices Based on Alternative Aggregation Methods



Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Figure A1 (continued): Kernel Density Estimates for Wealth Indices Based on Alternative Aggregation Methods



Data source: Integrated Public Use Microdata Series (IPUMS) - International.

Table A4a: Average Change in Literacy (persons ages 18-30) across Wealth Quintiles, for Alternative Aggregation Methods [1/]

	Mean	Average difference across quintiles			
		Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	93.4	4.21	4.30	4.33	3.73
Brazil 2010	96.8	1.88	1.98	1.98	1.71
Cambodia 1998	75.3	5.23	3.60	3.29	NA
Colombia 1973	87.7	7.01	6.75	6.77	4.56
Colombia 2005	95.0	2.71	2.74	2.87	NA
Dominican Republic 2002	92.5	4.03	4.00	4.02	2.64
Mexico 1970	79.8	9.28	9.33	9.38	8.43
Mexico 2000	95.9	3.10	3.11	3.13	2.57
Panama 1980	92.1	2.03	2.10	2.10	2.27
Panama 2010	97.1	1.42	1.42	1.44	1.24
Peru 1993	94.1	4.43	4.17	4.17	NA
South Africa 1996	NA	NA	NA	NA	NA
Simple average	***90.9***	***4.12***	***3.95***	***3.95***	***3.39***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. The average difference is calculated as the difference between top and bottom quintiles divided by four.

Table A4b: Logit Model for Literacy (Persons Ages 18-30) Wealth Index Coefficient (Odds-ratio) [1/]

	Asset count	Binary PCA	Polychoric PCA	Ordinal PCA	Log income per capita
Brazil 2000	**3.744**	**3.069**	**3.663**	**3.408**	**2.638**
	0.108	*0.126*	*0.140*	*0.134*	*0.081*
Brazil 2010	**2.410**	**1.769**	**2.050**	**1.831**	**1.886**
	0.047	*0.054*	*0.068*	*0.055*	*0.045*
Cambodia 1998	**NA**	**1.406**	**1.310**	**1.377**	**NA**
		0.037	*0.029*	*0.027*	
Colombia 1973	**NA**	**2.228**	**2.042**	**2.038**	**1.366**
		0.054	*0.041*	*0.042*	*0.031*
Colombia 2005	**2.475**	**2.949**	**3.104**	**2.882**	**NA**
	0.108	*0.236*	*0.235*	*0.206*	
Dominican Republic 2002	**2.260**	**2.667**	**2.862**	**2.800**	**1.545**
	0.082	*0.077*	*0.096*	*0.089*	*0.035*
Mexico 1970	**1.636**	**2.022**	**2.057**	**2.063**	**1.654**
	0.043	*0.079*	*0.083*	*0.084*	*0.065*
Mexico 2000	**3.401**	**3.493**	**3.419**	**3.135**	**1.590**
	0.187	*0.154*	*0.131*	*0.097*	*0.066*
Panama 1980	**1.812**	**1.701**	**1.813**	**1.760**	**1.428**
	0.307	*0.204*	*0.227*	*0.223*	*0.101*
Panama 2010	**2.432**	**1.957**	**2.259**	**2.065**	**1.142**
	0.209	*0.119*	*0.157*	*0.119*	*0.059*
Peru 1993	**1.625**	**2.011**	**1.911**	**1.877**	**NA**
	0.165	*0.218*	*0.222*	*0.244*	
South Africa 1996	**NA**	**NA**	**NA**	**NA**	**NA**

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. All estimates are statistically significant at the 1 percent level unless otherwise noted (### p>.01, ## p>.05, # p>.10). The table includes odds-ratio coefficients in bold and clustered standard errors in italic. Standard errors are clustered using mesoregions for Brazil, districts for Cambodia and Panama, municipalities for Colombia and Dominican Republic, states for Mexico, provinces for Peru, and magisterial districts for South Africa. The estimation sample is restricted to persons 18 to 30 years old that are not household heads.

Control variables: sex, age, and age-squared of the person; sex, age, age-squared, and educational attainment of household head (dummies for primary, secondary, and university); urban/rural.

