Measurement error in income and schooling, and the bias of linear estimators

Paul Bingley, *SFI*
Alessandro Martinello, *Lund University, KWC, and SFI*

August 20, 2016

Abstract

We propose a general framework for determining the extent of measurement error bias in OLS and IV estimators of linear models, while allowing for measurement error in the validation source. We apply this method by validating Survey of Health, Ageing and Retirement in Europe (SHARE) data with Danish administrative registers. Contrary to most validation studies, we find measurement error in income is classical, once we account for imperfect validation data. We find non-classical measurement error in schooling, causing a 38 percent amplification bias in IV estimators of the returns, with important implications for the program evaluation literature.

We are grateful for comments from James Banks, Martina Brandt, Arnaud Chevalier, Christian Dahl, Mette Ejrnæs, Anders Holm, Bertel Schjerning, Søren Leth-Petersen and Ian Walker. We thank Claus Thustrup Kreiner for graciously sharing with us results from his random tax audit dataset. Funding was provided by the Danish Strategic Research Council (DSF-09-065167, DSF-09-070295) and Danish Social Sciences Research Council (FSE-09-063859). The usual disclaimer applies.

This paper uses data from SHARE release 2.5.0, as of May 24th 2011. The SHARE data collection has been primarily funded by the European Commission through the 5th framework program (project QLK6-CT-2001- 00360 in the thematic program Quality of Life), through the 6th framework program (projects SHARE-I3, RII-CT- 2006-062193, COMPARE, CIT5-CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th framework programme (SHARE-PREP, 211909 and SHARE-LEAP, 227822). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064, IAG BSR06-11, R21 AG025169) as well as from various national sources is gratefully acknowledged (see `http://www.share-project.org` for a full list of funding institutions).

Contact the corresponding author at alessandro.martinello@nek.lu.se .

I. Introduction

Researchers have long known that measurement error in the data of interest can affect the consistency of parametric estimators of even the simplest linear model (Stefanski, 1985, 2000). Researchers often implicitly assume that measurement error is classical, i.e., an additive, independent error term with constant variance (Fuller, 1987). Such an assumption is particularly convenient in the estimation of linear models, as Instrumental Variable (IV) estimators are robust to classical measurement error in the explanatory variables. However, all the seminal validation studies of Mellow and Sider (1983); Duncan and Hill (1985); Bound and Krueger (1991); Bound et al. (1994); Barron et al. (1997); Bollinger (1998) and more recent studies such as Bricker and Engelhardt (2008) find that measurement error in labor market-related outcomes is non-classical and negatively correlated with the quantity of interest. All these studies focus on the consequences of measurement error for Ordinary Least Squares (OLS) estimators, and all assume that the validation data is measured without error.

Our paper expands this line of research not only by focusing on the consequences of non-classical measurement error for IV estimators of linear models, but also by challenging the notion that validation data is measured without error. We present novel strategies for the validation of economic surveys that allow for measurement error in the validation data and incorporate traditional validation analysis as a special case. We show that moderate measurement error in the validation data causes the negative correlation between annual gross income and measurement error found when we perform a traditional validation study. Moreover, as measurement error in bounded variables is generally non-classical (Kane et al., 1999), we also estimate the measurement error properties of length of schooling, and show that when we instrument such an imperfectly measured variable, we obtain inflated IV estimates[1]. This finding is particularly relevant for the program evaluation literature and contributes to explaining why IV estimates of returns to schooling are typically larger in magnitude than OLS estimates, regardless of the expected sign of omitted variable bias.

We proceed in four steps. First we characterize measurement error bias in linear models in a general framework. Second, we match Survey of Health, Ageing and Retirement in Europe (SHARE) data with Danish administrative registers. Third we build moment conditions to identify the sufficient statistics characterized in the first step. Fourth, we estimate the measurement error properties of gross income and length of schooling. While measurement error also entails an efficiency loss, this paper focuses only on the consequences of measurement error for consistency of OLS and IV estimators, consequences that we can assess without imposing distributional assumptions.

The paper begins by recalling and expanding the measurement error model of Bound et al. (1994). We distinguish between imperfectly measured vari-

[1] Hyslop and Imbens (2001) show that such results also apply to their Optimal Prediction Error model.

ables as the dependent variable in OLS, an explanatory variable in OLS, or an explanatory variable in IV estimation. The variances of the measurement error and of the quantity of interest and their covariance are sufficient for identifying the expected measurement error bias in any of these three cases.

