Stanford Center on Poverty and Inequality Inequality.com

Measuring Intergenerational Economic Mobility with Tax-Return Data

Towards an IRS Platform

David Grusky, Stanford University
Pablo Mitnik, Stanford University
Christopher Wimer, Stanford University

The Stanford Center on Poverty and Inequality is a program of the **Institute for Research in the Social Sciences (IRiSS)**. Support from the **Elfenworks Foundation** gratefully acknowledged.

This working paper series is partially supported by Grant Number AE00101 from the U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation (awarded by Substance Abuse Mental Health Service Administration).

Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation (awarded by Substance Abuse Mental Health Service Administration).

Measuring Intergenerational Economic Mobility with Tax-Return Data
Towards an IRS Platform

A Proposal to the Economic Mobility Project of the Pew Charitable Trusts

September, 2011

David B. Grusky
Pablo A. Mitnik
Christopher Wimer

Stanford Center for the Study of Poverty and Inequality

The United States purports to have an unusually strong commitment to equal opportunity, yet surprisingly it hasn't collected the mobility data needed to reliably monitor whether that commitment is being upheld. Although mobility and opportunity cannot of course be equated, it's widely understood that mobility data provide fundamental evidence on opportunity, which is why virtually all late industrial countries, save the U.S., have well-developed systems for monitoring mobility. It's not as if the U.S. is a more general laggard in developing social indicators. To the contrary, it has a well-developed program for measuring unemployment and related labor force statistics, a comprehensive set of indicators for monitoring educational outcomes, and an improving system for monitoring poverty and income inequality. The purpose of this project is to provide the first analyses of intergenerational mobility based on U.S. tax returns and to thereby assist in developing mobility indicators that are as reliable, sound, and comprehensive as those available in these other domains (see Grusky & Cumberworth, forthcoming, for an extended argument on the need for such indicators).

The main obstacle to developing a high-quality mobility monitoring system is the availability of appropriate data. The key problem with the PSID (Panel Study of Income Dynamics) and other surveys frequently used to study mobility is that the available samples are very small. As Björklund, Jäntti, & Solon (2007) concluded in a recent review, changes in intergenerational income mobility have been vexingly "difficult to estimate," mainly because of the small size of the PSID sample. Similarly, Lee & Solon (2009) recently noted that mobility estimates are "highly imprecise" in the U.S., again because the data are so sparse. This sample size problem has led to all manner of creative statistical fixes (e.g., Aaronson & Mazumder 2008, Hertz 2007, Lee & Solon 2009), but one wouldn't think that a country deeply committed to social mobility would have to rely so heavily on heroic fixes to monitor that commitment.[1]

There have been frequent calls to expand the size of the PSID sample. Although there are good reasons to enlarge the PSID, the strongest case for doing so isn't based on our need to build a better system for monitoring economic mobility. If we're truly committed to monitoring mobility, we would do far better exploiting the extraordinary resource that IRS tax returns represent. Because all tax filers have been required, starting in 1987, to report the Social Security numbers of dependent children, it is now possible to use tax records to go forward in time and identify the income, earnings, and occupations of dependents after they enter the labor force. Since 1997, the entire population of tax returns has been recorded in electronic form, with the important implication that the U.S. could in the not-so-distant future begin monitoring levels of economic mobility *with population data*.[2] We argue that now is the time to begin working toward bringing about that outcome.

[1] In our own analyses, we won't of course be able to eliminate all such statistical "fixes," indeed we'll be resorting to a fair number as well. But the tax-return approach that we are advocating will ultimately make a (nearly) model-free approach possible.

[2] It's merely a matter of waiting until the recorded dependents are old enough to be firmly attached to the labor force. In 2014, some of the dependents whose parents filed tax returns in 1997 will be 32 years old, and population-level analyses could therefore begin to be undertaken about two years later, when the corresponding data become available. These analyses could be expanded every year to cover successively older age groups (and, as an added

The purpose of our project is to lay the groundwork for this new tax-return approach to monitoring mobility by exploiting the so-called "OTA Panel," a panel of tax returns put together by the Department of the Treasury's Office of Tax Analysis, covering the period 1987-1996 (see Nunn et al. 2008). The first two authors of this proposal have recently signed an MOU (Memorandum of Understanding) with the IRS, covering up to March 2013, that allows us to conduct the first tax-return studies of intergenerational mobility in the U.S. The analyses will not just provide a possible template for future studies based on full-population IRS data but will also provide the best estimates to date of contemporary economic mobility in the U.S. among adults in their mid to late 30s. As explained in detail below, the proposed research will employ a data set much larger than those currently employed, which will allow us (a) to present new estimates of mobility that are more precise and reliable than those currently available, and (b) to carry out analyses of mobility and immobility at the top and bottom of the income and earnings distributions (which survey-based studies cannot, by virtue of sample size, well examine). Moreover, by using IRS data on pre-tax and after-tax income, we will able to provide the first analysis of the effects of the tax system on economic mobility.

The analyses that we propose will thus provide the richest evidence to date on the structure and patterning of intergenerational economic mobility in the U.S. This is reason enough to complete them. We are hopeful, however, that they will also demonstrate the potential of tax-return analyses and thus create strong interest in building a more permanent monitoring system based on such returns. If our collaboration with the IRS is successful, further MOUs may therefore be issued and allow us to expand the scope of our analyses.

What other types of analyses might be possible in the near future with tax data? There are at least three opportunities that are near-term feasible (but it should be stressed that we have no commitment, as yet, from the IRS to go beyond the currently-negotiated MOU):
Occupational mobility: The detailed occupations of both the filer and spouse are ascertained on Form 1040 and 1040A and may again be linked to those of their dependent children (and ultimately those children's spouses as well). If simultaneous analyses of economic and social mobility were carried out, we could monitor the total effects of economic *and* social background and establish the extent of cross-over effects between economic and social domains (i.e., the effects of economic resources on occupational outcomes and vice versa).
Wealth mobility: The wealth of filers may be imputed from information on dividends, mortgages, capital gains distributions, pensions and annuities, and (for the extremely wealthy) linked estate tax returns. Although monitoring wealth mobility has long been understood as

benefit, one could also use more years of income data for parents). It should be possible by 2036 to obtain estimates of lifetime intergenerational economic mobility in which statistical fixes play a modest role. Soon thereafter, the role of statistical modeling as a substitute for incomplete information on lifetime economic status would be reduced to a minimum, and a full-fledged tax-return monitoring system could be in place (assuming that tax data are integrated with SSA data on pensions). We further elaborate on this issue in the section titled "Why begin using IRS data now?" The main conclusion from this section is that, while early-career estimates could be available as early as 2014, it would only be possible to generate estimates approaching lifetime values by 2022.

critical to measuring opportunity, there are no good sources of data on such mobility. If IRS wealth measures were successfully developed, we could begin to study occupational, income, and wealth mobility simultaneously; and a truly comprehensive mobility accounting system would, for the first time, become possible.

Demographic effects: The structure of filing families can also be gleaned from changes in filing status. This is important because it's long been argued that one of the more important forces behind rising mobility is the emergence of one-parent and blended families (e.g., Biblarz, Raftery, & Bucur 1997). However, because of sample size problems, this hypothesis hasn't been tested in past studies of trends.

The upshot is that there's an important opportunity here for a fundamental breakthrough in monitoring mobility in the United States. Although it's not been exploited, the United States has administrative data in hand that are nearly as rich as the Nordic registers, long understood as the standard for mobility analysis. Insofar as the U.S. is serious about monitoring mobility, the best way forward is to carefully build on tax data in ways that neither compromise the charge of the IRS or, of course, the confidentiality of taxpayers.

