Guidelines to Achieve Cross-country Data Comparability

Dirk Krueger, Fabrizio Perri, Luigi Pistaferri, Gianluca Violante

1 List of updates

This Section describes the main changes we have made to the guidelines after the Penn conference, building on the feedback we got from you and from the other conference participants. The updates are explained in more detail in the relevant sections of the document.

- **Comparison with National Accounts:** authors should document how aggregates from micro data compare to corresponding macro aggregates from NIPA.
- **Equivalization:** we decided to stick with the OECD scale, but we encourage you to try to use regression analysis, or explore other ways to equivalize income and consumption in the second part of the paper.
- **Clarification on individual inequality statistics:** When computing measure of individual wages and earnings, one should do it for all working age males and females in the sample, not just heads and spouses.
- **Life cycle profiles:** we have changed the way we want you to extract the age-inequality profiles.
- **Inequality measures:** we have chosen to use the variance of the logs instead of the standard deviation, since it can be decomposed more naturally. We also ask you to compute measures of the college premium, the experience premium and the gender premium. If your data are not top-coded, we encourage you to examine the dynamics of income shares at the very top of the distribution.
- **Deflation:** we have decided to simply use the CPI to deflate income and consumption measures
- **Outliers:** we have slightly modified our suggested way to deal with outliers at the bottom of the distribution.
- **Earnings process:** We now encourage you to explore alternatives to the simple random walk plus i.i.d. earnings process in the second part of the paper.
- **Lognormality test:** in the second part of the paper, we encourage you to test lognormality of the income and consumption distributions.
- **Data documentation:** One goal of the project is to deposit the data used for each country-specific paper on the RED website. It is therefore useful to add an appendix to your paper with data documentation allowing external researchers to use the data without too much effort.
- **List of Figures:** To insure more comparability across country studies we specifically describe a minimal list of figures that each paper should contain (see section 9)

2 Framework of analysis

- Before we start discussing the details of the data, it is helpful to remember the two models we have in mind to organize the micro data (recall our introductory document).

- The first is the exogenous labor supply model. In this model we treat labor earnings of the household as exogenous. The budget constraint reads as

$$c + a' = y^L + (1 + r)\, a + b + T$$

where y^L is labor earnings of *all* household members, $y^A = ra$ is private asset income of the household, b captures net private inter-vivos transfers from other households, T denotes transfers minus taxes from the government to the household, c is consumption of the household and a' represents assets that are accumulated for the next period. In this set-up we will mainly focus on y^L, $y^{L+} = y + b$, $y = y^{L+} + ra$, $y^D = y + T$, on c, and on some components of a.

- The second model is one with endogenous labor supply, where y is no longer exogenous but can be written as

$$y = w_m l_m + w_f l_f$$

where (l_m, l_f) represent labor supply of the male and the female and (w_m, w_f) represent wages of the male and the female. In this model we will treat wages as exogenous and, in addition to a and c, we will also report statistics on wages and labor supply for both members. We recognize that in some countries there is a nontrivial fraction of households where additional household members, besides spouses, contribute to household labor supply. We encourage you to explore this issue in the second part of the paper.

3 Choice of Data Set

- For several countries, more than one data set is available (for example, for the US, income data is available from PSID, CEX and CPS, among other data sets). Use your best judgement in the choice of your data source. When making this choice, keep in mind the following criteria: 1) the quality of the measurement; 2) the size of the data set; 3) the number of variables of interest (income, earnings, wage, hours, consumption, and wealth) jointly contained in the data set. A large number of relevant variables allows you to calculate cross-sectional correlations between pairs of variables; 4) a panel dimension. Longitudinal data sets permit estimation of the wage/earnings process discussed in section 8 and allows to compute cross-sectional correlations between pairs of variables in growth rates, not just in levels.

- Obviously, we encourage you to use multiple data sets to ensure that the analysis covers as many variables of interest as possible with the highest possible quality. Also, the use of two data sets for the same variable enables robustness checks of the findings; to this end, it is of course imperative that household, sample and variables definitions used across the two data sets are identical.

4 Unit of analysis, household head and sample selection

- The unit of analysis is the household. All types of households (singles, couples, extended families, etc.) will be considered. "Independent" analyses in the second part of the paper could concentrate on how the level and trend in inequality is driven by changes in the composition of household types in the population (this is especially relevant for countries that

have experienced dramatic shifts in the composition of household types, such as the UK and the US).

- We want to make the definition of the head of the household uniform across countries to insure that the facts are comparable across countries. For single households, this is of course not an issue. For households formed by couples (i.e. households in which there is one member denoted as the head or the reference and another denoted as the spouse) treat the male (in the of case mixed sex couples), the oldest male (in the case of 2-males couples) or the oldest female (in the case of 2-females couples) as the head of the household. For non-couple households, please treat the oldest working age (i.e., age 25-60) male as the head of the household. If no working age male is present in the household, then the oldest working-age female of the household is treated as the household head. If there are no male nor females with age 25-60 the household is not included in the sample (see below).

- Please select only households headed (as defined above) by an individual aged 25-60. The purpose of this sample selection criterion is to select households already out of school and not yet retired. By age 25 most individuals in most countries are out of school. The actual (average) retirement age in most countries is 60 or later. However, in some European countries early retirement is very common. For this reason it would be useful to also document retirement patterns by age, and how these pattern have changed over time.

- When constructing the data set, for each individual in the household you should keep information on demographics, i.e., at the very minimum gender, race (if applicable), age, education (defined as numbers of years of schooling), marital status, number of adults and children in the household (needed to construct adult-equivalence scales) and household characteristics such as geographical location (i.e. state, region or city or/and urban/rural). Occupation, when available, is also of interest and can be used in the independent analysis conducted in the second part of the paper.

4.1 Top coding

- In various data sets some variables of interest will be top-coded. First, we recommend to report large changes in top-coding thresholds (like the one which took place in the US CPS in 1993). Furthermore, if you believe that top-coding is an issue, we strongly encourage you to replace the top-coded observation with an imputed value based on an estimation procedure of the upper tail of the distribution. The Appendix includes a detailed description of this procedure kindly prepared by David Domeij. David has agreed to make his code available, so you can contact him directly in case you decide to do apply this procedure.

4.2 Treatment of poor-quality data and outliers

- One important issue is to preserve data quality. Some data sets flag households with inferior data quality (i.e. the US CEX classifies some households as incomplete income respondents) and those should be excluded from the analysis.

- Some households report implausibly low wages. Here we suggest the following strategy. Exclude also households which have at least one working member whose hourly wage is below an extremely low threshold. For countries with a minimum wage legislation this threshold could be 1/2 the minimum wage, for countries who do not have a minimum wage (such as Italy or Germany), the threshold should be chosen using appropriate local knowledge to exclude observations that are obviously due to measurement error or misreporting. As a general rule

we recommend to use 1/2 of the wage paid for fully unskilled labor (e.g., half of the wage paid to the lowest-paid workers at the local McDonalds restaurant, see Ashenfelter-Juraida 2004). It would be useful to report what fraction of the sample is excluded due to this exclusion restriction and how this fraction changes over time.

- Some households report implausibly low consumption. Exclude households reporting food consumption expenditures (if available) below what you think is a reasonable threshold. For the US, we will use the "Mac Donald's $1 menu rule". For Russia, where home-grown food is prevalent, it is probably safe not to throw away any observation based on low food expenditures. In other countries, use your own judgement.

- Inequality measures, especially those based on logarithms, are very sensitive to very low values. We propose the following strategy to deal with these outliers. Compute the bottom 0.25, 0.50 and 1 percentile of the distribution for all your measures of income and expenditures, and analyze how sensitive the variance of the log is to trimming the bottom x percentile (with x equal to 0.25, 0.50 and 1). Our experience shows that eliminating the bottom 0.25% already makes the measures much more stable. Once you converge on a trimming threshold (which could differ by variable), set to "missing" all the values below the threshold and compute your inequality statistics. We recommend *not* to drop the observation, i.e., the entire household - since that household may provide valid data for the other variables of interest.