Table A5a: Average Change in Primary School Completion (Persons Ages 18-30) across Wealth Quintiles, for Alternative Aggregation Methods [1/]

	Mean	Average difference across quintiles			
		Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	63.3	15.90	16.01	16.02	14.26
Brazil 2010	84.9	6.96	7.28	7.29	6.69
Cambodia 1998	37.1	9.19	7.44	6.92	NA
Colombia 1973	47.1	16.76	16.59	16.56	13.90
Colombia 2005	86.4	9.20	9.30	9.44	NA
Dominican Republic 2002	75.6	10.21	10.11	10.15	7.21
Mexico 1970	35.6	16.76	17.40	17.36	15.70
Mexico 2000	87.0	8.40	8.42	8.44	6.52
Panama 1980	78.3	8.28	8.37	8.37	8.36
Panama 2010	92.3	3.97	3.99	4.04	3.31
Peru 1993	77.2	13.94	13.48	13.51	NA
South Africa 1996	84.5	6.45	6.68	6.64	5.21
Simple average	***70.8***	***10.50***	***10.42***	***10.39***	***9.02***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. The average difference is calculated as the difference between top and bottom quintiles divided by four.

Table A5b: Logit Model for Primary School Completion (Persons Ages 18-30) Wealth Index Coefficient (Odds-ratio) [1/]

	Asset count	Binary PCA	Polychoric PCA	Ordinal PCA	Log income per capita
Brazil 2000	**2.866**	**2.881**	**3.100**	**3.073**	**2.546**
	0.068	*0.114*	*0.102*	*0.106*	*0.059*
Brazil 2010	**1.913**	**1.714**	**1.915**	**1.782**	**1.740**
	0.028	*0.032*	*0.037*	*0.035*	*0.026*
Cambodia 1998	**NA**	**1.464**	**1.410**	**1.462**	**NA**
		0.036	*0.030*	*0.032*	
Colombia 1973	**NA**	**2.264**	**2.126**	**2.165**	**1.574**
		0.037	*0.035*	*0.034*	*0.028*
Colombia 2005	**2.514**	**2.450**	**2.643**	**2.493**	**NA**
	0.048	*0.068*	*0.072*	*0.064*	
Dominican Republic 2002	**1.851**	**2.268**	**2.354**	**2.341**	**1.550**
	0.046	*0.043*	*0.052*	*0.051*	*0.030*
Mexico 1970	**1.722**	**2.468**	**2.540**	**2.568**	**1.776**
	0.041	*0.119*	*0.106*	*0.109*	*0.078*
Mexico 2000	**2.576**	**2.778**	**2.770**	**2.641**	**1.490**
	0.113	*0.138*	*0.125*	*0.110*	*0.055*
Panama 1980	**1.921**	**2.065**	**2.161**	**2.125**	**1.689**
	0.188	*0.134*	*0.153*	*0.148*	*0.061*
Panama 2010	**2.336**	**1.989**	**2.299**	**2.120**	**1.169**
	0.131	*0.088*	*0.110*	*0.086*	*0.043*
Peru 1993	**1.848**	**2.450**	**2.411**	**2.400**	**NA**
	0.161	*0.203*	*0.254*	*0.295*	
South Africa 1996	**NA**	**1.945**	**1.996**	**1.955**	**1.463**
		0.043	*0.042*	*0.040*	*0.022*

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. All estimates are statistically significant at the 1 percent level unless otherwise noted (### p>.01, ## p>.05, # p>.10). The table includes odds-ratio coefficients in bold and clustered standard errors in italic. Standard errors are clustered using mesoregions for Brazil, districts for Cambodia and Panama, municipalities for Colombia and Dominican Republic, states for Mexico, provinces for Peru, and magisterial districts for South Africa. The estimation sample is restricted to persons 18 to 30 years old that are not household heads.

Control variables: sex, age, and age-squared of the person; sex, age, age-squared, and educational attainment of household head (dummies for primary, secondary, and university); urban/rural.