Proposed analyses

We propose to carry out economic mobility analyses based on the sample of approximately 1.3 million tax returns in the 1987-1996 OTA Panel. The records for these returns, providing 10 years of annual earnings and family income measures, will be linked using Social Security identifiers to the recent tax records of adults who appear as dependent children in the OTA Panel. The result of this procedure will be a sample from the 1972-78 birth cohorts (i.e., 9-15 years old in 1987) who were 31-37 years old in 2009 and for whom parental and self-reported measures of income or earnings are available. Although it is not possible to know in advance the exact number of observations that will be secured via matching, in cooperation with SOI we've produced an estimate that suggests a matched sample of approximately 4,000 per year of linked data in 2004-2009, and about half that many, on average, in each of the years 1998-2003. This will provide a sample of contemporary mobility records that is substantially larger than any prior sample of contemporary data.[3] Appendix 1 describes in detail how we've estimated the number of expected matches as well as the new Intergenerational Economic Mobility (IGEM) data set that we'll be constructing. We propose to conduct two sets of analyses, as described below, that use our IGEM

[3] The main sample in Solon's (1992) pathbreaking study had only 348 father-son pairs. Although the PSID has grown in size since then, the effective sample sizes it provides to measure trends in intergenerational elastiticities are still dismayingly small. When Mayer & Loopo (2008) recently studied those born between 1956 and 1970 (and with positive income at age 30), the average sample size for each cohort was just 154 after pooling men and women together. Even Lee & Solon (2009), who used an econometric approach that made substantially more efficient use of data than the Mayer- Loopo approach, ended up with very small year-specific samples. On average, their sample included 510 observations per year for men and 575 observations per year for women, after pooling all ages available each year. We expect, by contrast, to have close to 4,000 observations per year for people aged 26-32 in 2004, 27-33 in 2005, 28-34 in 2006, and so forth (up to 2009). We will also have approximately 2,000 observations per year for people aged 26-31 in the period 1998-2003. See Appendix 1 for details.

data set to provide more precise and detailed estimates of mobility than has heretofore been possible.

Intergenerational elasticity analyses

The study of economic mobility has historically been based on estimates of intergenerational elasticities (IGEs) for men and women. In our first set of analyses, we will also estimate IGEs, but we will improve on past estimates in the following three ways:

More precise estimates: As we've noted above, the OTA panel will not only yield a comparatively large data set, but will also allow us to collect many years of earnings and income data for parents. There's also good reason to believe that tax data are less prone to response error than are survey data. It follows that tax-return analyses will produce more reliable and precise estimates of the IGE as well as reduce the known downward biases associated with using limited information on the lifetime economic status of parents.[4]

After-tax measures: Whereas previous research has only used pre-tax measures of economic standing, we will analyze both pre-tax and after-tax measures and thus provide the first-ever estimates of the impact of the tax system on economic mobility. There's been much speculation about the biasing effects of resorting to pre-tax measures, but it hasn't been possible until now to assess those biases.

Correcting bias in elasticities: In the existing literature, IGEs have been estimated by running OLS regressions of the logarithm of the child's family income (or earnings) on the logarithm of the family's income (and several lifecycle controls). It has recently been shown that the parameters of log-linearized models estimated by OLS may lead to biased estimates of true elasticities (Silva & Tenreyro 2006). We will address this issue by (a) conducting specification tests after estimating elasticities with the usual approach, and (b) reestimating elasticities using the PML (pseudo maximum likelihood) estimator most efficient for our data (either the Poisson estimator advocated by Silva & Tenreyro [2006], or the Gamma estimator discussed by Manning & Mullahy [2001]).[5]

The upshot is that we'll provide gold standard estimates of IGEs in the United States by solving problems of sample size, information quality, downward biases associated with measurement error, income type (i.e., pre-tax), and model misspecification. The foregoing may all seem to be mere technical problems, and indeed they are, but nonetheless they have absolutely stymied efforts to date to provide a convincing read on economic mobility. Appendix 2 describes in detail the models we plan to use to estimate IGEs.

[4] It's well known (e.g., Solon 1992) that a very substantial downward bias arises from estimating IGEs with one year, or even a few years, of information on the family of origin's income (see section 4 of Solon [1999] for a review). Recent research (e.g., Mazumder 2003) has shown that a large bias persists even when several years of information are used (see Appendix 2 for more details). By using the many years of information available in the OTA panel and in the population of tax returns, we will be able to address this issue more successfully than in the existing literature.

[5] Unlike the OLS estimation of log-linearized models, which may produce biased estimates, both PML estimators are always consistent. They may, however, differ in terms of efficiency.

Mobility table analyses

The study of occupational mobility has typically been prosecuted by analyzing mobility tables (MT), whereas the study of economic mobility has typically been prosecuted by estimating intergenerational elasticities.[6] This difference in approach, which is partly a disciplinary artifact, has prevented us from uncovering important features of mobility regimes. In the U.S. context, one reason why MT analyses have not been featured in studies of economic mobility is that they demand relatively large samples, which (as we've discussed) have not been available. We will be able to overcome that problem and will therefore couple our IGE analyses with MT analyses.

The two approaches are nicely complementary. While IGE analyses provide a useful scalar measure of average mobility and immobility, MT analyses provide more detailed descriptions of absolute and relative mobility patterns and can, in particular, uncover possible nonlinearities in those patterns. This is crucial given previous research suggesting that, both in the United States and abroad, the strongest forms of immobility occur at the top and bottom of the income distribution (e.g., Björklund, Jantti, and Solon 2007; Herz, 2005). Because high-income filers were oversampled in the OTA panel (and indeed the *full population* of extremely high-income filers is available), we will be able to establish whether immobility is, as many have argued, effectively assured at the very top of the distribution (e.g., annual income over $1M).

Our MT analyses will be based on gender-specific mobility tables derived from the same IGEM data set described above (and introduced in more detail in Appendix 1). In constructing these tables, we will minimize measurement-error biases by calculating cross-year averages of annual income and earnings (instead of using single-return measures).[7] We will then transform our (averaged) measures of economic standing (pre-tax and after-tax earnings and family income) for parents and children into discrete categories using absolute dollar thresholds and threshold based on quintiles. The mobility tables we analyze will be the result of cross-tabulating these discretized measures for parents and children.

We will describe patterns of absolute mobility with the usual descriptive statistics (e.g., total mobility, inflow and outflow rates) and patterns of relative mobility and association with log-linear and log-multiplicative models (e.g., Goodman 1979; Hout 1983; Clogg 1995). We will be especially focused on teasing out differences across economic classes in the propensities of inheritance by fitting models with inheritance terms that differ by income or earnings categories. The off-diagonal association will either be scaled or freely estimated using RC association models. If the category scales are freely estimated, we can then test whether the metric assumed under an IGE approach is misleading. The resulting analyses will therefore provide the first description of absolute and relative economic mobility in the United States that is (a) based on reliable and

[6] But note that, as early as 1979, Atkinson et al. (1979) employed mobility tables to study income mobility. For a recent debate on whether economic mobility has decreased in England, conducted in terms of MT analysis, see Blanden et al. (2004) and Erikson & Goldthorpe (2010).

[7] We have also begun experimenting with adjustments that "purge" effects of differences in life cycle position. We anticipate applying these adjustments but haven't yet definitively settled on doing so.

precise estimates, (b) affected as little as possible by measurement-error induced downward bias, and (c) conducted with both pre-tax and after-tax measures of income and earnings.

The after-tax analyses will likely prove especially revealing. Although our IGE analyses will help us assess the average effect of the tax system on economic mobility, they are uninformative insofar as the goal is to assess its effects at particular levels of the family income distribution. The tax system may, for example, have especially large effects at the bottom of the distribution (by virtue of the Earned Income Tax Credit). If one is interested in the consumption opportunities and life chances of workers and their families, it is critical not only to measure after-tax income but also to use methodological approaches that can go beyond the measurement of average effects across the income distribution. Our MT analyses will allow us to do exactly that.

Why begin using IRS data now?

As we've noted above, all tax returns have been recorded in electronic form since 1997, which means that in the future it will be possible to monitor income mobility for the *entire* U.S. population. The question that we take on here is whether we might be well-advised to forego any IRS analyses, in particular those proposed here, until data on the full population become available.