We estimate these parameters by matching survey measures of gross income and length of schooling with corresponding validation measures, drawn from administrative registers. Crucially, surveys may be contaminated by non-classical measurement error arising from non-random recall error or a flawed interview process (Biemer et al., 2004). This type of error occurs, for example, whenever low-income individuals overstate their earnings or high-income individuals understate theirs. Pischke (1995) shows that the respondents to the Panel Study of Income Dynamics-Validation Study do not report temporary shocks in their annual income flows. Instead, they attempt to provide information about their "normal" level of income, thus giving responses that are biased towards the mean. Our approach is not limited to measurement error in surveys but also applies to other data sources, especially as we allow some contamination of our validation data with measurement error.

We match the Danish part of SHARE, a longitudinal survey that collects data across nineteen European countries on individuals aged 50 or more and their spouses, with administrative records. By wave four, SHARE reports information from 150,000 interviews of 86,000 persons across all waves (Börsch-Supan et al., 2013) and is one of the most extensive surveys of elderly populations worldwide. As the Health and Retirement Study (HRS) served as a template for the development of SHARE and other sister surveys such as the English Longitudinal Study of Ageing and the Japanese Study of Aging and Retirement, the data collection mechanism and the questions asked are similar across this family of surveys. Although we use SHARE Denmark data for our analysis, the goal of this paper is not solely to determine the quality of a specific survey. Instead, our goal is to both develop strategies that, while allowing for imperfect validation data, can estimate the measurement error properties in a survey, and assess the consequences of these properties for the general class of linear models.

To initiate the first wave, SHARE Denmark drew a random sample of individuals aged 50 and above from the Central Person Register, an administrative database that indexes each resident of Denmark by a unique personal identifier (CPR). Because of this sampling method, we are able to successfully match 97% of the SHARE Denmark sample with uncensored tax records and other administrative data of high quality (see Jensen and Rasmussen, 2011, for schooling and Browning and Leth-Petersen, 2003, for income data), creating an exceptional dataset for a validation study.

Labor economists have conducted several validation studies. Duncan and Hill (1985) and Bound et al. (1994) use employer-provided payroll data as a validation source for ad hoc surveys of labor market outcomes, replicating questions from the Panel Study of Income Dynamics. Bound and Krueger (1991) link information on labor market earnings from the Current Population Survey to censored (at the top tax threshold) employer-reports of earnings from the Social Security Administration (SSA). Similarly, Bricker and

Engelhardt (2008) link responses from the HRS to SSA records of earnings subject to the Medicare payroll tax (virtually uncensored and also available for public sector workers), and Kristensen and Westergaard-Nielsen (2007) link Danish administrative data on schooling and earnings to the European Community Household Panel. All these studies assume that the validation data is measured without error, and they consistently find evidence against classical measurement error, especially for male labor earnings. We find that administrative records are contaminated with measurement error and show that when we allow for imperfect validation data, we cannot reject the hypothesis of classical measurement error in gross income (in logarithms)[2] in SHARE Denmark.

If validation data is measured with error, a simple analysis of the difference between survey and register variables cannot identify the properties of measurement error in the survey. To estimate these properties while allowing for imperfect validation data, we build two systems of moment equations that identify the characteristics of measurement error in gross income and length of schooling. We then estimate the parameters of interest via generalized method of moments (GMM). Exclusion restrictions provide identification without the need for distributional assumptions.

Three recent papers allow for measurement error in validation data: Kapteyn and Ypma (2007), Kreiner et al. (2015) and Abowd and Stinson (2013). However, these papers provide only partial answers to the question of how survey measurement error affects the consistency of linear estimators. While Kapteyn and Ypma (2007) allow validation data to be imperfect only because of mismatching observations in the two sources, Kreiner et al. (2015) and Abowd and Stinson (2013) do not separately identify all the statistics needed for calculating the expected measurement error bias for both OLS and IV estimators. While, in comparison to previous studies, we provide less information on the general measurement error structure (and thus say little about efficiency), we provide precise information on the expected measurement error bias in the general class of linear models. Where we estimate the properties of measurement error in gross income, our paper is closest to Kapteyn and Ypma (2007) in terms of data and approach. As in that study, we question the assumption that validation data are error-free, and we match survey data with Scandinavian administrative registers.