5 Variable Definitions

5.1 Labor supply

- The analysis for labor supply and hourly wages should be conducted on individuals, not households. Therefore, from your original sample, select *all* males and females of working age (i.e., aged 25-60).

- Whenever possible, we suggest computing two measures of labor supply for both males and females. First, annual hours worked (l_m, l_f). Second, the fraction of the working-age population not working, working part-time and working full-time. Many surveys ask about "usual hours worked in a typical week". If such question is available, classify an individual as working part-time if her/his usual weekly hours worked are between 1 and 29 and working full time if her/his usual weekly hours are above or equal to 30 hours. If your data does not contain that information, but it contains a question about hours worked "last week", you could use that question. Alternatively, some surveys may directly ask the question whether the individual holds a "part-time or full-time job".

5.2 Net Wealth

- We suggest to compute two measures of household wealth: net financial wealth (a) and net total wealth (a^+). By net financial wealth we mean financial assets (e.g., checking/saving accounts, bonds, stocks, private pension funds, cash, unincorporated business holdings) net of liabilities (e.g., credit card debts, consumers loans). By net total wealth we mean net financial wealth plus the market value of all residential real estate owned (including the primary residence) minus the value of outstanding debt on mortgage and home equity lines.

- You may have noticed that we have excluded personally owned vehicles (cars, trucks, vans, motorcycles, boats, helicopters, planes, etc...) from the definition of wealth. The reason is that often it is very hard to recover market values for these items. In the second part of the

paper, we encourage you to add information on vehicles, should that information be available from your data set.

5.3 Wages, Earnings and Income

5.3.1 Wages and Earnings

- From your sample of all working-age individuals, compute individual wages (w_m, w_f) as average hourly earnings, i.e., the ratio of annual earnings to annual hours of work. Individual earnings should include wages and salaries (income from dependent labor) plus the labor part of business (i.e. self-employment) income. Do it separately for males and females.[1]

- When feasible, two definitions of household income should be considered:

  1. Pre-government household earnings as the sum of individual earnings before taxes $\left(y^L\right)$.
  2. Pre-government household earnings plus private transfers (alimony, child support, transfers from relatives, etc.) plus income from retirement plans, if present. This is our definition of pre-government household non-financial income $\left(y^{L+}\right)$.

5.3.2 Asset Income

- When possible, we suggest to compute (or, if available directly in the data, to report) two measures of asset income: net (as defined above for wealth) financial asset income $y^A = ra$ and net total asset income $y^{A+} = ra^+$.

- Define net financial income y^A as the sum of:

  (a) Dividends on stocks

  (b) Interests on bonds and bank accounts, net of interest paid on household financial debt (e.g. interest paid on credit card debt).

  (c) The asset part of business (i.e., self-employment) income (see the definition of labor income in Section 5.3.1).

- Define net total asset income y^{A+} as the sum of net financial income plus net rents from all owned real estate property, i.e., the imputed rent for the owned primary residence plus rental income from additional owned real estate (net of mortgage interest payments on that real estate).

- It may be interesting to study, for example, in the second part of the paper the correlation between asset income and labor income, because this statistic may help to empirically discriminate between models with different structures of financial markets.

- For Germany, the US, the UK, Canada and most recently Sweden, researchers at Cornell have spent a fair amount of time making labor and asset income definitions comparable for those countries.[2] See the attached table for details on how exactly each category is constructed. We think that for the five countries in question this exact classification could be used and

[1] Some datasets (like PSID) give the labor share of business income. If that is not directly available, use an external estimate of the labor share from National Accounts following the algorithm proposed by Cooley and Prescott in Frontiers of Business Cycle Research (1995, chapter 1, page 19). In any case, document what fraction you used.

[2] Documentation on CNEF is available here: http://www.human.cornell.edu/che/PAM/Research/Centers-Programs/German-Panel/cnef.cfm

for the remaining countries a similar classification should be adopted. Note that the CNEF data retain the individual/family ID's of the original data sets. Thus one can easily merge the original data set with the CNEF income variables record by record (at least for some of the years).

5.3.3 Disposable Income

- Compute pre-government household income (y) as the sum of household non-financial income $\left(y^{L+}\right)$ plus financial income $\left(y^{A}\right)$.
- Compute pre-government household income $\left(y^{L+}+y^{A}\right)$ plus public transfers (e.g., unemployment insurance, social security benefits, welfare payments, etc.) minus taxes paid by all members of the household. This is our definition of post-government (or disposable) household income $\left(y^{D}\right)$.
- You will note that as we move from w to y^{L} to y^{L+} to y to y^{D} we incorporate one by one the key insurance channels in the economy: labor supply, private transfers, financial markets, and government.

5.4 Consumption

- For those countries where household consumption data is available, we suggest that the analysis is conducted using two different consumption definitions:

  1. Non-durable expenditures (c). For data sets offering a breakdown into expenditure categories, such as CEX in the US or FES in the UK, this should ideally include:
     (a) Food, alcohol and tobacco;
     (b) Personal care items (i.e., personal hygiene items, etc.);
     (c) Fuels, utilities, and public services;
     (d) Household operations (i.e., maid and gardening services, but not furniture);
     (e) Public transportation;
     (f) Gasoline and motor oil;
     (g) Apparel;
     (h) Reading items;
     (i) Miscellaneous non-durable expenditures (i.e., attorney fees, etc.);
     (j) Entertainment (movie tickets, say, but not equipment, such as DVD players, etc.);
     (k) Lodging expenses (hotels etc.);
     (l) Vehicle expenses (maintenance and repairs);
     (m) Education expenditures;
     (n) Out of pocket health expenditures.
  2. Non-durable expenditure plus services from housing (rent paid for tenants, and imputed rent for homeowners) (c^{+}).

- For those countries where services from other consumer durables are available or can reliably be imputed feel free to use the second part of your paper to provide a more detailed analysis of these items.

6 Adjustments

- We should deflate income and consumption measures in a way that is consistent across data sets. We propose to deflate both income and consumption variables with the equivalent of the U.S. Consumers Price Index (CPI) for each country.

- Construct three definitions of consumption: raw, equivalized, and residual. Raw household consumption is the outcome of your calculations of Section (5.4), deflated by the price index. Equivalized consumption is raw household consumption equivalized by dividing it by the "OECD equivalence scale". This equivalence scale assigns a value of 1.0 to the first household member, a value of 0.7 to each additional adult and a value of 0.5 to each child (i.e. members 16 and younger). Thus, the scale for a single with no kids is normalized to 1.0, while real household consumption expenditures for a family of two adults and two children should be divided by 2.7 to be equivalized. Residual consumption is the residual of the regression (1), see below. We encourage you to verify the robustness of your calculations with respect to alternative ways to equivalize data in the second part of the paper.

- Construct three definitions of household earnings. raw, equivalized, and residual. Raw household earnings is the outcome of your calculations of Section (5.3), deflated by the price index described above. Equivalized household earnings is computed dividing raw earnings by the same equivalence scale we use for consumption. Residual earnings is the residual of equation (1), see below.

- For many questions it is the distribution of "offered wages", rather than of observed wages, that is of interest. While we suggest that in the first part of the paper you do the inequality analysis with the distribution of observed wages, in the second part of the paper you could investigate the biases induced by only observing a truncated wage distribution due to the selection problem (only accepted wages are observed), which may be especially severe for female wages. There are standard procedures in the labor literature to correct for this bias, see the Appendix for a detailed description and some references.

7 Statistics to be computed

7.1 Means

- Even though the objective of the study is to document the cross-sectional dispersion of variables, we should also report time series of the *means* for the key variables in levels: household earnings, income, consumption, and wealth, and individual hours and wages.

- We ask you to draw a comparison of the levels and trends of averages computed from micro data to the per-capita aggregates from National Accounts. It is well known, for example, that in the US, there are large discrepancies between the CEX and NIPA in the measurement of average real consumption expenditures; the same seems to be true for Germany. One contribution of the special issue will be to assess whether such discrepancies do also exist in other countries. For this exercise we suggest the following strategy:

  1. Keep every household in the sample, i.e. ignore all restrictions discussed above (i.e., age of the head, outliers, etc...).
  2. Compute average per capita earnings, per capita disposable income and per capita consumption from micro data and compare it with the closest possible definition of the same variable from National Accounts.