It's important to be clear about when credible mobility estimates will become available with full-population IRS data. If one insists on lifetime value estimates, the wait would in fact be a long one. We estimate that full-population analyses yielding lifetime values could begin in 2021 and that the mobility estimates themselves might then be secured by 2022. (As discussed in footnote 2, one could of course begin full-population analyses earlier, but only by settling for early-career estimates.) The wait for lifetime values is long because dependents who are 15 years old in 1997,[8] and thus appear on the first available electronic records for the population, wouldn't become 37 years old until 2019. The 2019 tax return data will, in turn, become available in 2021 and hence that's when mobility estimates approaching lifetime values could begin to be ascertained (see Appendix 2). This first set of analyses would pertain to a single birth cohort (i.e., the 1982 cohort), but data for younger birth cohorts would become available, one year at a time, as those cohorts aged up to 37 years old. Although one might be tempted to bring in more recent birth cohorts (e.g., the 1983 cohort), to do so would be problematic because it would increase the well-known

[8] If one decided to analyze children who were older than 15 in 1997, this would of course make it possible to begin the population-level analyses a few years earlier. We could, for example, start three years earlier insofar as we decided to analyze those who were 18 years old in 1997. The obvious problem with doing so would be that one cannot safely assume that 18 year-olds who are named as dependents in 1997 represent the full population of 18 year-olds in that year. That is, some of the 18 years-olds are already ensconced in the labor force and filing their own returns, and it's hardly plausible to assume that such early-leavers are a random draw from the full population of 18 year-olds. Although this problem would be reduced by analyzing 16 or 17 year-olds, the evidence on high-school dropout behavior in the U.S. suggests that only at age 15 are virtually *all* children claimed as dependents by their parents. We discuss in Appendix 2 additional reasons why parental measures that start when the children are 15 years old are preferable to measures that start when the children are older.

"lifecycle downward bias" arising whenever income or earnings is measured for workers under 40 years old (see Appendix 2).

The wait is therefore a rather long one. Is it too long? It's our view that we would indeed be ill-advised to wait until 2022 to secure tax-based mobility estimates. We review below the four main reasons behind this conclusion.

The policy costs of waiting. Over the course of the next 11 years, major policy decisions will have to be made on such matters as early childhood education, workforce training programs, college loans and financing, unemployment benefits, community college support, minimum wages, and tax policies, all of which directly impinge on income mobility. Because the U.S. is founded on a strong commitment to equal opportunity, we should make such decisions in light of the best possible information about how much mobility there is, who is most likely to experience it, and whether it's increasing or diminishing. The information that is now available is *not* the best possible and it's *not* nearly good enough for many policy purposes. The clearest route to securing information that is truly reliable and appropriate for policy-making purposes is by analyzing currently available IRS data. We believe that waiting another 11 years to produce highly-reliable income mobility estimates is almost as indefensible as proposing, for example, a 11-year hiatus in monitoring poverty, unemployment, or inequality.

Trend analysis. The analyses that we're proposing will establish a set of baseline mobility estimates that will then allow mobility scholars in 2022 to assess whether mobility is stable, increasing, or decreasing. Delaying any assessment of mobility until full population data are available is tantamount to deciding that the first data point for reliably monitoring trends in U.S. mobility will be 2019 and no sooner. Although one could compare the 2019 estimates with earlier estimates based on PSID data, doing so would not allow for comparisons based on after-tax measures, nor would they allow us to estimate trends in mobility at the upper end of the income distribution (because there are so few upper-end cases in the PSID). Worse yet, the differences in the method of data collection (i.e., survey vs. administrative) would complicate the comparison and, in particular, would raise the possibility that any seeming change might be an artifact of such method effects. The PSID estimates are also problematic because they are based on small samples and on just a few measurements of parental income (which downwardly biases the estimates). In some cases, one might be able to develop statistical fixes for these problems, but there's of course no statistical fix to be had that corrects for small samples. For these reasons, we suspect that future scholars would conclude that, for the purposes of studying trends, the data from the OTA Panel had best be exploited, just as we are proposing to exploit them *now*. This suggests that the only real decision of consequence is whether we should wait until 2022 to carry out the OTA analyses. The obvious advantage of doing so now rather than later is that they can then inform any relevant policy decisions in the years up to 2022.

Building an IRS infrastructure. The project that we're proposing to carry out will also assist in developing new methods and approaches for overcoming the problems endemic to analyzing administrative data. Although we've argued that administrative data are vastly superior to all alternatives, this is not to imply that there aren't a great many challenges and obstacles in using them. The main output of our analyses will of course be the mobility estimates themselves, but an important spinoff of our project will be new protocols for (a) dealing with nonfilers; (b) combining administrative data from other sources (e.g., W-2s) with data from income tax returns; (c) matching the income tax returns of parents and children; and (d) constructing matched data that can be used for a wide range of income mobility analyses. The research that we carry out with OTA panel data will thus put us well "ahead of the game" in terms of the data infrastructure needed to conduct future analyses with population data (when the latter becomes feasible in 2021).

Exploiting a known opportunity. The analyses proposed here are only possible because of the MOU that we've carefully negotiated with the SOI staff. Important though it is, the analysis of intergenerational mobility is obviously not a core SOI mission, and we rely therefore on the good will of SOI to undertake these analyses. For a host of reasons, we now have a window of opportunity to work with the SOI staff, but one cannot be assured that such opportunities will necessarily persist into the future. We think it's therefore crucial to exploit this opportunity now. If this relationship is carefully nurtured (and we're committed to doing so), this will not only make it possible to develop better estimates now but will also, by virtue of demonstrating the extraordinary value of tax-based analyses, make it more likely that IRS tax returns data can be used in the future.

The foregoing reasons make it clear, we think, that further delay comes at a very high price. We turn next to specifying in more detail the research products that we propose to deliver and the timeline for delivering them.

Research products and timeline

We're hopeful that our tax-return analyses will come to be viewed as the fundamental baseline estimates of contemporary mobility in the U.S. Because we've submitted our anticipated first paper to the NTA meeting on November 17-19, 2011 (per our MOU), we will need to complete much data analysis and writing in the early fall.[9] According to our arrangement with SOI, they will construct the IGEM data set in direct consultation with us, thus ensuring that it meets the requirements of the planned analyses. However, because of legal constraints, we cannot be provided with direct access to the micro-data. Instead, SOI will provide us with a "fake data set" (a data set with the same structure as the original but with fabricated data) that we will use to write our programs in SAS, and they will then run our programs for the IGE analyses in their own servers, per our request. This complicates the research process enormously, but is a cost we are willing to pay in order to carry out the project.

[9] This participation in the TNA conference is necessary because SOI researchers are only provided with research release time for projects that lead to presentations at tax-relevant conferences.

We anticipate receiving the fake data set by the end September, completing the IGE analyses for the NTA conference by the end of October, and finishing writing the NTA conference paper in November. We will then produce two academic papers out of the IGE analyses, one that focuses on family income and another that focuses on earnings. The task of writing these two papers, each of which will entail many additional empirical analyses beyond those carried out for the NTA meeting, will require another six months. We plan to allocate approximately three months per paper (and hence will finish the IGE analyses at the end of May, 2012).

We will turn thereafter to the MT analyses. Although we will still be using the IGEM data set for these analyses, there will be considerable work in writing the programs needed to generate the requisite tables. The latter will take approximately one month to complete (and hence will be finished at the end of June, 2012). The analyses of the resulting mobility tables, which won't require access to the microdata and can be estimated with our own computers, can then be completed by the end of July, 2012. We anticipate that the writeup of those analyses will be finished by October, 2012. As with the IGE analyses, we plan on two academic papers coming out of the MT analyses, one focusing on family income and the other on earnings.

The first PEW report, which will be finished in June 2012, will describe the results from our IGE analyses. Likewise, as soon as we finish our two MT papers, we plan to write a second report for PEW summarizing the MT results. We anticipate finishing the project by November 2012.