However, several key differences exist between our approach and that of Kapteyn and Ypma (2007). First, while Kapteyn and Ypma (2007) allow for mismatch between the survey and validation data, they otherwise assume the validation data is perfectly measured. We relax the assumption of perfect validation data, and our register-based survey sampling frame matches observations between the two data sources with a high degree of certainty.[3] Second, they separately examine earnings and pension income, while we focus

[2]To avoid cumbersome repetition, hereafter we simply use the term "gross income". Unless otherwise noted, this term refers to the log transformation of gross income.

[3]In a robustness check, we replicate our results, excluding (19) individuals reporting an age inconsistent with the registered date of birth.

on gross income. Third, they estimate the rich error structure in their model by imposing distributional assumptions on unobservables for identification. In contrast, we estimate a more parsimonious model only through first- and second-order moment equations, thus allowing for non-normal distributions of unobservables.

As the nature of our validation data differs for schooling and income, we develop separate identification strategies. Although we develop these strategies specifically for our data sources, they are applicable in general situations: one in which measurement error in at least one of two noisy independent measures is classical, and another in which only a subsample of a noisy variable is matched with a perfect validation source.

Because earnings, capital income, and pension income are third-party reported in Denmark (by the employer, bank, and the state or pension fund), we assume that measurement error in administrative income records is not correlated with true gross income. This assumption is also imposed by Kreiner et al. (2015) on the same tax register data, and is justified if first-party response error is the only cause of non-classical measurement error in continuous and unbounded variables. A separate analysis of the same sample of random tax audits studied in Kleven et al. (2011) supports this assumption by revealing that tax audit adjustments are not statistically correlated with the post-audit measure of gross income. We do not impose any further distributional assumptions other than finite and constant mean and variance for the error components. For identification, we use exclusion restrictions provided by third-party-reported variables that, while correlated with income, we assume to be otherwise uncorrelated with the first-party measurement error process in SHARE.

The assumption of classical measurement error in the validation data is not justified for bounded variables, such as length of schooling. Kane et al. (1999) point out that if schooling, which is an ordered categorical variable, is measured with error, then this error is in general negatively correlated with the true value of schooling[4]. As an extreme example, measurement error in a binary variable will always be negatively correlated with the true quantity of interest: if the true value is one, measurement error can only be non-positive; If zero, non-negative. Kane et al. (1999) show that measurement error in length of schooling is non-classical in both survey responses in the National Longitudinal Study of the High School Class of 1972 and their validation source, a selected sub-sample of data from the Post-secondary Education Transcript Survey.

Moreover, our original source of administrative information for schooling are third-party institutional reports for only a quarter of our sample, whereas information for the remainder of the sample is drawn from the last population census in 1970. Therefore, both SHARE survey and census reports of length of schooling can be contaminated with first-party response error.

[4]Examples of misclassification error non-negatively correlated with the true value of the quantity of interest can be constructed for non-binary categorical variables. However, such examples typically require asymmetric distributions.

However, because schooling seldom changes in late adulthood, and because the variable definition is the same in both the survey and the administrative data, we assume that the institutional reports we have for 25% of our sample are error free.

We then estimate the properties of measurement error in length of schooling through the three-way comparison among SHARE survey data, institutional reports, and census responses. We split our data into two samples: those for whom we assume administrative information to be precisely measured (institutional reports) and those for whom validation data is of similar nature to the survey measure (census reports). We characterize measurement error through a comparison of the differences between survey and validation data in the two samples. Intuitively, we perform an external validation study within an internal validation study. This is a novel approach to the validation of economic survey data. Similarly to the analysis of gross income, for schooling we do not impose any distributional assumptions on the measurement error components of the model.

Once we allow for imperfect validation data, we find that measurement error in SHARE is classical for annual gross income but non-classical for years of schooling. Entering years of schooling in a model as an explanatory variable causes a bias in both the OLS and IV estimators. More specifically, IV estimators suffer from amplification bias, leading to a 38% overestimation of the true returns to schooling.

The remainder of the paper is organized as follows. Section II recalls the general measurement error problem and identifies the sufficient statistics for measurement error that determine the bias for OLS and IV estimation of linear models. Section III presents our survey (SHARE Denmark) and validation (Danish administrative registers) data. Section IV presents our empirical strategies for estimating measurement error properties. Section V presents our results, and section VI concludes.

