

The Global Consumption and Income Project (GCIP): An Overview

Rahul Lahoti¹, Arjun Jayadev², and Sanjay G. Reddy³

This Version: March 27th, 2016. Future versions and related materials will be made available on www.gcip.info⁴

We introduce two separate datasets (The Global Consumption Dataset (GCD) and The Global Income Dataset (GID)) making possible an unprecedented portrait of consumption and income of persons over time, within and across countries, around the world. The current benchmark version of the dataset presents estimates of monthly real consumption and income for every percentile of the population (a ‘consumption/income profile’) for more than 160 countries and more than half a century (1960-2015). We describe the construction of the datasets and demonstrate possible uses by presenting some sample results concerning the distribution of consumption, poverty and inequality in the world.

Keywords: Consumption, Income, Growth, Global Income Distribution, Poverty, Inclusive Growth, Inequality

JEL Classification: B41, C80, I30, I32, O10, O15

¹ Dept. of Economics, Georg-August-Universität Göttingen; rahul.lahoti@gmail.com

² Dept. of Economics, University of Massachusetts at Boston and Azim Premji University; arjunjayadev@gmail.com

³ Dept. of Economics, The New School for Social Research and Initiative for Policy Dialogue; [reddysanjayg@gmail.com](mailto:redmysanjayg@gmail.com)

⁴ We are thankful for the important contribution to this project made by Michalis Nikiforos, who among other things, executed much of the work required to construct an earlier version of the database. We are also most grateful to Ingrid Kvangraven, Gibran Mian, Ibrahim Shikaki, Shenuque Tissera and Brandt Weathers for helpful research assistance. We thank participants in seminars at the United Nations Department of Economic and Social Affairs and University of Goettingen for their suggestions. We acknowledge support for this project from Azim Premji University, the CUNY Graduate Center Advanced Research Collaborative, the T.A.J. Residency (SKE Projects, Bangalore) and the New School for Social Research. We have also benefitted from indirect support from various other institutions with which we have been associated over the last years, which we do not individually name here.

1. Introduction: Aims of the Project

Increases in mean per capita income are often used as an index of a society's economic development. However, it is a metric that is widely recognized to be quite insufficient. In recent years, public debate has been concerned with whether growth experiences are 'delivering' by enhancing well-being. Some recent work has focused on broadening the indicators which are used to assess social progress (see for example Stiglitz, Sen, & Fitoussi, 2010) while other work has been concerned with the highly unequal distribution of gains, whether accompanied by sizable improvements in the level of income and reductions in poverty (as in China) or by relative stagnation in the incomes of a considerable portion of the population (as in the United States). In the last two decades the increased availability of high-quality data has enabled researchers to provide an integrated portrait of inequalities within and between countries. Such studies of inequality have, however, not generally been integrated with analyses of income growth.

We describe below an effort to create resources that can help address a range of questions, related to absolute levels, gains and relative distribution, by offering plausible estimates of the income and consumption enjoyed by different portions of the population within countries and in the world as a whole over a reasonably long time period. Specifically, we introduce the Global Consumption and Income Project (GCIP), which has as its foundation the creation of two separate datasets (The Global Consumption Dataset (GCD) and The Global Income Dataset (GID)) containing a portrait of consumption and income of persons over time, within and across countries, around the world. The project aims not only to construct but also to analyze these data in future work. The datasets present estimates of monthly real consumption and income of various quantiles of the population (a 'consumption/income profile') for the vast majority of countries in the world (more than 150) for every year for more than half a century (1960-2015). The methodology of construction of the dataset allows for comparable data to be presented for an arbitrary number of quantiles (e.g. percentiles, ventiles, deciles, quintiles or other choices). The benchmark versions that we intend to make initially available for public use will report data in terms of mean levels of income and consumption by decile and in terms of 2005 and 2011 PPP dollars.⁵

Using the GCIP one can estimate a Lorenz curve, mean and consumption and income profile for any given year and country or aggregate of countries. This enables us to create a synthetic population⁶ from which any poverty measure (headcount ratio, poverty gap ratio, FGT measure etc.), inequality measure (Gini coefficient, ratio of mean to median, Palma ratio, Theil index etc.) or measure of inclusiveness in growth and development (for example measures of how widely shared growth or pro-poor growth has been) can be calculated.

The resulting nearly continuous portrait of the evolution of the world consumption and income pattern is unique. It goes beyond the Penn World Tables in presenting estimates of the distribution within countries and it goes beyond recent analyses of the world distribution both in greatly extending the period covered and in presenting estimates for every year as well as for both income and consumption. Whereas with rare exceptions (for example Lakner & Milanovic, 2013) such databases and studies based upon them have focused on relative inequalities alone, we provide data on levels of consumption and income so as to enable assessment of level and distribution together, as is required for analyses of topics such as the

⁵ The summary statistics and the methods for the databases (Version 0.1) that we report here reflect their versions as of March 14th, 2015 and the secondary data for this version was downloaded on or before that date, The databases are, however, being continuously updated.

⁶ For the GCIP we create synthetic populations that consists of 100 'persons', each representing a percentile in the distribution, but we can generate such a population of any size. Indeed, a separate concept that we employ is that of a 'model population' in which each representative individuals stands in for a certain number of persons (e.g. 10,000) from a certain country and segment of that country's income distribution.

inclusivity of growth and development. We have also developed, and intend to provide publicly, in-built tools for filling in missing data, enhancing data reliability, and creating portraits of aggregates of countries. Our intent is that the GCIP should meet a high standard of transparency, allowing for third-party replication, modification and updating and the adoption of alternate assumptions for the selection and treatment of data from the underlying universe, unlike any of the current databases. Among the benefits of such an approach is likely to be that the database can eventually be kept up-to-date through the involvement of multiple users, ensuring that it remains current. The fact that inferences often depend greatly on very detailed data choices makes such transparency indispensable⁷.

Constructing the data set involves undertaking several decisions with regard to the selection of data as well as with regard to the manner in which estimates are generated for country-years in which no household survey was undertaken. Here we document the process of construction and specific choices concerning data in greater detail. Some of the other methods we have developed (e.g. for Lorenz curve estimation and aggregation) and software programs will be provided online at the project website (www.gcip.info). We briefly describe the methods we have employed in the construction of the benchmark version of the database and presents results for a few countries and aggregates. Extensions of the primary database (for instance involving quintiles or ventiles rather than deciles or different PPP concepts and base years) are created using analogous methods.

2. Comparison with Existing Databases

Estimates based on per-capita income of countries have been present since the 1950s and have been used to estimate global inequality (see for example Nurkse (1953) for an early estimate of the world income distribution on this basis, drawing on data collected by the League of Nations and the still nascent United Nations). Since the mid-1990s, when the Deininger and Squire dataset (Deininger & Squire, 1996) was released, economists have had data on the distribution of income across many countries, if often in summary form. This availability in turn has led to greater efforts to try and extend the data (for example, through the World Income Inequality Database (WIID)⁸ developed by WIDER, to ‘harmonize’ it by taking measures to ensure its greater comparability, as for example with the Standardized World Income Inequality Database (Solt 2009) and to extend the data backwards in time (see e.g. Pinkovskiy & Sala-i-Martin, 2009, which forms estimates for as early as 1970). The World Bank has been developing global poverty estimates on the basis of its own collection of data since the late 1970s, and the World Bank’s Povcalnet database has been available to the general public since 2001 as a result of demands for greater data access and transparency. This institutional collection of data has also been the basis for the influential work of Milanovic (2002, 2005).⁹

Our work seeks to go beyond these earlier efforts in at least four ways. First, we construct estimates of both consumption and of income. It is well-known that consumption and income not only have different levels for individuals but different distributions for populations. They are moreover of independent interest, both because they represent concepts of advantage which are of evaluative concern for distinct reasons and because they provide different bases for empirical inference concerning material living standards. The level and distribution of the difference between the two (i.e. of savings or dissavings) may also be informative. We therefore create separate income and consumption estimates for each country-

⁷ For the example of the dependence of global poverty estimates on such choices, the implications of which are often obscured, see Reddy and Lahoti (2015)).

⁸ World Income Inequality Database Version 3.3: <https://www.wider.unu.edu/project/wiid-%E2%80%93-world-income-inequality-database>

⁹ Recently, the World Bank has made available a Global Consumption Database, which provides a detailed household-survey based picture of consumption patterns within countries, but this is available only for a very recent comparison year. (presently, 2010). See <http://datatopics.worldbank.org/consumption/>.

year observation and quantile in the database. Second, we aim to create a complete time-space tableau, interpolating where necessary in order to estimate mean level of income or consumption for every country and year as well as for distinct quantiles of the population. Third, we allow for the aggregation of estimates of the level and distribution of income for user-defined regions and groups of countries. This capability relies on our having previously created estimates that are aligned in time in a given year, through interpolation where necessary. This aspect of our effort therefore builds on the preceding one. We have developed our own software and methods to merge distributions for these user-defined aggregates, providing a flexible capability for researchers and policy analysts. Fourth, we aim to provide documentation of our methods and tools that is as complete as practicable so as to permit the adoption of alternate assumptions in order to construct other versions of the databases and to promote ongoing improvement of methods, tools and data through suitable engagement of specialists and the general public.

How does the GCIP compare to more recent efforts?

Lakner & Milanovic (2013) build upon Milanovic (2005) and seek to describe the global income distribution between 1988 and 2008. They analyze the evolutions of levels of income as well as the distribution of income over time. They choose a few benchmark years and describe the change in the global distribution over the period using surveys based on observations at or near to those years. Whereas they pool income and consumption data without adjustment we employ a ‘standardized’ income concept (drawing on a broader universe of both consumption and income surveys and estimating income from consumption surveys or vice versa), and employ a much longer time series, in addition to the features of the project that allow for additional dimensions of flexibility, as mentioned above. Although we adopt this standardized approach because we believe it to enhance comparability, our data can also be used in ‘pooled’ fashion if desired, in keeping with their procedure and that adopted more recently by the World Bank.¹⁰

In another recent exercise Dykstra, Dykstra and Sandefur (2014) queried the Povcalnet database using automated methods to create a cumulative distribution of income or consumption (pooled together in that database) for a large number of survey-years (from each of 942 surveys spanning 127 countries over the period 1977 to 2012). The resulting database can (as with the GCIP) be used for diverse purposes, some of which would have been very difficult without downloading the data in this comprehensive way. The exercise highlights the difficulty in accessing even nominally public data for research and replication in view of the restrictive format in which it is often presented, the prevalence of poor documentation and the contrasting value of fully publicly accessible datasets. In creating an earlier version of the GCIP we undertook a very similar exercise. However, we abandoned that effort because (a) the computational effort for the exercise was very high and the cumulative distribution could simply be replicated for the entire distribution for as many points as desired, and more flexibly and transparently, by replicating the reported parametric regressions that underlay the data, (b) the Povcalnet database is largely confined to developing countries and to years from the early 1980s onwards and (c) there was no reason to privilege Povcalnet as a source of survey data even for developing countries, for which there are other sources of data too. The GCIP has been constructed to differ in key respects. The GCIP has wider area and time coverage (due to inclusion of surveys from other sources, largely secondary but sometimes primary), it incorporates a standardized welfare concept (consumption or income, with one estimated from the other where necessary) making within and cross-country comparisons more meaningful, it allows for the estimation of all measures for every year (not just the survey year or a reference year around which

¹⁰ See Ferreira et al (2015) for details on the pooling method. There are questions however as to whether such pooling is sensible (Reddy and Lahoti (2015)) which is why we endeavor to separate consumption and income estimates.

surveys are grouped), it provides tools for creating user-defined composites of countries in any given year, it provides flexibility in choices as to how to construct and update the dataset, and in choosing specific estimation methods for the Lorenz curve (as opposed to accepting the version which happens to be chosen by Povcalnet, which may reflect not only variable methods but sometimes generate invalid estimates of Lorenz curves). One of the key goals of GCIP is transparency, realized by providing documentation that is as complete as possible and access to all data and code to the extent feasible, in order to facilitate application of alternative assumptions in database creation or analysis.

Edward and Sumner (2013) have created a database closest in spirit and construction to ours. The Edward and Sumner GrIP ('Gr'owth, 'I'nequality and 'P'overty) model (version 1.0) takes distribution (quintile and decile) data and combines this with data on national population and on the mean consumption per capita in internationally comparable PPP \$ to develop a database with similar aims to ours. However, the GCIP includes information before 1990, provides both consumption and income levels for each decile and allows for different PPP concepts as well as for market exchange rates. In this outline, we focus, however, on the present benchmark version which provides data in 2005 PPP dollars.