This is an aggressive timetable, but we're keen on moving quickly and demonstrating the payoff to the tax-return approach. If it proves to be as successful as we anticipate, we'd like to negotiate a second MOU with the IRS that can take on all or some of the supplementary analyses that we described above (i.e., occupational mobility analysis, wealth mobility analysis, family structure effects).

Timetable of key events

June – October, 2011: Creation of matched data set.
October – November, 2011: First round of IGE analyses are carried out.
November, 2011: Writing of NTA paper.
November 17-19, 2011: Presentation of paper at NTA meeting.
December, 2011 – March, 2012: Second round of IGE analyses are carried out.
December– May, 2012: Completion of two journal-quality papers from IGE analyses.
June, 2012: Production of PEW report summarizing results of IGE papers.
June, 2012: Generation of tables for MT analysis.
July, 2012: MT analyses are carried out.
August –October, 2012: Completion of two journal-quality papers from MT analyses.
November, 2012: Production of PEW report summarizing results of MT papers.

Appendix 1
Description and construction of the IGEM data set

This appendix describes the IGEM data set that we will use in our analyses, explains how it will be constructed, and provides additional information on various methodological issues not covered elsewhere.

Description of the IGEM data set

The IGEM data set, which SOI is assembling in direct consultation with us, will make it possible to study intergenerational mobility with the OTA panel. In its “long form” representation, the observations in the IGEM data set are person-years. The persons may be understood as “children” (in the parlance of intergenerational mobility analysis) and are representative of the 1972 to 1978 birth cohorts.

We summarize the structure of the IGEM data set in Figure 1.1. As shown there, the years included are 1998-2009, and each child first appears in the data set when she or he is 26 years old. The children in the 1972 cohort appear in 1998; the children in the 1973 cohort appear in 1999; the children in the 1974 cohort appear in 2000; and so forth up to the 1978 cohort (which appears in 2004). Each child is last observed in 2009. It’s shown in Figure 1.1 that the 1972 cohort was 37 years old in 2009; the 1973 cohort was 36 years old in 2009; the 1974 cohort was 35 years old in 2009; and so forth. Each cell in brown represents a subset of the child-years included in the dataset.

For each child-year in the data set, much information from the child’s tax return is available, including gender, age, pre-tax family income, after-tax family income, individual earnings, and various other items from Form 1040 for the corresponding tax year. As indicated in Figure 1.1, this information is taken from the full population of tax returns that, since 1997, has been electronically available.

In addition, for each child-year in the dataset, the following information from the tax returns of their parents is available: gender, age, pre-tax family income, after-tax family income, individual earnings, and various other items from Form 1040 *for each year between 1987 and the year in which the corresponding child became 29 years old*. Figure 1.2 offers a visual summary of this structure. Each cell in blue represents a child-year for which information about parents is included in the records (with the rows pertaining to the child’s cohort). This information is taken from the OTA panel for years 1987-1996 and from the full-population electronic tax returns for years 1997-2007.

Construction of the IGEM data set

The OTA panel, from which the IGEM will be constructed, was designed to represent the U.S. population in each year from 1987-1996 (see Cilke & Nunns 2008 for more details on the OTA panel). It combines data from three sources: the SOI Family Panel (or "cohort panel"), the refreshment panel, and the PSID panel. We will discuss each of these panels in turn.

SOI Family Panel. The SOI Family Panel, which is the foundation of the OTA panel, is based on a sample of 1987 tax returns. The panel includes all individuals who were listed on U.S. tax returns in 1987, either as taxpayers (including spouses on joint returns) or as dependents. The records for this 1987 "filing population" were then combined with additional information on the filers (and their dependents) that appeared in the 1988-1996 tax returns. If information from tax returns was not available, other administrative sources (e.g., Social Security earnings files) were used instead. The use of these additional administrative sources is important because it allowed the OTA to impute income and earnings, whenever tax return information was not available, using different but still highly reliable information.

The Refreshment Panel. But what about individuals who didn't file in 1987? In many cases, such individuals fell under the income threshold for filing, and hence they're drawn disproportionately from the lower end of the income distribution. We of course want that lower end of the distribution to be properly represented in our sample. The OTA used the "refreshment panel" to represent those in the 1987 non-filing population who appeared in a return in at least one year between 1988 and 1996. The procedures used to identify such nonfilers were well thought out and implemented, but we won't rehearse the details behind those procedures here (see Cilke & Nunns 2008, pp. 12-13). As was the case with the cohort panel, there were some people in the refreshment panel who (a) did not file in one or more years between 1987 and 1996, or (b) did indeed file in all years but, for any number or reasons (e.g., clerical error), one or more of the records was not located or recorded. The resulting missing data were nonetheless imputed by the OTA by combining information from other years in which a return was available with information from other administrative sources (e.g., Social Security earnings files). The most important point for our purposes is that the SOI Family Panel and the Refreshment Panel, taken together, are representative of everyone in the 1987 population save those who did not appear in any tax return between 1987 and 1996.

The PSID Cases. The only unrepresented persons, then, are those who never appeared in a tax return between 1987 and 1996. This residual category, referred to as the "permanent nonfiling population," is undoubtedly far smaller and less problematic than the corresponding residual category in surveys. The residual category in surveys is troubling because it includes not only (a) those who didn't appear in the population lists from which the sample was drawn, but also (b) the vast legion of individuals who did appear on the lists and were drawn into the sample yet could never be located by the interviewers (or were located but refused to participate in the survey or to answer the items of interest). Although the permanent nonfiling population is likely very small (both in absolute terms and relative to those who are unrepresented in surveys), the OTA did nonetheless use information from the PSID to attempt to represent these individuals. We haven't yet ascertained the size of this PSID supplement, but it will likely be ignorably small. If we are wrong and the permanent nonfiling population proves to be large (a result that will come out of our analyses of the OTA panel), we can attempt to use PSID mobility data to represent that population. But our strong prior in this regard is that the permanent nonfiling population will be small and that any attempt to represent it by combining PSID data with data from the OTA and refreshment panels is likely to cause more harm than good.

The main conclusion to be had is that our IGEM data set should come very close to representing all those who were between 9 and 15 years old in 1987 (i.e., the 1972-1978 birth cohorts). In our comparisons of (partial and preliminary) IGEM data against U.S. Census marginal distributions, the results have been encouraging and suggest that the IGEM data will indeed well represent the target population. We are still in the midst of carrying out further tests, evaluating the size of the permanent nonfiling population, and building our matching protocols. The process of building the final IGEM data set is nonetheless in place and will involve (a) identifying OTA panel members who were 9-15 years old in 1987, (b) identifying the parents of those panel members (by virtue of their listing of social security numbers for dependents), (c) locating the children's tax returns in 1998-2009 (via their social security numbers and using the electronic database of 1998-2009 returns), (d) using other sources of administrative data (e.g., Social Security earnings files) to generate measures of income and earnings for those years between 1998 and 2009 in which children did not file (or filed but their returns cannot be identified), (e) using the social security numbers of parents to locate their tax returns between 1987 and that year in which their dependent became 29 years old (using either the OTA panel for pre-1997 returns or the electronic database for 1998-2009 returns), and (f) applying other sources of administrative data (e.g., Social Security earnings files) whenever parents did not file (or filed but their returns cannot be identified).

Other methodological issues

We conclude Appendix 1 by briefly discussing (a) how we calculated expected sample sizes, (b) how we will define "parents," and (c) how we plan to address the complications of marital breakup and remarriage (during the period in which parental income is measured).