II. Measurement error bias in linear models

The consequences of measurement error depend on the structure of the model of interest, on whether the variable measured with error is a dependent or an explanatory one, and on the properties of the error (Hausman, 2001; Bound et al., 2001). In practice, there can be several sources of measurement error. We might have access to a variable that is simply a proxy of an unobserved quantity. In a survey, respondents might not recall precisely the information requested, or interviewers might transcribe the wrong answer. While theoretically different, these types of measurement error can be similarly modeled.

We focus on the consequences of additive measurement error for the consistency of OLS and IV estimators. Linear models are widely used in empirical microeconomics for their robustness and simplicity and are the backbone of most studies in the program evaluation literature. Program evaluation strategies such as difference-in-differences or regression discontinuity designs ultimately require the computation of OLS or IV estimators, both of which may suffer from measurement error biases.

We are interested in estimating the relationship between a dependent variable y and an explanatory variable x. However, we observe not one or both variables in the model, but a measure

$$\begin{aligned} m_s &= m + \underbrace{\kappa_s + \rho_s (m - \mu_m) + \varepsilon_s}_{\text{measurement error}} \\ &= \mu_m + \kappa_s + (1 + \rho_s)(m - \mu_m) + \varepsilon_s, \quad E\left[\varepsilon_s^2\right] = \sigma_s^2 \end{aligned} \tag{1}$$

where $m \in \{y, x\}$ and the subscript $\cdot_s$ indicates that we are validating survey information. Measurement error consists of three components: κ_s is a constant representing non-zero average measurement error; ε_s is an independent and identically distributed error term with mean zero and variance σ_s^2; and ρ_s represents the dependence between measurement error and the quantity of interest.

The classical measurement error model is a special case of the model described in equation (1), for $\rho_s = \kappa_s = 0$ (Stefanski, 2000, 1985). In this case, measurement error $m_s - m$ has mean zero and is independent of the quantity of interest. We do not model the dependence between other covariates and measurement error in m here, as the simple univariate model is sufficient for motivating our analysis.

We distinguish between three types of bias occurring in linear models: bias in OLS estimators due to measurement error in the explanatory variables (bias in $\widehat{\beta}^{OLS}$); bias in IV estimators due to measurement error in the explanatory variables (bias in $\widehat{\beta}^{IV}$); and bias due to measurement error in the dependent variable (bias in $\widehat{\beta}^{LHS}$). A well-known result is that, using the notation introduced in equation (1),

$$E\left(\widehat{\beta}^{OLS}\right) = \frac{Cov\left(x_s, y\right)}{Var\left(x_s\right)} = \beta \frac{(1+\rho_s)\,\sigma_x^2}{(1+\rho_s)^2\,\sigma_m^2 + \sigma_s^2} \tag{2}$$

where the multiplicative bias in the right-hand side of the equation collapses to the reliability ratio λ_s if the measurement error is classical (Fuller, 1987).

A less well-known result is that, while IV estimators are robust to classical measurement error in the explanatory variable[5], this property does not hold for non-classical measurement error, as $\widehat{\beta}^{IV}$ converges to

$$E\left(\widehat{\beta}^{IV}\right) = \beta \frac{1}{1+\rho_s}. \tag{3}$$

For negative ρ_s, equation (3) implies that the IV estimator on average overestimates the coefficient of interest β. The effect of non-classical measurement error on the consistency of IV estimators is often overlooked in the

[5] This property has been often cited as an explanation for IV estimates usually being larger than their OLS counterparts, even when we might expect omitted variable bias to increase the magnitude of the OLS estimate. Such findings are common in the labor economics literature, especially in studies estimating earnings returns to schooling (Card, 2001).

empirical microeconomics literature and is especially relevant for bounded and discrete variables. Kane et al. (1999) stress that, as length of schooling is typically an ordered categorical variable, measurement error (misclassification) in schooling is typically non-classical and negatively correlated with the true value of the variable. Thus IV estimates of average returns to schooling tend to be inflated.

