Two-Sample Two-Stage Least Squares (TSTSLS) estimates of earnings mobility: how consistent are they?

John Jerrim [1]

Álvaro Choi [2]

Rosa Simancas Rodríguez[3]

1 Institute of Education, University of London

2 Institut d'Economia de Barcelona, University of Barcelona

3 University of Extremadura

October 2014

Abstract

Academics and policymakers have shown great interest in cross-national comparisons of intergenerational earnings mobility. However, producing consistent and comparable estimates of earnings mobility is not a trivial task. In most countries researchers are unable to observe earnings information for two generations. They are thus forced to rely upon imputed data instead. This paper builds upon previous work by considering the consistency of the intergenerational correlation (ρ) as well as the elasticity (β), how this changes when using a range of different instrumental (imputer) variables, and highlighting an important but infrequently discussed measurement issue. Our key finding is that, while TSTSLS estimates of β and ρ are both likely to be inconsistent, the magnitude of this problem is much greater for the former than it is for the latter. We conclude by offering advice on estimating earnings mobility using this methodology.

Key Words: Earnings mobility, two sample two stage least squares.

JEL codes: I20, I21, I28.

Contact Details: John Jerrim (J.Jerrim@ioe.ac.uk) Department of Quantitative Social Science, Institute of Education, University of London, 20 Bedford Way London, WC1H 0AL

Acknowledgements: We would like to thank seminar participants at Nuffield College (University of Oxford), John Micklewright and Lindsey Macmillan for their helpful comments.

1. Introduction

Over the last twenty years, academics and policymakers have shown great interest in intergenerational mobility – the strength of the association between individuals' social origin and social destination. Economists have added much to this debate, particularly through their examinations of the link between the earnings (or incomes) of fathers and sons. However, due to data limitations, obtaining consistent estimates of earnings mobility remains a non-trivial task (Solon 1992; Black and Devereux 2011; Blanden 2013). The contribution of this paper is to present new evidence on the consistency of Two-Sample Two-Stage Least Squares (TSTSLS) estimates of earnings mobility; a methodology now widely applied in this literature (Appendix A reviews almost 30 papers where it has been used). Indeed, TSTSLS has proven to be the only way to estimate earnings mobility in a number of countries, including Australia, France, Italy, Spain, Switzerland, Japan, China and South Africa. Figure 1 illustrates the particularly prominent role it has therefore played in cross-national comparisons of earnings mobility; of the 20 countries included in Corak (2012), TSTSLS has been used in more than half (those with white bars).

<< Figure 1 >>

Yet, despite the important work of Björklund and Jäntti (1997) and Nicoletti and Ermisch (2008), more needs to be known about the consistency of TSTSLS estimates of earnings mobility. We therefore build upon the aforementioned authors' work by extending their framework from the intergenerational elasticity (β) to the intergenerational correlation (ρ), quantifying the inconsistency of TSTSLS estimates when using a range of different instrumental (imputer) variables, and considering a potentially important (yet little discussed) measurement issue.

The TSTSLS estimation procedure can be summarised as follows. Ideally, earnings mobility would be estimated via the following Ordinary Least Squares (OLS) regression model:

$$Y_{True} = \alpha + \beta . X_{True} + u \qquad (1)$$

Where:

Y_{True} = (Log) permanent earnings of sons

X_{True} = (Log) permanent earnings of fathers

Two different measures of earnings mobility would then typically be produced: the intergenerational earnings elasticity (β_{OLS}):

$$\beta_{OLS} = \frac{\sigma_{X,Y}}{\sigma_X{}^2} \quad (2)$$

Where:

$\sigma_{X,Y}$ = The covariance between father's and son's permanent earnings

$\sigma_X{}^2$= The variance of father's earnings

and the intergenerational correlation (ρ_{OLS}):

$$\rho_{OLS} = \frac{\sigma_{X,Y}}{\sigma_X{}^2} . \frac{\sigma_X}{\sigma_Y} = \frac{\sigma_{X,Y}}{\sigma_X . \sigma_Y} \quad (3)$$

Where:

σ_X= The standard deviation of father's earnings

σ_Y= The standard deviation of son's earnings

The measure of X_{True} preferred in the literature is a time-average of father's annual earnings across several years (X_{AVG})[1]. However, in many countries, earnings data cannot be linked across generations – i.e. there is no dataset where both father's and son's earnings can be observed. The TSTSLS approach attempts to overcome this problem via imputation – predictions of father's earnings are made based upon other observable characteristics (e.g. their occupation and education level). Equation 1 is then estimated using these predictions of father's earnings ($\hat{X}$) instead of a measure that has been directly observed (e.g. X_{AVG}). This is often described as an instrumental variable technique in the earnings mobility literature (e.g. Lefranc and Trannoy 2005; Nuñez and Miranda 2011), though it can alternatively be viewed as a cold-deck imputation procedure (Nicoletti and Ermisch 2008) or a 'generated regressor' approach (Murphy and Topel 1985; Wooldridge 2002:115; Inoue and Solon 2010).

[1] Although five consecutive years of father's earnings is often used (Solon 1992; Vogel 2008; Björklund and Chadwick 2003; Hussein et al 2008; Corak and Heisz 1999), more than ten may be needed if there is substantial auto-correlation in the transitory component of earnings over time (Björklund and Jäntti 2009; Mazumder 2005).

Solon (1992), Björklund and Jäntti (1997) and Nicoletti and Ermisch (2008) consider the properties of TSTSLS estimates of the intergenerational elasticity (β_{TSTSLS}). They show that consistent estimates can be obtained if either:

- The instrumental (imputer) variables have no direct effect upon son's earnings
- The R^2 of the equation used to predict father's earnings equals one

Yet, as father's education and occupation are the instruments (imputer variables) usually available, it is widely recognised that neither of these conditions hold. (Father's education and social class are likely to independently influence offspring's earnings, while also not being perfect predictors of father's permanent earnings). It is thus often stated that β_{TSTSLS} will be upward inconsistent as a result[2].

The key issue thus becomes the *magnitude* of this upward inconsistency. It is small enough to be safely ignored, or is it so large that TSTSLS estimates of earnings mobility become problematic? Likewise, if more detail is added to the model predicting father's earnings, does this significantly reduce the upward inconsistency? Unfortunately, little is currently known about these important issues. Indeed, the only study to quantify the inconsistency of β_{TSTSLS} is Björklund and Jäntti (1997). For one particular imputation model, containing a specific set of predictor variables, they find upward inconsistency of around 30 percent.

We contribute to this evidence base in multiple ways. First, the framework of Björklund and Jäntti (1997) and Nicoletti and Ermisch (2008) is extended from the intergenerational elasticity to the intergenerational correlation (ρ_{TSTSLS}). We use this to explain why ρ_{TSTSLS} is *downward* inconsistent in our empirical analysis (i.e. in the opposite direction of the inconsistency of β_{TSTSLS}). Second, new evidence is provided on the inconsistency of β_{TSTSLS} and ρ_{TSTSLS} using a range of different imputer variables, and thus the extent to which this problem can be reduced through use of a more detailed first-stage prediction model. Third, we divide β_{TSTSLS} and ρ_{TSTSLS} into components to demonstrate what is driving their inconsistency, and show how this changes when different prediction models are specified. It is also hoped that this will resolve some confusion in the applied literature, where it is often stated that the goal is to 'choose the instruments in order for the R^2 of the [father's earnings prediction]

[2] The following section will present a framework which illustrates why this is the case.

regression to be as high as possible' (Cervini-Pla 2012:9)[3]. Finally, we note how most studies make predictions of father's current earnings (X_{single}), whereas permanent earnings (X_{True}) is the actual unobserved variable of interest. We argue that, in this situation, more general expressions for the inconsistency of β_{TSTSLS} and ρ_{TSTSLS} are needed. Our empirical analysis then illustrates how conventional wisdom (e.g. β_{TSTSLS} always being upward inconsistent) no longer holds.

The paper now proceeds as follows. Properties of TSTSLS earnings mobility estimates are reviewed in section 2. This is followed by an overview of the Panel Survey of Income Dynamics (PSID) dataset and our empirical methodology in section 3. Results are presented in section 4, and conclusions in section 5.

2. TSTSLS estimates of earnings mobility

Our starting point is the framework of Nicoletti and Ermisch (2008). As noted in the introduction, the model of interest is:

$$Y_{True} = \beta . X_{True} + \mu \quad (4)$$

Where:

Y_{True} = Log son's permanent earnings

X_{True} = Log father's permanent earnings

X_{True} is unobserved in the 'main' dataset, but it does contain additional characteristics (Z), such as father's education and occupation, likely to be associated with X_{True}.

Now say a second 'auxiliary' sample (i) contains a measure of respondents' permanent earnings[4] (ii) is drawn from the same population and (iii) contains the same Z variables. The following OLS regression model can be estimated:

$$X_{True} = \delta . Z + v \quad (5)$$

[3] In our empirical analysis we show that adding variables to increase the first-stage R^2 can actually increase the inconsistency of β_{TSTSLS} and ρ_{TSTSLS}.

[4] Time-average earnings would be the preferred measure within the auxiliary dataset. Unfortunately, this is rarely available, and so current earnings are often used as the 'first-stage' dependent variable instead. We illustrate how this influences TSTSLS estimates in section 4.