Calculation of sample sizes

We first outline how we calculated the expected sample sizes that have been presented throughout the proposal and are further discussed in Appendix 2. In making these calculations, we assumed that (a) each tax return represents one family, and (b) tax returns have approximately the same composition as families in the March supplement of the Current Population Survey (CPS). We then used the CPS to compute the number of families that had children aged 9-15 in 1987 (approximately 17 million) and the average number of children of these ages within these families (approximately 1.4 such children per family). Given a tax return sampling rate of approximately 1/5,000 for the OTA Panel (in 1987), we calculated the expected number of children in the OTA panel in 1987 as:

Number of children = 1.4 x (17,000,000 / 5,000) = 4,760

Finally, assuming a matching rate of 0.9 in year 2009 (in which children from all our birth cohorts are represented), the expected number of children for whom we will have 2009 information is 4,284.

We further assumed that the matched children will be uniformly distributed by age. If one then (conservatively) rounds down, the expected number of matched children of each age in 2009 will be 600. The latter estimate was assumed to apply to all years between 1998 to 2008. We have used this calculation, coupled with the assumption that there are an equal number of men and women, to produce the figures reported in the body of the text and in estimating the sample sizes available for the models of Appendix 2.

Parents, divorce, and remarriage

The analyses presented here will rest on a definition of "parents" as the primary taxpayer and spouse (when there is one). The parents so defined may or may not be the biological parents, but they are likely the social parents and, as such, are typically assumed to be more consequential. We have no choice but to follow the convention in this regard.

We can, however, do better than convention in representing the more complicated family effects that arise out of divorce and remarriage. This opportunity arises because, in the event of a post-1987 divorce, we will have access to parental income and earnings for the family that continues to claim the child as a dependent (i.e., the "resident family") as well as the family that no longer claims the child as a dependent (i.e., the "separated family"). We can therefore estimate models (a) that use the income of the resident family alone to measure parental income, (b) that take the average of the income of the resident and separated family as the measure of parental income, and (c) that control for the income of the separated family (when there is one).

Appendix 2
Models and Subsamples for IGE estimation

This appendix introduces some of the models and subsamples we plan to use to estimate IGEs. Although we don't attempt to be exhaustive here, we do list the core models upon which we'll be building our analyses. We will estimate IGEs pertaining to (a) the income of origin and destination families, (b) the income of origin families and the earnings of children, and (c) the earnings of fathers, mothers, and their children. However, for the purpose of keeping the presentation simple and tractable, we focus in this appendix on models pertaining to the income of origin and destination families. The models and subsample selection rules will be identical for the other cases. In all cases, estimation will be conducted separately for male and female children, thus (implicitly) allowing for a full set of gender interactions.

Methodological desiderata

Over the last several decades, there has been sustained and substantial methodological progress in estimating IGEs, with the contributions of Behrman & Taubman (1990), Solon (1992), and Zimmerman (1992) marking the beginning of what can be characterized as a methodological revolution (for more recent important contributions, see Baker & Solon 2003; Grawe 2003; Haider & Solon 2004; Hertz 2007; Lee & Solon 2009; Mazumder 2001, 2005; Solon 1999). As a result of this methodological work, researchers estimating IGEs today must bear in mind many different and often conflicting methodological desiderata, the most important of which are the following:

Precision. As with all estimates, we of course want IGE estimates to be as precise as possible, which translates into a preference, all else equal, for large samples and independent observations.

Robustness. The mobility researcher often attempts to exploit imperfect data (e.g., children's income measurements obtained when children are different ages) by deploying model-based corrections that rest on untestable assumptions, difficult-to-test assumptions, or known-to-be-false idealizations. We'd of course prefer to minimize the extent to which we rely on such assumptions.

Reduction of downward measurement-error bias. Before 1990, the convention in estimating IGEs was to resort to single-year measures of fathers' income or earnings, the latter being understood as acceptable proxies for lifetime income or earnings. By the early 1990s, however, it had become widely accepted that, due to transitory fluctuations in income and earnings, this approach involved classical errors-in-variables downward bias and greatly underestimated IGEs. To address this bias, it then became conventional to proxy lifetime income or earnings by averaging income or earnings data, such averages based typically on a few years. Unfortunately, because transitory shocks are highly serially correlated, an average based on only a few years of parental income data will reduce but not eliminate the bias. There is also evidence that IGE estimates are sensitive to the ages at which parental income is measured because the variance of transitory shocks decreases up to age 40 and then increases again. Additionally, because parental income changes as parents age, measuring it at different parental ages across children will downwardly bias estimates as well. Finally, our estimates will be downwardly biased if they are taken when the children are of different ages, as we know that family income is especially consequential at certain key points in children's lives. The foregoing sources of downward bias mean that researchers should (a) average parental income and earnings over as many years as possible; (b) measure parental income and earnings when parents are as close as possible to 40 years old as well as control for differences in parental age; and (c) measure parental income at the same point in the child's lifecycle (or, alternatively, control for differences in children's ages at the time of measurement).[10]

Reduction of life-cycle bias. Because age-earnings profiles not only have a well-known curvature but also are heterogeneous across family-of-origin income levels, neither the association between current and lifetime income nor measured IGE are constant over the lifecycle of children. As a result, using one-year measures of children's income as a proxy for lifetime income may generate biases unrelated to the classical errors-in-variables bias discussed above. Indeed, we now know that the bias is negative when income is measured early in children's careers, positive when it is measured late in their careers, and close to zero if it is measured when children are 40 years old. It follows that researchers should not only (a) measure children's income at the same age (or control for differences in children's age), but also (b) measure children's income as close as possible to when they are 40 years old.

Empirical strategies

The obvious problem with these methodological desiderata is that they cannot be satisfied simultaneously and hence different strategies will be better in some respects and worse in others (as regards methodological tradeoffs). How, then, to proceed? Our approach will be to estimate IGEs using several empirical strategies that differ in their strengths and weaknesses. Although it will be time consuming to do so, we won't be able to convince ourselves and others that our estimates are indeed gold-standard without such robustness checks.

The four strategies that we outline below will, however, share various common features, such as (a) measuring parental income at exactly the same children's ages; (b) controlling for differences in the age at which parental income is observed by including either a quartic on parental age (variant 1) or a quartic on father's age (variant 2), with age measured at the midpoint of the period in which parents' or father's income was observed;[11] and (c) controlling for the period in which income is measured in all models in which more than one year of data on children's income is used. We describe below each of the main empirical strategies that we plan to employ.

One-year one-age models

[10] The former strategy is the one commonly used in the literature.

[11] In variant 2, whenever the father is not present, the mother's age (or the age of the primary taxpayer) will be substituted for the father's age. For the sake of brevity, we assume in our writeup that variant 1 is employed.

Under this strategy, the income of children is measured at one common age, and as a result the children are in the sample for just one year. We plan to estimate three models of this type. As shown in Figure 1, Models A1 to A3 only differ by virtue of (a) the ages at which children's income is measured, and (b) the number of years of children data used. The cells in brown indicate the combination of years and children's ages in each sample. For instance, the sample for Model A1 includes children who were 37 years old in 2009, while the sample for Model A2 includes children who were 36 years old either in 2008 or 2009.

We will use between 9 and 15 years of income data for parents. The cells in darker blue in Figure 1 indicate the years for which parental income will definitely be employed in computing parental average income. As shown there, we will include income measures from the years in which children were between 15 and 23 years old, meaning that our parental measure will be based on *at least* 9 years of data. We may also use measures of income obtained when the children were yet older and thus employ up to 16 years of parental income data (by including measurements represented by the light blue cells in Figure 1). The decision of how many years of income data to use will depend on how much older (in terms of their age at the midpoint of the period covered) the parents in the sample become as we add additional years. Although on one hand we would like to include as many years of data as possible, on the other hand we want our measurements to be taken when parents are as close as possible to being 40 years old. Because we do not have income measures before 1987, there will necessarily be a tradeoff between these two goals, a tradeoff that is best resolved when the IGEM data set is ready for analysis and the specifics of the tradeoff can be assessed.