3. Document discrepancies both in levels and trends over time and try to explain the extent to which these are due to different definitions in micro data and National Accounts.

7.2 Premia

- For each year in your sample period, compute the following "premia" from the wage data described above.

  1. Education premium: the average wage of college educated males (or males with at least 16 years of education) divided by the average wage of non college-educated males (or workers with less than 16 years of education).
  2. Gender premium: the average wage of males divided by the average wage of females.
  3. Experience premium: the average wage of males aged 45-55 divided by the average wage of males 25-35 years old.

7.3 Cross-sectional dispersion measures

- The primary analyses should be conducted using logarithms of the variables. This poses some problems for variables that can take non-positive values (such as hours, earnings, wages, or wealth). This choice mechanically excludes from the analysis individuals (or households) reporting a non-positive value for the variable of interest, but it is nevertheless commonly used in labor economics and macroeconomics. We discussed above some of the potential sample selection problems associated with this choice.

- Our preferred measure of inequality is the variance of the logarithm of the variable because it allows for a cardinal interpretation of changes in inequality and it can be easily decomposed.

- Given the presence of outliers for some of the variables, we also ask you to also construct measures of inequality based on the levels of the variable In particular we recommend the following measures of dispersion:

  1. The 90th-50th and the 50th-10th percentile ratio (which have the added advantage of not being affected by top-coding issues and outliers).
  2. The Gini coefficient (which is commonly used in studies of wealth inequality and is also less affected by outliers at the bottom of the distribution).

  Note that these statistics should be computed on the same sample on which the variance of the logs is computed (i.e. the sample for which zeros are excluded and for which the trimming described in section 4.2 was done) so that the reader can assess separately the importance of a different inequality measure (as opposed to a different sample selection). This is the point of figure 4,8, and 11 described in section 9.

- For some variables (for example wealth or female labor supply) the presence of zeros or negative values is quite important. For describing inequality in those variables you might also want to use alternative measures such as the coefficient of variation or Gini coefficient (but now with the zeros included).

- If very good measures of income are available for the very top of the distribution (i.e., very big sample and no top-coding), then it would be interesting to report some measures of income or wealth concentration at the top of the distribution, say, the top 1% and compare these measures with what Piketty and Saez found for the US and France.

- We will focus on two dimensions of cross-sectional inequality (in earnings, wages, hours, asset income, consumption, and wealth): one is inequality over time (like the standard inequality measures reported by the CENSUS, for example). We ask you to compute the inequality indices for every year, and to report the longest time series possible. The other is inequality over the life-cycle, i.e. cross-sectional inequality by age group.

- For the time dimension, we suggest to compute the variance of the logs at three levels. First, on the raw data. Second, on the equivalized data. Third on the residual data. The residual data are defined as the *equivalized* data after taking out (through a simple regression analysis) the effects on the variable of interest (in logs) of year dummies, a polynomial in the age of the household head to control for compositional changes in the age distribution, dummies for sex/family composition (e.g. single male, single female, couple without children, couples with children, non-couple households, etc.) to control for changing household composition, and a control for the education level of husband and wife (e.g., college-college, college-high school, high-school-college, high school-high school). In the process of doing so it might be very helpful to assess the impact on overall inequality of each component. If the sample size allows it, the coefficient on the controls should be time varying, i.e. the regression should be performed separately year by year. More precisely, run the following regression (say, on household earnings y)

$$\ln y_{i,t} = D_t^y + \beta_{1,t} D_{i,t}^f + f_t\left(A_{i,t}\right) + \beta_{2,t} D_{i,t}^e + \beta_{3,t} D_{i,t}^r + \varepsilon_{i,t}^y \tag{1}$$

  where D_t^y is a year dummy, $D_{i,t}^f$ is a set of dummies for family composition, $f_t\left(A_{i,t}\right)$ is a polynomial in the household heads' age (say quartic), and $D_{i,t}^e$ is a set of dummies for educational attainment of the household, and $D_{i,t}^r$ is a set of race dummies. Note that all the regression coefficients are allowed to vary year by year. To measure the level and change in earnings inequality accounted for by, say education, it is enough to compute the cross-sectional variance of the $\beta_{2,t} D_{i,t}^e$ component, year by year.

- For household earnings and household consumption, we request you prepare two pictures that can be compared across countries. In one, plot the raw, equivalized, and residual variance of log of the variable, year by year (3 lines). In the second, plot year by year, the cross-sectional variance of each "observable" component of equation (1) (4 lines) where the dependent variable in the regression is equivalized. These two pictures are described in section 9 as figures 7 and 10.

7.4 Life cycle profiles

- For the life-cycle dimension, we have to take a stand on what to do about standard problem of lack of separate identification of time, cohort, and age effects. We have decided to remain agnostic and do the analysis in two ways.

  1. Let the typical cross-sectional moment for age group a at time t be $M\left(a,t\right)$ (e.g., the cross-sectional variance of log consumption for age group 25-35 in year 2000). Regress $M\left(a,t\right)$ on a full set of age group and year dummies. Compute the age profile from the predicted age-portion of the age-time regression.
  2. Let the typical cross-sectional moment for age group a in cohort k be $M\left(a,k\right)$ (e.g., the cross-sectional variance of log consumption for age group 25-35 of the cohort born in 1967). Regress $M\left(a,k\right)$ on a full set of age group and cohort dummies. Compute the age profile from the predicted age-portion of the age-cohort regression. See Heathcote, Storesletten and Violante (JEEA, 2005) for more details.

7.5 Cross-sectional co-movement measures

- We suggest to use the Pearson correlation coefficient to construct measures of cross-sectional co-movement among the variables of interest. Clearly, the correlation analysis is limited by the extent to which several variables are available in the *same* data set.

- We suggest to compute these correlations for every year that is feasible, and then to report the longest possible time series of the correlations.

7.6 List of moments to compute

7.6.1 Time-series

- For the longest possible time series available in your data set, compute the following cross-sectional variances (σ) of log variables:

$$\sigma\left(\log w_h\right), \sigma\left(\log w_s\right), \sigma\left(\log l_h\right), \sigma\left(\log l_s\right), \sigma\left(\log y^L\right), \sigma\left(\log y^{L+}\right),$$
$$\sigma\left(\log y^A\right), \sigma\left(\log y\right), \sigma\left(\log y^D\right), \sigma\left(\log c\right), \sigma\left(\log c^+\right), \sigma\left(\log a\right), \sigma\left(\log a^+\right)$$

  and do it both on the raw data and on the data after controlling for education, family composition, etc. as explained above, see equation (1).

- For comparison, for the same variables compute the other measures of inequality (Gini, CV, 90-10 and 50-10).

- If data allow it, exploit the panel dimension to compute variance of the changes in the log between $t-1$ and t for the variables of interest, for example $\sigma\left(\Delta \log y\right)$ where $\Delta \log y_{i,t} = \log y_{i,t} - \log y_{i,t-1}$.

- Compute as many cross-sectional correlations between variables of interest (and first-differences) as possible, and investigate whether there are interesting patterns when controlling for age, education, family composition, etc.

- Section 9 contains the minimal list of figures that you should report in your paper.

7.6.2 Life Cycle

- Plot figures 13 and 14, described in section 9.

7.7 Lognormality

- In the second part of the paper, you may want to run lognormality tests for the distribution of wages, income and consumption and report the results. These results are useful for informing inputs and evaluating outputs, for example, when simulating quantitative models (e.g., we know that using a lognormal distribution for the shocks is a good approximation of the data).