The extreme example of a misclassified binary variable, which is a special case of the OLS model discussed by Aigner (1973), helps clarify why bounded variables are particularly subject to non-classical measurement error. Consider a linear model where x follows a Bernoulli distribution with probability p, a common situation in the program evaluation literature, where the interest often lies in the effect of a discrete treatment variable. Assume that we observe x with probability π; with probability $1-\pi$ the respondent misunderstands the question and gives the wrong answer $1-x$. In this case the expected value of the parameter ρ_s will be negative and equal to $-2(1-\pi)$. Equation (3) shows that, for an error probability of 10%, the IV estimator $\widehat{\beta}^{IV}$ converges to an estimate of β inflated by 25%.

Similarly, $\widehat{\beta}^{LHS}$ converges to

$$E\left(\widehat{\beta}^{LHS}\right) = \beta\left(1+\rho_s\right) \tag{4}$$

and will therefore suffer from a bias factor equal to the linear coefficient of a regression of the measurement error on the true quantity of interest (Bound et al., 1994).

Equations (2), (3), and (4) show that the biases of linear estimators depend on only three parameters defined in our measurement error model: ρ_s, σ_m^2, and σ_s^2. In practice, such parameters are defined by the second moments of the distributions of the quantity of interest and its measurement error. Once ρ_s, σ_m^2, and σ_s^2 are known, computing the expected bias for any linear model is straightforward.[6]

III. Data

We study measurement error in total annual gross income and years of schooling as recorded in SHARE. We focus on the first wave of SHARE Denmark, which in 2004 interviewed a representative sample of residents of Denmark aged 50 and above (main respondents) and their spouses, for a total of 1,707 individual respondents. Our validation source is administrative register data, which provides official demographic information for January 1, 2004 and tax reports for the year 2003.

[6]Equations (2), (3), and (4) refer to a model in which the single mismeasured variable is independent of the others—a univariate model being the simplest case. While this simple model is sufficient for motivating our analysis, one can generalize the bias factors for the general case of m mismeasured variables $\boldsymbol{X_m} = \boldsymbol{XA} + \boldsymbol{\Upsilon}$, where $\boldsymbol{A}$ is an invertible $m \times m$ matrix. For example, the 2SLS estimator that uses $\boldsymbol{Z}$ as instruments for $\boldsymbol{X}$ converges to $\left(\boldsymbol{\Delta}'\boldsymbol{\Delta}\boldsymbol{A}\right)^{-1}\boldsymbol{\Delta}'\boldsymbol{Y}$, where $\boldsymbol{\Delta} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}$.

The strength of our dataset is that SHARE Denmark selected the sample using our validation data; and thus a CPR number linking survey and registry data exists in principle for all the sampled respondents. We retrieve this linkage information from the data collection agency (SFI-survey) and Statistics Denmark[7].

We retrieve administrative records for 1,670 of the 1,707 individual respondents, corresponding to 97% of the first wave of SHARE Denmark. Of the 37 observations we cannot match, 21 are interviews whose main respondent appears as single in the registers[8]. In the remaining 16 observations (14 households), we cannot identify the main respondent. Of the 1,670 successfully matched respondents, only 19 report a year of birth different than that in the register data, and 12 out of 19 report a year of birth within one year of that recorded in the registers. Excluding these 19 observations has no impact on our results (see appendix). We are thus confident that mismatch is not an issue for our analysis and that we can ignore it as an error component.

SHARE collects a wide array of information, from health to employment status. This paper focuses on total gross income and years of schooling. We choose to study total gross income instead of earnings for three reasons. First, the age composition in our sample is such that a large fraction of the respondents has either zero earnings or zero pension income. Thus studying total income helps us increase our sample size. Second, in Denmark income from employment and pensions are received in the same way, i.e., both are automatically transferred to the resident's primary bank account. Therefore, the two income sources share many of the characteristics (e.g., salience, recollection bias) affecting first-party response error.

Third, both SHARE and the Danish tax authorities collect information on various income sources, which they then sum together. Summing across different income components is robust to income classification discrepancies (survey and validation data classifying sources of income in different ways). For example, although Danish tax authorities separately record bonuses, professional fees, and employment earnings, some respondents might consider them all as employment earnings. By focusing on gross income, we use consistent measures in both survey and register data. Both measures are in Danish Kroner (DKK) in the original datasets.

[7]The data collection agency (SFI-Survey) sent the survey responses together with the CPRs of main SHARE respondents to Statistics Denmark. Using these CPRs, Statistics Denmark linked the survey data to validation data from administrative registers, and replaced the CPRs with scrambled individual identifiers for confidentiality. Only the data collection agency and Statistics Denmark know CPRs, whereas we only observe the scrambled individual identifiers.