Where:

Z = The instrumental (imputer) variables

And then used to predict log permanent father's earnings:

$$\hat{X} = \hat{\delta}.Z \tag{6}$$

Where:

$\hat{X}$ = Predicted log father's permanent earnings

$\hat{\delta}$= Estimated regression coefficients from the first-stage prediction model

Hence (7) can now be estimated rather than (1):

$$Y_{True} = \beta.\hat{X} + u \tag{7}$$

Estimates of β_{TSTSLS} and ρ_{TSTSLS} then follow from equations 2 and 3 (substituting $\hat{X}$ for X).

The two most commonly used Z variables are father's education and occupation (see Appendix A). However, both are likely to directly influence son's earnings (i.e. they are likely to be endogenous)[5]. Consequently, son's log earnings will actually be given by:

$$Y_{True} = \lambda_1.X_{True} + \lambda_2.\hat{X} + u \tag{8}$$

With λ_1 being the direct impact of the father's *actual* permanent earnings on son's earnings and λ_2 the effect of father's *predicted* earnings on son's earnings. (From this point forward, we drop the 'True' subscript for notational convenience). Solon (1992) and Björklund and Jäntti (1997) show that β_{TSTSLS} thus converges in probability to:

$$plim\ \beta_{TSTSLS} = \lambda_1 + \lambda_2.\frac{\sigma_{\hat{X}}}{\eta.\sigma_X}$$

$$= \beta + \lambda_2.\sigma_{\hat{X}}.\frac{(1-\eta^2)}{\eta.\sigma_X} \tag{9}$$

Where:

[5] One way to think about this is that father's education and social class influences their children's labour market outcomes, over and above the impact the greater earnings that highly educated, professional father's generate.

$\sigma_{\hat{X}}$ = The standard deviation of father's *predicted* earnings

σ_X= The standard deviation of father's *actual* long-run earnings

With:

$$\eta = \frac{\sigma_{\hat{X},X}}{\sigma_{\hat{X}}.\sigma_X}$$

Where:

$\sigma_{\hat{X},X}$= The covariance between predicted and actual log father's earnings

Under the assumption that the covariance between predicted and actual log father's earnings is equal to the covariance between predicted father's earnings and itself:

$$\sigma_{\hat{X},X} = \sigma_{\hat{X},\hat{X}} \tag{10}$$

η becomes[6]:

$$\eta = \frac{\sigma_{\hat{X},X}}{\sigma_{\hat{X}}.\sigma_X} = \frac{\sigma_{\hat{X},\hat{X}}}{\sigma_{\hat{X}}.\sigma_X} = \frac{{\sigma_{\hat{X}}}^2}{\sigma_{\hat{X}}.\sigma_X} = \frac{\sigma_{\hat{X}}}{\sigma_X} = R \tag{11}$$

Where:

R = The square root of the variance explained (R^2) in the first-stage prediction model (i.e. of equation 5).

The probability limit of β_{TSTSLS} then becomes:

$$= \beta + \lambda_2.\sigma_{\hat{X}}.\frac{(1-\frac{{\sigma_{\hat{X}}}^2}{{\sigma_X}^2})}{\frac{\sigma_{\hat{X}}}{\sigma_X}.\sigma_X}$$

$$= \beta + \lambda_2.\sigma_{\hat{X}}.\frac{(1-\frac{{\sigma_{\hat{X}}}^2}{{\sigma_X}^2})}{\sigma_{\hat{X}}.}$$

$$= \beta + \lambda_2.(1 - \frac{{\sigma_{\hat{X}}}^2}{{\sigma_X}^2})$$

[6] The covariance between a variable and itself is equal to the variance of that variable. Hence $\sigma_{\hat{X},\hat{X}}$becomes ${\sigma_{\hat{X}}}^2$.

$$= \beta + \lambda_2.(1 - R^2) \quad (12)$$

With the inconsistency of β_{TSTSLS} therefore:

$$\lambda_2.(1 - R^2) \quad (13)$$

There are a number of important points to note about (11), (12) and (13). First, as $0 \leq R^2 \leq 1$, the variance of father's predicted earnings must be less than or equal to the variance of actual father's earnings:

$$0 < \sigma_{\hat{X}}{}^2 \leq \sigma_X{}^2$$

Second, if the variance of father's predicted earnings ($\sigma_{\hat{X}}{}^2$) were equal to the variance of father's actual earnings ($\sigma_X{}^2$), then R^2=1 and the inconsistency of β_{TSTSLS} reduces to zero. Hence, in this framework, the inconsistency of β_{TSTSLS} is driven by incorrect estimation of the variability in father's predicted earnings. Third, if the Z variables are indeed exogenous with respect to son's earnings, then λ_2 equals 0, and β_{TSTSLS} is consistent. However, if parental education and occupation are the Z chosen, λ_2 will almost certainly be positive (λ_2>0)[7]. Thus, under the reasonable assumption that λ_2>0, and given $R^2 \leq 1$, β_{TSTSLS} will be upwardly inconsistent. Fourth, if everything else remains unchanged, the magnitude of this upward inconsistency will decrease as the variance explained in the first-stage prediction equation increases. Or, to put this another way, the upward inconsistency will decrease as the variance of father's predicted earnings tends towards the variance of father's actual earnings ($\sigma_{\hat{X}}{}^2 \rightarrow \sigma_X{}^2$). Fifth, it is important to recognise, however, that including additional variables to increase the R^2 of the first-stage prediction equation may simultaneously influence λ_2. Consequently, adding a particularly endogenous Z variable could increase λ_2 to such an extent that it more than offsets the benefits of any change to the first-stage R^2. Whether adding variables to the prediction equation reduces the inconsistency of β_{TSTSLS} is therefore an (underexplored) empirical issue, representing a gap in the literature that this paper attempts to fill.

Next, we extend the framework of Björklund and Jäntti (1997) and Nicoletti and Ermisch (2008) to the intergenerational correlation (ρ_{TSTSLS}). If one could observe X_{True} and Y_{True}, ρ would simply be:

[7] In other words, offspring with more educated parents from higher social classes are likely to earn more than offspring from less advantaged backgrounds, even after father's actual long-run earnings have been taken into account.

$$\rho = \beta.\frac{\sigma_X}{\sigma_Y} \tag{14}$$

Replacing σ_X with $\sigma_{\hat{X}}$, and β with β_{TSTSLS}, ρ_{TSTSLS} converges in probability to:

$$Plim\ \rho_{TSTSLS} = \beta_{TSTSLS}.\frac{\sigma_{\hat{X}}}{\sigma_Y}$$

$$= [\beta + \lambda_2.(1 - R^2)].\frac{\sigma_{\hat{X}}}{\sigma_Y}$$

$$= \left[\beta.\frac{\sigma_{\hat{X}}}{\sigma_Y} + \lambda_2.\frac{\sigma_{\hat{X}}}{\sigma_Y}.(1 - R^2)\right] \tag{15}$$

The inconsistency of ρ_{TSTSLS} is then given by (15) – (14):

$$\left[\beta.\frac{\sigma_{\hat{X}}}{\sigma_Y} + \lambda_2.\frac{\sigma_{\hat{X}}}{\sigma_Y}.(1 - R^2)\right] - \beta.\frac{\sigma_X}{\sigma_Y}$$

$$= \beta.\left[\frac{\sigma_{\hat{X}}}{\sigma_Y} - \frac{\sigma_X}{\sigma_Y}\right] + \lambda_2.\frac{\sigma_{\hat{X}}}{\sigma_Y}.(1 - R^2) \tag{16}$$

Now define A as the left-hand side of (16) and B as the right-hand side:

$$A = \beta.\left[\frac{\sigma_{\hat{X}}}{\sigma_Y} - \frac{\sigma_X}{\sigma_Y}\right] \tag{17}$$

$$B = \lambda_2.\frac{\sigma_{\hat{X}}}{\sigma_Y}.(1 - R^2) \tag{18}$$

Under the previously stated assumption that ${\sigma_{\hat{X}}}^2 \leq {\sigma_X}^2$, then A ≤ 0 (i.e. this will lead to *downward* inconsistency in ρ_{TSTSLS}). In contrast, assuming that λ_2>0 then, as $R^2 \leq 1$, B ≥ 0 (i.e. this will lead to *upward* inconsistency in ρ_{TSTSLS}). Therefore, unlike β_{TSTSLS}, one does not know the direction of the inconsistency in ρ_{TSTSLS}. Rather, it depends upon the relative magnitudes of A and B. This is again an empirical issue, which we provide the first evidence upon in our analysis.

The derivations presented above have all relied upon the following assumptions:

- The main and auxiliary datasets are random samples from the same population
- The Z variables are independent and identically distributed across the two datasets
- That X_{True} is the first-stage dependent variable, and it is this quantity that we wish to impute into the main dataset.

To meet these assumptions, it would be ideal for the main and auxiliary datasets to be identical (with the exception, of course, that the former does not include X_{True}). In this situation, the consistency of β_{TSTSLS} and ρ_{TSTSLS} is driven solely by the choice of imputer variables (Z) as set out above.

In reality, these assumptions may not be met. For instance, Björklund and Jäntti (1997) note it is common for respondents to report their own education and occupation (Z) in the auxiliary dataset, but for offspring's proxy reports of their father's characteristics to be available in the main dataset. The impact this has upon the consistency of β_{TSTSLS} and ρ_{TSTSLS} will depend upon the nature and extent of this measurement error. We therefore also consider this issue in our empirical analysis.