8 Estimation of wage dynamics

- When a panel dimension on individual wages and household earnings is available, we suggest to estimate a statistical model which is the sum of a permanent (unit root) and a transitory component. If the panel is long enough, we should allow the variances of the innovations to

the two components to be time varying, with loading factors denoted by λ_t. More specifically, for example for household earnings, we have:

$$\ln y_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\psi}_t + \alpha_{i,t} + \varepsilon_{i,t}, \text{ with } \varepsilon_{i,t} \sim N\left(0, \sigma^2_{\varepsilon,t}\right) \tag{2}$$

where the term $\mathbf{x}_{i,t}$ represents exactly the same vector of controls as in equation (1) (year, age, race, family composition, education dummies) with time-varying coefficients $\boldsymbol{\psi}_t$, and where

$$\alpha_{i,t} = \alpha_{i,t-1} + \eta_{i,t}, \text{ with } \eta_{i,t} \sim N\left(0, \sigma^2_{\eta,t}\right) \tag{3}$$

and similarly for head and spouse wages. Thus $\alpha_{i,t}$ is the permanent component and $\varepsilon_{i,t}$ the transitory component. Assume that $\eta_{i,t}$ and $\varepsilon_{i,t}$ are i.i.d. across agents, that $E\left(\varepsilon_{i,s}\eta_{i,t}\right) = 0$ for all s, t and that $E\left(\varepsilon_{i,s}\varepsilon_{i,t}\right) = E\left(\eta_{i,s}\eta_{i,t}\right) = 0$ for all $s \neq t$. Note that we allow both the variance of the transitory component and the variance of the permanent shock to be time-varying.[3]

- We recognize that this specification is restrictive in that we do not allow for the autocorrelation of the shocks $\alpha_{i,t}$ to be lower than unit root. The permanent-transitory model we propose has several advantages: 1) it fits the micro data very well, and 2) it is easy to estimate (see below). Moreover, it is well-known that given the short time dimension available in typical panel data sets, estimates of the autocorrelation coefficient in the $\alpha_{i,t}$ are downward biased. Thus estimating an autocorrelation coefficient of 0.95 in a short sample can well be consistent with the unit root model in the actual data-generating process. Having said this, we encourage you to pursue this issue further in the second part of the paper.

- The estimation of the permanent-transitory model we propose can be easily performed by minimum-distance estimation. In the Appendix we outline in detail the procedure to follow.

- In the second part you should try to experiment with more complex earnings dynamic process (not necessarily involving a unit root component). For example, some researchers find that a unit root process for the permanent component plus an MA(2) process for the transitory component fits the data better (Gottschalk and Moffitt, 1995)

9 List of figures

Here we outline a minimal list of plots that all papers should contain. By going over the slides from the conference we understand that for some countries it might be impossible to produces some of the figures we describe below due to strong data constraints. For those cases, we leave it to the authors to try to get as close as possible to the list. This is obviously only a minimal list and you should feel free to add plots or to add lines to a plot (for example if the same statistic can be computed on two data sets) if for a particular country you think it is instructive to do so. Also the ordering of the plots we present here is a proposal. If you feel that it is better for your paper to follow a different order feel free to do so. The data used to plot those figures (i.e. the summary statistics, not the raw data used to produce the statistics) should be made available on the web.

[3] With respect to the permanent component it is useful to generalize this model by allowing for an initial condition, i.e. the variance of $\alpha_{i,t}$ for an entrant cohort at date t. This parameter is very useful in OLG or "perpetual youth" models where we want household entering the model economy to draw their permanent component for productivity from a *non-degenerate* distribution. We call this variance σ^2_α and in the Appendix we explain how to estimate it.

- **Figure 1.** *Comparison with NIPA I.* Mean income per capita in the main survey you use for income inequality and income per capita from NIPA. Particular care should be given to make the two measures comparable: means from the micro survey should be per capita (as opposed to per household) and mean statistics in the micro survey should be computed including also households with heads younger than 25 and older than 60. Also attempt to choose a definition of income which is comparable (i.e. contains the same components) in NIPA and in the survey. For example in the US one income definition which is almost perfectly comparable in CPS and in NIPA is total wages and salaries. Instead the definition of total money income from CPS and personal income from NIPA are not comparable.

- **Figure 2.** *Comparison with NIPA II.* Mean consumption per capita in the main survey you use for consumption inequality and consumption per capita from NIPA. The same remarks from figure 1 apply here.

- **Figure 3.** *Comparison with NIPA III.* Employment to population ratio in the main survey you use for labor supply and employment population ratio in NIPA (by NIPA here we mean the official macro-employment data). For more on figures 1 through 3 see subsection 7.1.

- **Figure 4.** *Basic inequality in wages.* This picture should have the 4 panel format (i.e. four panels on the same figure). The x-axis on each panel should be time, the y-axis should be var-logs on the first panel, 90/50 on the second, 50/10 on the third and Gini on the fourth. The inequality measures in this picture should refer to wages for all men and women. See figure 6 for more data on men v/s women

- **Figure 5.** *Wage premia* (see section 7.2). This picture should have the 4 panel format. The x-axis on each panel should be time, the y-axis on the first panel should be the education premium, the y-axis on the second panel should be the gender premium, the y-axis on third panel should be the experience premium and the y-axis on the fourth panel should be the variance of the residual wage (as computed from regression 1).

- **Figure 6.** *Inequality in labour supply.* This picture should have the 4 panel format. The x-axis on each panel is time. The first panel should report the variance of log wages for men and women. The second panel should report the variance of log hours for men and women. The third panel should report the correlation between hours and wages for men and the fourth panel should report the correlation between hours and wages for women. Notice here that the individuals that are used to construct the statistics in this picture are the ones who report both positive hours and a wage which is above a minimum threshold (see section 4.2).

- **Figure 7.** *Earnings inequality and its decomposition.* This picture should have a 2 panel format. In each panel the x-axis is time. The first panel should contain the variance of log raw household pre-government earnings, the variance of log equivalized pre-government earnings and the variance of the residuals from equation (1). The second panel should contain the variance of each observable component (using the estimated coefficients) of equation 1 (see also the last bullet of section 7.3).

- **Figure 8.** *Basic inequality in equivalized earnings.* This picture should have the 4 panel format. The x-axis on each panel should be time, the y-axis should be var-logs on the first panel, 90/50 on the second, 50/10 on the third and the Gini on the fourth. The inequality measures should refer to equivalized household pre-government earnings. If pre-tax earnings are not available (as for example in the case of Italy) use after-tax earnings.

- **Figure 9.** *From wages to disposable income.* This picture should have only 1 panel. The x-axis is time and it should describe how inequality evolves as we move from the narrowest definition of resources available to a household (wage of the head) to the widest (disposable income). The maximum amount of lines would describe the time evolution of the variance of the logs for the following *equivalized* variables: wage (w) of the head, pre-government earnings of the head, pre-government earnings of the household (y^L), pre-government non-financial income (y^{L+}), pre-government household income (y), total disposable income of the household (y^D). We understand that for very few countries it will be possible to plot all these lines, but we expect that for all countries should be able to plot at least some of them.

- **Figure 10.** *Consumption inequality and its decomposition.* This picture should have a 2 panel format. In each panel the x-axis is time. The first panel should contain the variance of log raw nondurable consumption, the variance of log equivalized nondurable consumption and the variance of the residuals from equation (1). The second panel should contain the variance of each observable component (using the estimated coefficients) of equation 1 (see also the last bullet of section 7.3).

- **Figure 11.** *Basic inequality in equivalized non-durable consumption.* This picture should have the 4 panel format (i.e. four panels on the same figure, the x-axis on each panel should be time, the y-axis should be var-logs on the first panel, 90/50 on the second, 50/10 on the third and Gini on the fourth) and the inequality measures should refer to equivalized nondurable consumption.

- **Figure 12.** *From disposable income to consumption.* This picture should have the 4 panel format. The x-axis on each panel should be time. The first panel should contain variance of the logs of equivalized disposable income and the variance of the logs of nondurable consumption. The other three panels should report 90-50,90-10 and Gini for the same variables.

- **Figures 13.** *Inequality over the life-cycle I* (i.e. controlling for time effects, see section 7.4). This picture should have the 4 panel format. The x-axis on each panel should be age (of the various groups 25-30, 30-35...). The first panel should plot (against age) the variance of log wages, the second panel should plot the variance of some measure of log raw earnings, the third panel should plot the variance of log equivalized earnings (the same measure used in the previous panel) while the fourth panel should plot a measure of the variance of log equivalized consumption. Note that here we only focus on the variance of the logs. If you find that different inequality measures (inter-quintile ratios, Ginis) paint a different picture of inequality over the life-cycle please feel free to add to this picture.