[8]SHARE surveys both the sampled individuals and their spouses, if relevant. However, while the data collection agency knows the CPR number of the main respondent, it does not know the spouse's. We retrieve information on the spouses through a cohabitor identification number (CNR). This number is unique for adults who share the same street address at the time of the interview and who are married to one another or in a registered partnership. Non-registered cohabiting couples share a single CNR only if they are of the opposite gender, if their age difference is less than 15 years, and if no other adult lives at the same address.

As is standard practice, we exclude 21 outliers reporting zero income in either the survey (19 observations) or the validation (2 observations) data. While including these observations does not change our broad conclusions, estimates of the measurement error variances and standard errors increase. This selection reduces our sample to 1,649 observations (96.6% of the sample interviewed in the first wave of SHARE Denmark).

Not all data collected in SHARE is directly first-party reported. Proxy interviews and multiple imputations are used whenever the respondent is unable to respond or is uncertain about a source of income. In terms of incidence of proxy interviews, the Danish portion of SHARE is representative of the overall SHARE sample, as Denmark is the median SHARE country by aggregated first-party response rate in both the demographic module and employment and pensions module (96% and 93%, respectively). While in our main analysis we do not distinguish between respondent and proxy interviews, we show in the Appendix that our findings do not change when we restrict the sample to first-party responses.

Whenever the respondent cannot provide a precise assessment of income for the previous year, an unfolding sequence of bracketed response categories starts. Given this information, SHARE provides multiple imputations for each source of the respondent's income (for details on the imputation procedure, see Christelis, 2011). Often only a small portion of a respondent's total income is imputed. While 35% of the matched observations have at least some imputed income, only a quarter of their income is imputed on average (the unconditional proportion of imputed income is roughly 9%).

One of the strengths of our study is the almost complete match of survey respondents with the registers. Thus, to avoid introducing selection on unobservables and to maintain a sample size as large as possible, in the analysis on income measurement error we aggregate multiple imputations by respondent and use their average as if it were a non-imputed response. The consequences of this approach are that the standard errors of our estimators may be downward biased, making our tests more liberal, and that our main analysis corrects jointly for both first-party response error and imputation error. While our preferred model accounts for both types of errors, if imputations introduce additional non-classical measurement error, our sample could be of particularly low quality in terms of mean reversion. To deal with these concerns, in the Appendix we replicate our estimation for income excluding all respondents subject to any form of imputation. However, when we focus on this selected sample, the results from our preferred model do not change.

Our validation data for income are 2003 tax authority records. The 2003 tax year corresponds to the period that the respondents were asked to recall during the SHARE Denmark interviews in March 2004. In Denmark, employment earnings, pensions, and other social transfers are third-party reported, by the employer, pension fund, or the state. Capital income from stocks, bonds, or mutual funds owned through a Danish institution are also third-party reported, thus making the Danish tax register a reliable validation source (Browning and Leth-Petersen, 2003; Kristensen and Westergaard-Nielsen, 2007). Tax returns are posted in April, to be returned with corrections by the end of the month. Tax evasion motives can affect only reports of self-employment income or income from undisclosed second jobs (Kleven et al., 2011).

Our validation data for schooling are official registers used by the Danish government. These are in part self-reports from a census of 1970 and in part updates through institutional reports of qualifications. (For additional information on the census data, see the appendix). We cannot retrieve information about education level in our validation data for all respondents, especially for individuals born before 1921. Therefore, we constrain our analysis to respondents born after 1920 and for whom we observe their education level in the administrative registers. We then exclude non-respondents to the education questions in SHARE Denmark and the single outlier reporting no education (i.e., zero years of schooling). The selection process leaves us with a sample of 1,538 validated observations.

After the census was conducted, information on education qualifications obtained subsequently was reported by a third party. For qualifications obtained in Denmark, the institution providing the education and granting the qualification had to record it and report to the Ministry of Education. The validation data to which we have access contains the most recent education record, implying that any difference in means between survey and validation measures is not driven by missing qualifications obtained after the 1970 census.