Moreover, there is the additional complication of how father's earnings are measured in the auxiliary dataset. Returning to equation (5), it has thus far been implicitly assumed that X_{True} (permanent father's earnings) is available within the auxiliary dataset. Yet, in practise, this is almost never the case. Rather, researchers typically have access to data for a cross-section of adults whose earnings are recorded for one particular year (X_{SINGLE}). A common choice is a labour force survey, for example. Therefore the prediction model is often specified as (19) rather than (5):

$$X_{SINGLE} = \delta . Z + \gamma . A + v \tag{19}$$

where:

X_{SINGLE} = Earnings in a single year for a cross-section of adults

Z = The imputation variables

A = Age group dummy variables

Estimates from (19) are then used to generate predictions of father's earnings in the main dataset instead of equation (6), with age set to around 40 (as the approximate point when annual earnings reach their peak):

$$\hat{X}_{single} = \hat{\delta} . Z + \hat{\gamma} . Age40 \tag{20}$$

Yet little is known about the consistency of TSTSLS estimates in such situations, where the first-stage dependent variable (X_{Single}) differs from the unobserved construct of interest

(X_{True}). Indeed, this issue was not explicitly considered by Björklund and Jäntti (1997) or Nicoletti and Ermisch (2008), and should not be assumed to be an innocuous change to the framework presented above.

We illustrate this point with an example. First, suppose that X_{true} is contained within the auxiliary dataset, along with a sufficiently rich set of Z so that the first-stage R^2 equals one. Consequently, $\hat{X}$ will be identical to X_{True}, thus resulting in consistent estimates of β_{TSTSLS} and ρ_{TSTSLS} (e.g. recall equation 13). Now consider the same scenario, but where X_{Single} is the first-stage dependent variable. A first-stage R^2 of one would imply that $\hat{X} = X_{single}$, resulting in rather different estimates of β_{TSTSLS} and ρ_{TSTSLS} (i.e. it is well established in the literature that $X_{Single} \neq X_{True}$). Specifically, the use of X_{Single} would lead to *downwardly* inconsistent estimates of β_{TSTSLS} and ρ_{TSTSLS}. This highlights how the corollaries presented within the framework above (e.g. β_{TSTSLS} always being upward inconsistent) do not necessarily hold when the first-stage variable being imputed (X_{Single}) differs from the construct actually of interest (X_{True}).

More general expressions for the inconsistency of β_{TSTSLS} and ρ_{TSTSLS} are therefore required, which hold whether either X_{single} or X_{true} are used as the first-stage dependent variable. First, consistent estimates of β_{OLS} from equation (1) converge in probability to:

$$Plim\ \beta = \frac{\sigma_{X,Y}}{{\sigma_X}^2} \tag{21}$$

Under TSTSLS, as X is unavailable, $\hat{X}$ enters in its place:

$$Plim\ \beta_{TSTSLS} = \frac{\sigma_{\hat{X},Y}}{{\sigma_{\hat{X}}}^2} \tag{22}$$

The inconsistency of β_{TSTSLS} is now given by (22) minus (21):

$$\frac{\sigma_{\hat{X},Y}}{{\sigma_{\hat{X}}}^2} - \frac{\sigma_{X,Y}}{{\sigma_X}^2} \tag{23}$$

Note that, in this more general framework, β_{TSTSLS} can be either upwards or downwards inconsistent. Indeed, the direction and magnitude of the inconsistency depends upon one's ability to correctly estimate the ratio of the covariance between father's and son's earnings ($\sigma_{X,Y}$) to the variance of father's earnings (${\sigma_X}^2$).

Equations (24) to (26) provide analogous expressions for ρ_{TSTSLS}. If X_{true} and Y_{true} were available in the main dataset, ρ could be consistently estimated by:

$$Plim\ \rho = \frac{\sigma_{X,Y}}{\sigma_X{}^2} . \frac{\sigma_X}{\sigma_Y} = \frac{\sigma_{X,Y}}{\sigma_X . \sigma_Y} \tag{24}$$

Replacing, X with $\hat{X}$, ρ_{TSTSLS} converges in probability to:

$$Plim\ \rho_{TSTSLS} = \frac{\sigma_{\hat{X},Y}}{\sigma_{\hat{X}}{}^2} . \frac{\sigma_{\hat{X}}}{\sigma_Y} = \frac{\sigma_{\hat{X},Y}}{\sigma_{\hat{X}} . \sigma_Y} \tag{25}$$

with the inconsistency of ρ_{TSTSLS} now given by (25) minus (24):

$$\frac{\sigma_{\hat{X},Y}}{\sigma_{\hat{X}} . \sigma_Y} - \frac{\sigma_{X,Y}}{\sigma_X . \sigma_Y} \tag{26}$$

In our empirical analysis we illustrate how the inconsistency of β_{TSTSLS} and ρ_{TSTSLS} can vary substantially depending on whether X_{Avg} (as a measure of X_{True}) or X_{single} is used as the first-stage dependent variable.

To conclude, we note that generated regressors (e.g. $\hat{X}$) are also subject to sampling variation. Consequently, second stage standard errors will be underestimated unless this additional uncertainty is taken into account. Murphy and Topel (1985), Wooldridge (2002) and Inoue and Solon (2010) provide formulae to make an appropriate adjustment to the estimated standard errors, while Björklund and Jäntti (1997), Inoue and Solon (2010) and Piraino (2014) suggest bootstrapping as a viable (if computer intensive) alternative. We do not dwell on this issue in this paper, and focus upon the inconsistency of TSTSLS point estimates. Nevertheless, this additional source of sampling uncertainty should always be taken into account when applying such generated regressor techniques[8].

[8] In our empirical application, we report bootstrapped standard errors. However, as our auxiliary dataset is set to contain 500,000 observations, sampling uncertainty in our generated regressor(s) is only a minor issue.

3. Data

The Panel Survey of Income Dynamics (PSID) is a nationally representative sample of US households. It began in 1968, with annual follow-ups to 1997, and bi-annual interviews thereafter. Detailed information has been collected at each sweep from the household head and their partner. Offspring are tracked as they leave the initially sampled household. Consequently, the PSID contains earnings information across multiple years for both fathers and sons. Throughout our analysis we restrict the sample to include sons who were household heads aged between 30 and 60 in 2011, and who reported their earnings for the previous year. Moreover, we only include sons whose father can be identified, has reported annual earnings on at least five occasions during their prime working years (between ages 30 and 60), and where both parent and offspring reports of father's education, occupation and industry are available.

After making these restrictions, our working sample equals 1,024 observations. Table 1 illustrates that approximately 80 percent of these individuals have at least 15 reports of father's annual earnings available, with 60 percent having 20 or more. A 'permanent' measure of father's earnings is created by averaging across all available reports for each sample member. We call this X_{AVG}, the closest measure to X_{True} available in the PSID. All earnings data have been adjusted to 2010 prices.

<< Table 1 >>

As part of each PSID sweep, fathers were asked detailed questions about their educational attainment, occupation and industry (we label father's reports of these variables as Z_{FA}). Education has been recorded using the highest grade ever completed, which we have converted into eight groups (see Table 2). Occupation and industry have been recorded using three digit census codes. These are finely defined categories – separating occupations and industries into approximately 200 groups. We use this detailed information on father's occupation and industry (taken from the year their offspring turned age 15[9]) as the key imputer variables (Z). At times, we also use more broadly defined '1 digit' occupation and industry groups (as presented in Table 2).

[9] Thus the occupation and industry of the average father included in the sample was taken from the 1983 PSID wave, where they were (on average) approximately 40 years old.

<< Table 2 >>

Sons also reported similar information about their father's education, occupation and industry (denoted Z_{CH}). For instance, in the 2011 sweep, sons were asked:

How much education did your father complete?

What was your father's usual occupation when you were growing up?

What kind of business or industry was that in?

Information on Z is thus available both directly from fathers (Z_{FA}) and indirectly via their sons (Z_{CH}). We exploit this in the following section to examine the robustness of TSTSLS mobility estimates to who reports the Z characteristics.

<u>*Creating an auxiliary dataset*</u>

The 1,024 PSID observations described above form our 'main' dataset (PSID-MAIN). To create an auxiliary dataset, we sample *with replacement* from these individuals. This generates an auxiliary sample containing 500,000 observations. (Henceforth PSID-AUX). The intuition behind this approach is similar to creating a single bootstrap re-sample[10]. Specifically, by randomly re-sampling from PSID-MAIN, we create a second random draw of individuals who belong to the same population[11]. This approach has three important advantages. First, one can guarantee that the main and auxiliary datasets are drawn from the same population. Second, the main and auxiliary datasets contain exactly the same variables measured in exactly the same way. Third, the size of the auxiliary dataset is under our control.

We exploit these advantages to produce TSTSLS mobility estimates under 'ideal conditions' (i.e. large auxiliary dataset, identical measurement of key variables across datasets, samples drawn from the same population). This enables us to investigate the consistency of β_{TSTSLS} and ρ_{TSTSLS} under different choices of the Z (imputer) variables. We then add additional complicating factors into the analysis (e.g. measurement of Z differing across datasets) to investigate the robustness of TSTSLS estimates to other challenges researchers face.

[10] Indeed, if we were to create an auxiliary dataset of size 1,024, then this would be equivalent to us taking a single bootstrap re-sample.

[11] A random number seed has been set to ensure results are replicable. We have experimented with different random number seeds and found little substantive change to our results.

Methodology

PSID-AUX is used to impute father's earnings ($\hat{X}$) into PSID-MAIN following the TSTSLS approach. The twist, of course, is that PSID-MAIN also contains an actual observed measure of father's long-run earnings (X_{AVG}). One can therefore investigate how intergenerational mobility estimates change when using $\hat{X}$ to measure father's earnings rather than X_{AVG}.