- **Figures 14.** *Inequality over the life-cycle II* (i.e. controlling for cohort effects, see section 7.4). Same figure as before, with the only difference that age profiles for dispersion are now obtained controlling for cohort effects.

- **Figure 15.** *Wealth.* This picture should have the 4 panel format. The x-axis on each panel is time. The first panel should contain the total (across all households) net financial wealth over total (across all households) disposable income ratio, and the second panel should contain the Gini of equivalized net financial wealth (a). Panels 3 and 4 should report the same information for net total wealth (a^+). The sample for this picture should be the same sample used for computing earnings inequality.

- **Figure 15.** *Estimated coefficient of the stochastic wage and income process.* This picture should have the 2 panel format. The x-axis on each panel is time. The first panel should contain two lines with the estimates from (2) of both $\sigma^2_{\varepsilon,t}$ and $\sigma^2_{\eta,t}$ for the wage process for the

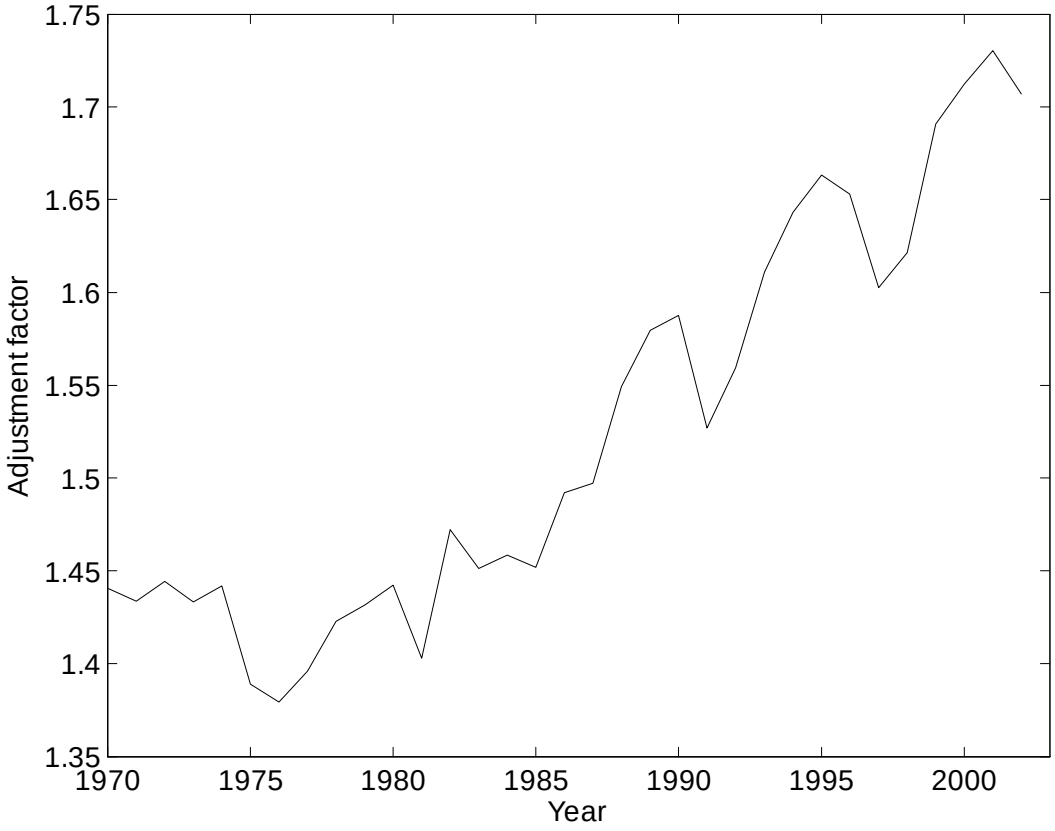

Figure 1: The adjustment factor $a_t/(a_t - 1)$ that topcoded earnings are multiplied by under the assumption that the top decile of the earnings distribution is Pareto distributed.

household head. The second panel should contain the same, but for the household earnings process.

A Appendix: Correcting for top-coding

Under the Pareto assumption, the cumulative distribution function is given by $F(y) = 1 - (b/y)^a$, where y is earnings, a is a shape parameter and b is a scale parameter. This implies that the ratio between the average income above y and y is equal to $a/(a-1)$. That is, top-coded earnings should be multiplied by the adjustment factor $a/(a-1)$. Hence, knowledge of the shape parameter a is sufficient for constructing an adjustment factor. One way of obtaining year-specific estimates of a_t is as follows. Let $\vartheta_t(y)$ denote the fraction of earners with income greater than y in year t, i.e., $\vartheta_t(y) = (b_t/y)^{a_t}$, and after taking logarithms, $\log \vartheta_t(y) = constant_t - a_t \log y$. Based on this relationship between the fraction of earners with income above y and y, compute OLS estimates of a_t. The data points $\{\vartheta_t(y_{i,t}), y_{i,t}\}$ are constructed from the earners in the top decile who are not top-coded. David Domeij has used this approach on CPS data between 1970 and 2002 (which is the same data used by Katz and Murphy, 1992, for the US) to illustrate this method. The R^2 values in these regressions were above 0.98. This figure shows the implied annual estimates of the adjustment factor, $a_t/(a_t - 1)$. The adjustment factor is approximately 1.45 until 1985 and then gradually increases towards 1.7. Please email David Domeij if you would like to use his program implementing this correction.

B Appendix: Estimation of the stochastic process for wage/earnings residuals

We first discuss the construction of empirical moments, then the construction of theoretical moments, then identification, estimation and, finally, inference.**Empirical moments:** Define the residual *in first differences* of our regression of $\ln y_{i,t}$ onto observable characteristics $\mathbf{x}_{i,t}$ as $g_{i,t} \equiv \Delta\left(\ln y_{i,t} - \mathbf{x}'_{i,t}\boldsymbol{\psi}\right)$. The full vector of interest for individual i is:

$$\mathbf{g}_i = \begin{pmatrix} g_{i,1} \\ g_{i,2} \\ ... \\ g_{i,T} \end{pmatrix}$$

where, for simplicity, we indicate with $t = 0$ the first year in the panel (i.e., the first observation on g is in period $t = 1$) and with T the last. If the individual was not interviewed in year t (i.e., if the panel is unbalanced) or if the observation is missing for that year, we replace the unobservable $g_{i,t}$ with a zero. Conformably with the vectors above, we define:

$$\mathbf{d}_i = \begin{pmatrix} d_{i,1} \\ d_{i,2} \\ ... \\ d_{i,T} \end{pmatrix}$$

where $d_{i,t} = 1\{g_{i,t} \text{ is not missing}\}$. Now we can derive:

$$\mathbf{m} = vech\left\{\left(\sum_{i=1}^{N} \mathbf{g}_i\mathbf{g}'_i\right) \oslash \left(\sum_{i=1}^{N} \mathbf{d}_i\mathbf{d}'_i\right)\right\} = \begin{pmatrix} \text{var}(g_1) \\ \text{cov}(g_1, g_2) \\ ... \\ \text{cov}(g_1, g_T) \\ ... \end{pmatrix}$$

where $\oslash$ denotes an element-wise division, and where we denote a row vector with the $'$ symbol. The vector $\mathbf{m}$ therefore contains the empirical estimates of $cov(g_t, g_{t+s})$, a total of $\frac{T(T+1)}{2}$ unique empirical moments. To obtain the empirical variance-covariance matrix of $\mathbf{m}$, define conformably with $\mathbf{m}$ the individual vector:

$$\mathbf{m}_i = vech\left\{\mathbf{g}_i\mathbf{g}'_i\right\}$$

The variance-covariance matrix of $\mathbf{m}$ that can be used for inference is:

$$\mathbf{V} = \left[\sum_{i=1}^{N} \left((\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})'\right) \circledast \left(\mathbf{D}_i\mathbf{D}'_i\right)\right] \oslash \left(\sum_{i=1}^{N} \left(\mathbf{D}_i\mathbf{D}'_i\right)\right)$$

where $D_i = vech\left\{\mathbf{d}_i\mathbf{d}'_i\right\}$ and $\circledast$ denotes an element-wise product. The square roots of the elements in the main diagonal of $\mathbf{V}$ provide the standard errors of the corresponding elements in $\mathbf{m}$.**Theoretical moments:** In equations (2) and (3), we have posited our statistical model for $g_{i,t}$ as the sum of a permanent plus transitory component. Based on this model, we can construct our theoretical moments (i.e. the model equivalent of $\mathbf{m}$) as a function of the models' parameters. Let $\boldsymbol{\Lambda}$ be the vector of parameters we are interested in (i.e., the year-specific variances of the permanent shock and the transitory shock) and let $f(\boldsymbol{\Lambda})$ be the vector of theoretical moments (i.e., the model

equivalents of the vector $\mathbf{m}$) which we index with a *. Under our statistical model, $f(\mathbf{\Lambda})$ is given by :

$$f(\mathbf{\Lambda}) = \begin{pmatrix} \text{var}^*(g_1) \\ \text{cov}^*(g_1, g_2) \\ \text{cov}^*(g_1, g_3) \\ ... \\ \text{cov}^*(g_1, g_T) \\ ... \\ \text{cov}^*(g_{T-1}, g_T) \\ \text{var}^*(g_T) \end{pmatrix} = \begin{pmatrix} \sigma^2_{\eta,1} + \sigma^2_{\varepsilon,1} + \sigma^2_{\varepsilon,0} \\ -\sigma^2_{\varepsilon,1} \\ 0 \\ ... \\ 0 \\ ... \\ -\sigma^2_{\varepsilon,T-1} \\ \sigma^2_{\eta,T} + \sigma^2_{\varepsilon,T} + \sigma^2_{\varepsilon,T-1} \end{pmatrix}$$

and more in general:

$$\begin{aligned} \text{cov}^*(g_t, g_{t+s}) &= \begin{cases} -\sigma^2_{\varepsilon,t+s} & \text{if } s = -1 \\ -\sigma^2_{\varepsilon,t} & \text{if } s = 1 \\ 0 & \text{otherwise} \end{cases} \\ \text{var}^*(g_t) &= \sigma^2_{\eta,t} + \sigma^2_{\varepsilon,t} + \sigma^2_{\varepsilon,t-1} \end{aligned}$$

Clearly, with an MA(2) process for the transitory component the covariance structure of earnings changes is different. For example, if

$$\varepsilon_{it} = v_{it} + \theta_1 v_{it-1} + \theta_2 v_{it-2}$$

then,

$$\begin{aligned} \text{var}^*(g_t) &= \sigma^2_{\eta,t} + \sigma^2_{v,t} + (1-\theta_1)^2 \sigma^2_{v,t-1} + (\theta_1 - \theta_2)^2 \sigma^2_{v,t-2} + \theta_2^2 \sigma^2_{v,t-3} \\ \text{cov}^*(g_t, g_{t-1}) &= -(1-\theta_1)\sigma^2_{v,t-1} + (1-\theta_1)(\theta_1 - \theta_2)\sigma^2_{v,t-2} + \theta_2(\theta_1 - \theta_2)\sigma^2_{v,t-3} \\ \text{cov}^*(g_t, g_{t-2}) &= -(\theta_1 - \theta_2)\sigma^2_{v,t-2} + \theta_2(1-\theta_1)\sigma^2_{v,t-3} \\ \text{cov}^*(g_t, g_{t-3}) &= -\theta_2 \sigma^2_{v,t-3} \end{aligned}$$

In what follows we describe identification in the simplest (MA(0)) case. **Identification:** One important point is that two restrictions (normalizations) must be imposed for identification purposes. First, $\sigma^2_{\varepsilon,0}$ is not separately identified from $\sigma^2_{\eta,1}$ because $\text{var}^*(g_1)$ has to pin down both parameters, while $\sigma_{\varepsilon,1}$ is being identified by $\text{cov}^*(g_1, g_2)$. Second, $\sigma^2_{\eta,T}$ is not separately identified from $\sigma^2_{\varepsilon,T}$ because in the last period we cannot assess from observing $\text{var}^*(g_T)$ whether a shock is transient or persistent . There are several normalizations possible to solve these identification problems. One common restriction (and the one we ask you to use) is to impose equality of the first two variances of transitory shock as well as equality of the last two, i.e. $\sigma^2_{\varepsilon,1} = \sigma^2_{\varepsilon,0}$ and $\sigma^2_{\varepsilon,T} = \sigma^2_{\varepsilon,T-1}$. Thus, our vector of parameters becomes $\mathbf{\Lambda} = \left\{\sigma^2_{\varepsilon,1}, \sigma^2_{\varepsilon,2}, ..., \sigma^2_{\varepsilon,T-1}; \sigma^2_{\eta,1}, .., \sigma^2_{\eta,T}\right\}$ with size $2T-1$.[4] **Estimation:** We solve the problem of estimating $\mathbf{\Lambda}$ by minimum distance, i.e. minimizing the distance between empirical and theoretical moments:

$$\min_{\mathbf{\Lambda}} (\mathbf{m} - f(\mathbf{\Lambda}))' \mathbf{A} (\mathbf{m} - f(\mathbf{\Lambda}))$$

[4] Note that the vector $\mathbf{\Lambda}$ does not include the parameter σ^2_α, i.e. the "initial" variance of the permanent component α. The reason is that the estimation in first-differences we have outlined cannot identify σ^2_α. However, there is a simple way of identifying and estimating σ^2_α which can be implemented after having done the estimation in first-differences. Consider the variance of residual wages or earnings (in *levels*) for the entrant cohort (age 25) at date t, which can be easily computed in the data. In our model, this moment equals to $\sigma^2_\alpha + \sigma^2_{\varepsilon,t}$. Since we know how to identify and estimate $\sigma^2_{\varepsilon,t}$, we can recover σ^2_α residually from this additional moment. One could also explore whether there are significant cohort effects in σ^2_α, i.e. if σ^2_α varies with t. We ask you to document the σ^2_α so estimated.

where $\mathbf{A}$ is a weighting matrix. There are several possible choices for the weighting matrix. The optimal minimum distance (OMD) imposes $\mathbf{A} = \mathbf{V}^{-1}$; the equally weighted minimum distance (EWMD) imposes $\mathbf{A} = \mathbf{I}$; and the diagonally-weighted minimum distance (DWMD) requires that $\mathbf{A}$ is a diagonal matrix with the elements in the main diagonal given by $diag\left(\mathbf{V}^{-1}\right)$. The last two have proved to perform better in small samples. We suggest to use the identity matrix, i.e. $\mathbf{A} = \mathbf{I}$. You may want to verify the robustness of your estimates to alternative assumptions. **Inference:**

For inference purposes we require the computation of standard errors of the estimated parameters. Chamberlain (1984) shows that the asymptotic standard errors can be obtained as:

$$\widehat{var\left(\widehat{\mathbf{\Lambda}}\right)} = \left(\mathbf{G}'\mathbf{A}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{A}\mathbf{V}\mathbf{A}\mathbf{G}\left(\mathbf{G}'\mathbf{A}\mathbf{G}\right)^{-1}$$

where $\mathbf{G} = \frac{\partial f(\mathbf{\Lambda})}{\partial \mathbf{\Lambda}}\Big|_{\mathbf{\Lambda}=\widehat{\mathbf{\Lambda}}}$ is the Jacobian matrix evaluated at the estimated parameters $\widehat{\mathbf{\Lambda}}$. The (square root of the) diagonal of this matrix yields the standard errors.