Table A6a: Average Change in Secondary School Completion (Persons Ages 18-30) across Wealth Quintiles, for Alternative Aggregation Methods [1/]

	Mean	Average difference across quintiles			
		Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	29.4	16.52	16.58	16.46	16.32
Brazil 2010	50.1	14.80	15.46	15.45	15.54
Cambodia 1998	4.3	3.14	3.08	3.01	NA
Colombia 1973	9.2	6.30	6.38	6.39	7.22
Colombia 2005	53.4	17.82	18.09	18.15	NA
Dominican Republic 2002	29.0	13.91	13.92	14.01	11.29
Mexico 1970	4.2	3.37	3.46	3.48	3.26
Mexico 2000	29.2	15.19	15.19	15.17	13.18
Panama 1980	27.4	14.55	15.23	15.07	14.76
Panama 2010	54.0	16.18	16.15	16.21	10.88
Peru 1993	52.4	18.56	18.26	18.31	NA
South Africa 1996	30.2	15.75	15.60	15.04	15.82
Simple average	***31.1***	***13.01***	***13.12***	***13.06***	***12.03***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. The average difference is calculated as the difference between top and bottom quintiles divided by four.

Table A6b: Logit Model for Secondary School Completion (Persons Ages 18-30) Wealth Index Coefficient (Odds-ratio) [1/]

	Asset count	Binary PCA	Polychoric PCA	Ordinal PCA	Log income per capita
Brazil 2000	**2.637**	**3.406**	**3.139**	**3.246**	**3.011**
	0.044	*0.097*	*0.068*	*0.072*	*0.051*
Brazil 2010	**1.991**	**2.167**	**2.372**	**2.403**	**2.147**
	0.031	*0.062*	*0.058*	*0.073*	*0.037*
Cambodia 1998	**NA**	**1.700**	**1.767**	**1.750**	**NA**
		0.058	*0.055*	*0.055*	
Colombia 1973	**NA**	**3.481**	**3.189**	**3.321**	**2.066**
		0.311	*0.210*	*0.243*	*0.050*
Colombia 2005	**2.230**	**2.495**	**2.629**	**2.630**	**NA**
	0.024	*0.055*	*0.045*	*0.053*	
Dominican Republic 2002	**1.792**	**2.327**	**2.301**	**2.290**	**1.777**
	0.031	*0.058*	*0.043*	*0.048*	*0.033*
Mexico 1970	**1.762**	**2.685**	**2.782**	**2.816**	**1.519**
	0.066	*0.313*	*0.293*	*0.306*	*0.067*
Mexico 2000	**2.223**	**2.900**	**2.887**	**3.008**	**2.000**
	0.064	*0.193*	*0.162*	*0.192*	*0.036*
Panama 1980	**1.864**	**2.641**	**2.609**	**2.750**	**2.564**
	0.121	*0.148*	*0.173*	*0.151*	*0.155*
Panama 2010	**2.480**	**2.627**	**2.666**	**2.650**	**1.241**
	0.080	*0.135*	*0.103*	*0.114*	*0.023*
Peru 1993	**1.750**	**2.396**	**2.368**	**2.303**	**NA**
	0.131	*0.178*	*0.242*	*2.303*	
South Africa 1996	**NA**	**2.308**	**2.392**	**2.338**	**2.047**
		0.060	*0.063*	*0.061*	*0.047*

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. All estimates are statistically significant at the 1 percent level unless otherwise noted (### p>.01, ## p>.05, # p>.10). The table includes odds-ratio coefficients in bold and clustered standard errors in italic. Standard errors are clustered using mesoregions for Brazil, districts for Cambodia and Panama, municipalities for Colombia and Dominican Republic, states for Mexico, provinces for Peru, and magisterial districts for South Africa. The estimation sample is restricted to persons 18 to 30 years old that are not household heads.

Control variables: sex, age, and age-squared of the person; sex, age, age-squared, and educational attainment of household head (dummies for primary, secondary, and university); urban/rural.