As indicated in the body of the text, we plan to estimate IGEs both by OLS and by using PML (pseudo maximum likelihood) estimators that, unlike the OLS estimator typically used in the literature, are always consistent. These PML estimators involve generalized linear models with a log link function and either a Gamma or Poisson distribution for the dependent variable. We will estimate Models A1 to A3 using the following three specifications:

<u>OLS</u>

$$\ln Y_{it} = \alpha + \rho \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 + \gamma_1 D_t^{2008} + \gamma_2 D_t^{2007} + \varepsilon_i$$

<u>Poisson PML</u>

$$Y_{it} \sim Poisson(\mu_{it})$$
$$\mu_{it} = E(Y_{it}) = \exp(\alpha + \rho \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 + \gamma_1 D_t^{2008} + \gamma_2 D_t^{2007})$$

<u>Gamma PML</u>

$$Y_{it} \sim Gamma(\mu_{it})$$
$$\mu_{it} = E(Y_{it}) = \exp(\alpha + \rho \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 + \gamma_1 D_t^{2008} + \gamma_2 D_t^{2007})$$

where Y_{it} is the family income of child *i* in year *t; X_i* is the average income of the parents for child *i*; ρ is the IGE of family income; A_i is the measure of the age of the parents for child *i*; and

D_t^{2008} and D_t^{2007} are year dummies, included as necessary.[12] In all cases, we will compute robust standard errors, and we will use population weights to estimate the models. The standard errors will also take into account that the sample is stratified.

The main comparative advantages of Models A1-A3 are the following. First, they are robust because they obviate the need to model the children's lifecycle, which always involves making assumptions that are difficult or impossible to test. Second, because each son or daughter is in the sample only once, the observations are independent (which favors precision).[13] Lastly, the income of the children is measured close to when the lifecycle bias is expected to vanish, especially so in the case of Model A1. The key disadvantage, however, is that the sample sizes are smaller than in any other approach we'll take. Based on the calculations discussed in Appendix 1, we expect samples of approximately 300, 600, and 900 children for Models A1, A2, and A3 respectively (such samples being available for men and women alike). It follows that precision may be an issue.

Three-year three-age model, without life cycle adjustment for children

The next model that we'll estimate is a natural extension of the previous ones. As shown in Figure 2, Model B measures the income of children when they are 33, 34, or 35 years old in 2007, 2008, or 2009. This approach implies that some children will appear in the sample two or even three times. We will then estimate IGEs using specifications identical to those employed to estimate Model A3. As before, we will compute robust standards errors, but in this case observations are not independent and thus our computations will have to take into account clustering. We will again use population weights in estimating this model.

Unlike Models A1-A3, which are based on sons or daughters of exactly the same age (either 35, 36, or 37 years old), here we have sons or daughters between the ages of 33 and 35, which means our estimates may be understood as the average elasticity over such ages. As with Model A3, the estimated elasticity is also a period average, an average pertaining to the years 2007 to 2009.

There are many advantages to Model B. Like Models A1-A3, it is comparatively robust, as no modeling of the children's lifecycle is required. But in this case, unlike that of Models A1 to A3, precision should not be an issue. Indeed, here we expect to have available samples of size 2,700 (in children-years), for both men and women. Although the need to compute cluster-corrected standard errors entails some loss of precision (relative to the case in which all observations are independent), in balance we expect a large increase in precision. The main disadvantage of this model, as compared to Models A1-A3, is that the children's income is measured further away from

[12] While D_t^{2008} and D_t^{2007} are both included in Model A3, only the former is included in Model A2. Neither is included in Model A1.

[13] It's possible, albeit unlikely, that the sample will include a few cases of children from the same family of origin (e.g., twins, children born 11 months apart). If this is the case, we can adjust standards errors to take into account such clustering, but the effect of doing so on the precision of our estimates should be negligible (in light of how rare such clustering will be).

age 40 than in any of those models, which means that lifecycle downward biases are likely to be somewhat larger.

Multi-year multi-age models with lifecycle adjustments for children and n-year average IGEs

In the "Class C" models, the income of children is measured at several different ages, and all sons or daughters appear in the sample several times. We plan to estimate three models of this general type that vary with respect to (a) sample size, (b) the number of years of parental income used to compute average income, and (c) the children's ages available to model the children's lifecycle. These differences are represented in Figure 3.[14]

The sample for Model C1 includes all children-years for children who were 9-15 years old in 1987 and for all years in which they were at least 26 years old. As in Models A1-A3 and Model B, we will use between 9 and 15 years of income data for parents. In Models C2 and C3, the samples include fewer children-years, but more years of parental income data can be used in computing average income. The sample for Model C3, for instance, includes children-years for children who were 9-11 years old in 1987 and at least 26 years old in 2002 and beyond. This approach allows us to extend the number of years of parental income data (i.e., 13-19 years of data), but of course at the cost of a smaller sample size and relatively early measurements of children's income.

The econometric specification for this model is very similar to that used by Lee & Solon (2009). The only important difference is that, instead of estimating year-specific IGEs, we will estimate the average IGE for the period covered in each sample. We will include period and parental age controls similar to those used in Model B, but now will also control for the child's lifecycle position. To do so, we will include a quartic on children's age, along with interactions between these four age variables and parental average income.[15] We will again estimate IGEs by OLS and with Gamma and Poisson PML estimators. This yields the following regression models:

<u>OLS</u>

$$\ln Y_{it} = \alpha' D_t + \rho \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 + \delta_1 Z_{it} + \delta_2 {Z_{it}}^2 + \delta_3 {Z_{it}}^3 + \delta_4 {Z_{it}}^4 + \theta_1 Z_{it} \ln X_i + \theta_2 {Z_{it}}^2 \ln X_i + \theta_3 {Z_{it}}^3 \ln X_i + \theta_4 {Z_{it}}^4 \ln X_i + \varepsilon_{ict}$$

<u>Poisson PML</u>

$$Y_{it} \sim Poisson(\mu_{it})$$

$$\mu_{it} = E(Y_{it}) = \exp\big(\alpha' D_t + \rho \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 + \delta_1 Z_{it} + \delta_2 {Z_{it}}^2 + \delta_3 {Z_{it}}^3 + \delta_4 {Z_{it}}^4 + \theta_1 Z_{it} \ln X_i + \theta_2 {Z_{it}}^2 \ln X_i + \theta_3 {Z_{it}}^3 \ln X_i + \theta_4 {Z_{it}}^4 \ln X_i\big)$$

[14] This figure doesn't identify the additional years of parental income data (i.e., when their children are older than 23) that we may use. This is because the cells for these additional years overlap with the cells identifying the children-years that will be included in each model.

[15] These interactions are needed because life-cycle trajectories are heterogeneous across family-of-origin income levels.

Gamma PML

$$Y_{it} \sim Gamma(\mu_{it})$$

$$\mu_{it} = E(Y_{it}) = \exp\big(\alpha' D_t + \ \rho \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 \ + \ \delta_1 Z_{it} + \ \delta_2 {Z_{it}}^2 + \delta_3 \, {Z_{it}}^3 + \delta_4 \, {Z_{it}}^4 + \theta_1 Z_{it} \ln X_i + \ \theta_2 {Z_{it}}^2 \ln X_i + \theta_3 \, {Z_{it}}^3 \ln X_i + \theta_4 \, {Z_{it}}^4 \ln X_i\big)$$

with $Y_{it}, X_i, \ \rho$, and A_i defined as before, D_t a vector containing year dummy variables for all years indicated in Figure 3, $Z_{it} = W_{it} - 37$ (or 35 or 33, depending on the model), and W_{it} the age of child *i* in year *t*.