We do not attempt to discuss comprehensively the merits and demerits of previous efforts but instead seek to focus on the distinguishing features of the GCIP dataset. It is nevertheless useful to attempt to summarize the differences between our approach and existing efforts (see Table 1). We believe that the GCIP provides data for a wider set of countries, aggregates of countries, years and concepts, as well as tools for their analysis, than do other existing databases.

3. Construction of Global Consumption and Income Datasets

Constructing a consumption (or income) profile for a given country-year requires two distinct pieces of information: the relative distribution and the mean in that year. These two are sufficient to create a unique profile of actual consumption (or income) levels of each decile in the country-year. We thus divide the process of creating the database into four distinct steps.

In the first step, we collect data on relative distributions and mean levels for each country from various existing sources. Where there is more than one survey for a country-year we select one, preferring consumption data sources for the consumption database and income data sources for the income database (Other choices are of course also possible, including to pool the income and consumption data without preferring one concept of advantage). Second, we standardize the distributions by converting all distributions that are not already in the required format (consumption or income distributions depending on the database) into estimated equivalents. The selected surveys for country-years consist of both consumption and income surveys. Where surveys of both kinds are available they differ, as the share of income tends to be higher for lower quantiles and the share of income lower for higher quantiles for income as compared to consumption distributions. Hence to make any meaningful comparison among distributions across and within countries and over time, we must transform the distributions. Although the conceptual case for doing so is strong this is rarely if ever done in international comparisons. In the third step, where necessary we estimate a consumption mean for the GCD (Global Consumption Database) for survey-years where we have only an income mean and we estimate an income mean for the GID (Global Income Database) for survey-years where we have only a consumption mean so as to place the means too in more comparable units. We also attempt to detect means that are extreme outliers so as to enhance data reliability. Fourthly, using the mean and distributional data previously generated, we estimate a Lorenz curve for the survey years (using standard parametric methods that have been found to perform acceptably in recovering underlying true distributions, although other methods are available in case these fail). Finally for non-survey years we estimate the consumption/income profile by interpolation or extrapolation by using the appropriate per capita growth rate figures from the World Development

Indicators (WDI)¹¹ to create a time-weighted average of the ‘perspectives’ on the estimation year that are associated with the nearest survey-years. This set of procedures gives rise to a complete time-space tableau covering the world between 1960 and as near as we can come to the present. We describe each step in detail below.

3.1. Creating the Universe of Surveys

The GCIP draws data on relative distributions from diverse sources, such as the EU-SILC database (for European countries), the LIS (previously the Luxembourg Income Study), the SEDLAC database (for Latin American countries), UNU-WIDER World Income Inequality Database (henceforth WIID), the World Bank’s Povcalnet database, and Branko Milanovic’s WYD database¹². We are committed in principle to an ecumenical approach that integrates historical and contemporary data from *all* relevant sources, including country statistical offices, UN agencies, academic studies and private sector sources.¹³. Povcalnet is a collection of surveys starting from the early 1980s. Until recently, it covered only developing countries but now incorporates surveys for a number of developed countries, largely building on data from LIS. WIID is a collection of surveys from various secondary sources, covers both developed and developing countries and spans the period 1960-2012. Our third major source, the LIS, has harmonized data according to its chosen protocols from primary surveys for over 40 countries mostly from upper and middle-income countries. Although it provides data in household equivalence-scale adjusted form we extract the data we use from the underlying databases in per capita form.

Our first step is to generate a ‘union’ of all available distributional and level data for all the country-years of interest. The initial database thus constructed sometimes contains more than one observation for a country-year since multiple household surveys were undertaken in certain country-years and the same survey (in several instances with conflicting mean or distribution information) might be reported in multiple sources. The first task is therefore to refine the observations so as to arrive at one observation for each country and year. Surveys contained in GCIP may be reported as having a certain source, coverage of geographical area (national, or only urban areas), population and age, a certain assigned quality rating as stated in the underlying secondary source, concept of advantage (income vs. consumption, and specific income definition) and unit of analysis (household, individual, etc.). To choose one observation for country-years for which there are multiple we apply a lexicographic ordering to a set of selection criteria, which we discuss further below. The criteria and their sequence in the ordering are based on what we consider important considerations for common usage scenarios for the database. These can be altered if other usage scenarios are envisioned or indeed if users’ judgments as to the relevance and importance of specific selection criteria differ from our own.

Before applying the various criteria, we restrict the universe of surveys to per capita surveys. This has the disadvantage of causing some loss of surveys and thus a reduction in the number of observations in our dataset, although much less than if we had chosen any other specific equivalence scale concept. For example, we are in the process of including the European Union Survey of Living conditions (EU-SILC) data. As this distributional data is reported an OECD-based equivalence scale we must recalculate the distributions in per-capita terms before including it. We prefer per-capita distributions for a number of

¹¹ World Development Indicators. Accessed Feb 1st, 2014. Retrieved from <http://data.worldbank.org/data-catalog/world-development-indicators>.

¹² www.lisdatacenter.org (accessed June 2015).

¹³ GCIP also includes surveys for Cyprus, Hong Kong, Singapore and New Zealand from Branko Milanovic’s World Income Database (WYD) as surveys for these countries were not available in the other secondary sources. We have also employed our own country research on specific individual cases to supplement our major sources, through correspondence with statistical agencies, identification of relevant historical documents etc. We list specific sources in our online appendix of country assumptions (see gcip.info).

reasons, and in keeping with the practice of other researchers (including the World Bank's Povcalnet, Milanovic and Lakner and others). Per capita surveys are simpler to analyze and to understand and correspond more directly to concepts in the national accounts. They are also the most common form of survey in the secondary data sources. The drawback of using only per-capita information is that differences in the real value of resources arising from variations in household size and composition are not taken registered. On the other hand, limiting our focus to per capita surveys greatly aids comparability. A variety of studies have shown that portraits of poverty, inequality, household consumption behavior and other facts greatly depend on the equivalence scale chosen.¹⁴ There is moreover reason to believe that even if the same equivalence-scales are being compared the extent and character of this dependence would vary greatly between country-years due to differences between country-years in the demographic composition of households belonging to different parts of the distribution. Whereas the exact nature of the dependence can be explored when the household level data is available, that is not possible when only summary results using a specific equivalence scale are reported, as is generally the case in the collections of data that we use. Rather than making our conclusions dependent in an unknown but very likely substantial way on the specific equivalence scales used we think it more sensible to use per-capita surveys. When it is reported that a survey uses an equivalence scale, typically insufficient detail is presented about the method that was used, making it difficult or impossible to compare distinct surveys meaningfully. As noted above, for LIS surveys, which report data using an equivalence scale, we obtain data in per capita terms using micro-data¹⁵.

The lexicographic ordering of various criteria which we employ is as follows: whether a survey mean is reported, type of survey (consumption/income), the nature of the income/consumption definition, database source (e.g. EU-SILC, LIS, Povcalnet, SEDLAC, WIID, WYD, or primary source), area coverage, population coverage, quality as defined in the source database, source of the data as reported in the secondary database (e.g. source of a WIID observation) currency unit and survey series (as defined by statistical authority, e.g. German Socio-Economic Panel). As we are interested in both levels and distributions we prefer surveys with mean information over ones for which means are not reported. For the GCD, which focuses on consumption estimates, we prefer consumption surveys to income surveys (and vice-versa for the GID). Among income definition concepts we prefer concepts that are closer to arriving at total income net of taxes and transfers. The order of preference of income definition concepts appearing in the underlying databases (for which we draw upon the classification scheme and related definitions presented in the WIID) is as follows, from most preferred to least preferred: disposable income, disposable monetary income, gross income, gross monetary income, taxable disposable income, primary income, net earnings, gross earnings and finally a residual category for concepts that are not fully specified, i.e., we don't know if the reported data refers to net, gross or disposable income. Although it would be desirable in principle to make adjustments to the data based on relationships between the estimates corresponding to these distinct categories, in order to make them more comparable, we do not do so as we not have sufficient data corresponding to the distinct concepts but the same countries or survey-years to establish these relationships.

Our order of preference of data by source employs the following ordering (earlier preferred to later): LIS, SEDLAC, EU-SILC, Povcalnet, WYD, WIID, primary source. This ordering reflects a number of judgments. Reported Povcalnet and LIS survey results are often compiled from primary data, while WIID is a collection of secondary data. We judge that Povcalnet and LIS may be more rigorously scrutinized and have a smaller probability of transcription or other errors as compared to WIID surveys and hence among global sources we prefer these two to the WIID. We view SEDLAC and the EU-SILC

¹⁴ See e.g. Buhmann et al (1988), Blaylock (1991), Coulter, Cowell and Jenkins (1992a and 1992b), Banks and Johnson (1994), Anand and Morduch (1996), Aaberg and Melby (1998), Cowell and Mercader-Prats (1999) or Sefil (2015) and the more recent literature cited therein.

¹⁵ We use LIS's remote-execution system called LISSY to gain access to LIS's micro-datasets.

as being high quality sources of regional data (for Latin America and Europe respectively) and thus give high preference to them. Since LIS surveys have until recently included few if any developing countries and Povcalnet (as of recently) includes only select developed countries (corresponding to LIS countries) the overlap in terms of country-years covered by these is in proportional terms small. However, when there is an overlap we prefer LIS to Povcalnet for the reasons that LIS makes unit-level data available to us, and that LIS aims at achieving a higher degree of internal comparability among its surveys through specific effort at harmonization. The availability of unit-level data allows direct verification of the per-capita distributions calculated (which do in fact appear to coincide with the Povcalnet distributions for developed countries calculated from the same source). Due to this preference ordering, the external comparability of our estimates with Povcalnet based estimates for developing countries that derive to a larger extent from other sources, in particular, World Bank poverty estimates, is diminished (although we are to a large degree able to replicate these). The WYD overlaps heavily with Povcalnet but includes a few additional sources. To ensure greater comparability of GCIP with Povcalnet we place WYD after Povcalnet in our ordering.

We prefer surveys with broader area and population coverage and surveys deemed higher quality by the source database to others. WIID surveys report a quality rating but Povcalnet and LIS surveys do not report any quality rating. Given that Povcalnet and LIS are constructed using primary data and have stricter inclusion requirements we assign them the highest quality rating (but it must be remembered that this is only an ordinal characterization). Among sources in WIID (or in principle any other secondary source) survey data reported as originally from LIS or from the Deininger and Squire database are preferred over other sources.¹⁶ We prefer surveys that report means in local currency units over those which are reported in other units because the method of conversion into international units by the source can often be non-transparent. We also prefer surveys in which the survey series is known over those for which it is missing. Even after applying all of these criteria we find that some country-years have multiple surveys. At this stage we choose among these that survey which leads to the survey source being more compatible with the portrait presented by other years' observations for the same country (especially the nearest survey years for which data are available) or apply other criteria¹⁷. In certain instances, we exercise our judgment and drop certain surveys or prefer a survey to which the lexicographic ordering would not have led¹⁸.

¹⁶ For earlier years in particular, WIID draws on other sources, such as Jain (1975).

¹⁷ After applying the lexicographic ordering we observe multiple surveys for same country-year in four instances in the present version of the GCIP, which we resolve as follows. In the case of Barbados (1970) we use the survey that refers to the economically active population over one which covers only 'income recipients'. For China 1995, Brazil 1970 and Colombia 1964 we keep the survey that allows for a more consistent data series across the years for the country. This is an exercise of judgment and users might prefer the dropped surveys to be part of the database in which case they can make that choice.

¹⁸ We modify our lexicographic ordering in the rare instances where there are known issues of comparability of the survey with other surveys for that country. One example of this is Indian consumption survey for 1999. The Indian consumption survey in 1999 used a shorter recall period of 7 days as opposed to usual practice of using a thirty-day recall period in other survey years for India making comparisons with other surveys difficult. Similarly, in the universe of surveys in the GCIP, Russia has consumption survey data reported by Povcalnet and income survey data from the LIS. The surveys from these sources for Russia exhibit vast differences in means information. Applying our lexicographic ordering we might have picked the LIS reported surveys in the instances when the both are present for the same survey-year. But in this case to maintain consistency of information across the time series and to keep as large a number of compatible observations as possible we choose to use Povcalnet surveys over those from LIS. We provide a list in our online appendix of country assumptions (see gcip.info) of all the cases in which we exercise our judgment over and above applying the rules described earlier.

3.2 Standardizing the Distributions

Surveys vary widely by their focus (e.g. type of advantage, such as consumption or income), as well as details of their method (e.g. length of recall period, level of detail in surveys, whether unobserved costs and benefits are imputed (such as the value of rent for self-owned residences) and survey frame as well as timing) making comparability between countries difficult (For a discussion of this point see Smith, Dupriez and Troubat (2014)). Of particular interest to us is that the definition of income varies widely between surveys. Some report gross income, others after-tax income and others still wider or narrower categories, often with somewhat obscure definitions. Table 2 reports the various income/consumption concepts used in surveys included in the GCD and GID, along with their frequencies, adopting the classification used in the WIID.