The first-stage prediction model, estimated using PSID-AUX, takes the form:

$$X_{AVG} = \alpha + \gamma . Z_{FA} + u \tag{27}$$

X_{AVG}= Father's observed time-average earnings

Z_{FA}= Father's reports of the imputer variables

The key decision is then which variables to include in Z_{FA}. Appendix A provides an overview of those typically used in the literature. There are four common choices:

(i) broad education level - e.g. Dunn (2007)
(ii) broad education and broad occupation - e.g. Björklund and Jäntti (1997)
(iii) broad education, occupation and industry - e.g. Piraino (2007)
(iv) broad education and detailed (3 digit) occupation - e.g. Leigh (2007)

This guides the combination of Z used in this paper. Table 3 illustrates the variables we include in five different first-stage model specifications (henceforth M1 to M5).

<< **Table 3** >>

Parameter estimates from these first-stage models are presented in Appendix B. These are used to impute father's earnings ($\hat{X}$) into PSID-MAIN:

$$\hat{X} = \hat{\alpha} + \hat{\gamma} . Z_{FA} \tag{28}$$

The following regression model is then estimated six times within PSID-MAIN - once using X_{AVG} to measure father's earnings and five times using the different predictions of $\hat{X}$:

$$Y_{2010} = \alpha + \beta . X + \varepsilon \tag{29}$$

Where:

Y_{2010} = Log annual earnings of sons in 2010

X = Father's earnings (measured using either X_{AVG} or $\hat{X}$)

We then compare estimates of β_{TSTSLS} and ρ_{TSTSLS} (obtained using $\hat{X}$) to β_{OLS} and ρ_{OLS} (obtained using X_{AVG}).

In our main analysis, son's earnings (Y) are taken from a single year (2010), when they are aged between 30 and 60. Ideally, to minimize the impact of 'life-cycle bias' (Haider and Solon 2006), a tighter age restriction would have been used (e.g. 35 to 45 year old sons only)[12]. Unfortunately, making such a restriction here would result in a significant reduction in sample size. We nevertheless appreciate the importance of this issue, and have hence investigated the sensitivity of our results to (a) restricting the sample of sons to 35 to 45 year olds only (b) using a five-year average of son's earnings. Although there is some evidence of lifecycle bias in our estimates, conclusions regarding the consistency of β_{TSTSLS} and ρ_{TSTSLS} remain largely unchanged. (All estimates available from the authors upon request).

4. Results

This section presents results from our empirical analysis of the PSID. Sub-section 4.1 focuses upon the choice of the instrumental (imputer) variables. Sub-section 4.2 turns to the issue of who reports the information on these Z characteristics (fathers or their sons). Finally, sub-section 4.3 considers the impact of how earnings are measured within the auxiliary dataset.

4.1 The choice of instrumental (imputer) variables

Table 4 compares estimates of β_{TSTSLS} and ρ_{TSTSLS} to β_{OLS} and ρ_{OLS}. Whereas β_{OLS} stands at 0.568[13], TSTSLS estimate M1 equals 0.753, M2 equals 0.767 and M3 0.717. β_{TSTSLS} is thus upward inconsistent by approximately 30 percent. β_{TSTSLS} declines under M4 and M5 (≈0.65) though the upward inconsistency remains non-trivial (15 percent).

<<Table 4>>

[12] Bohlmark and Lindquist (2006) suggest that lifecycle bias is approximately zero when sons are age 38 in the United States.

[13] Using a five-year average of father's earnings, Solon (1992) and Björklund and Jäntti (1997) estimate β_{OLS} to be approximately 0.40. However, Mazumder (2005) argues that a five-year average of father's earnings may be insufficient to eliminate problems of measurement error and transitory fluctuations. These estimates may therefore be downward inconsistent. Indeed, Mazumder obtains substantially higher values of β_{OLS} (0.61) when averaging father's earnings over 16 years. The fact that we obtain a higher estimate of β_{OLS} (0.56) than Solon and Björklund and Jäntti is therefore likely to be due to father's earnings having been averaged over more than 20 years (see Table 1).

To provide further insight into these results, Table 5 panel A presents the components of the inconsistency of β_{TSTSLS}, corresponding to equations (9) to (12) in section 2. For example, why is the upward inconsistency of β_{TSTSLS} not reduced between M1 and M2, despite the notable increase in the first-stage R^2? Table 5 illustrates that the addition of father's occupation (M2) also influences the direct effect of predicted father's earnings on son's earnings (λ_2); it increases from 0.30 to 0.36 as the R^2 moves from 0.38 to 0.45. In terms of consistency, losses due to the former are not offset by gains from the latter. Consequently, the upward inconsistency of β_{TSTSLS} increases from 0.185 to 0.199. This illustrates how simply choosing '*the instruments in order for the R^2 of the* [first-stage] *regression be as high as possible'* (Cervini-Pla 2012:9) will not necessarily reduce the inconsistency of β_{TSTSLS}. Indeed, the addition of variables which influence λ_2 as well as the first-stage R^2 can actually do more harm than good.

Table 5 Panel A also reveals that two factors drive the big reduction in the inconsistency between M3 and M4. The first is the large increase in the standard deviation of father's predicted earnings ($\sigma_{\hat{X}}$) from 0.385 to 0.449. This, via equation (11), substantially increases the first-stage R^2. The second is the decrease in λ_2, which falls from 0.29 to 0.21. Why is there then no further reduction of the inconsistency between M4 and M5? Table 5 reveals that although $\sigma_{\hat{X}}$ (and thus R^2) increase, λ_2 approximately returns back to its level under M3 (0.29). The effect of the former cancels out the latter, meaning no net gain regarding the consistency of β_{TSTSLS}.

<<Table 5>>

Returning to Table 4, ρ_{OLS} equals is 0.316. The TSTSLS M1 estimate is 0.259; *downward* inconsistency of approximately 18 percent. However ρ_{TSTSLS} increases as additional Z_{FA} variables are added to the prediction model, with the downward inconsistency standing at 12 percent using M3 (ρ_{TSTSLS} = 0.277), and essentially zero using M5 (ρ_{TSTSLS} = 0.307). The inconsistency of ρ_{TSTSLS} therefore tends to be (a) in the opposite direction (b) smaller in magnitude and (c) less sensitive to the combination of the Z variables than the inconsistency of β_{TSTSLS}. Indeed, Table 4 illustrates how ρ_{TSTSLS} is not usually too far from ρ_{OLS}. This is important given that, of the near 30 studies applying TSTSLS reviewed in Appendix A, only Björklund and Jäntti (1997) report the intergenerational correlation.

Table 5 Panel B splits the inconsistency of ρ_{TSTSLS} into two components: part A (corresponding to equation 17) and part B (corresponding to equation 18). Recall how the former induces downward inconsistency in ρ_{TSTSLS}, while the latter leads to upward inconsistency. It becomes clear that the comparatively small inconsistency of ρ_{TSTSLS} (relative to the inconsistency of β_{TSTSLS}) is due to these two components partially cancelling one another out. However, the downward pressure induced by part A is always slightly greater than the upward pressure from part B, leading to the overall downward inconsistency of ρ_{TSTSLS}.

What happens as additional variables are added to the prediction model? First, the downward pressure induced by part A is always reduced. This is because the standard deviation of father's predicted earnings ($\sigma_{\hat{X}}$) is the only term within equation A that changes (see equation 17), and can only increase towards the 'true' value (σ_X) as variables are added to the prediction model. In contrast, part B includes ($\frac{\sigma_{\hat{X}}}{\sigma_Y}$) and $(1 - R^2)$[14], with a greater value of $\sigma_{\hat{X}}$ increasing the former but decreasing the latter. Moreover, λ_2 is also found in component B, which fluctuates in value between M1 and M5. Thus, whereas adding information to the prediction model clearly reduces the inconsistency induced by part A, the influence on part B is hard to predict. Our empirical analysis does suggest, however, that gains from the former more than offset any losses from the latter. Consequently, the inconsistency of ρ_{TSTSLS} does generally decline when information is added to the first-stage prediction model.

To conclude this sub-section, we present estimates of rank order mobility (i.e. father's and son's relative position in the earnings distribution). Figure 2 provides a selection of findings from estimated transition matrices. (See Appendix C for full results). The top set of bars illustrate the percent of sons in each earnings quartile given that their father is in the *top* earnings quartile. The bottom set of bars presents analogous results for the sons of fathers in the *bottom* earnings quartile. Interestingly, TSTSLS estimates compare relatively well. For instance, using time-average father's earnings, 40 percent of sons with fathers in the bottom earnings quartile remain in the bottom quartile (white segment of the bottom set of bars), while just 11 percent rise to the top quartile (black segment bottom set of bars). The TSTSLS estimates produce very similar results – even when the imputation model is relatively sparse (e.g. 38 percent and 10 percent respectively using prediction model M1). Moreover, although there is slight underestimation of the probability that sons of high earning fathers will remain

[14] See equation (11) for the relationship between $\sigma_{\hat{X}}$ and R.

in the top quartile (black segments in the top set of bars), there nevertheless remains a high degree of consistency between the TSTSLS and time-average results. In additional analyses (available upon request) we follow Chetty et al (2014) and Gregg, MacMillian and Vittori (2014) and produce rank-rank mobility estimates using TSTSLS. Key findings are very similar to those for the transition matrices presented above.