By using a "displaced" measure of children's age (e.g., actual age – 37) instead of actual age, we avoid any difficulties in interpretation due to the presence of interactions between the age terms and parental income. Indeed, with age entered in the regressions in this way, the interaction terms vanish at age 37 (or 35 or 33), and ρ can be directly interpreted as the average IGE over the *n* years of children's income data included in the analysis (i.e., 12 in C1, 10 in C2, and 8 in C3), for sons or daughters 37 (or 35 or 33) years old.[16]

Why are Models C1-C3 so attractive? First, the sample sizes are in this case much larger than in any previous model, indeed we expect to have approximately 18,900, 12,000, and 6,300 children-years for Models C1, C2, and C3 respectively (for men and women alike). The cluster corrections that are required in such models will of course entail some loss of precision, but we again expect, in balance, a large increase in precision compared to previous models. Second, for Models C2 and C3, our measures of parental income will be based on more years of data than in any previous model. The number of averaged years for Model C1 will be the same as in previous models (i.e., 9-15 years), but for Model C2 the measure will be based on 11-17 averaged years and for Model C3 it will be based on 13-19 averaged years. Third, these models make it possible to estimate *n*-year average IGEs for children aged 33, 35 and 37 (depending on the model), or even at 40, although the latter would require engaging in out-of-sample prediction. Under Models A1-A3 and B, it was also possible to estimate IGEs for these ages, yet Models C1-C3 should produce more precise estimates (and, for Models C2 and C3, the errors-in-variables downward bias will be smaller). Of course, these advantages are not obtained without an important cost, in particular the need to assume that we're correctly modeling the children's lifecycle. The latter is of course the all-important comparative disadvantage of this class of models.

[16] Because we are modeling the children's lifecycle over a relatively short period (8-12 years), our approach is less vulnerable to the criticism (Hertz 2007) that income-age profiles tend to change over the long run (see Lee & Solon 2009). In his own alternative approach, Hertz estimated separate income-age profiles for each birth cohort between 1952 and 1965, but due to data limitations he was obliged to estimate just one income-age profile for the 1966 through 1975 cohorts.

Multi-year multi-age model with lifecycle adjustments for children and IGE trend over 1998-2009

The specification for this model is just like that of Model C1, but now a separate IGE is estimated for each year between 1998-2009. The econometric specification is identical in the OLS case to that used by Lee & Solon (2009). The resulting models are as follows:

<u>OLS</u>

$$\ln Y_{it} = \alpha' D_t + \rho_t \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 + \delta_1 Z_{it} + \delta_2 {Z_{it}}^2 + \delta_3 {Z_{it}}^3 + \delta_4 {Z_{it}}^4 + \theta_1 Z_{it} \ln X_i + \theta_2 {Z_{it}}^2 \ln X_i + \theta_3 {Z_{it}}^3 \ln X_i + \theta_4 {Z_{it}}^4 \ln X_i + \varepsilon_{ict}$$

<u>Poisson PML</u>

$$Y_{it} \sim Poisson(\mu_{it})$$

$$\mu_{it} = E(Y_{it}) = \exp\left(\alpha' D_t + \rho_t \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 + \delta_1 Z_{it} + \delta_2 {Z_{it}}^2 + \delta_3 {Z_{it}}^3 + \delta_4 {Z_{it}}^4 + \theta_1 Z_{it} \ln X_i + \theta_2 {Z_{it}}^2 \ln X_i + \theta_3 {Z_{it}}^3 \ln X_i + \theta_4 {Z_{it}}^4 \ln X_i\right)$$

<u>Gamma PML</u>

$$Y_{it} \sim Gamma(\mu_{it})$$

$$\mu_{it} = E(Y_{it}) = \exp\left(\alpha' D_t + \rho_t \ln X_i + \beta_1 A_i + \beta_2 A_i^2 + \beta_3 A_i^3 + \beta_4 A_i^4 + \delta_1 Z_{it} + \delta_2 {Z_{it}}^2 + \delta_3 {Z_{it}}^3 + \delta_4 {Z_{it}}^4 + \theta_1 Z_{it} \ln X_i + \theta_2 {Z_{it}}^2 \ln X_i + \theta_3 {Z_{it}}^3 \ln X_i + \theta_4 {Z_{it}}^4 \ln X_i\right)$$

The only difference between these models and those of Class C is that the IGE parameter now has a time-period index. Hence, ρ_t is the IGE in year *t,* for sons or daughters who are 37 years old. The comparative advantages and disadvantages of this model are similar to those of model C1, but with the exception that estimates are very likely to be less precise.

References

Aaronson, David and Bhashkar Mazumder. 2008. “Intergenerational Economic Mobility in the U.S., 1940 to 2000, ” *Journal of Human Resources* 43(1):139-172.

Atkinson, A. B., C. G. Trinder, and A. K. Maynard. 1979. “Evidence on Intergenerational Income Mobility in Britain,” *Economic Letters* 1:383-388.

Baker, Michael and Gary Solon. 2003. ““Earnings Dynamics and Inequality among Canadian Men, 1976–1992: Evidence from Longitudinal Tax Records,” Journal of Labor Economics 21(2):289–321.

Biblarz, Timothy J., Andrian E. Raftery, and Alexander Bucur. 1997. “Family Structure and Social Mobility,” *Social Forces* 75 (4): 1319-39.

Björklund, Anders and Marianne Sundström. 2002. "Parental Separation and Children’s Educational Attainment: A Siblings Approach," IZA Discussion Papers 643, Institute for the Study of Labor (IZA).

Björklund, Anders, Markus Jantti, and Gary Solon. 2007. ”Nature and Nurture in the Intergenerational Transmission of Socioeconomic Status: Evidence from Swedish Children and Their Biological and Rearing Parents.” NBER Working Paper 12985, National Bureau of Economic Research.

Blanden, J., A. Goodman, P. Gregg, and S. Machin. 2004. ”Changes in Intergenerational Mobility in Britain,” in M. Corak (ed.): *Generational Income Mobility in North America and Europe*. Cambridge: Cambridge University Press.

Cilke, James and James Nunns. 2008. ”Basic Data,” in James Nunns, Deena Ackerman, James Cilke, Julie-Anne Cronin, Janet Holtzblatt, Gillian Hunter Emily Lin, and Janet McCubbin: Treasury Panel Model for Tax Analysis, Office of Tax Analysis Technical Working Paper 3.

Clogg, Clifford. C. 1995. “Latent Class Models,” in Arminger, Gerhard, Clifford C. Clogg and Michael E. Sobel (eds.) Handbook of Statistical Modeling for the Social and Behavioral Sciences. Chapter 6.

Grawe, Nathan. 2003. ”Life Cycle Bias in the Estimation of Intergenerational Earnings Persistence,” Statistics Canada Analytical Studies Branch Research Paper 207.

Erikson, Robert and John Goldthorpe. 2010. ”Has Social Mobility in Britain Decreased? Reconciling Divergent Findings on Income and Class Mobility,” *The British Journal of Sociology* 61(2):211-230.

Goodman, L. A. 1979. ”Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories,” Journal of the American Statistical Association 74: 537-52.

Grusky, David and Erin Cumberworth. Forthcoming. “A National Protocol for Measuring Intergenerational Mobility?” National Academy of Science.

Haider, Steven and Gary Solon. 2006. Life-Cycle Variation in the Association between Current and Lifetime Earnings, NBER Working Paper 11943.

Hertz, Tom. 2005. ”Rags, Riches and Race: The Intergenerational Economic Mobility of Black and White Families in the United States,” in Samuel Bowles, Herbert Gintis and Melissa Osborne Groves (eds.): *Unequal Chances. Family Background and Economic Succcess*. New York: Princeton University Press.

Hertz, Tom. 2007. “Trends in the Intergenerational Elasticity of Family Income in the United States,” *Industrial Relations* 46:22-50.

Hout, Michael. 1983. *Mobility Tables. Newbury Park, CA: Sage.*

Lee, Chul-In and Gary Solon. 2009. “Trends in Intergenerational Income Mobility,” *Review of Economics and Statistics* 91:766-772.

Manning, W. and J. Mullahy. 2001. “Estimating Log models: To Transform or Not to Transform?” *Journal of Health Economics* 20:461–494.

Mayer, Susan and Leonard Lopoo. 2008. “Government Spending and Intergenerational Mobility,” *Journal of Public Economics* 92:139–158.