As is well known, the distribution of consumption is expected to be less unequal than the distribution of income. Those concerned with estimating global inequality or poverty almost universally recognize this concern but do not generally correct for it¹⁹ (Ferreira et al (2015); Lakner and Milanovic (2013)). Comparing measures of inequality or poverty across countries can therefore be highly misleading. Similarly, aggregating information for groups of countries to obtain, a measure of poverty or inequality, for say, Sub-Saharan Africa becomes difficult and results obtained from combining income and consumption based surveys may lead to misleading results.

One effort to address this issue is the work of Solt (2009) who makes the assumption (plausible at least for developed countries) that the LIS may be treated a ‘gold standard’ and then tries to adjust other surveys using a regression based method to estimate a ‘standardized’ summary measure of the distribution of income (the Gini coefficient that would be expected to result from counterfactual and missing LIS surveys) in other countries. His database is confined to measures of inequality. Niño-Zarazúa, Roope, and Tarp (2014) also estimate standardized consumption distributions, by adjusting the share of each consumption decile by the average difference between income and consumption decile shares for a set of country-years which had both type of surveys.

We take a different approach here. As it turns out there exist in the WIID and the LIS a total of 204 instances across 71 countries in which there is both consumption and an income survey reported by the same statistical agency for the same country-year. For most of these (more than ninety percent) information on consumption and income for the survey year was collected from the same survey. These survey countries are spread across all geographical regions of the world and across various country income groupings (Table 3).

We use this information to estimate the expected relationship between income and consumption. Our purpose was to identify a regression relationship between consumption and income for each quintile²⁰. (We use quintiles rather than deciles in order to maximize the number of observations, as in earlier years often only quintile data is reported) Given that the errors across the five regressions might be correlated (and indeed, the Breusch-Pagan test suggested so), we employed a Seemingly Unrelated Regressions approach to estimate the relationship²¹. When we wish to estimate consumption shares from income

¹⁹ Deininger and Squire (1996), in the context of their dataset, suggest adding 6.6 Gini points to Gini coefficients based on consumption to obtain the corresponding income Gini coefficients based on average difference across their dataset.

²⁰ As noted earlier there are various income concepts collected by surveys and the choice to employ them might have affected estimates of mean levels and distributions. We do not standardize among these various income concepts in GCIP.

²¹ We experimented with several different specifications and also used a (more theoretically appropriate) Dirichlet regression to estimate our equation. The Dirichlet distribution is the multivariate generalization of the beta

shares the regression formula we use is:

$$SC_{ij} = \alpha_i + \beta_i SI_{ij} + \gamma_i X + \varepsilon \quad (1)$$

Where SC is the share of consumption of quintile i , SI is the share of income of quintile i , X refers to a set of controls for country income level, region, income concept used in the survey and time. Finally i and j are subscripts for country and quintile respectively.

When we wish to obtain the income share, we redo this exercise and reverse the regressor and regressand to obtain

$$SI_{ij} = \alpha_i + \beta_i SC_{ij} + \gamma_i X + \varepsilon \quad (2)$$

Table 4a and 4b provides the results of these regressions. In both sets of regressions the r-squared is moderately high, ranging from 0.47 to 0.76.

Table 5 provides an indication of the performance of this regression by reporting the results of an in-sample prediction analysis. The ability to predict consumption shares with a degree of reliability gives us confidence as to its general applicability.

We use these regression formulae to obtain a derived implied consumption distribution when one has only an income distribution available for a country and a derived implied income distribution when one only has information on the consumption distribution. We undertake this exercise for the whole dataset so that every country can be assigned an income and consumption distribution (at least one original and at most one derived) for every survey year.

However, prior to the final assignment we must make an adjustment for the adding-up constraint that the sum of percentage shares in the derived distribution must sum to one hundred. Typically, one is left with income or consumption that is unaccounted for by the simple application of the regression coefficients, for the reason that the regressions were undertaken independently. The sum of shares might be above or below 100. We think it reasonable that the unaccounted for income may be added or subtracted (depending on the direction of the error in the total) proportionally equally across quintiles. This is admittedly only one possible choice: we could apply another rule of apportionment. However, in the absence of compelling reasons to do otherwise, we think this a sound choice. Because of the independence of the quintile-specific regressions it is also possible that the derived implied consumption or income distribution might break the monotonicity restriction i.e. that the share assigned to a lower quintile might be greater than the share assigned to a higher quintile for the same country-year's estimated distribution. We check for this and in our preferred specification have not encountered any instances of this issue²². However, if non-monotonicity were encountered there would be ways of

distribution. Thanks to its properties, it is a convenient parameterization for compositional data. First, the dependent variables are restricted to the $[0,1]$ interval. Second, it ensures that the shares sum up to unity. Hence, it is a valid distribution for estimating quantile shares, i.e. $SC_{ij} \sim D(\alpha_i)$. However, the results do not differ significantly between the two estimations. Moreover, the Dirichlet regression assumption that all shares are negatively correlated is violated in our data. We therefore use the more standard SUR approach. See Emmenegger, 2015 for an analysis of the difference between the estimation techniques.

²² Users who would prefer not to do this standardization or replace our method with their own could, once the GCIP code is released, easily turn off or replace the standardization method.

addressing this.²³

An example of the application of this method is provided by Brazil in 1996. The GCIP has an income survey for Brazil for 1996, which we convert to an estimated “equivalent” consumption distribution based on our cross-country regression procedure. After application of the regression coefficients the sum of the shares of quintiles is 99.95. The deficit of 0.05 points is assigned proportionally to all the quintiles so that each quintile’s share is increased by the same percentage. The shares at various stages of the process are shown in Table 6.

3.3 Standardizing the means

While there has been substantial interest among researchers in the variance between survey and national accounts means (see for example Deaton 2005), there has been little or no examination to the best of our knowledge of the variance *between* means from surveys carried out in the same year for a given country. Our initial examination suggests that the differences can be extremely wide. For example, Bolivia has two surveys in WIID for 1997 which report monetary income means that differ by 30 percentage points (414 vs. 538 Bolivianos per month). This in turn means that although our lexicographic ordering gives us a particular mean, a slightly different ordering might have led us to choose a dataset with a very different level of income or consumption. This problem will plague any attempt to choose surveys. The mean number of surveys per country-year is 2.95 and the country-years with more than one survey have on average 3.78 surveys, and only thirty percent of country-years have only one survey (although as we noted at the outset this can be due to the same survey being reported by multiple secondary sources).

In future work, we hope to provide a more comprehensive examination of the issue of disparate survey means and their implication for such concerns as the global income or consumption distribution. For now, we simply note the problem and attempt to standardize the means for the surveys that our ordering leads us to. As noted before, the universe of surveys provides various definitions of income and consumption. Furthermore, these are often reported in non-comparable units (for example by providing the information in real or nominal terms, in local currency or international currency units, and for different time periods). Our next task is therefore to construct a consumption and income mean for every country-year in comparable units. In order to do this, we seek to generate an estimate of the consumption or income mean for each country-year for which we have an observation. Whenever an estimate of the mean was available from the survey with which we obtained the relative distribution, this was the preferred source of data²⁴. This mean, usually expressed in local currency units (LCUs) of the survey year²⁵, was then converted to 2005 LCUs using local consumer price indices²⁶ wherever available (and in rare cases,

²³ In particular, we would propose to apply a second regression to the independently estimated shares and assign estimates based on this regression to the quintiles, after adjusting to satisfy the adding-up constraint.

²⁴ Lakner and Milanovic (2013) and World Bank’s Povcalnet database also prefer survey means over national account means (see Anand and Segal (2008) for a discussion on the choice of means). Though we use survey means for our estimations, we will aim also to provide data on national account based means (GDP per capita and Household Final Consumption Expenditure (HFCE)) in the released version of GCIP. Mongolia is the only case for which we do not have any means from surveys and as a result we use means from national accounts as an alternative.

²⁵ We also attempt to adjust the means for any currency redenomination or change in currency that the country might have experienced. This is a non-trivial task as detailed historical knowledge of the country and its data sources is sometimes needed to do this.

²⁶ All our survey data is at national level and hence we use national CPI’s unlike Povcalnet, which uses separate rural and urban survey components and inflation rates for India and China (Ferreira et al, 2015). It is not obvious whether using sector-specific data is superior because of the lack of uncontroversial inter-sectoral and sector-specific price indices, as discussed in Reddy and Lahoti (2015).

where unavailable, the GDP deflator)²⁷.

In order to make the estimates comparable across countries, we then converted them into common units by applying 2005 PPP exchange rates²⁸ and converting all data into monthly per capita units (for example if the survey estimate of consumption is for a weekly amount, we multiply it by 30/7). GCIP also includes country-specific conversion factors for other ICP PPP base years and other PPP concepts (e.g. PPPs for food) which could be used to obtain data in alternate PPP units and market exchange rates for all country-years. Note that in all cases we use the unitary country-wide PPP. This contrasts with, for example, the World Bank approach, which uses sectoral PPPs for urban and rural areas for India, Indonesia and China, based on back-of-the-envelope assumptions about likely inter-sectoral differences. A fuller discussion on the issues involved in deciding to use unitary or sectoral PPPs for these major countries and the impact of specific assumptions in the case of poverty estimates can be found in Reddy and Lahoti (2015)²⁹.

Outlier Detection

Despite our best attempts at selecting the data carefully, the survey mean data that we are left with contain outliers. These are means that are implausible *prima facie* given other existing data. In many cases we are unsure of the source of the discrepancies, especially in light of the fact that we draw extensively on secondary data. We identify outliers using two criteria described below. A survey mean that is identified as an outlier by *both* the criteria is marked as an outlier and adjusted according to a procedure we will describe.

To identify outliers, we first run a separate regression for each country to identify the time trend in survey means for that country. In this step, we regress the survey mean with respect to time (years elapsed since 1960). If the survey mean is above or below two studentized residuals from the regression line we mark it as a potential outlier. We find that about 8% of our observations are marked as potential outliers using this criterion. Applying this ‘internal’ criterion in isolation would mark cases in which a country’s economy *actually* experienced sudden growth spurts or severe and sharp declines as outliers since a linear time trend may not be able to account for sudden transitions. To avoid this we impose a second ‘external’ condition, namely that the annualized survey mean growth rate is within certain bounds of the national accounts based growth rate in per capita gross domestic product. The acceptable band for the survey mean growth rate is defined by the growth rate of GDP per capita plus or minus twice the growth rate. (For instance, if the GDP per capita growth rate is 10% then the band is -10% to +30%). This criterion, while hardly restrictive, helps us to anchor the outlier detection criteria to a measure of validation external to the survey data, provided by the economy’s growth rate. About sixty observations (5% of surveys with means data) are marked as outliers using both the criteria. Instead of completing discarding the outliers we view them as still providing relevant information and therefore instead adjust and retain them. The outlier means are adjusted (decreased or increased) up to the acceptable outer bounds of the time trend line. For

²⁷ Our source for inflation data is World Development Indicators (WDI). This contrasts with Povcalnet, which for some countries uses alternate CPI indices. For Taiwan, for which WDI does not maintain any data, we obtain data from Taiwan’s National Statistical Office <http://eng.stat.gov.tw/>.

²⁸ We use 2005 EKS PPPs for ‘individual consumption expenditure household’ concept obtained from the International Comparison Program (ICP) website. Even though we use 2005 PPP exchange rate for the benchmark version of the database, this is not because we necessarily prefer it to other exchange rates. The choice of exchange rate depends on the research question. We are aware that exchange rates used have a substantial impact on the levels and also on global and regional inequality and have presented some alternate estimates using 2011 PPPs and market exchange rates in Jayadev, Lahoti and Reddy (2015). In Reddy and Lahoti (2015) GCIP is used to calculate poverty estimates using 2011 PPPs and in Jayadev, Lahoti and Reddy (2015c) estimates of the evolution of the global middle class based on distinct concepts are explored using alternate exchange rates.

²⁹ Details of the PPP conversion factors and the ways in which they are implemented are available at www.gcip.info

example, outliers that are higher than the trend line are adjusted so that they have a value equal to the trend-line plus two studentized residuals. Our reasoning for doing so is that if we were to adjust the means to a higher level they would remain outliers according to our criteria, which would not serve the purpose of adjustment. At the same time, adjusting them to a level lower than the bounds would lead to treating outliers as requiring adjustment to a level lower than for reported survey means which are above the adjusted value of the mean but below the outlier detection bounds. It is hard to understand what would be the rationale for such a difference in treatment.