<< **Figure 2** >>

Why is this relevant to our discussion of β_{TSTSLS} and ρ_{TSTSLS}? It illustrates how TSTSLS captures *rank order* mobility (i.e. father's and son's position in the earnings distribution) remarkably well. It thus provides further evidence that the inconsistency of β_{TSTSLS} and ρ_{TSTSLS} is largely being driven by scale miss-measurement (e.g. difficulties in accurately capturing the variance of father's earnings) rather than fathers being placed in the wrong part of the earnings distribution.

4.2 Measurement of imputer variables (Z)

The above investigation took place under 'ideal conditions', with identical measurement of key variables across main and auxiliary datasets. We now investigate the impact of the imputer variables being measured using son's recall of their father's characteristics (Z_{CH}) in the main dataset, while individuals own reports are used within the auxiliary dataset (Z_{FA}).

First, we investigate the uniformity of parent (Z_{FA}) and offspring (Z_{CH}) reports of father's education, occupation and industry. Appendix D provides full cross-tabulations, with summary results in Table 6. This includes the percentage of occasions where father's and son's report the same category ('percentage correct') and Kappa statistics of inter-rater reliability (a statistic which adjusts for agreement occurring by chance). Kappa statistics range from -1 (complete disagreement) to +1 (complete agreement) with Landis and Koch (1977) providing the following rules of thumb:

- 0-0.20 'Slight' agreement (between parent and child reports)
- 0.21–0.40 'fair' agreement
- 0.41–0.60 'moderate' agreement
- 0.61–0.80 'substantial' agreement
- 0.81–0.99 'almost perfect' agreement

<< **Table 6** >>

Fathers and sons report the same education and industry on more than 60 percent of occasions. Kappa statistics (0.52 and 0.55) are towards the top end of Landis and Koch's 'moderate' agreement category, with 'substantial agreement' when weighted Kappa is used (0.72 and 0.67)[15]. In contrast, just 27 percent of father's and son's report the same category for father's occupation, with Kappa statistics suggesting agreement is 'slight' (0.16) to 'fair' (0.28). One potential explanation is sons were asked about their father's occupation at a vague time point ('*what was your father's occupation when you were growing up*?') which we have compared to the job father's reported holding when sons were age 15. Consequently, we are unable to establish whether this lack of agreement is due to son's inability to accurately recall their father's occupation, or different interpretation of the questions asked (e.g. son's recalling their father's occupation at a different age).

Table 7 illustrates how switching to offspring reports of the imputer variables (Z_{CH}) influences estimates of β_{TSTSLS} and ρ_{TSTSLS}. Overall, this has relatively little impact upon our results. For instance, β_{TSTSLS} (ρ_{TSTSLS}) is estimated to be 0.767 (0.286) when using imputation model M2 and *father's* reports (Z_{FA}). This changes to 0.858 (0.291) when using son's reports instead (Z_{CH}). Similarly, under imputation model M5, estimates of β_{TSTSLS} (ρ_{TSTSLS}) stand at 0.642 (0.307) using father's reports, and 0.662 (0.292) using son's reports. Differences are therefore usually quite small, though on certain occasions are non-trivial. Nevertheless, our empirical analysis overall suggests that TSTSLS estimates are fairly robust to this particular measurement issue.

<u>*4.3 Imputation of current versus time-average father's earnings*</u>

Does changing the first-stage dependent variable from X_{Avg} to X_{single} influence β_{TSTSLS} or ρ_{TSTSLS}? Table 8 provides results, with X_{single} measured using father's earnings in 1980 (or the closest available year)[16].

<< Table 8>>

Key findings remain largely unaltered under M1, M2 and M3; large upward inconsistency in β_{TSTSLS} remains, with slight downward inconsistency in ρ_{TSTSLS}. However, β_{TSTSLS} is now much smaller under M4 and M5. For instance, under M5 β_{TSTSLS} was 0.642 when using X_{AVG}

[15] See Table 6 notes for how weighted Kappa is defined.

[16] Everything else is left unaltered. We have experimented with altering the year used to measure father's occupation and industry and found little change to the results. Father's reports of the imputer variables (Z_{FA}) are used within both datasets.

(upward inconsistency of 15 percent). But, after changing the first-stage dependent variable to X_{single}, β_{TSTSLS} falls to 0.415 (*downward* inconsistency of 25 percent). Similarly, ρ_{TSTSLS} using M5 is now 0.230 (downward inconsistency of 25 percent) having previously stood at 0.307 (downward inconsistency of one percent).

Table 9 breaks these TSTSLS estimates down into their respective components (corresponding to equation 22 for β_{TSTSLS} and equation 25 for ρ_{TSTSLS}). To begin, the covariance between father's and son's earnings (i.e. the common numerator of β_{TSTSLS} and ρ_{TSTSLS}) is similar – although always marginally smaller – using X_{single}. For instance, $\sigma_{\hat{x},y}$ under M3 falls from 0.106 using X_{AVG} to 0.098 using X_{single}. Likewise, under M1, M2 and M3, the variance of predicted father's earnings (${\sigma_{\hat{x}}}^2$) does not seem sensitive to the choice of the first-stage dependent variable (e.g. for M3, ${\sigma_{\hat{x}}}^2$ is 0.148 using X_{AVG} and 0.145 using X_{single}). Consequently, none of the key components of β_{TSTSLS} or ρ_{TSTSLS} are particularly influenced by the use of X_{single} rather than X_{AVG} when the first-stage prediction model is relatively sparse. Hence estimates of β_{TSTSLS} and ρ_{TSTSLS} are similar whichever earnings measure (X_{AVG} or X_{single}) is used.

<<Table 9>>

The same does not hold true, however, under M4 and M5. Specifically, the variance of father's earnings (${\sigma_{\hat{x}}}^2$) is significantly bigger when the first-stage dependent variable is X_{Single}. In contrast, the covariance between father's predicted earnings and son's earnings tends to be slightly smaller. Using M5 as an example, ${\sigma_{\hat{x}}}^2$ rises from 0.228 (X_{AVG}) to 0.305 (X_{Single}), while $\sigma_{\hat{x},y}$ falls from 0.146 (X_{AVG}) to 0.127 (X_{Single}). Thus, while the denominator of β_{TSTSLS} (${\sigma_{\hat{x}}}^2$) has substantially *increased* (and is now almost identical to the denominator of β_{OLS}) the numerator ($\sigma_{\hat{x},y}$) has slightly *decreased* (and remains 26 percent below the numerator of β_{OLS}). This causes β_{TSTSLS} to become *downwardly* inconsistent. Whether one uses X_{AVG} or X_{Single} as the first-stage dependent variable therefore seems to have much more influence upon the key components of β_{TSTSLS} when a detailed set of Z characteristics are included in the first-stage prediction model.

Building upon the intuition above, the *standard deviation* of father's predicted earnings ($\sigma_{\hat{x}}$) also enters the denominator of ρ_{TSTSLS}. The increase in $\sigma_{\hat{x}}$ from using X_{Single} as the first-stage dependent variable (as opposed to X_{AVG}) therefore also puts downward pressure on

ρ_{TSTSLS}[17]. Indeed, when using X_{Single}, ρ_{TSTSLS} actually moves further away from ρ_{OLS} as Z variables are added to the first-stage prediction model. For instance, the TSTSLS M2 estimate of ρ (0.271) is much closer to the OLS value (0.315) than the estimate obtained under M5 (0.230). In other words, the inconsistency of ρ_{TSTSLS} has *increased* in absolute magnitude, driven by the greater variability in father's predicted earnings. This is in direct contrast to results using X_{AVG} (presented on the left hand side of Table 9) where adding Z variables to the prediction model almost always brought ρ_{TSTSLS} and ρ_{OLS} closer together (i.e. *decreased* the inconsistency).

These results have important implications. First, changing the first-stage dependent variable can lead to rather different estimates of earnings mobility. Second, it is only safe to assume β_{TSTSLS} is upward inconsistent if X_{True} is the first-stage dependent variable (i..e. the earnings measure being imputed into the main dataset). Third, this strengthens the empirical evidence that TSTSLS estimates of the intergenerational correlation are typically downward inconsistent. Finally, even subtle changes to the imputation model can make important differences to β_{TSTSLS} and ρ_{TSTSLS}.

<< Table 8 >>

5. Conclusions

Intergenerational earnings mobility is a topic of great academic and policy concern. However, producing consistent estimates of earnings mobility is not a trivial task. In many countries earnings data cannot be linked across generations. Consequently, several studies estimate earnings mobility using TSTSLS instead. This paper has presented new evidence on the consistency of earnings mobility estimates based upon this methodology.

A summary of our results can be found in Table 10. This illustrates the sensitivity of β_{TSTSLS} and ρ_{TSTSLS} to using different first-stage imputation models and measurement of key variables. Column 1 indicates whether X_{Single} (1980) or X_{AVG} (AVG) is the first-stage dependent variable. Column 2 indicates whether father's (FA) or son's (CH) reports of Z are used, while column 3 provides the specification of the prediction model (to be cross-referenced

[17] The impact is less pronounced than for β_{TSTSLS} due to the *standard deviation* of father's predicted earnings being the key term rather than the variance.

with Table 3). Columns 4 and 5 provide estimates of β_{TSTSLS} and ρ_{TSTSLS}, with shading illustrating the absolute degree of inconsistency. The following findings emerge:

- β_{TSTSLS} is often (although not always) upwardly inconsistent.
- β_{TSTSLS} is particularly sensitive to the choice and measurement of the first stage imputation model. Estimates are up to 50 percent upwardly inconsistent or 30 percent downwardly inconsistent.
- Estimates of ρ_{TSTSLS} tend to be more stable and suffer less inconsistency. Of the 20 estimates in Table 10, 14 lie within 20 percent of ρ_{OLS}, with five within ten percent.
- Although the inconsistency of ρ_{TSTSLS} can in theory be in either direction, our empirical analysis suggests that, in practise, they tend to be below ρ_{OLS}.