Mazumder, Bhashkar. 2001. “The Mis-measurement of Permanent Earnings: New Evidence from Social Security Earnings Data,” Federal Reserve Bank of Chicago Working Paper no. 2001-24.

Mazumder, Bhashkar. 2003. “Revised Estimates of Intergenerational Income Mobility in the United States,” Federal Reserve Bank of Chicago Working Paper 2003-16.

Mazumder, Bhashkar. 2005. Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data, *The Review of Economics and Statistics*, 87(2): 235–255.

Nunns, James, Deena Ackerman, James Cilke, Julie-Anne Cronin, Janet Holtzblatt, Gillian Hunter, Emily Lin, and Janet McCubbin. 2008. *Treasury Panel Model for Tax Analysis*, Office of Tax Analysis Technical Working Paper 3.

Solon, Gary. 1992. “Intergenerational Income Mobility in the United States,” *The American Economic Review* 82(3):393-408.

Solon, Gary. 1999. “Intergenerational Mobility in the Labor Market,” in Ashenfelter, Orley and David Card (eds.): *Handbook of Labor Economics*, Volume 3, Part 1. New York: Elsevier.

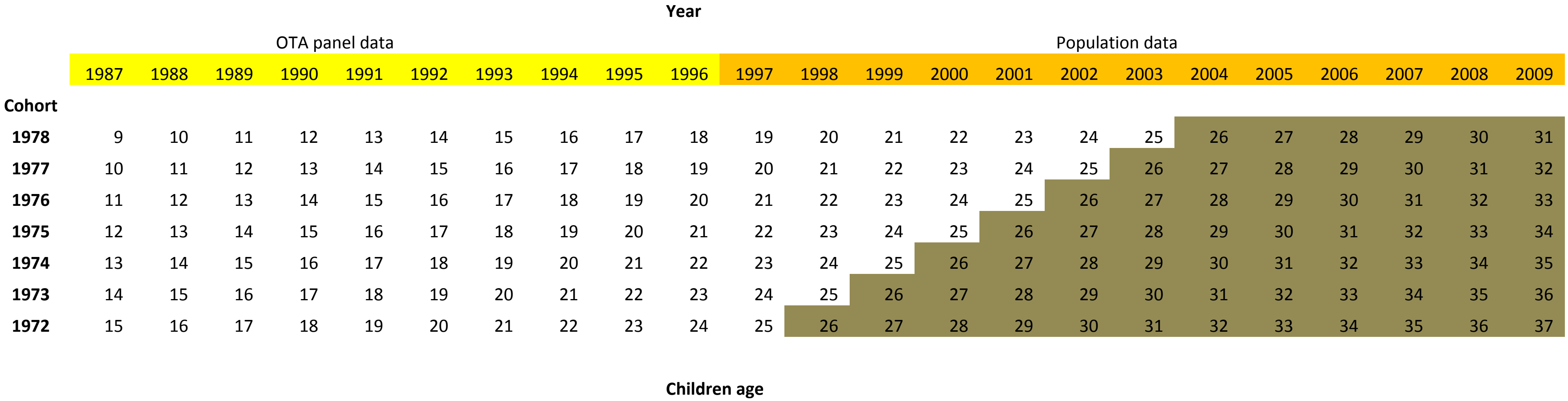
Solon, Gary S. 2002. “Cross-Country Differences in Intergenerational Earnings Mobility,” *Journal of Economic Perspectives* 16:59-66.

Solon, Gary S. 2008. “Intergenerational Income Mobility,” in David B. Grusky (ed.): *Social Stratification: Class, Race, and Gender in Sociological Perspective*, 3rdedition. Boulder: Westview Press.

Solon, Gary, Mary Corcoran, Roger Gordon, and Deborah Laren. 1991. “A Longitudinal Analysis of Sibling Correlations in Economic Status,” *Journal of Human Resources* 26 (Summer): 509-534.

Zimmerman, David. 1992. “Regression toward Mediocrity in Economic Stature,” American Economic Review 82: 409–429.

Figure 1.1: Observations in the IGEM dataset



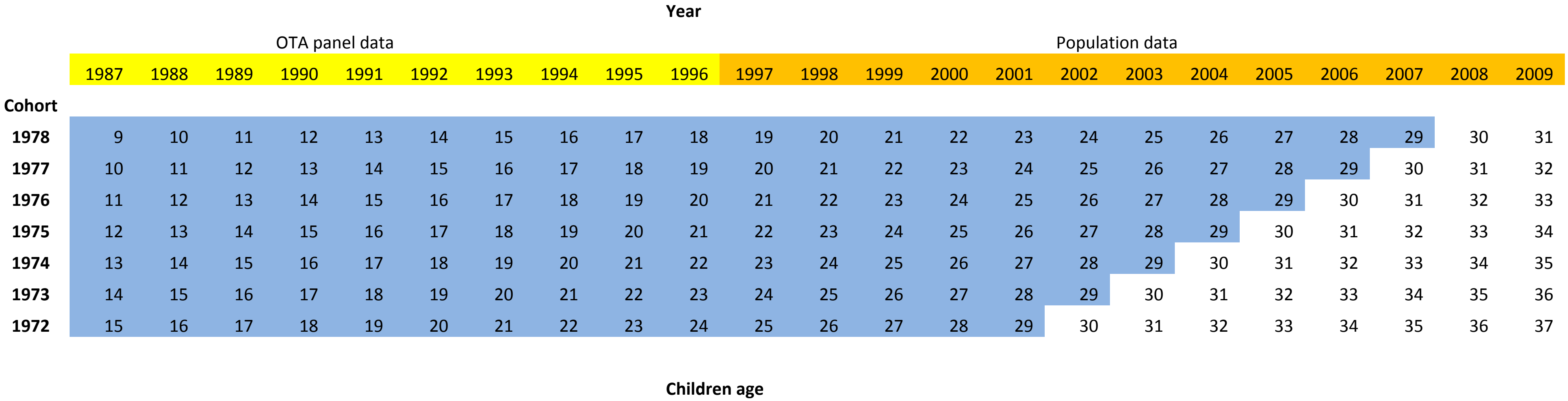
Year

	OTA panel data										Population data												
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Cohort																							
1978	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1977	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1976	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1975	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1974	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1973	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1972	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37

Children age

Each cell in brown represents a sub-set of children-years in the dataset. All together, the cells in brown represent all children-years in the dataset.

Figure 1.2: Parental information available by cohort



Year

	OTA panel data										Population data												
Cohort	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1978	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1977	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1976	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1975	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1974	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1973	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1972	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37

Children age

Each cell in blue represents a year-child age for which information about parents is included in the records for the children belonging to the corresponding cohort

Figure 2.1: One-year one-age models

Model A1: Children 37 year old in 2009, between 9 and 15 years of data for parents

Year

				OTA Panel data												Population data							
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Children	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Age	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37

Model A2: Children 36 year old in 2008 and 2009, between 9 and 15 years of data for parents

Year

				OTA Panel data												Population data							
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Children	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Age	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37

Model A3: Children 35 year old in 2007, 2008 and 2009, between 9 and 15 years of data for parents

Year

	OTA Panel data										Population data												
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Children	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Age	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37

Figure 2.2: Three-year three-age model without life cycle adjustment for children

Model B: Children 33-35 years old in 2007-2009, between 9 and 16 years of data for parents

Year

	OTA panel data										Population data												
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Children	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Age	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37

Figure 2.3: Multi-year multi-age models with sophisticated life-cycle adjustments for children, and n-year average IGEs

Model C1: Children 26-37 years old in 1998-2009, between 9 and 16 years of data for parents

Year

	OTA panel										Population data												
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Children	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Age	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37

Model C2: Children 26-35 years old in 2000-2009, between 11 and 17 years of data for parents

Year

	OTA panel										Population data												
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Children	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Age	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37

Model C3: Children 26-33 years old in 2002-2009, between 13 and 19 years of data for parents

	Year																						
	OTA panel										Population data												
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Children	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Age	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37