3.4 Generating a Lorenz Curve and Consumption/Income Profile

Having obtained or constructed means and distributional data for every survey year chosen, we estimate a Lorenz curve in parametric form using a widely employed regression framework (see Datt (1998); Miniou & Reddy (2009) for some discussion of the methods, also employed by Povcalnet). We prefer the generalized quadratic Lorenz curve estimation of Villasenor and Arnold (1989) for its theoretical properties but when the procedure fails to generate a valid estimated Lorenz curve we utilize the Beta Lorenz curve estimation due to Kakwani (1980) applied to quintiles³⁰. When both of these methods fail to generate a valid Lorenz curve, which happens occasionally, we move to a third parametric approach due to Rasche et al (1980)³¹. If this were to fail (which it does not for *any* of the current distributions in the GCIP database) we would use a fourth parametric method due to Chotikapanich, D. (1993). Finally, in case all of these fail we create a piecewise linear consumption profile based upon ‘connecting the dots’ defined by the quantile means, following a method we have developed and tested (after which we can also calculate the associated Lorenz curve, which is strictly convex, as required for its validity). We chose to use the generalized quadratic Lorenz curve and the beta Lorenz curve in part because these are the parametric estimation methods used by the World Bank, and this would facilitate comparison of estimates, but one could equally use the Rasche method, which provides very similar results, based on our comparative examination of the methods for subsets of the data.

Once we arrive at an estimated Lorenz curve, we use it in combination with the estimate of the mean to generate a consumption profile consisting of an estimated mean income or consumption level for each decile of the country-year³². Specifically, the mean income of each decile is calculated by taking the share of total income accounted for by that decile, and multiplying it by the survey mean times the number of deciles (10). For example if the Lorenz ordinates for the first 2 deciles are 0.02 and 0.05 respectively and the mean income is \$15, then the mean income of the first decile is $\$15 \cdot 10 \cdot .02 = \3 , while the mean income of the second decile is $\$15 \cdot 10 \cdot (.05 - .02) = \4.5 . We estimate decile means for the survey years in order to generate a lattice that can serve as a basis for interpolation and extrapolation of decile means to non-survey years. There are no deep-seated reasons for the use of decile means specifically for this purpose and we could have made a different choice.

Our goal is to estimate the consumption/income profile or set of quantile means for every country-year for

³⁰ In practice, when generating a valid Lorenz Curve, both procedures typically provide a reasonably good and similar fit to the data as captured by the sum of squared errors or other criteria. The Beta Lorenz curve fails the test of giving rise to a valid Lorenz Curve more often. The conditions of validity of these LC’s are discussed in Datt (1998).

³¹ To test the accuracy of Lorenz curves derived from the various parametric methods, we have compared income/consumption shares of various quintiles of the distribution obtained from parametric methods and that from unit-level data for a few LIS countries. Our initial findings indicate that all three methods perform very well in predicting the actual shares (within 1 percentage point in most cases). We hope to expand this analysis to all countries where we have access to unit-level data and report the results in a subsequent analysis.

³² In the case of the piecewise linear method for the estimation of the consumption profile, we would not need to generate a Lorenz curve at all.

the entire period covered by our database in order to obtain a complete ‘consumption/income profile tableau’. In order to attempt to fill in the consumption/income profile tableau, we estimate the profile for intermediate years using growth rate figures from the World Development Indicators (WDI) or other sources where necessary, such as the Economist Intelligence Unit for most recent years, in order to interpolate or extrapolate consumption or income profiles for non-survey years. As noted below, the survey coverage is very limited before 1980. This is one reason why several researchers may have preferred to begin their empirical efforts after that date. Moreover, whether before or after that date they typically confine themselves to survey-year estimates, which may not be temporally aligned across countries, thus limiting the possibilities for comparison and aggregation across countries. We are contrastingly interested in trying to extend coverage as fully as possible in order to facilitate these tasks, not only by increasing the span of time covered but by filling in missing years. We fully recognize the concerns that such extension may raise, and accordingly try to do so according to carefully chosen assumptions. A larger amount of the data before 1980 is interpolated or extrapolated due to sparse survey information and thus has to be treated with greater caution. Users who would prefer to not use the interpolation/extrapolation techniques below we employ can always chose to restrict their analysis to only survey years³³, which are clearly marked in the GCIP.

There are two methods used to estimate a consumption/income profile for the non-survey year, viz.:

Extrapolation

If the non-survey year lies before or after the first/last survey year for which we have a consumption or income profile, then the consumption or income profile of that year is extrapolated (forward or backward) based on the survey year and the relevant per-capita growth rates. For example, if we want to estimate the consumption or income profile for a country and the last survey-year happens to be in a given prior year, then for the subsequent years, we extrapolate the consumption profile using the following formula iteratively:

$$M_t = M_{t-1} * (1 + g) \quad (3)$$

where M is the estimated mean consumption/income of a decile, t is the year and g is the growth rate of mean consumption/income per capita between the two years.

Interpolation

If the non-survey year lies between two survey years for which we have a consumption or income profile, the consumption or income profile for this non-survey year is treated as a time-weighted average of the growth-adjusted consumption or income profiles (arrived at by extrapolating respectively backwards and forwards through applying the observed growth rates of mean per capita consumption or income) of the two survey years. This procedure is the same as described in Chen and Ravallion (2004) to impute means for non-survey years except that we extend the procedure to the overall distribution and estimate decile means in an analogous manner. Ferreira et al (2015) describe a procedure adopted by the World Bank more recently, in which the growth rates used in this process are adjusted by a ratio reflecting a presumed relationship between the growth rates from surveys and from national accounts in developing countries, and are thus lower than those in the national accounts. Specifically, they employ a multiplier of 0.87 to

³³ The estimates for survey years are not affected in any way by the interpolation/extrapolation. However, restricting the analysis to survey years would constrain cross-country comparison for a particular year as surveys are not generally lined up across countries. It would also limit the possibilities for aggregation if surveys for the countries being aggregated are missing for the year in question. Interpolation or extrapolation and aggregation are in this way intimately connected.

growth rates in all countries, except China and India where multiplier is 0.72 and 0.51 respectively, to make this adjustment³⁴. This is unlikely to make very much difference for interpolation insofar as the adjustment applies to projections from both of the nearest survey years but it may make some difference to extrapolation. We do not employ such a further adjustment but it would be a straightforward matter to construct a variant of GCIP that does so, either using the Bank's adjustment factors or another set of assumption (which for instance, distinguishes between the ratio applicable to countries of different kinds).

Since the consumption/income profiles for survey years are already expressed in comparable units (\$2005 PPP in the benchmark version of the database) we therefore use the growth rates of real (inflation adjusted) per capita consumption to arrive at an estimated consumption or income profile for each non-survey year. Since growth rate information is available from different sources, we must establish a preference ordering for the growth rate data type and source, which is as follows, from most preferred to least: growth rate of household final per capita expenditure from the WDI, growth rate of 'per capita final consumption expenditure etc' from the WDI, GDP per capita growth rate from the WDI, consumption per capita growth rate from the Penn World Tables, GDP per capita growth rate from the Penn World Tables, GDP per capita growth rate from the Total Economy Database (TED) (The Conference Board Total Economy Database 2010³⁵) and finally GDP per capita growth rates from Angus Maddison 's historical statistics³⁶.

The earliest year to which we extrapolate our data backwards is 1960. This is because annual growth rates of mean consumption from national accounts for a wide variety of countries are available only starting then. There are some instances in which the growth rate data for the very earliest years is missing and we restrict the extrapolation to the first year when the data is available for these countries. The result in all of these cases is that there are gaps in the tableau. This not only affects the ability to define trends over the entire period but also to construct regional or global aggregates which are fully comparable over time. We hope to fill these gaps over time, in part by drawing on broad expert and public participation, or adopting other assumptions (such as extending trend growth rates backward or forward). In the meantime, one option is to discard from consideration those entities for which we do not have data over a sufficient period and another is to restrict the temporal scope of the analysis. For certain purposes, it may be tenable to compare alternatives which both do and do not contain certain countries, but one must be aware of the potential distortions that could arise as a result of specific countries dropping in or dropping out of the portraits of aggregates over time. The empirical examples of aggregates we provide in this paper do not include any adjustments for such non-comparability over time but that could be done in more careful subsequent work.

3.5 Coverage of the Surveys

Tables 7a, 7b, 8a and 8b present summary statistics for the surveys in the GCD and GID. Tables 7a and 7b describe the number of surveys according to various criteria (e.g. source of data, decade, region, income group), Tables 8a and 8b describe the number of countries in the databases according to these same criteria. Table 9 provides information on the density of the surveys³⁷ by decades, region and income group, i.e. the percentage of the total potential surveys (defined as the total number of country-years in the

³⁴ Refer to footnote 48 in Ferreira et al. (2015) for more details on the adjustment factors used.

³⁵ Available at <http://www.conference-board.org/data/economydatabase/>

³⁶ The World Bank's Povcalnet also uses consumption per capita growth rates and GDP per capita growth rate data for interpolation/extrapolation. For more details refer to Footnote 48 in Ferreira et al (2015).

³⁷ The density of surveys is very similar in GCD and GID. This is because for a given country-year if there is only one survey available it will be used by both databases, and where there is more than one survey available a single survey will be chosen by each database (and hence the country-year coverage will be the same, even if the specific surveys selected are different). In Table 9 we therefore report the results for the GCD alone for simplicity.

decade and region).

There are a total of 1946 surveys in GCD over the fifty-five year period (1960-2015), from 161 countries. About forty three percent of the surveys are consumption surveys and more than ninety percent are nationally representative and cover the entire population. The coverage of surveys is sparse in the 1960s and 1970s with just over forty countries with surveys in each of these decades. The number of countries with at least one survey and the number of surveys with information on means both increase steadily in each decade, with rapid growth from the 1970s through the 1990s. Povcalnet is our biggest source of survey information, accounting for forty two percent of surveys in the GCD, followed by WIID (twenty nine percent), SEDLAC (thirteen percent) and LIS (thirteen percent). However, Povcalnet has almost no surveys in the first two decades, for which we instead rely heavily on WIID and to a lesser extent on LIS.

As can be seen from Table 9, there are very few actual surveys available for earlier decades. For example, coverage is between zero and twenty five percent in the 1960s depending on the region. As a result, many estimates for years before 1980 in our databases are based on backwards extrapolation. When that is the case, users are cautioned to exercise judgment in their analyses.

Additionally, one aspect of the paucity of coverage in earlier years is that there are several countries for which the only distributional data we have are interpolated backward from later surveys. There are several examples of countries for which we do not have any or sufficiently reliable data for prior years. For a number of cases, e.g. Bahamas, Cuba, Germany³⁸, Israel, Kosovo, Puerto Rico, Somalia, former Soviet Republics, former Yugoslavia, Malta, Mongolia, Myanmar, West Bank and Gaza, etc. we have had to make specialized assumptions after undertaking research on the country's available data from historical records as well as economics analyses, to address issues related to splits and unifications, data gaps, or conflicting observations from distinct sources that are not otherwise resolved. The procedure we adopt to deal with cases of country splits and unifications is described in the appendix to this paper. The special assumptions made in the case of specific countries and years are identified in the online appendix of country assumptions (see gcp.info). We hope to try and acquire such information in future versions of the database, by eliciting the engagement of specialists and the general public.

Although we have until now restricted ourselves to survey-based information, we intend in the future to explore the use of census data where feasible, although issues of comparability will have to be considered carefully,

3.6 Aggregation Module

We have developed a module that can be used to obtain readily a consumption or income profile for an arbitrary grouping of countries. This can help to determine trends in poverty, inequality, growth in median consumption or income or other statistics of interest for any set of countries defined by region, income level, association membership or indeed any other criteria of interest³⁹. These patterns can be juxtaposed with individual country experiences to understand how the set of countries is performing. We can perform

³⁸ For example, for Germany prior to unification (1990), West Germany's distribution, mean and population information are used for Germany. We do not presently have any information on East Germany so prior to unification it is not included as part of Germany for these earlier years.

³⁹ Existing global datasets are not generally accompanied by an aggregation tool. In Povcalnet, users can obtain poverty headcounts for any arbitrary set of countries examined as a group but it does not provide a single aggregate distribution and hence cannot be used for inequality or inclusive growth analysis for the grouping of countries. Edward and Sumner (2013) present a notable exception.

various analytical exercises with data aggregated in this way such as decomposing contributions to levels (or changes in) inequality, poverty or other statistics into within-country and between-country components. The evolution of a group of countries can be surprising as it necessarily reflects the relative growth performance of different countries as well as their internal distributional dynamics. For instance, the evolution of inequality within a region (such as Latin America in recent years) may for this reason be different from what might be suggested by the evolution of inequality within individual countries. Several Latin American countries have experienced a dramatic decline in inequality in recent years: between 2000 and 2010 according to our estimates, the income Gini coefficient for Brazil has dropped 8 points (from 0.58 to 0.46). Chile's income Gini coefficient has dropped by 7 points (from 57 to 40). Despite the dramatic developments within a number of countries, our estimates indicate that in the same period the overall income Gini coefficient for Latin America and Caribbean has dropped too, but only by 4 points (from 0.55 to 0.51). This is because of the contribution of differential growth rates of different countries, which might not be apparent at first. This is an insight only made possible by looking at the composite of countries, as we are able to do. A few illustrative examples of applications of the aggregation module are discussed in Jayadev, Lahoti and Reddy (2015).