<< Table 10 >>

Based upon our findings, we provide the following guidance to researchers wishing to estimate earnings mobility using TSTSLS. First, ρ_{TSTSLS} and β_{TSTSLS} should both be reported where possible. But, if a choice has to be made, our empirical analysis suggests there may be reasons to prefer the former over the latter[18]. Second, the auxiliary and main datasets should contain information on educational attainment and *detailed* (3 digit) occupation as a minimum. This means that at least two first-stage specifications can be estimated – a 'broad' specification (as per our model M2 or M3) and a 'detailed' specification (as per our model M4 or M5). One can then investigate how this changes estimates of ρ_{TSTSLS} and β_{TSTSLS}, including a breakdown into their separate components (as per our Table 9). Third, the auxiliary dataset should ideally contain information on respondents' time-average earnings (X_{AVG}). The use of cross-sectional data with respondents' earnings reported at a single time-point (e.g. a labour force survey) should be considered a second-best alternative. Fourth, as briefly discussed in section 2, standard errors should be corrected to account for the sampling variation in the predictions of father's earnings. This can be done via a Murphy-Topel correction (Murphy and Topel 1985) or appropriate application of a bootstrap technique (Inoue and Solon 2010; Björklund and Jäntti 1997)[19]. Fifth, researchers should note that their estimates of earnings mobility may differ from other studies due to methodological rather than substantive reasons. This includes instances

[18] At the same time, it is important to recognise that 'classical' measurement error in son's earnings will lead to inconsistent estimates of ρ but not β (Black and Devereux 2011). Counter-arguments can therefore be made as to why one may prefer β over ρ – hence our advice that both should be reported whenever possible.
[19] Hardin (2002) and Hole (2006) illustrate how this can be implemented in Stata.

where TSTSLS has been used in rather different ways (e.g. different combinations, definitions and measurement of key variables). Finally, we urge great care to be taken when comparing mobility estimates across studies – and across countries - where different methodologies have been applied.

References

Björklund, A. and Chadwick, L. 2003. 'Intergenerational income mobility in permanent and separated families.' *Economics Letters* 80: 239–246.

Björklund, A. and Jäntti, M. 1997. 'Intergenerational income mobility in Sweden compared to the United States.' *The American Economic Review* 87(5): 1009 –1018.

Björklund, A. and Jäntti, M. 2009. 'Intergenerational economic inequality.' Pp. 491 - 521 in The Oxford Handbook of Economic Inequality, edited by Wiemer Salverda, Brian Nolan and Timothy Smeeding. Oxford: Oxford University Press.

Black, S. and Devereux, P. 2011. 'Recent developments in intergenerational mobility.' In The Handbook of Labor Economics, edited by Orley Ashenfelter and David Card. North-Holland: Elsevier.

Blanden, J. 2013. 'Cross-national rankings of intergenerational mobility: A comparison of approaches from economics and sociology.' *Journal of Economic Surveys* 27(1):38–73.

Böhlmark, A. and Lindquist, M. 2006. 'Life-cycle variations in the association between current and lifetime income: replication and extension for Sweden.' *Journal of Labor Economics* 24(4): 879–896.

Cervini – Pla, M. 2012. 'Exploring the sources of earnings transmission in Spain.' MPRA working paper 34942. Accessed 09/10/14 from http://mpra.ub.uni-muenchen.de/36093/

Chetty, R.; Hendren, N.; Kline, P.; Saez, E. and Turner, N. 2014. 'Is the United States still a land of opportunity? Recent trends in intergenerational mobility'. NBER Working Paper 19844. Accessed 15/08/2014 from http://www.nber.org/papers/w19844.

Corak, M. and Heisz, A. 1999. 'The intergenerational income mobility of Canadian men: evidence from longitudinal tax data.' *Journal of Human Resources* 34(3): 504–533.

Corak, M. 2012. 'Inequality from generation to generation: the United States in comparison', University of Ottawa working paper. Accessed 13/12/12 from http://milescorak.files.wordpress.com/2012/01/inequality-from-generation-to-generation-the-united-states-in-comparison-v3.pdf

Dunn, C. 2007. 'The intergenerational transmission of lifetime earnings: evidence from Brazil.' *B.E. Journal of Economic Analysis and Policy* 7(2): article 2.

Gregg, P.; MacMillian, L. and Vittori, C. 2014. 'Moving towards estimating lifetime intergenerational economic mobility in the UK.' *DoQSS working paper 14-12*. Accessed 15/08/2014 from http://repec.ioe.ac.uk/REPEc/pdf/qsswp1412.pdf

Haider, S. and Solon, G. 2006. 'Life-cycle variation in the association between current and lifetime earnings.' *American Economic Review* 96(4): 1308-1320.

Hardin, J. 2002. 'The robust variance estimator for two-stage models'. *Stata Journal* 2: 253–266.

Hole, A. 2006. 'Calculating Murphy–Topel variance estimates in Stata: A simplified procedure.' *The Stata Journal* 6(4):521-29.

Hussain, M; Munk, M. and Bonke, J. 2009. 'Intergenerational earnings mobilities: how sensitive are they to income measures?' *Journal of Income Distribution* 18(3):79-92.

Inoue, A., and Solon, G. 2010. 'Two-sample instrumental variables estimators.' *The Review of Economics and Statistics* 92(3): 557 - 561.

Landis, J. and Koch, G. 1977. 'The measurement of observer agreement for categorical data.' *Biometrics* 33(1): 159-174.

Lefranc, A. and Trannoy, A. 2005. 'Intergenerational earnings mobility in France: is France more mobile than the US?' *Annales d'Economie et Statistique* 78(2): 57 – 77.

Leigh, A. 2007. 'Intergenerational economic mobility in Australia.' *The B.E. Journal of Economic Analysis & Policy*: 7(2): article 6.

Mazumder, B. 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.

Murphy, K. and Topel, R. 1985. 'Estimation and inference in two-step econometric models.' *Journal of Business and Economic Statistics* (3)4: 370 – 379.

Nicoletti, C. and Ermisch, J. 2008. 'Intergenerational earnings mobility: changes across cohorts in Britain.' *B.E. Journal of Economic Analysis and Policy* 7(2): article 9.

Nuñez, J. and Miranda, L. 2011. 'Intergenerational income and educational mobility in urban Chile.' *Estudios de Economia* 38(1): 195 – 221.

Piraino, P. 2007. 'Comparable estimates of intergenerational income mobility in Italy.' *The B.E. Journal of Economic Analysis & Policy* 7(2): article 1.

Piraino, P. 2014. 'Intergenerational earnings mobility and equality of opportunity in South Africa.' Southern Africa Labour and Development Research Unit working paper 131. Accessed 04/07/2014 from http://opensaldru.uct.ac.za/bitstream/handle/11090/696/2014_131_Saldruwp.pdf?sequence=1

Solon, G. 1992. 'Intergenerational income mobility in the United States.' *The American Economic Review* 83(3): 393 – 408.

Vogel, T. 2008. Reassessing intergenerational mobility in Germany and the United States: the impact of differences in lifecycle earnings patterns. SFB Discussion Paper 2006-055, Humbolt University, Berlin. Accessed 30/04/2013 from http://www.wiwi.hu-berlin.de/professuren-en/vwl/wtm2/Employees/vogel/vogelmobility

Wooldridge, J. 2002. *Econometric analysis of cross-section and panel data*. Massachusetts: MIT Press.

Table 1. Number of father's earnings observations available

Number of father's earnings observations	%	Cumulative %
6	1	1
7	1	1
8	1	3
9	2	4
10	1	6
11	3	8
12	2	11
13	3	14
14	3	17
15	2	19
16	4	23
17	3	26
18	4	30
19	5	35
20	6	41
21	8	49
22	7	56
23	7	63
24	6	70
25	8	78
26	7	84
27	5	89
28	5	94
29	4	97
30	3	100
n	**1,024**	

Notes: Author calculations using the PSID dataset.

Table 2. Education, broad (1 digit) occupation and broad (1 digit) industry categories

Education	Occupation	Industry
No education = 0 grades completed	Professional	Agriculture
Grades 1 to 5	Managers / senior administrators	Mining
Grades 6 to 8	Sales workers	Construction
Grades 9 to 12	Clerical	Manufacturing
High school = 12 grades	Craftsman	Transport and communication
Some college = grades 13 to 15	Operatives	Wholesale and retail
College degree = grade 16	Transport	Finance
Advanced college degree = grade 17	Laborers	Business services
	Farmers	Personal services
	Service workers	Entertainment
		Professional services
		Public administration

Notes: Refers to information on father's education, broad occupation and broad industry available within the PSID.

Table 3. The imputer (Z) variables used in the first-stage prediction models

	M1	M2	M3	M4	M5
Race	√	√	√	√	√
Education	√	√	√	√	√
Occupation (1 digit)	-	√	√	-	-
Occupation (3 digit)	-	-	-	√	√
Industry (1 digit)	-	-	√	√	-
Industry (3 digit)	-	-	-	-	√

Notes: M1 to M5 refers to the five different specifications of the first stage prediction model. All variables refer to characteristics of PSID fathers.