In sum, 75% of our sample have the original census record and 25% have an updated institutional report. We compare this data with the 2004 SHARE response. All data sources (SHARE, 1970 census and institution reports) provide the highest achieved level of education, which both Statistics Denmark and SHARE convert to years of schooling as the least number of years in which a specific level of education can be obtained. Because information collected by Statistics Denmark is typically more refined than the one available in SHARE, we recode years of schooling according to the SHARE conversion table, with the exception of vocational education. For vocational education, one cannot map the information collected by Statistics Denmark into SHARE-level data. Therefore, we impute 13 years of schooling (the median years imputed by Statistics Denmark) for all vocational education. (See the appendix for the full conversion table.)

IV. Identifying measurement error parameters

In a typical validation study, we observe two measures of a quantity m, m_s, and m_r, from a survey and a validation source (register data), respectively. A standard assumption is that m_r precisely measures the quantity m, that is, applying the notation defined in equation (1), $\kappa_r = \rho_r = \sigma_r^2 = 0$. If this assumption holds, measurement error in the survey is precisely defined as the difference between the survey measure m_s and the validation measure $m_r = m$. We could then simply regress this difference on the validation measure to identify κ_r, ρ_r, σ_r^2, and σ_m^2, where the latter is the true variance of the

Table 1
Studies assuming no error in the validation data source

Study	BK, 1991		BBDR, 1994		BE, 2008	
Survey	CPS		PSID-VS		HRS	
Val. source	SSA (federal tax)		Employer payrolls		SSA (Medicare)	
Year	1976	1977	1982	1986	1991	2003
$\widehat{\kappa}_s$	-	0.04**	0.007	0.003	0.059**	0.089**
λ_s	0.82	0.84	0.70	0.85	0.68	0.72
$\widehat{\rho}_s$	-0.194	-0.197	-0.172	-0.104	-0.304	-0.173
N	2924	2924	422	320	2670	635

NOTE—* $p < 0.1$, ** $p < 0.05$ for the hypothesis of $\widehat{\kappa}_s = 0$. All reported $\widehat{\rho}_s$ are significantly different than zero at the 5% confidence level. The symbols κ_s, λ_s and ρ_s refer to the notation introduced in equation (1).
SOURCES.—Bound and Krueger (1991) (BK, 1991), Bound et al. (1994) (BBDR, 1994) and Bricker and Engelhardt (2008) (BE, 2008).

quantity of interest and the rest are the parameters characterizing measurement error.

Most validation studies maintain the assumption of perfectly measured validation data. Table 1 lists some key results from three such studies, validating surveys collected over three decades. The first row reports the estimated average differences between survey and validation measures (identifying κ_s under the assumption of perfect validation data) for each of the validation studies, with the exception of 1976 earnings (not reported in Bound and Krueger, 1991). The second row of Table 1 reports the reliability ratios λ_s, estimated in each of these three validation studies. Under the assumption of classical measurement error, λ_s provides an estimate of $\beta/\widehat{\beta}^{OLS}$. However, the third row of the table shows that all these studies find evidence of non-classical, mean-reverting response error.

Bollinger (1998), using the same dataset as in Bound and Krueger (1991), finds that the negative correlation between the cross-sectional difference of survey and validation measures originates primarily from individuals with low earnings (according to the SSA) who report higher values in the survey. A natural question is whether such low earnings reflect the true value that interests the econometrician, or whether the validation dataset lacks information on certain types of unreported earnings, which are instead correctly reported in the survey data (Abowd and Stinson, 2013).

Unreported earnings in the validation data can produce spurious evidence of mean-reverting response error. Even assuming a simple measurement error model such as

$$\begin{aligned} m_s' &= m + \varepsilon_s \\ m_r' &= m + \varepsilon_r \end{aligned} \tag{5}$$

where ρ_s is equal to zero, measurement error calculated as $m_s' - m_r' = \varepsilon_s - \varepsilon_r$ is negatively correlated with m_r', because of the term ε_r appears on both variables.

While the studies in Table 1 acknowledge the possibility that evidence of mean-reverting error can derive from measurement error in the validation data, their authors argue that this should not be the case for the examined data (see Bound and Krueger, 1991 for a discussion). Kapteyn and Ypma (2007) challenge these arguments by estimating a rich error structure that, while maintaining the assumption that the validation data is exactly measured, allows for mismatch between administrative and survey data. They find no evidence of mean-reverting measurement error in labor earnings (and weak evidence in pension income) for a sample of Swedish respondents. Their estimated linear relationships ρ_s between measurement error and the variable of interest are equal to -0.013 and -0.131 for earnings and pension income respectively.