We confine ourselves here to briefly describing the method used to combine countries and to obtain a single consumption/income profile for the set of countries. We first obtain a consumption or income profile for all the individual countries within the grouping of countries and for a given year using the procedure described in previous sections. Next, employing a 'poverty-line sweep' method, we obtain consumption levels for the 0.5 and 1.5 percentiles of the group. Specifically, we start at an arbitrary but low income/consumption level and calculate the percentage of population of each country that has income/consumption below this level. Then, using the population share of each country in the aggregate grouping we obtain the percentage of the group population at this level of the poverty line. We then iteratively adjust the level of the poverty line until we obtain the income/consumption level under which the desired percentage of the group population (i.e. 0.5 or 1.5) lies almost exactly, to a specified level of tolerance. Using these 0.5 and 1.5 percentile income/consumption levels for the group as starting points, we then raise the income/consumption level progressively in steps so as to obtain income/consumption levels at just over 100 points along the distribution, using continuous error corrections to adjust the size of the steps as we proceed so as to arrive at points within every or nearly every percentile interval. The resulting set of percentile points and the corresponding income/consumption levels are then connected linearly to obtain a consumption profile and to create a 'synthetic population', i.e. a representative population with the requisite profile, ordinarily consisting of one 'synthetic person' to represent the income or consumption level of each percentile. Using the synthetic population we can calculate any poverty or inequality measure, measure of inclusivity of growth etc. that we may wish to calculate for the group with complete flexibility⁴⁰.

3.7 Comparability of Survey Data

Atkinson and Brandolini (2001, 2009) provide an account of problems of comparability of surveys across countries and discuss issues with existing databases. They emphasize the need for better documentation of the existing datasets, in order that there may be understanding of the methods used in constructing the datasets and greater clarity as to what kinds of comparisons are meaningful and to what degree. However, they also place a considerable burden of responsibility on the researcher for understanding the data and for determining the extent of comparability. Some recent reviews of existing databases (e.g. Jenkins (2015); Smeeding and Latner (2015)) highlight the need for accessibility, replicability, transparency, quality of documentation and user-friendliness. One of the major motivations for developing GCIP was indeed to be able to provide transparent and replicable data. We plan to release the

⁴⁰ The computer code detailing the aggregation procedure will ultimately be provided on www.gcip.info.

data, the underlying code and documentation of both methods used and of the various choices made (in particular through our websites www.gcip.info or www.globalconsumptionandincomeproject.org)

In addition, as already pointed out, GCIP is intended to provide a platform that is flexible. Researchers may choose to make different choices in data selection or processing and to generate their own version of the database (with suitable acknowledgment). Researchers could for instance easily change the exchange rate used for converting levels of income or consumption within countries to common currency units using the alternate conversion factors provided with the data or choose to use National Account means over survey means (both are also included). They could also easily modify the code we provide to turn off or even replace the procedure we employ to ‘standardize’ distributions or means, choosing to work only with income data, only with consumption data, or by pooling both. They may choose to change the order of priority we employ when selecting distributional or mean data for a given country-year or not to make such a selection at all and to work with the entire universe of data that we have collected. They may choose to work with survey year data alone, avoiding any interpolation or extrapolation to non-survey years.

In GCIP we at present select only per-capita data, standardize the welfare concept (by focusing on either consumption or income) and select surveys where multiple are available in order to improve quality and comparability according to our best judgment. We do not, and in several cases cannot, address other comparability issues due to a host of other reasons such as the use of diverse income concepts in different income surveys, varying survey designs across countries etc. However, in all instances we intend to provide the user with all the relevant details in accessible format to make the right decision for the purpose at hand.

4. Results⁴¹

In this section, by way of conclusion we provide a few figures and tables that offer more specific indications of the kinds of analysis that are possible with the dataset. These are only illustrative examples. GCIP has been used to investigate poverty, inequality, inclusive growth, the profile of the middle class in Jayadev, Lahoti and Reddy (2015), Reddy and Lahoti (2015) and Jayadev, Lahoti and Reddy (2015b) and is being used to investigate these and other topics in a number of working papers under preparation. For the particular examples that follow we choose to focus on the global consumption distribution.

Evolution of World Consumption Distribution

Figure 1 shows the evolution of world consumption distribution from 1980 to 2010 with the contribution of China and the rest of the world. It shows “twin peaks” in the 1980s (perhaps first identified by Quah (1996)). However, the period since then has seen the transformation of world consumption from a bimodal to a unimodal distribution and one in which the overall distribution has narrowed (also observed by Milanovic, 2005, Weisbrod et al, 2007 and noted in our own recent work, e.g. Jayadev, Lahoti and Reddy (2015a); see also Jayadev Lahoti, and Reddy (2015b) which use GCIP to describe the evolution of the global middle class for various understandings of the term.). This is primarily due to China’s growth in the period and to some extent the recent rapid growth in some other developing countries, especially after 2000. The Chinese consumption median increased by 644 percent between 1980-2010 while the rest of world experienced only a slight increase in the consumption median (nine percent). The corresponding change in means for the two groups is higher but a vast differential is here too apparent (885 percent vs. 26 percent). The world consumption distribution for 2010 might give a false impression of a symmetrical distribution of consumption across the world with little inequality but that is partly a result of using a log-

⁴¹ The results presented in this section are based on the datasets version dated Feb 1, 2016.

scale for consumption expenditure. If we plot consumption instead of log-consumption the distribution seems highly skewed, as shown in Figure 2.

The factors underlining the changing world distribution are underlined rather dramatically if one looks at the evolution of the global consumption Gini coefficient and the Theil index when including and excluding China, as in Figure 3. (This sort of analysis is straightforward when using the GCIP due its flexible aggregation feature). Rapid Chinese growth and its large population have meant that the global consumption Gini coefficient has fallen monotonically from its peak of .71 in the 1970s and 1980, to a low of 0.64 in 2013. However, excluding China from the picture suggests an altogether different picture. Consumption inequality for the world without China was increasing from 1970 to 2000 and has since declined slightly. Over the period of 1970 to 2013 it has increased slightly from .64 in 1970 to .66 in 2013. Given the multitude of sources of uncertainty affecting these aggregate estimates – sampling, measurement and coverage errors in surveys, widely varying and inconsistent survey concepts and methods, difficulties with estimating and interpreting PPPs, poorly measured growth rates and other sources – it's not clear if these modest changes should be viewed as statistically or economically significant (see more on the sources of errors in global inequality estimates in Anand and Segal (2008)). It is interesting to note that the world excluding China has become more unequal (after 2000-2005) than the world including China, whereas the opposite was true earlier. China's inclusion has clearly also had a significant moderating influence on world inequality, in particular in the period 1980-2000. This was presumably because rapid Chinese growth served to reduce average gaps between the Chinese people and those richer than them. This effect overwhelmed the increase in inequality within the country and indeed increases in inequality elsewhere.

Poverty Headcount Ratios

Figure 4 presents the evolution of the number of poor across the world (with and without China) defined by various poverty lines. Here again we see the huge impact of China in global poverty reduction. The rest of the world has seen the number of poor increase (or almost no change for the World Bank's \$1.25 (2005 PPP) poverty line in recent years) from 1960 to 2010, with a slight decline after 2010, while China has seen a huge decline in number of poor. Future decline in poverty for this poverty line will need to be brought about as a result of a sizable contribution from other developing countries as the estimates poverty headcount ratio in 2013 for China for this poverty line is only eleven percent. Using GCIP, Reddy and Lahoti (2015) present alternate poverty estimates using 2011 PPPs and alternate poverty lines and compare them with the World Bank's latest estimates.

These are only examples of the range of analyses to which the GCIP can be put. As we discuss in Jayadev, Lahoti and Reddy (2015a) the portrayal of how successful the development process worldwide has been over the last fifty years depends very much on the criteria that we adopt (e.g. how high we set a poverty line). Such an insight is only possible because of the long-range and global analysis that the GCIP makes possible. Applications to monitoring and forecasting are also potentially interesting, since with continually updated and forecasted growth data, income and consumption profiles of national, regional and world populations can be estimated, so as to assess the impact of recent events or to examine the possible future consequences of trends.

5. Conclusion

The lottery of birth -- to whom one is born, when and where -- accounts for the majority of variation in the resources and opportunities available to human beings. Within nations, other influences -- one's gender, ethnic or racial category and other such factors -- serve to disadvantage some individuals in myriad, often invisible, ways from before they are born until their deaths. These patterns of inequality can

be reinforced over generations through the effects of structural barriers, differences in political power or social discrimination, limiting the potential of persons to flourish equally or indeed to do so at all. One recent estimate suggests that the richest eight percent of individuals in the world enjoy the same income as the other ninety two percent of the population (Milanovic 2013) and this is likely an underestimate as the incomes of the rich are poorly reflected in household surveys and even in tax records⁴².

In recent years, prominent social movements across the world (from the Indignados in Spain to the Movimento dos Trabalhadores Rurais Sem Terra (MST) in Brazil to the Occupy movements and the Arab Spring protests) have all been at least partly driven by the the perceived illegitimacy of existing (and often growing) economic and political inequalities, and these are only the most well-known such instances. Governments in many parts of the world, it seems, are faced with dissatisfied or disaffected citizenries. At the same time, the middle class is also burgeoning in many countries and, especially if modestly defined, arguably in the world as a whole. Poverty appears to have fallen by certain measures although in a very geographically uneven way. Rapid economic growth appears to have been very unevenly shared within and across countries with some countries doing far better than others at generating inclusive growth, regardless of how this is understood. These diverse facts give rise to a complex picture of a changing global reality. Better research and data is needed to begin to capture the gross contrasts as well as the necessary nuances. Such data must be used not only for purposes of description but in order to better understand and influence the determinants of the changing relative and absolute fortunes of people.

We have described the Global Consumption and Income Project and presented some examples of the sorts of inferences that might be drawn from it. There are myriad applications that can be imagined. We present a work in progress that offers possibilities for a deeper understanding of the evolution of material well-being both within and across countries, for geographical regions and groups of countries and for the world as a whole, and that extends from description to explanation. It is to this end that we introduce our project, and invite interested specialists and the world public to help us build upon and improve it.

⁴² Some recent attempts have been made to try and include additional data from alternative sources such as tax and administrative records generally (most notable in this regard is the work of the authors of the Top Incomes Database). We hope to supplement future versions of our database with information from such sources and have begun initial work in this regard.