Table 4. Estimates of the intergenerational elasticity (β) and correlation (ρ) using different TSTSLS imputation models

TSTSLS model	First-stage R^2	Elasticity		Correlation	
		β_{TSTSLS}	SE	ρ_{TSTSLS}	SE
M1	0.385	0.753	0.087	0.259	0.030
M2	0.449	0.767	0.080	0.286	0.030
M3	0.483	0.717	0.077	0.277	0.030
M4	0.658	0.641	0.066	0.289	0.030
M5	0.742	0.642	0.062	0.307	0.030
OLS	-	**0.568**	**0.053**	**0.316**	**0.029**

Notes: Authors' calculations using the PSID dataset. Sample restricted to the same 1,024 individuals across all specification. SE stands for standard error. M1 to M5 indicate which first-stage TSTSLS imputation model has been used (see Table 3). Estimates using observed time-average father's earnings (OLS) reported in the bottom row. A full set of first-stage parameter estimates can be found in Appendix B.

Table 5. Estimates of the inconsistency of TSTSLS earnings mobility estimates

(a) Elasticity

	M1	M2	M3	M4	M5
λ_2	0.300	0.361	0.287	0.213	0.283
$\sigma_{\hat{X}}$	0.343	0.371	0.385	0.449	0.477
σ_X	0.554	0.554	0.554	0.554	0.554
$\sigma_{\hat{X},X}$	0.118	0.138	0.148	0.202	0.228
R^2	0.385	0.449	0.483	0.658	0.742
β_{OLS}	0.568	0.568	0.568	0.568	0.568
β_{TSTSLS}	0.753	0.767	0.717	0.641	0.642
Inconsistency	0.185	0.199	0.149	0.073	0.073

(b) Correlation

	M1	M2	M3	M4	M5
β_{OLS}	0.568	0.568	0.568	0.568	0.568
$\sigma_{\hat{X}}$	0.343	0.371	0.385	0.449	0.477
σ_X	0.554	0.554	0.554	0.554	0.554
σ_Y	0.995	0.995	0.995	0.995	0.995
Inconsistency part A	**-0.120**	**-0.105**	**-0.097**	**-0.060**	**-0.044**
λ_2	0.300	0.361	0.287	0.213	0.283
$\sigma_{\hat{X}}$	0.343	0.371	0.385	0.449	0.477
σ_Y	0.995	0.995	0.995	0.995	0.995
R^2	0.385	0.449	0.483	0.658	0.742
Inconsistency part B	**0.064**	**0.075**	**0.058**	**0.033**	**0.035**
ρ_{OLS}	0.316	0.316	0.316	0.316	0.316
ρ_{TSTSLS}	0.259	0.286	0.277	0.289	0.307
Inconsistency	**-0.057**	**-0.030**	**-0.039**	**-0.027**	**-0.009**

Notes: Authors' calculations using the PSID dataset. M1 to M5 refer to the TSTSLS imputation model specification used (see Table 3). See equation (11) and (12) for the components of the intergenerational elasticity and equations (16) to (18) for the components of the intergenerational correlation.

Table 6. The agreement between parent and offspring reports of father's education, occupation and industry

	Education	Occupation (broad groups)	Industry (broad groups)
Percent agreement	62	27	61
Kappa	0.52	0.16	0.55
Weighted Kappa	0.72	0.28	0.67

Notes: Authors' calculations using the PSID dataset. See Appendix D for full cross-tabulations. The Kappa statistic is a measure of inter-rater reliability that adjusts for agreement occurring by chance. It ranges from -1 (complete disagreement) to +1 (complete agreement) with 0 indicating no agreement. Landis and Koch (1977) provide rules of thumb for interpreting levels of agreement using Kappa: 0.01–0.20 'slight', 0.21–0.40 'fair', 0.41–0.60 'moderate', 0.61–0.80 'substantial', and 0.81–0.99 'almost perfect'. Weighted Kappa is where categories further apart (e.g. father reports high school and offspring reports bachelor degree) are considered to show greater levels of disagreement than categories closer together (e.g. father reports associates degree and offspring reports bachelor degree).

Table 7. TSTSLS estimates of the intergenerational correlation and elasticity when son's reports of father's Z characteristics

	β_{TSTSLS}		ρ_{TSTSLS}	
TSTSLS model	**Father's reports (Z_{FA})**	**Son's reports (Z_{CH})**	**Father's reports (Z_{FA})**	**Son's reports (Z_{CH})**
M1	0.753	0.800	0.259	0.264
M2	0.767	0.858	0.286	0.291
M3	0.717	0.815	0.277	0.292
M4	0.641	0.689	0.289	0.276
M5	0.642	0.662	0.307	0.292
OLS	**0.568**		**0.316**	

Notes: Authors' calculations using the PSID dataset. Sample restricted to the same 1,024 individuals across all specification. Table illustrates how β_{TSTSLS} and ρ_{TSTSLS} differ when using son's reports (Z_{CH}) of their father's characteristics (e.g. education, occupation and industry) rather than using father's own reports (Z_{FA}). M1 to M5 refer to the specification of the TSTSLS imputation model used (see Table 3). Estimates using observed time-average father's earnings (OLS) reported in the bottom row.

Table 8. Estimates of the intergenerational correlation and elasticity using different first stage dependent variables

	β_{TSTSLS}		ρ_{TSTSLS}	
	X_{AVG}	X_{Single}	X_{AVG}	X_{Single}
M1	0.753	0.798	0.259	0.251
M2	0.767	0.741	0.286	0.270
M3	0.712	0.678	0.277	0.259
M4	0.641	0.476	0.289	0.236
M5	0.642	0.415	0.307	0.230
OLS	**0.568**		**0.316**	

Notes: Authors' calculations using the PSID dataset. Sample restricted to the same 1,024 individuals across all specification. X_{AVG} where time-average father's earnings is the dependent variable in the first stage imputation model (i.e. 'ideal conditions'). X_{Single} where father's 1980 earnings is the dependent variable in the first stage imputation model. M1 to M5 refer to the specification of the TSTSLS imputation model used (see Table 3).

Table 9. The numerator and denominator of β_{TSTSLS} and ρ_{TSTSLS} when 'current' earnings used as the first-stage dependent variable

(a) Intergenerational elasticity

	First-stage dependent variable = X_{AVG}					First-stage dependent variable = X_{Single}				
		$\sigma_{\hat{x},y}$		${\sigma_{\hat{x}}}^2$			$\sigma_{\hat{x},y}$		${\sigma_{\hat{x}}}^2$	
	β_{TSTSLS}	Value	%	Value	%	β_{TSTSLS}	Value	%	Value	%
M1	0.753	0.089	-49	0.118	-62	0.798	0.078	-55	0.098	-68
M2	0.767	0.106	-39	0.138	-55	0.741	0.098	-44	0.132	-57
M3	0.717	0.106	-39	0.148	-52	0.678	0.098	-44	0.145	-53
M4	0.641	0.129	-26	0.202	-34	0.476	0.116	-33	0.244	-20
M5	0.642	0.146	-16	0.228	-26	0.415	0.127	-27	0.305	-1
OLS	**0.568**	**0.175**	**-**	**0.307**	**-**	**0.568**	**0.175**	**-**	**0.307**	**-**

(b) Intergenerational correlation

	First-stage dependent variable = X_{AVG}						First-stage dependent variable = X_{Single}					
		$\sigma_{\hat{x},y}$		$\sigma_{\hat{x}}$		σ_y		$\sigma_{\hat{x},y}$		$\sigma_{\hat{x}}$		σ_y
	ρ_{TSTSLS}	Value	%	Value	%		ρ_{TSTSLS}	Value	%	Value	%	
M1	0.260	0.089	-49	0.343	-38	1.00	0.251	0.078	-55	0.312	-44	1.00
M2	0.286	0.106	-39	0.371	-33	1.00	0.271	0.098	-44	0.364	-34	1.00
M3	0.277	0.106	-39	0.385	-31	1.00	0.259	0.098	-44	0.381	-31	1.00
M4	0.289	0.129	-26	0.449	-19	1.00	0.236	0.116	-33	0.494	-11	1.00
M5	0.308	0.146	-16	0.477	-14	1.00	0.230	0.127	-27	0.553	0	1.00
OLS	**0.317**	**0.175**	**-**	**0.554**	**-**	**1.00**	**0.315**	**0.175**	**-**	**0.554**	**-**	**1.00**

Notes: Authors' calculations using the PSID dataset. M1 to M5 refer to the TSTSLS imputation model specification used (see Table 3). X_{AVG} where time-average father's earnings is the dependent variable in the first stage imputation model. X_{Single} where father's 1980 earnings is the dependent variable in the first stage imputation model. 'Value' presents the value of the statistic in question. '%' illustrates percentage underestimation relative to OLS results.