C Appendix: Offered Wages vs. Observed Wages

Wages are observed conditional on individuals working. If employment decisions are endogenous, we risk biasing the estimates of the variances of the underlying distribution of offered wages. That is, suppose that we are interested in estimating the variance of the log of offered wages, $var\left(\ln w_{i,t}\right) = \sigma^2$. Suppose that individuals work only if their offered wage is above a certain threshold (say, α_t - in search models, this would be people's reservation wage). Given that people who don't work do not provide a wage, estimating the variance of offered wages using only workers is going to give the impression that the distribution of offered wages is more compressed than it actually is (because the lower tail is being cut off). Formally, we would be estimating the variance of the log of offered wages using $var(\ln w_{i,t} | \ln w_{i,t} > \alpha_t) < var(\ln w_{i,t})$. If α_t moves over time, so does the bias. For example, suppose that the reservation wage increases over time (for example because of asset accumulation). Even if $var(\ln w_{i,t})$ is constant, we will have the impression that it is actually falling because $var(\ln w_{i,t} | \ln w_{i,t} > \alpha_t)$ declines if α_t increases (you keep chopping off parts of the lower tail of the offered wage distribution - people would look more and more alike). The procedure we outline below tries to obtain bias-free estimates of the variance component of wages in the transitory-permanent shock model outlined in section 8. Assume logged offered wages are given by the following process:

$$\ln w_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\psi} + \alpha_{i,t} + \varepsilon_{i,t} \tag{4}$$

where

$$\alpha_{i,t} = \alpha_{i,t-1} + \eta_{i,t} \tag{5}$$

(the variance of shocks $\varepsilon_{i,t}$ and $\eta_{i,t}$ can be time-varying as suggested above). To address the selection problem one could follow the approach of Low, Meghir and Pistaferri (2006), to which we refer for technical details. The idea is as follows. First, model the selection process into and out of employment, i.e., the decision to work. Then, construct sample selection terms and estimate wage growth equations conditioning on these terms. Finally, obtain the estimates of the variances of interest by modelling the first and second moments of unexplained wage growth. We simplify the problem by assuming normality of all error terms. Define the latent utility from labor market participation as $P^*_{i,t} = z'_{i,t}\varphi + \pi_{i,t}$. The associated labor market participation index is $P_{i,t} = \mathbf{1}\left\{P^*_{i,t} > 0\right\}$, which is unity for participants. We assume: $\left(\begin{array}{cc} \pi_{i,t} & \pi_{i,t-1} \end{array}\right)' \sim N\left(\mathbf{0}, \mathbf{I}\right)$.

Taking first differences of the wage equation (4), using the process for permanent shocks (5), we obtain:

$$\Delta \ln w_{it} = \Delta x'_{i,t}\psi + \eta_{i,t} + \Delta\varepsilon_{i,t}$$

Wage growth between $t-1$ and t is only observed for those who work in both periods. To achieve identification of the relevant parameters, make the following assumptions:

1. Denote $\sigma_\eta^2 = E\left(\eta_{i,t}^2\right)$ and $\sigma_\varepsilon^2 = E\left(\varepsilon_{i,t}^2\right)$ (for all i,t) the variances of the permanent productivity shock and transitory shock, respectively.
2. Denote $E\left(\eta_{i,t}\pi_{i,s}\right) = \sigma_\eta\rho_{\eta\pi}$ if $s=t$ and assume it to be zero otherwise.
3. Denote $E\left(\varepsilon_{i,t}\pi_{i,s}\right) = \sigma_\varepsilon\rho_{\varepsilon\pi}$ if $s=t$ and assume it to be zero otherwise.
4. Assume $E\left(\varepsilon_{i,t}\pi_{i,t}\right) = \sigma_\varepsilon\rho_{\varepsilon\pi}$ for all t.
5. Assume $E\left(\varepsilon_{i,s}\eta_{i,t}\right) = 0$ for all s,t conditional on π. Also assume $E\left(\varepsilon_{i,s}\varepsilon_{i,t}\right) = E\left(\eta_{i,s}\eta_{i,t}\right) = 0$ for all $s \neq t$ conditional on π.

Suppose now that we select only those who work at t and $t-1$ ($P_{i,t}=1, P_{i,t-1}=1$). It is easy to show that:

$$\begin{aligned} E\left(\Delta \ln w_{i,t} | P_{i,t}=1, P_{i,t-1}=1\right) &= \Delta x'_{i,t}\psi + E\left(\eta_{i,t} + \Delta\varepsilon_{i,t} | P_{i,t}=1, P_{i,t-1}=1\right) \\ &= \Delta x'_{i,t}\psi + G_{i,t} \end{aligned} \tag{6}$$

where $G_{i,t}$ is a "selection" term induced by labor market participation in both periods. In particular, one can show using the assumptions 1.-5. above that

$$G_{i,t} = \rho_{\eta\pi}\sigma_\eta\lambda_{P_t=1} + \rho_{\varepsilon\pi}\sigma_\varepsilon\left(\lambda_{P_t=1} - \lambda_{P_{t-1}=1}\right)$$

with $\lambda_{P_t=1} = \frac{\phi\left(z'_{i,t}\varphi\right)}{\Phi\left(z'_{i,t}\varphi\right)}$, $\lambda_{P_{t-1}=1} = \frac{\phi\left(z'_{i,t-1}\varphi\right)}{\Phi\left(z'_{i,t-1}\varphi\right)}$, and $\phi(\cdot)$ [$\Phi(\cdot)$] being the p.d.f. [c.d.f., respectively] of the standard normal distribution.[5] Controlling for $G_{i,t}$ in (6) implies that the estimate of ψ will be consistent. Note that if there was no selection ($\rho_{\eta\pi} = \rho_{\varepsilon\pi} = 0$, i.e., if people's working decisions were independent of the realizations of their stochastic wage components), G_{it} would be zero and we could obtain consistent estimates of ψ (and hence of the residuals, which we then use to obtain estimates of σ_η^2 and σ_ε^2) by using only the wage observations of those who work. Define at this point unexplained wage growth (observed only for participants in both periods):

$$g_{it} = \Delta\left(\ln w_{it} - x'_{it}\psi\right) = \eta_{it} + \Delta\varepsilon_{it} \tag{7}$$

>From the estimation of (6), and given consistency of ψ, the residual so constructed is a point-by-point consistent estimate of true unexplained wage growth. One can now use a method of moments procedure to identify the underlying stochastic process. The key parameters that we need to identify are the variance of the permanent shocks and the variance of the transitory shock. This is achieved by using the first and second moments of the residuals, as well as the first-order autocovariance. In the process not only the two variances of interest but also all the relevant correlations that drive selection can be estimated. The relevant formulae come from Tallis (1961).[6] Of course, the first moment is just G_{it} (because $G_{it} = E\left(g_{it} | P_{i,t}=1, P_{i,t-1}=1\right)$) The second moment is given by

$$E\left(g_{it}^2 | P_{it} = P_{it-1} = 1\right) = \sigma_\eta^2\left(1 - \rho_{\eta\pi}^2 z'_{it}\varphi\lambda_{P=1}\right) + \sigma_\varepsilon^2\left[2 - \rho_{\varepsilon\pi}^2\left(z'_{it}\varphi\lambda_{P=1} + z'_{it-1}\varphi\lambda_{P_{-1}=1}\right)\right]$$

[5] In estimation we do not use the restrictions on the parameters of interest imposed by (6). This only results in a loss of efficiency, but it does not affect consistency. We estimate the standard errors by the block bootstrap.

[6] See Tallis G.M. (1961), "The Moment Generating Function of the Truncated Multi-normal Distribution", Journal of the Royal Statistical Society. Series B (Methodological), 23(1), 223-229.