References:

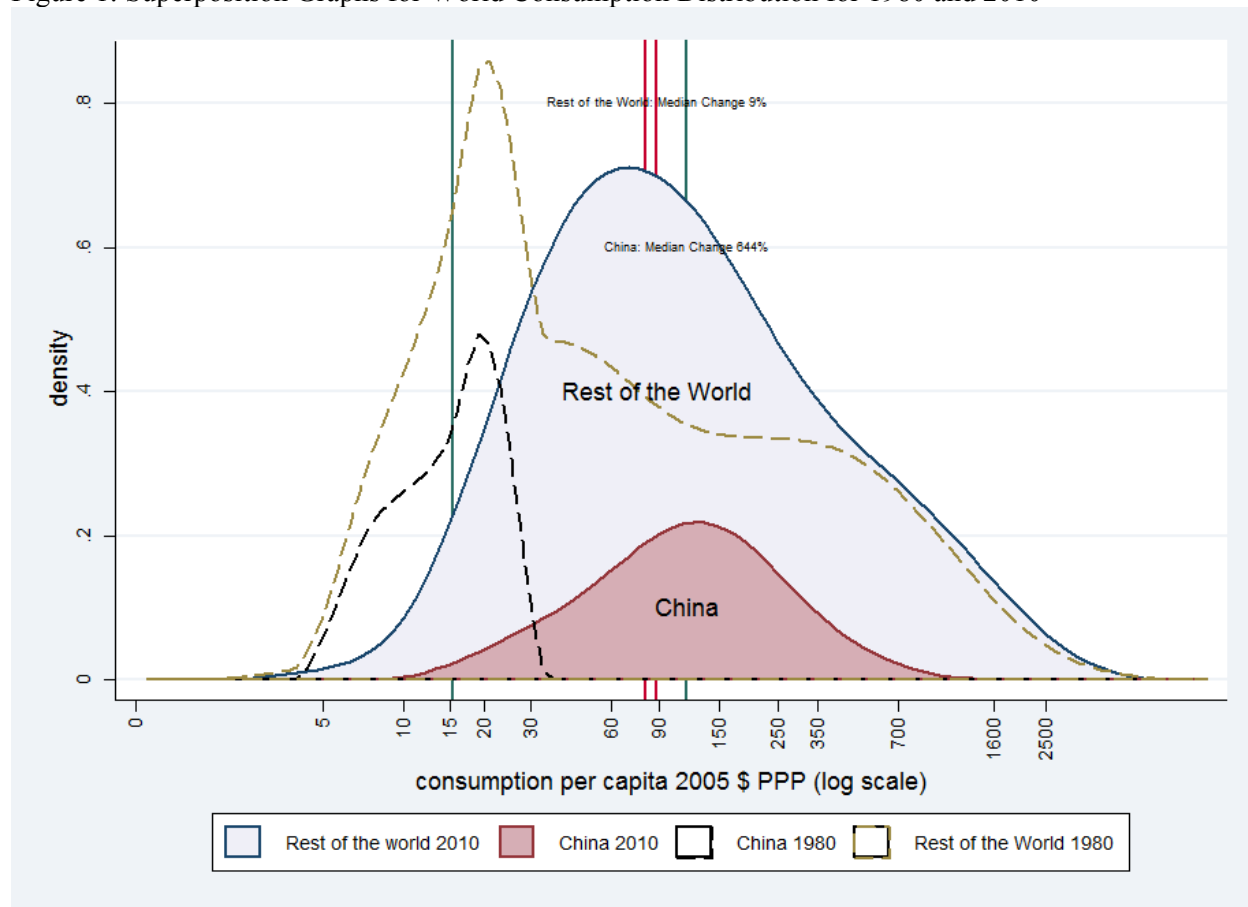
- Aaberg, R. and I. Melby (1998), "The Sensitivity of Income Inequality to Choice of Equivalence Scales", *Review of Income and Wealth*, Vol. 44, No. 4
- Atkinson, A, & A. Brandolini, (2001). "Promise and Pitfalls in the Use of "Secondary" Data-Sets: Income Inequality in OECD Countries As a Case Study," *Journal of Economic Literature*, Vol. 39, No.3, pages 771-799, September.
- Atkinson, A. B., & A. Brandolini, (2009). "On data: a case study of the evolution of income inequality across time and across countries" *Cambridge Journal of Economics*, Vol. 33 No. 3, 381-404.
- Anand, S. and P. Morduch, (1996), "Poverty and the Population Problem: Evidence from Bangladesh" , Harvard Institute for International Development, Development Discussion Paper No. 559, available on <http://www.cid.harvard.edu/hiid/559.pdf> .
- Anand, S., & P. Segal, (2008). "What do we Know about Global Income Inequality?" *Journal of Economic Literature*, 57-94.
- Banks, J. and P. Johnson (1994), "Equivalence Scale Relativities Revisited", *Economic Journal*, Vol. 104, pp. 883-890.
- Blaylock (1991), "The Impact of Equivalence Scales on the Analysis of Income and Food Spending Distributions", *Western Journal of Agricultural Economics*, Vol. 16, No 1, pp. 11-20.
- Buhmann, B, L. Rainwater, G. Schmaus and T. Smeeding (1988), "Equivalence Scales, Well-Being, Inequality, And Poverty: Sensitivity Estimates Across Ten Countries Using The Luxembourg Income Study (Lis) Database", *Review of Income and Wealth*, Vol.34, No.2,pp. 115-142.
- Chen, S., and M. Ravallion (2004). "How Have the World's Poorest Fared Since the Early 1980s?", *World Bank Research Observer*, Vol. 19, No. 2, pp. 141-169.
- Chotikapanich, D. (1993)., "A comparison of Alternative Functional Forms for the Lorenz Curve", *Economics Letters*, Vol. 41 No. 2, 129-138.
- Coulter, F. A. E., Cowell, F. A. and Jenkins, S. P. (1992a) "Differences in Needs and Assessment of Income Distributions", *Bulletin of Economic Research*, Vol. 44, pp. 77-124
- Coulter, F. A. E., Cowell, F. A. and Jenkins, S. P. (1992b) "Equivalence Scale Relativities and the Extent of Inequality and Poverty", *Economic Journal*, Vol. 102, pp.1067-1082.
- Cowell, F.A. and M. Mercader-Prats (1999), "Equivalence Scales and Inequality", LSE STICERD Working Paper No. 27, available on http://eprints.lse.ac.uk/2190/1/Equivalence_of_Scales_and_Inequality.pdf
- Datt, G. (1998). "Computational Tools for Poverty Measurement and Analysis". Washington, D.C.: International Food Policy Research Institute.
- Deaton, A. (2005). "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)", *Review of Economics and Statistics*, 87(2), 395-395.

- Deininger, K., and L. Squire (1996). "A New Data Set Measuring Income Inequality". *The World Bank Economic Review*, 10(3), 565-591.
- Dykstra, S., B. Dykstra, B. and J. Sandefur, (2014). "We Just Ran 23 Million Queries of the World Bank's Website", Center for Global Development, Working Paper 362.
- Edward, P and Sumner A (2013), "The Geography of Inequality: Where and by How Much Has Income Distribution Changed since 1990?", Center for Global Development, Working Paper 362
- Ferreira, F. H. G., Chen, S., Dabalen, A. L., Dikhanov, Y M., Hamadeh, N, Jolliffe, D. M., Narayan, A., Prydz, E.B., Revenga, A. L., Sangraula, P., Serajuddin, U., Yoshida, N. (2015), "A global count of the extreme poor in 2012 : data issues, methodology and initial results". Policy Research working paper; no. WPS 7432. Washington, D.C. World Bank Group.
<http://documents.worldbank.org/curated/en/2015/10/25114899/global-count-extreme-poor-2012-data-issues-methodology-initial-results>
- Jain, S. (1975). "*Size distribution of income: A compilation of data*" (pp. 101-104). Washington, DC: World Bank.
- Jayadev, A., Lahoti, R., and Reddy, S. G. (2015a), "Who Got What, Then and Now? A Fifty Years Overview from the Global Consumption and Income Project.: working paper, May 4th.
- Jayadev, A., Lahoti, R., and Reddy, S. G. (2015b). "The Middle Muddle: Conceptualizing and Measuring the Global Middle Class", working paper.
- Jenkins, S. P. (2015). "World Income Inequality Databases: an Assessment of WIID and SWIID", *The Journal of Economic Inequality*
- Kakwani, N.C. (1980). "Functional Forms for Estimating the Lorenz Curve: A Reply". *Econometrica*, Vol. 48, Issue 4, pp 1063-1064
- Lakner, C. and B. Milanovic (2013). "Global Income Distribution: from the Fall of the Berlin Wall to the Great Recession". *World Bank Working Paper* No. 6719, December.
- Luxembourg Income Study Database (LIS), www.lisdatacenter.org (multiple countries; June 1st- June 3rd 2014). Luxembourg: LIS.
- Milanovic, B. (2002). "True World Income Distribution, 1988 And 1993: First Calculation Based On Household Surveys Alone". *The Economic Journal*, 112(476), 51-92.
- Milanovic, B. (2005). *Worlds Apart: Measuring International and Global Inequality*. Princeton, N.J.: Princeton University Press.
- Milanovic, B. (2012). "Global Income Inequality by the Numbers in History and Now - an Overview", Washington, D.C.: World Bank, Development Research Group, Poverty and Inequality Team.
- Minoiu, C. and S.G. Reddy (2009). "The Estimation of Poverty and Inequality through Parametric Estimation of Lorenz Curves: An Evaluation". *Journal of Income Distribution*, vol. 18(2), pages 160-178, June.

- Niño Zarazúa, M., Roope, L., & Tarp, F. (2014). "Global Interpersonal Inequality: Trends and Measurement"
- Nurkse, R. (1953). The Problem of capital Formation in less-developed Countries. *Oxford University Press*, 33, 1-337.
- Pinkovskiy, M., & Sala-i-Martin, X. (2009). "Parametric Estimations of the World Distribution of Income". NBER Working Paper 15433.
- Quah, D.T. (1996). "Twin Peaks: Growth and Convergence in Models of Distribution Dynamics". *Economic Journal*, vol. 106(437), pages 1045-55, July.
- Rasche, R. H, Gaffney, J, A.Y.C Koo, and N Obst (1980). "[Functional Forms for Estimating the Lorenz Curve](#)", *Econometrica* Vol. 48, No. 4 (May, 1980) , pp. 1061-1062
- Reddy, S. G., and Lahoti, R. (2015). "\$1.90 per day: What does it say?" (No. 189). Courant Research Centre: Poverty, Equity and Growth-Discussion Papers.
- [Sefil, S. \(2015\), "Sensitivity Of Turkish Income Distributions to Choice of Equivalence Scale", *Topics in Middle Eastern and African Economies*, Vol. 17, Issue No. 1.](#)
- Serajuddin, Umar; Uematsu, Hiroki; Wieser, Christina; Yoshida, Nobuo; Dabalén, Andrew L.. 2015. *Data Deprivation : Another Deprivation to End*. Policy Research Working Paper; no. WPS 7252. Washington, D.C. : World Bank Group.
<http://documents.worldbank.org/curated/en/2015/04/24426058/data-deprivation-another-deprivation-end>
- Smith, L. C., O. Dupriez, and N. Troubat, (2014). "Assessment Of The Reliability And Relevance Of The Food Data Collected In National Household Consumption And Expenditure Surveys." *International Household Survey Network* working paper.
- Solt, F. (2009). "Standardizing the World Income Inequality Database". *Social Science Quarterly*, 90(2), 231-242.
- Stiglitz, J. E., A. Sen, and J-F. Fitoussi (2010). *Mismeasuring Our Lives: Why GDP Doesn't Add Up*. New York: New Press.
- Smeeding, T., and J.P. Latner, (2015). "PovcalNet, WDI and 'All the Ginis': a Critical Review" *The Journal of Economic Inequality*, 1-26.
- Villaseñor, J., & Arnold, B. C. (1989). "Elliptical Lorenz Curves". *Journal of Econometrics*, 40(2), 327-338.
- Weisbrod, J., S. Vollmer, Sebastian and H. Holzmann (2007). "Perspectives on the World Income Distribution: Beyond Twin Peaks Towards Welfare Conclusions".

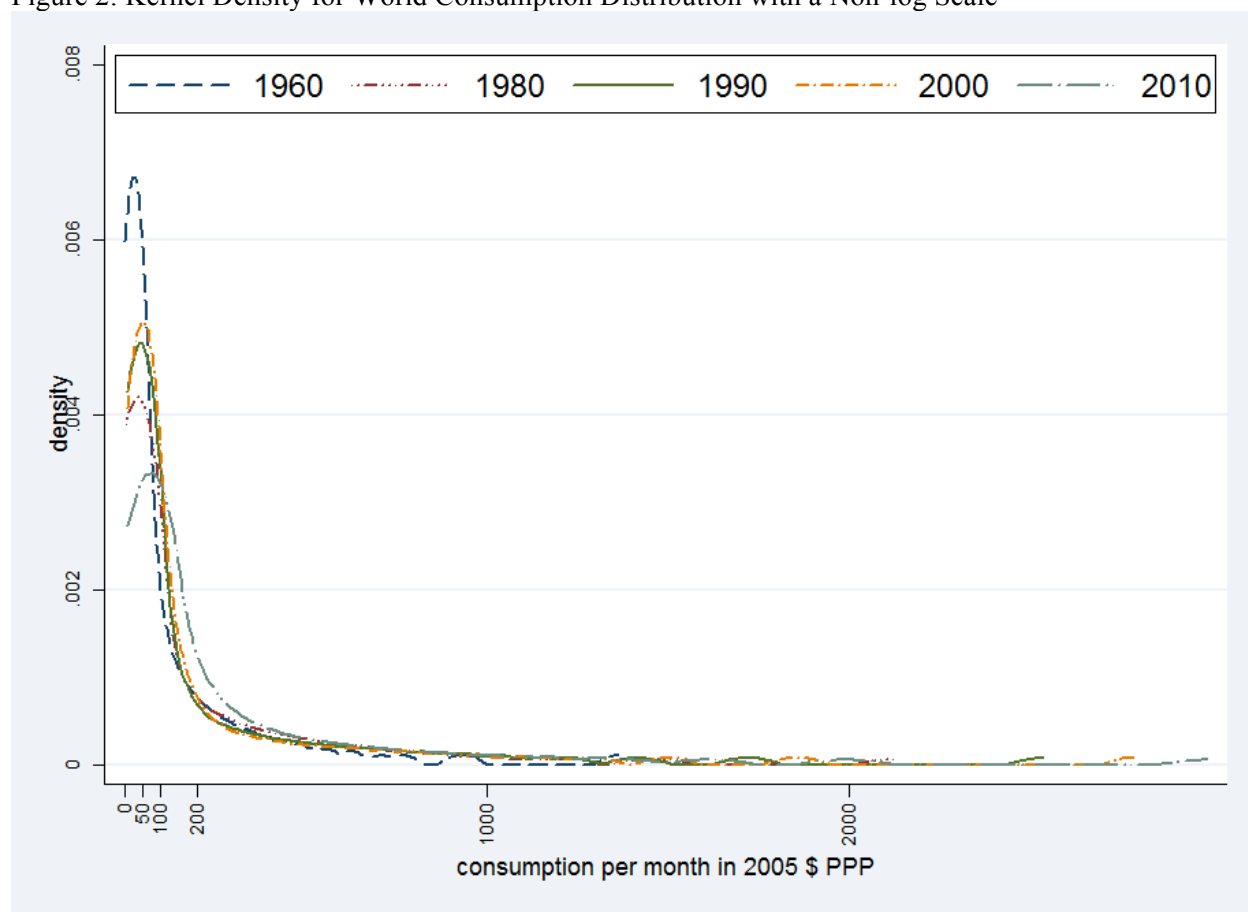
FIGURES AND TABLES

Figure 1: Superposition Graphs for World Consumption Distribution for 1980 and 2010