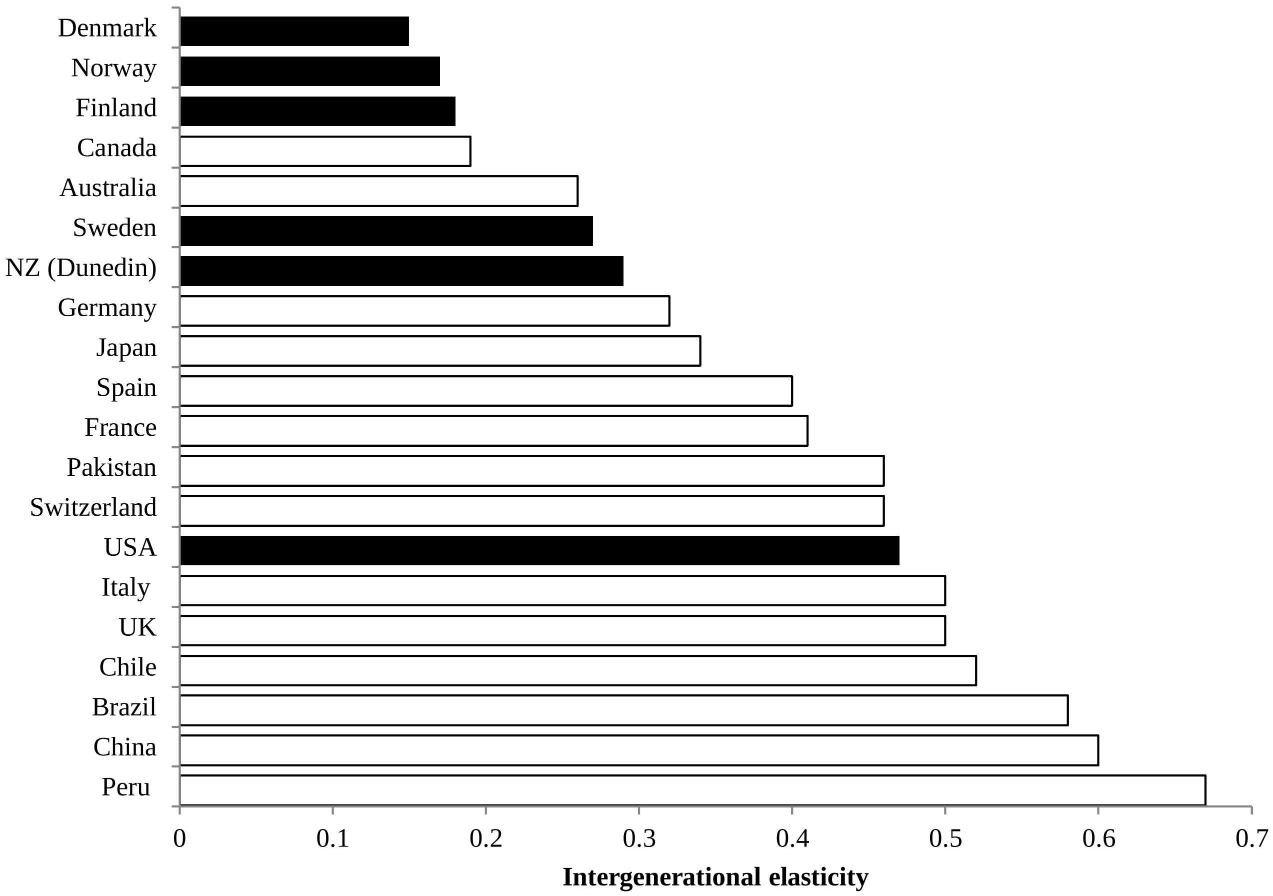
Table 10. A comparison of TSTSLS estimates using different measures of key variables and different imputation model specifications

(1) First-stage dependent variable	(2) Father / son reports of Z	(3) Imputer variables (Z)	(4) $\boldsymbol{\beta_{TSTSLS}}$	(5) $\boldsymbol{\rho_{TSTSLS}}$
AVG	CH	M2	0.858	0.291
AVG	CH	M3	0.815	0.292
1980	CH	M1	0.807	0.248
1980	CH	M2	0.806	0.259
AVG	CH	M1	0.800	0.264
1980	FA	M1	0.798	0.250
AVG	FA	M2	0.767	0.286
AVG	FA	M1	0.752	0.259
1980	FA	M2	0.741	0.270
1980	CH	M3	0.731	0.263
AVG	FA	M3	0.717	0.277
AVG	CH	M4	0.689	0.276
1980	FA	M3	0.678	0.259
AVG	CH	M5	0.662	0.292
AVG	FA	M5	0.642	0.307
AVG	FA	M4	0.641	0.289
OLS benchmark			**0.568**	**0.316**
1980	FA	M4	0.476	0.236
1980	CH	M4	0.472	0.213
1980	FA	M5	0.415	0.230
1980	CH	M5	0.349	0.183

Notes: Authors' calculations using the PSID dataset. Auxiliary dataset sample size set to 500,000 observations. 'Imputer variables' refers to the Z variables used to predict father's earnings (see Table 3). AVG / 1980 refers to the first-stage dependent variable (AVG = time-average; 1980 = single measure of father's earnings in 1980). FA/CH indicates whether father's or son's reports of the Z characteristics used in the main dataset.

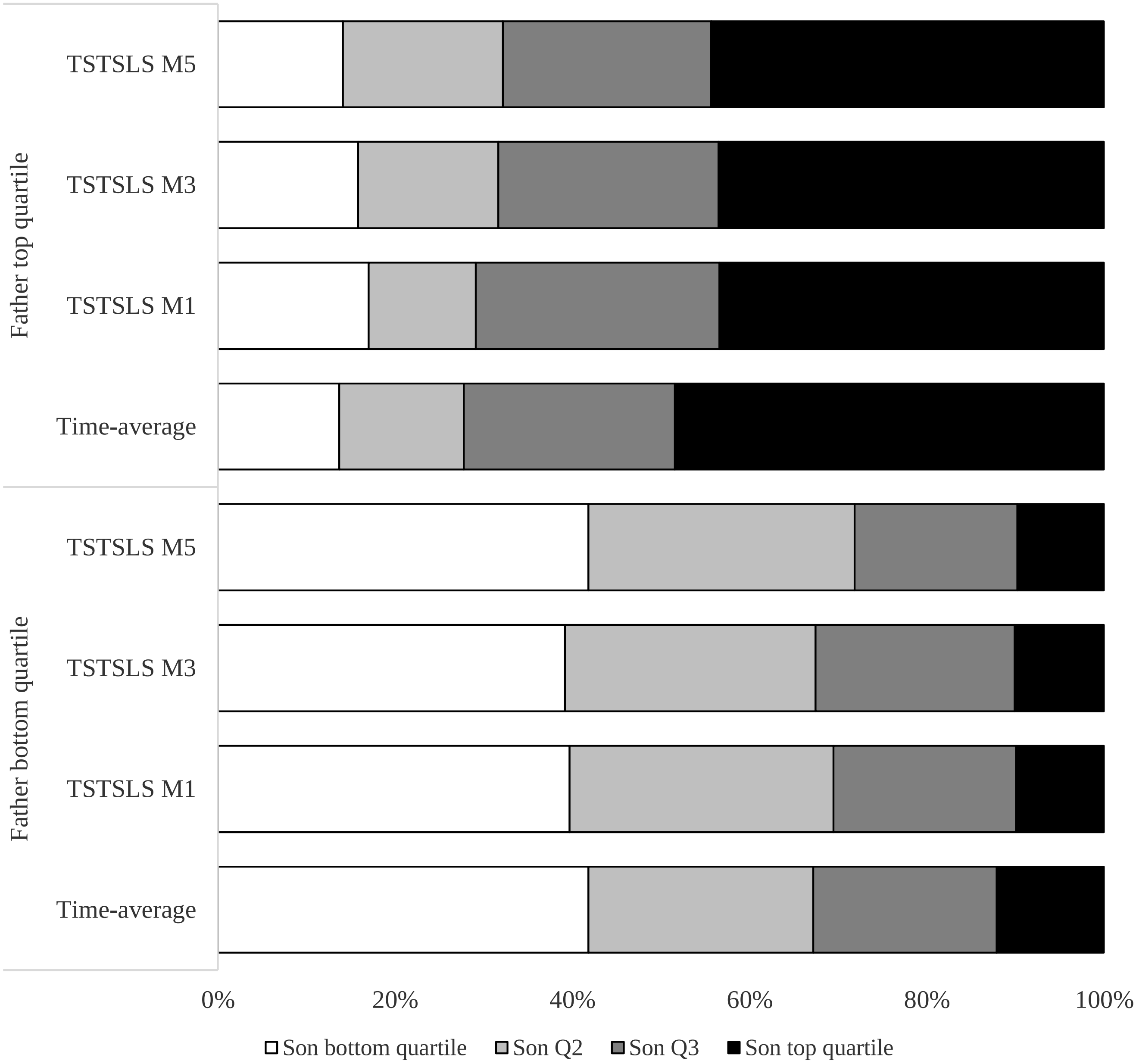
(white)	Absolute difference relative to time-average benchmark less than 10%
(light grey)	Difference relative to time-average benchmark 10% to 20%
(grey)	Difference relative to time-average benchmark 20% to 30%
(medium grey)	Difference relative to time-average benchmark 30% to 40%
(dark grey)	Difference relative to time-average benchmark 40% to 50%
(darkest grey)	Difference relative to time-average benchmark >50%

Figure 1. An international comparison of intergenerational earnings mobility



Notes: Estimates drawn from Corak (2012). New Zealand (NZ) data based upon a sample born in Dunedin and is not nationally representative. The colour of the bar indicates the estimation strategy used. Black bars indicate where OLS regression with time-average parental earnings has been used. White bars indicate where TSTSLS has been applied.

Figure 2. The estimated earnings quartile of sons, conditional upon their father being in the top (bottom) earnings quartile



Notes: Authors' calculations using the PSID dataset. Top set of bars illustrates the percent of sons in each earnings quartile given that their father is in the *top* earnings quartile. Bottom set of bars present analogous estimates for sons whose fathers are in the *bottom* earnings quartile. 'Time-average' where father's (observed) time-average earnings used to produce mobility estimates. TSTSLS estimates presented for imputation model specifications M1, M3 and M5 (see Table 3 for further details). Full cross-tabulations are presented in Appendix C.