Table A7a: Average Change in Motherhood (Women Ages 18-30) across Wealth Quintiles, for Alternative Aggregation Methods [1/]

	Mean	Average difference across quintiles			
		Binary PCA	Ordinal PCA	Polychoric PCA	Log income per capita
Brazil 2000	53.0	-9.93	-10.30	-10.18	-11.21
Brazil 2010	46.9	-9.83	-10.42	-10.47	-12.53
Cambodia 1998	55.1	-0.74	-3.07	-2.71	NA
Colombia 1973	64.7	-7.16	-7.99	-7.90	-7.14
Colombia 2005	55.6	-9.10	-9.63	-9.59	NA
Dominican Republic 2002	67.7	-8.68	-9.06	-9.11	-6.71
Mexico 1970	58.4	-6.71	-7.18	-7.12	-7.55
Mexico 2000	56.1	-7.90	-8.63	-8.51	-7.59
Panama 1980	69.2	-7.03	-8.48	-8.33	-7.94
Panama 2010	56.9	-9.76	-10.42	-10.29	-7.97
Peru 1993	57.5	-11.57	-11.83	-11.87	NA
South Africa 1996	63.5	-6.40	-6.03	-5.75	-5.82
Simple average	***58.7***	***-7.90***	***-8.59***	***-8.49***	***-8.27***

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. The average difference is calculated as the difference between top and bottom quintiles divided by four.

Table A7b: Logit Model for Motherhood (Women Ages 18-30) Wealth Index Coefficient (Odds-ratio) [1/]

	Asset count	Binary PCA	Polychoric PCA	Ordinal PCA	Log income per capita
Brazil 2000	**0.787**	**0.762**	**0.758**	**0.751**	**0.525**
	0.009	*0.009*	*0.009*	*0.009*	*0.010*
Brazil 2010	**0.692**	**0.754**	**0.693**	**0.720**	**0.472**
	0.009	*0.014*	*0.013*	*0.015*	*0.012*
Cambodia 1998	**NA**	**0.810**	**0.831**	**0.826**	**NA**
		0.015	*0.017*	*0.015*	
Colombia 1973	**NA**	**0.710**	**0.679**	**0.678**	**0.641**
		0.024	*0.023*	*0.024*	*0.016*
Colombia 2005	**0.803**	**0.923**	**0.851**	**0.871**	**NA**
	0.012	*0.028*	*0.025*	*0.027*	
Dominican Republic 2002	**0.775**	**0.739**	**0.718**	**0.723**	**0.744**
	0.014	*0.021*	*0.014*	*0.014*	*0.016*
Mexico 1970	**0.935**	**0.927**	**0.911**	**0.912**	**0.768**
	0.029	*0.031*	*0.031*	*0.031*	*0.029*
Mexico 2000	**0.809**	**0.803**	**0.776**	**0.776**	**0.675**
	0.020	*0.021*	*0.022*	*0.022*	*0.035*
Panama 1980	**0.806**	**0.791**	**0.765**	**0.770**	**0.552**
	0.035	*0.051*	*0.049*	*0.054*	*0.035*
Panama 2010	**0.754**	**0.757**	**0.715**	**0.722**	**0.796**
	0.043	*0.059*	*0.048*	*0.053*	*0.020*
Peru 1993	**0.723**	**0.638**	**0.646**	**0.658**	**NA**
	0.023	*0.026*	*0.033*	*0.036*	
South Africa 1996	**NA**	**0.797**	**0.805**	**0.830**	**0.715**
		0.016	*0.016*	*0.016*	*0.018*

Data source: Integrated Public Use Microdata Series (IPUMS) - International. NA = Not available

1. All estimates are statistically significant at the 1 percent level unless otherwise noted (### p>.01, ## p>.05, # p>.10). The table includes odds-ratio coefficients in bold and clustered standard errors in italic. Standard errors are clustered using mesoregions for Brazil, districts for Cambodia and Panama, municipalities for Colombia and Dominican Republic, states for Mexico, provinces for Peru, and magisterial districts for South Africa.

Control variables: age and age-squared, marital status, educational attainment (dummies for primary, secondary, and university), family size, and urban/rural.