Source: Authors' calculations using GCD

Figure 2: Kernel Density for World Consumption Distribution with a Non-log Scale



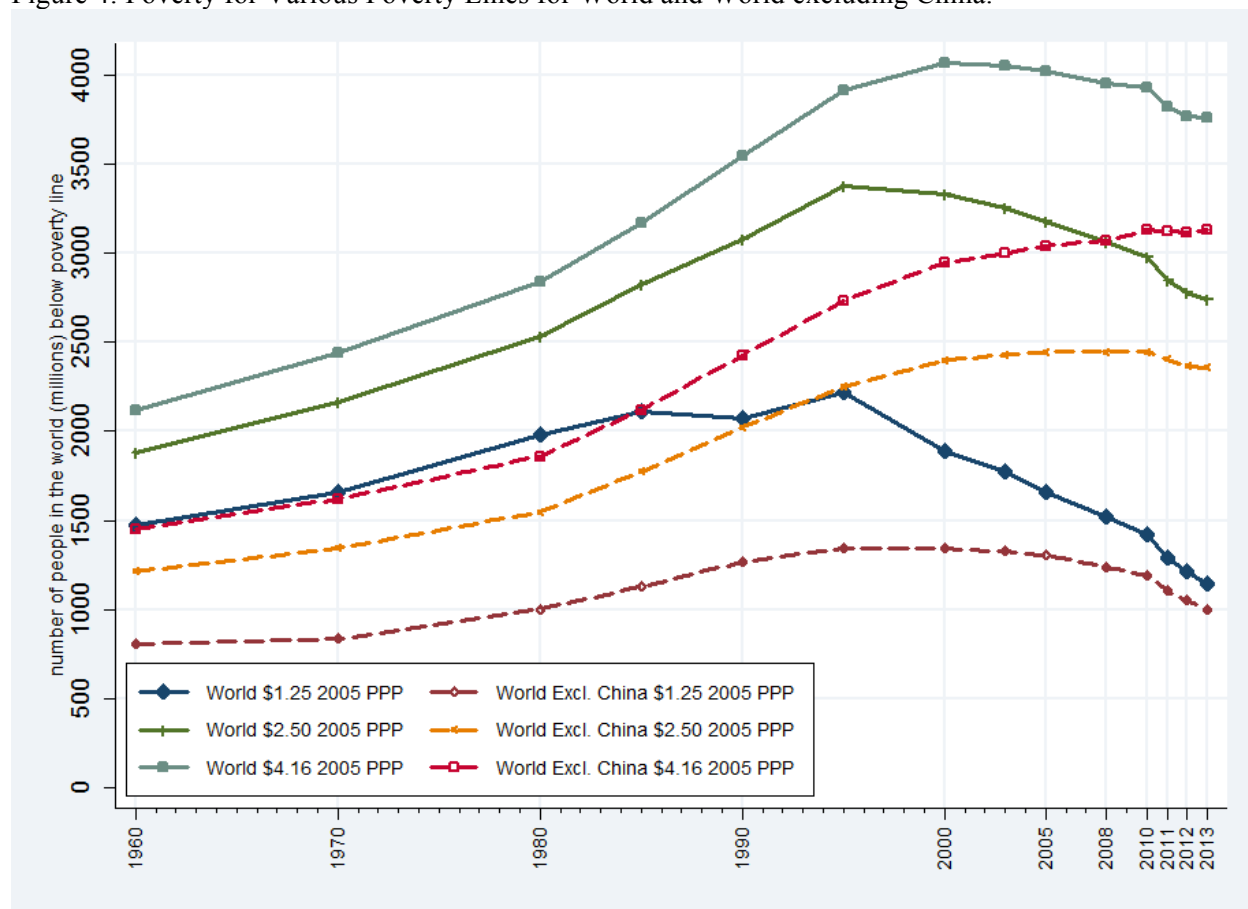
Source: Authors' calculations using GCD

Figure 3: Inequality for World with and without China



Source: Authors' calculations using GCD

Figure 4: Poverty for Various Poverty Lines for World and World excluding China.



Source: Authors' calculations using GCIP.

Note: The \$4.16 poverty line corresponds to the minimum per-capita expenditure required to obtain a home-cooked nutritious meal in the US as estimated in the Thrifty Food Plan of the United States Department of Agriculture (USDA). More details on the poverty line can be found in Reddy and Lahoti (2015).

Table 1: Comparison of Various Global Datasets.

Database Features	Penn World Tables	WIID (Version 2.0C)	SWIID	Povcalnet	GrIP	GCIP
Coverage by type of country	Both developing and developed countries	Both developing and developed countries	Both developing and developed countries	Both developing and developed countries	Both developing and developed countries	Both developing and developed countries
Temporal coverage	1950-2011(not all countries)	1960-2008	1960-2005	1980-2012	1990-2012	1960-2012
Level of consumption/income	Both, based on national accounts	Only one of consumption or income, and not for all surveys.	Neither	Consumption or income only, based on surveys wherever possible	Consumption or income only, based on both NA and surveys	Both, based on surveys and national accounts.
Distribution by quantile	No	Yes	No	Yes, only survey years	Yes	Yes, all years
Adjustment of data on distributions to achieve greater comparability	N/A	No	Yes (through econometric estimation of Gini coefficients, adopting LIS as 'gold standard')	No	No	Yes (through econometrically estimated quintile-specific consumption-income ratio)
Interpolation for non-survey years	Yes	No	No	Not of reported quantiles or means, but implicit in reported poverty estimates	Yes- using distributions from surrounding region	Yes

Flexibility in modifying database according to alternate assumptions	No	No	In certain respects	In certain respects	In certain respects	Transparent about sources and methods so as to be flexible
Inequality measures	No	Gini only	Gini only	Selected, for survey years only	Potentially all	Potentially All
Aggregate over countries	Yes	No	No	Yes, but only for Poverty Measures	Yes for poverty	Yes, for poverty, inequality and the complete consumption or income profile (arbitrary number of quantile means)

Table 2: Welfare Concepts in GCIP Datasets

Global Consumption Database (GCD)		
Welfare concept measured	Number of Surveys	Percentage
Consumption	830	42.7
Earnings, Gross	20	1
Earnings, Net	18	0.9
Factor Income	1	0.1
Income	285	14.6
Income, Disposable	588	30.2
Income, Gross	162	8.3
Monetary Income, Disposable	32	1.6
Monetary Income, Gross	8	0.4
Taxable Income, Gross	2	0.1
Total	1946	100
Global Income Database (GID)		
Consumption	575	29.5
Earnings, Gross	20	1
Earnings, Net	18	0.9
Factor Income	1	0.1
Income	307	15.8
Income, Disposable	758	38.9
Income, Gross	217	11.1
Monetary Income	1	0.1
Monetary Income, Disposable	35	1.8
Monetary Income, Gross	12	0.6
Taxable Income, Gross	2	0.1
Total	1946	100

Source: Author's calculation

Table 3: Summary Statistics for Surveys used in Standardization

	LIS	WIID	Total
# Surveys	87	117	204
# of Surveys by Regions			
Africa & Middle East	10	32	42
East Asia & Pacific	12	21	33
Europe	45	20	65
Latin America & Caribbean	18	19	37
Post Soviet	1	17	18
South Asia	1	8	9
# of Countries from each region			
Africa & Middle East	2	19	21
East Asia & Pacific	3	5	8
Europe	12	7	19
Latin America & Caribbean	4	6	10
Post Soviet	1	8	9
South Asia	1	3	4
# of Surveys before and after 1995			
1995 or Later	65	65	130
Before 1995	22	52	74
# Types of Income Surveys			
Income, Disposable	87	36	123
Income, Gross	0	81	81
# of Surveys by Income Grouping			
Low income	0	20	20
Lower middle income	4	44	48
Upper middle income	42	39	81
High income	41	14	55

Table 4a: Seemingly Unrelated Regression (SUR) for converting Income distribution into equivalent consumption distribution for Global Consumption Database

	Quintile				
	1	2	3	4	5
Income Quintile Share	0.363*** 0.031	0.334*** 0.029	0.342*** 0.028	0.317*** 0.032	0.340*** 0.027
Income Grouping (Base: Low Income)					
Lower Middle Income	-0.661* 0.353	-0.481 0.403	-0.183 0.402	0.436 0.38	0.894 1.381
Upper Middle Income	0.099 0.396	0.493 0.453	0.628 0.452	1.071** 0.426	-2.196 1.553
High Income	0.144 0.44	0.513 0.505	0.723 0.505	1.180** 0.478	-2.817 1.733
Region Grouping (Base: Africa & Middle East)					
East Asia & Pacific	2.276*** 0.316	1.954*** 0.363	1.371*** 0.363	0.276 0.348	-5.929*** 1.25
Europe & Central Asia	1.219*** 0.328	1.527*** 0.382	1.422*** 0.38	0.887** 0.353	-4.814*** 1.297
Latin America & Caribbean	-0.009 0.331	-0.13 0.378	-0.011 0.377	0.057 0.358	-0.015 1.293
Post Soviet	0.710* 0.389	1.370*** 0.447	1.496*** 0.447	1.258*** 0.425	-4.894*** 1.534
South Asia	2.511*** 0.473	2.460*** 0.545	1.981*** 0.542	0.922* 0.509	-7.896*** 1.86
Welfare Concept (Base: Income Disposable)					
Income, Gross	0.227 0.253	0.406 0.288	0.533* 0.287	0.767*** 0.271	-2.054** 0.987
Time (Base: 1995 or Later)					
Pre 1995	-0.131 0.191	0.075 0.218	0.312 0.217	0.413** 0.205	-0.58 0.747
Constant	3.773*** 0.408	6.230*** 0.492	8.739*** 0.539	13.164*** 0.706	34.561*** 2.277
R2	0.663	0.649	0.621	0.47	0.634
N	204				

Table 4b: Seemingly Unrelated Regression (SUR) for converting Consumption distribution into equivalent income distribution for Global Income Database

	1	2	3	4	5
Consumption Quintile Share	0.325*** 0.034	0.318*** 0.032	0.345*** 0.031	0.293*** 0.034	0.324*** 0.031
Income Grouping (Base: Low Income)					
Lower Middle Income	0.197 0.392	0.273 0.436	0.486 0.431	0.557 0.456	-1.491 1.499
Upper Middle Income	1.430*** 0.421	1.668*** 0.47	1.684*** 0.465	1.206** 0.492	-5.990*** 1.616
High Income	2.112*** 0.455	2.836*** 0.508	3.114*** 0.503	2.921*** 0.533	-10.873*** 1.75
Region Grouping (Base: Africa & Middle East)					
East Asia & Pacific	1.002*** 0.361	1.821*** 0.396	2.219*** 0.389	2.810*** 0.405	-7.846*** 1.359
Europe	2.286*** 0.352	3.602*** 0.395	3.556*** 0.389	2.616*** 0.405	-12.140*** 1.349
Latin America & Caribbean	-0.485 0.364	0.294 0.406	0.899** 0.401	1.539*** 0.424	-2.197 1.395
Post Soviet	1.830*** 0.418	2.452*** 0.469	2.733*** 0.464	2.887*** 0.489	-9.879*** 1.611
South Asia	1.965*** 0.532	3.243*** 0.591	3.308*** 0.581	2.758*** 0.606	-11.272*** 2.023
Time (Base: 1995 or Later)					
Pre 1995	0.32 0.209	-0.177 0.233	-0.484** 0.231	-0.583** 0.244	0.895 0.801
Constant	0.475 0.392	2.553*** 0.48	4.988*** 0.551	11.068*** 0.779	48.899*** 2.192
R-square	0.686	0.753	0.761	0.631	0.755
N	204				

Table 5: In-Sample Predictions of Income to Consumption Regressions.