Note again that if there was no selection $\rho_{\eta\pi} = \rho_{\varepsilon\pi} = 0$, the second moment of the residual in first differences would be $E\left(g_{it}^2 \,|P_{it} = P_{it-1} = 1\right) = \sigma_\eta^2 + 2\sigma_\varepsilon^2 = E\left(g_{it}^2\right)$, and conditioning on participation would not matter. The first-order autocovariance is given by

$$E\left(g_{it}g_{it-1}\,|P_{it} = P_{it-1} = P_{it-2} = 1\right) = -\sigma_\varepsilon^2\left(1 - \rho_{\varepsilon\pi}^2 z'_{it-1}\varphi^g \lambda_{P_{-1}=1}\right)$$

Here is the algorithm in detail. a) Run a Probit regression for participation. This gives you estimates of φ, say $\widehat{\varphi}$. This allows you to construct consistent estimates of $\lambda_{P_t=1}$ and , $\lambda_{P_{t-1}=1}$, i.e. , $\widehat{\lambda}_{P_t=1} = \frac{\phi\left(z'_{i,t}\widehat{\varphi}\right)}{\Phi\left(z'_{i,t}\widehat{\varphi}\right)}$ and $\widehat{\lambda}_{P_{t-1}=1} = \frac{\phi\left(z'_{i,t-1}\widehat{\varphi}\right)}{\Phi\left(z'_{i,t-1}\widehat{\varphi}\right)}$. b) Run an OLS regression of $\Delta \ln w_{i,t}$ onto $\Delta x'_{i,t}$, $\widehat{\lambda}_{P_t=1}$ and $\widehat{\lambda}_{P_{t-1}=1}$ using only participants at both t and $t-1$ (for whom wage growth is observed). This gives you consistent estimates of ψ, say $\widehat{\psi}$. It also allows you to construct a consistent estimate of g_{it}, say $\widehat{g}_{it} = \Delta \ln w_{it} - \Delta x'_{i,t}\widehat{\psi}$, c) Estimate by NLS the system of three equations

$$\begin{aligned}
E\left(g_{it}|P_{i,t} = 1, P_{i,t-1} = 1\right) &= \rho_{\eta\pi}\sigma_\eta \lambda_{P_t=1} + \rho_{\varepsilon\pi}\sigma_\varepsilon\left(\lambda_{P_t=1} - \lambda_{P_{t-1}=1}\right) \\
E\left(g_{it}^2|P_{i,t} = 1, P_{i,t-1} = 1\right) &= \sigma_\eta^2\left(1 - \rho_{\eta\pi}^2 z'_{it}\varphi\lambda_{P=1}\right) + \sigma_\varepsilon^2\left[2 - \rho_{\varepsilon\pi}^2\left(z'_{it}\varphi\lambda_{P=1} + z'_{it-1}\varphi\lambda_{P_{-1}=1}\right)\right] \\
E\left(g_{it}g_{it-1}\,|P_{it} = P_{it-1} = P_{it-2} = 1\right) &= -\sigma_\varepsilon^2\left(1 - \rho_{\varepsilon\pi}^2 z'_{it-1}\varphi^g\lambda_{P_{-1}=1}\right)
\end{aligned}$$

imposing constraints across equations (this can be done in Stata "stacking" the equations). This would provide estimates of σ_η, σ_ε, $\rho_{\eta\pi}$ and $\rho_{\varepsilon\pi}$. Testing whether $\rho_{\eta\pi} = 0$ and/or $\rho_{\varepsilon\pi} = 0$ is implicitly a test for sample selection being important. Standard errors can be computed using the block-bootstrap procedure suggested by Horowitz (2002). In this way one can account for serial correlation of arbitrary form, heteroskedasticity, as well as for the fact that one is using a multi-step estimation procedure, pre-estimated residuals and selection terms. This procedure is likely conservative, since it allows for more serial correlation than that implied by the moment conditions used. Thus p-values are likely upward biased.

Table 1. Components of Income Categories

Income Category	United States	Germany	Great Britain	Canada
Private sources				
Labor income	Includes -wages and salaries -75% of positive farm income -75% of business income -reported earnings of self-employed	Includes -wages and salaries -reported earnings of self-employed	Includes -wages and salaries -reported earnings of self-employed	Includes -wages and salaries -net income of farm owners-operators -net income of owner-operators of unincorporated businesses
Husband	Labor earnings of the husband in the years before is death			
Survivor	Labor earnings of the widow			
Others[1]	Labor earnings of all other household members			
Private transfers	Income of the husband and wife from: -child support -help from relatives -other transfer income	Income from persons not in the household in the previous year	Income of all household members from: -education grants -sickness insurance -maintenance payments -foster allowance -payments from trade unions/friendly societies -non resident family members	Income of all household members from: -alimony and child support (including court-ordered) -other taxable transfer income
Retirement plans	Income of all household members from: -Veterans' pensions -other retirement income -employer pensions -annuity income	Income of all household members from: -Supplementary pensions for public sector employees (not civil servants) -Company pensions -all other pension income	Income of all household members from: -pensions from previous employer -pensions from spouse's ex-employer -private pension or annuity -widow or war widows pension -widowed mothers allowance	Income of all household members from: -employer pensions -annuities from Registered Retirement Savings Plans (RRSP) -withdrawals from Registered Retirement Income Funds (RRIF)

Table 1. Continued

Income Category	United States	Germany	Great Britain	Canada
Income from assets	The sum of income of the husband and wife's: -asset portion of farm income -asset portion of income from unincorporated business -asset portion of income from farming or market gardening -asset portion of income from roomers -rent, and income of all household members from: -dividends, interest, trust funds, and royalties	Household income from: -Dividends -Interest -Rent (minus operating and maintenance costs)	Income of all household members from: -Interest, dividends, annuities -Rent from boarders or lodgers -Rent from any other property	Income of all household members from: -Interest -net dividends -other investment income
Public sources				
Social Security	Income of all household members from: -Old-Age Insurance -Disability Insurance -Survivors Insurance	Income of all household members from the mandatory retirement insurance program (Gesetzliche Rentenversicherung) and related programs: -Old-Age pensions -Invalidity pensions -Miner pension -Farmer pension -War victim pension -Survivors pensions (widows and orphans) -Civil servant pensions -Worker accident pensions	Income of all household members from: -National Insurance retirement pension	Income of all household members from: -Old-Age Security -Guaranteed Income Supplement -Survivors Allowance -Spouse's Allowance -Canada/Quebec Pension Plan

Table 1. Continued

Income Category	United States	Germany	Great Britain	Canada
Other Cash Transfers	Income of all household members from: -Unemployment Insurance -Worker's Compensation -Aid to Families with Dependent Children (AFDC)/Temporary Assistance to Needy Families (TANF) -Supplemental Security Income (SSI) -Bonus value of Food Stamps -Other welfare income	Income of all household members from: -Unemployment Insurance -Unemployment relief -Student assistance -Maternity allowance -Subsistence allowance -Early retirement subsidy -Housing subsidy -Child allowance -Support for the care of sick family members -Nursing home allowance	Income of all household members from: -Severe disablement allowance -Industrial Injury allowance -Attendance allowance -Mobility allowance -Invalid care allowance -War disability pension -Disability living allowance -Disability working allowance -Incapacity benefit -Disability living allowance -Income support (IS) -Unemployment benefit (UB) -National Insurance sickness benefit (not employer's sick pay) -Child benefit -One parent benefit -Family credit -Maternity allowance -Housing benefit (rent rebate or rent allowance) -Council tax benefit (community charge benefit) -Other state benefit -Job Seekers Allowance -Educational grant -Foster allowance -Invalidity pension	Income of all household members from: -Canada Child Tax Benefit -Social Assistance -Employment Insurance -Worker's Compensation -Goods and Services Tax Credit -Provincial Tax Credits

Table 1. Continued

Income Category	United States	Germany	Great Britain	Canada
Taxes	Estimated total household taxes, including: -Social Security contributions (payroll taxes) -State taxes -Federal taxes	Estimated total household taxes, including: -Annual social security contributions -The sum of annual individual taxes for all household members -Annual solidarity surplus tax	Estimated total household taxes, including: -Income tax (local taxes not estimated) -National insurance contributions -pension contributions	Actual total household taxes, including: -Federal taxes -Provincial taxes
Net-of-Tax Household Income	Sum of all income components - taxes			

Sources: Disaggregated by the authors based on data from the Cross-National Equivalent File Codebook 1980-1998, Panel Study of Income Dynamics Users Manuals 1980-1997, German Socio-Economic Panel SOEPINFO 1984-1998, British Household Panel Survey User Manual Volumes A-H, Codebook prepared for Canadian Survey of Labour and Income Dynamics portion of Cross-National Equivalent File Codebook, 1998.