	Mean of Income Share	Mean of Consumption Share	Mean of Predicted Consumption Share	Difference between Actual Consumption and Predicted Consumption Shares	Difference between Consumption and Income Shares
1st Quintile	4.93	6.45	6.45	0.005	1.53
2nd Quintile	9.26	10.73	10.74	0.007	1.4
3rd Quintile	13.67	15.02	15.03	0.01	1.36
4th Quintile	20.49	21.43	21.45	0.014	0.95
5th Quintile	51.69	46.36	46.39	0.027	-5.32

Table 6: Conversion of Income Survey Distribution for Brazil, 1996 to Equivalent Consumption Distribution

	Consumption Shares			
	Original Income Share	After Regression	After Adjustment for Adding- up Constraint	Difference in Equivalent Consumption and Income Shares
Q1	2.5	4.67	4.67	2.17
Q2	5.8	8.04	8.04	2.24
Q3	10.1	12.18	12.19	2.09
Q4	18.2	18.98	18.99	.80
Q5	63.4	56.07	56.10	-7.30
Total		99.95	100	
Gini	57.7		48.7	
Poverty Rate (\$1.25 2005 PPP)	10.23		0.83	

Table 7a: Summary Statistics for Surveys included in Global Consumption Database (GCD)

	1960-69	1970-79	1980-89	1990-99	2000-09	2010-13	Total
Number of Surveys	79	82	288	496	716	285	1946
% Consumption Surveys	14	12	24	48	54	41	43
% Surveys Covering Complete Population	56	68	78	98	100	100	93
% Surveys Covering all Areas in the Country	95	94	93	96	98	98	96
% Surveys with Mean Levels Information	41	49	65	90	96	100	86
Source of Surveys (%)							
EU-SILC	0	0	0	0	0	10	1
LIS	3	15	13	13	13	18	13
Povcalnet	0	0	15	35	61	59	42
Data from Statistical Office	0	0	0	0	0.4	0	0.1
SEDLAC	0	1	6	15	17	11	13
WYD	0	0	1	2	1	0	1
WIId	97	84	65	34	8	3	29
Regions (%)							
East Asia & Pacific	5	13	12	10	9	8	10
Europe & Central Asia	38	44	53	40	52	57	49
Latin America & Caribbean	23	20	19	28	22	20	23
Middle East & North Africa	5	6	3	4	3	2	3
North America	0	5	1	1	1	1	1
South Asia	19	7	4	3	3	2	4
Sub-Saharan Africa	10	5	8	14	11	9	11
Income Grouping (%)							
Low income	9	5	4	8	15	13	8
Lower middle income	31	16	23	23	13	7	22
Upper middle income	39	44	44	36	37	53	40
High income	21	34	28	33	35	27	30

Table 7b: Summary Statistics for Surveys included in Global Income Database (GID)

	1960- 69	1970- 79	1980- 89	1990- 99	2000- 09	2010- 13	Total
Number of Surveys	79	82	288	496	716	285	1946
% Consumption Surveys	13	7	15	27	41	31	30
% Surveys Covering Complete Population	55	68	77	98	100	100	93
% Surveys Covering all Areas in the Country	95	94	93	95	97	98	96
% Surveys with Mean Levels Information	41	49	65	90	96	100	86
Source of Data (%)							
EU-SILC	0	0	0	0	0	10	2
LIS	2	15	13	14	14	18	14
Povcalnet	0	0	12	23	56	59	37
Data from Statistical Office	0	0	0	0	0.4	0	0.1
SEDLAC	0	1	6	17	18	10	13
WYD	0	0	1	2	1	0	1
WIID	97	84	67	44	12	4	34
Regions (%)							
East Asia & Pacific	5	13	12	10	9	8	10
Europe & Central Asia	38	44	53	40	52	57	49
Latin America & Caribbean	23	20	19	28	22	20	23
Middle East & North Africa	5	6	3	4	3	2	3
North America	0	5	1	1	1	1	1
South Asia	19	7	4	3	3	2	4
Sub-Saharan Africa	10	5	8	14	11	9	11
Income Grouping (%)							
Low income	9	5	4	8	15	13	8
Lower middle income	31	16	23	23	13	7	22
Upper middle income	39	44	44	36	37	53	40
High income	21	34	28	33	35	27	30

Table 8a: Summary Statistics on Countries included in Global Consumption Database (GCD)

	1960-69	1970-79	1980-89	1990-99	2000-09	2010-13	Total
Number of Countries	44	44	98	133	151	111	161
Source of Data							
EU-SILC	0	0	0	0	0	9	29
LIS	2	6	18	22	17	17	45
Povcalnet	0	0	22	62	106	73	140
Primary Data	0	0	0	0	2	0	2
SEDLAC	0	0	5	12	14	7	23
WYD	0	0	3	4	2	0	5
WIID	42	38	50	33	10	5	116
Region							
East Asia & Pacific	2	6	11	17	20	11	21
Europe & Central Asia	16	14	44	48	49	44	50
Latin America & Caribbean	12	10	18	24	20	18	27
Middle East & North Africa	3	4	6	8	9	6	10
North America	0	2	2	2	2	2	2
South Asia	4	4	5	6	8	7	8
Sub-Saharan Africa	7	4	12	28	43	23	43

Table 8b: Summary Statistics on Countries Included in Global Income Database (GID)

	1960-69	1970-79	1980-89	1990-99	2000-09	2010-13	Total
Number of Countries	44	44	98	133	151	111	161
Source of Data							
EU-SILC	0	0	0	0	0	1	29
LIS	1	5	17	21	26	34	45
Povcalnet	0	0	17	45	76	60	136
SEDLAC	0	0	3	11	16	9	22
WYD	0	0	2	4	3	0	5
WIID	42	39	59	51	27	6	119
Region							
East Asia & Pacific	2	6	11	17	19	11	20
Europe & Central Asia	16	14	44	48	48	44	50
Latin America & Caribbean	11	10	18	23	19	17	26
Middle East & North Africa	3	4	6	8	9	6	10
North America	0	2	2	2	2	2	2
South Asia		4	4	5	6	8	7
Sub-Saharan Africa	7	4	12	28	43	23	43

Table 9: Density of Surveys in Global Consumption Database (GCD) by Decade, Region and Income Group (% of all country-years that have survey information)

Region	1960-69	1970-79	1980-89	1990-99	2000-09	2010-13	1960-2013
East Asia & Pacific	2	6	19	25	29	23	17
Europe & Central Asia	7	9	18	36	55	29	29
Latin America & Caribbean	7	5	20	52	58	40	30
Middle East & North Africa	3	4	8	19	18	13	11
North America	0	20	15	35	30	38	21
South Asia	25	8	14	21	25	22	19
Sub-Saharan Africa	2	1	5	16	17	13	8
Income Group in 2010							
Low income	2	1	4	16	20	17	10
Lower middle income	6	3	11	29	38	23	18
Upper middle income	7	7	20	37	48	29	26
High income	5	8	18	33	39	26	22
Overall	5	5	14	30	37	24	19

Appendix: Country Splits and Unifications

I. Countries undergoing Splits:

There are countries in our database which experienced splits and for which we have data pertaining to the entities both before and after this event. Examples of the affected pre-existing entities include Czechoslovakia, Ethiopia, Indonesia, Malaysia, Pakistan, the USSR and Yugoslavia. Other countries which experienced splits over the period, such as Sudan, are ones for which we do not yet have sufficient data to incorporate them separately.

I.1 Former Socialist Economies:

For Czechoslovakia, the USSR and Yugoslavia we apply the following procedure. In most of these cases, we have data on means and distributions for the countries emerging from the split from near to the year in which they were formed and some distributional data on the combined country for the earlier period. As some of the countries undergoing splits were also formerly socialist economies, problems of available of distributional and mean information are compounded by the often lesser availability of household data for such countries.

Our approach to handling these countries is to attempt to assign a pseudo-mean and pseudo-distribution to the constituent countries even prior to the split, while recognizing that the assigned values may in fact be more characteristic of the unified country, and thus recommending appropriate interpretative caution. Such an approach allows us to estimate trends for the individual constituent units over a longer period, as well as to construct and report aggregates (whether prior to or after the split). Each of these possible choices has its benefits and costs in terms of statistical meaningfulness vs. inter-temporal comparability.

For the affected countries mentioned above, we use distributional data from the combined country to create a pseudo-distribution of each of the countries undergoing splits for years when the countries were one, as we do not possess distributional data for the individual constituent countries. We recognize that this is an inadequate assumption, in part because the distribution for the unified country reflects differences in income between successor countries as well as within them.

Although we do possess some mean estimates for the combined countries from national accounts or independent academic and institutional estimates, we do not use these for the constituent countries both because this would mean using the same mean for all constituent countries in the earlier years and because we don't have reliable national-level inflation data for the pre-split period to convert these estimates into units of a common PPP base year after the split. We therefore instead estimate means for the countries undergoing splits by extrapolating backward using real per-capita-income growth rate data, which is available in many instances. Where growth rate information for the constituent countries is not available, we use the growth rate of the combined country as a proxy for the growth rate of each of the countries resulting from the split. In particular, for the three countries undergoing splits mentioned above, we employ data on growth rates for the constituent countries from 1980 onwards from the Total Economy Database. Prior to 1980 we use the growth rates for the combined entities as provided by Angus Maddison.⁴³

⁴³ Maddison provides estimates of the level of real per-capita income in 1973 and in 1990 for a number of constituent countries (e.g. Soviet Republics) which could permit determining an average annual growth rate over the period, but we refrain from employing a smoothed estimate in favour of using TED annual estimates for the period from 1980 to 1990.

Table A1 has details on the data used for each of the combined countries undergoing splits.

Table A1:

	Survey Years	Growth Data	Source of Growth rate data	Inflation Data
Yugoslavia	1968-1990	1961-2006	Angus Maddison	
Bosnia and Herzegovina	2001-2007	1991-2015	TED and WDI	1994-
Croatia	1988-2008	1981-2015	TED and WDI	1985-
Montenegro	2005-2011	1991-2013	TED and WDI	2000-
Serbia	2002-2010	1991-2013	TED and WDI	1994-
Slovenia	1987-2011	1981-2015	TED and WDI	1992-
Macedonia, FYR	1994-2008	2015	TED and WDI	1990-
Kosovo	2003-2006			2000-
USSR	1980-1989	1961-1990	Angus Maddison	
Armenia	1996-2012	1981-2015	TED and WDI	1990-
Azerbaijan	1995-2008	1981-2015	TED and WDI	1990-
Belarus	1988-2011	1981-2015	TED and WDI	1990-
Estonia	1988-2011	1981-2015	TED and WDI	1992-
Georgia	1996-2012	1966-2015	TED and WDI	1965-
Kazakhstan	1988-2010	1981-2015	TED and WDI	1990-
Kyrgyz Republic	1988-2011	1981-2015	TED and WDI	1987-
Latvia	1988-2011	1966-2015	TED and WDI	1965-
Lithuania	1988-2011	1981-2015	TED and WDI	1990-
Moldova	1988-2011	1981-2015	TED and WDI	1989-
Russia	1988-2010	1961-2015	TED and WDI	1989-
Tajikistan	1999-2009	1981-2015	TED and WDI	1985-
Turkmenistan	1988-1998	1981-2015	TED and WDI	1987-
Ukraine	1985-2010	1981-2015	TED and WDI	1987-
Uzbekistan	1988-2003	1981-2015	TED and WDI	1987-
Czechoslovakia	1964-1992	1961-2006	Angus Maddison	1975-
Czech Republic	1988-2011	1971-2015	TED and WDI	1990-
Slovakia	1988-2011	1986-2015	TED and WDI	1992-

I.1 Other Countries Undergoing Splits

For the other countries undergoing splits, we have made specific assumptions, for instance about the coverage of surveys from before and after the split. These are mentioned in the online appendix on country assumptions (see gcip.info).

II. Countries Undergoing Unification:

Among the countries undergoing unification during the time interval covered by the database are Germany and Yemen. We do not have sufficient information for the constituent parts of Yemen prior to or posterior to its unification to form a picture of the country at this time.

For Germany, we use West Germany's distribution and mean for all of Germany prior to unification. We are actively interested in finding and integrating East German data from prior to unification so as to improve upon this inadequate approach.