As Section III explains, the structure of our data allows us to ignore the possibility of mismatch. However, although the Danish registers are likely to be at least as precise as the validation datasets used in other validation studies, we relax the assumption of perfect validation data, and we allow for errors in the registry reports of length of schooling and gross annual income. Without this assumption the parameters characterizing measurement error in the survey cannot be identified by a simple comparison of the two measures. We provide additional conditions for identification of those parameters separately for gross income and length of schooling.

A. Gross income

In the Danish tax system, earnings, capital income, and pensions are electronically third-party reported by the employer, a financial institution, or the public administration. These institutions have no incentive for misreporting income by introducing mean-reverting errors (Hyslop and Imbens, 2001). Moreover, as the register measure is high-stakes, individuals have a strong incentive for checking its accuracy. While occasional mistakes can and do occur—e.g., whenever employers or financial institutions incorrectly impute bonuses or income from investments matured in the first or the last months of 2003 to the wrong fiscal year—these errors are not correlated with true income levels. Kleven et al. (2011) study a sample of random tax audits in Denmark. In their sample of over 10,000 audits, a regression of tax audit adjustments on post-audit gross income gives a coefficient of 0.004, with a t-statistic of 0.63 (all variables in logarithms).[9] This result shows that measurement error in Danish administrative income data is classical.

We thus adopt the measurement error structure assumed by Kreiner et al. (2015) i.e., relaxing the assumption of perfect validation data by allowing for classical measurement error. Using the notation in equation (1) we can then write

$$\begin{aligned} m_s &= \mu_m + \kappa_s + (1+\rho_s)(m-\mu_m) + \varepsilon_s, \quad & E\left[\varepsilon_s^2\right] = \sigma_s^2 \\ m_r &= \mu_m + \kappa_r + (m-\mu_m) + \varepsilon_r & E\left[\varepsilon_r^2\right] = \sigma_r^2 \end{aligned} \tag{6}$$

[9] We are grateful to Claus Thustrup Kreiner for running this auxiliary regression for us on the tax audit data used in Kleven et al. (2011).

As Section II shows, measurement error bias in linear models depends only on the parameters ρ_s, σ_s^2, and σ_m^2. While the covariance between m_r and m_s is equal to $\sigma_m^2 (1+\rho_s)$, we cannot separately identify ρ_s and σ_m^2 because of the classical error component in m_r. To identify parameters of the model, we need additional information. Section II shows that IV estimation does not suffer from classical measurement error, which contaminates m_r. Thus IV estimation of the linear relationship between m_s and m_r identifies $1+\rho_s$.

A suitable instrument z_m for this estimation needs only to satisfy the condition

$$Cov\left(z_m, m_r\right) \neq 0 \ \wedge \ z_m \perp \varepsilon_s, \tag{7}$$

because the unobservable component of m_s not explained by m, is simply ε_s. In other words, any instrument that is correlated with m but is otherwise uncorrelated with the measurement error in the survey except through m is a valid instrument. Because our validation data is reported by a third party, we consider any variable measured in the registers correlated with gross income to be a potentially valid instrument. In other words, we fully exploit the assumption that measurement error in the registers (third party) and in the survey are independent of each other conditional on m. Under the assumption that ρ_r is equal to zero (classical measurement error in the registers), this exclusion restriction identifies ρ_s.

Once ρ_s is identified, the second order moments of m_s and m_r identify σ_s^2, σ_r^2, and σ_m^2. Without imposing additional assumptions, we can use GMM to estimate the parameters of interest, using

$$\begin{array}{ll}
1\text{: EXPV}_r & E\left[m_r - \widetilde{\mu}_r\right] = 0 \\
2\text{: EXPV}_s & E\left[m_s - \widetilde{\mu}_s\right] = 0 \\
3\text{: VAR}_r & E\left[\left(m_r - \widetilde{\mu}_r\right)^2 - \sigma_r^2 - \sigma_m^2\right] = 0 \\
4\text{: VAR}_s & E\left[\left(m_s - \widetilde{\mu}_s\right)^2 - \left(1+\rho_s\right)^2 \sigma_m^2 - \sigma_s^2\right] = 0 \\
5\text{: COV}_{sr} & E\left[\left(m_r - \widetilde{\mu}_r\right)\left(m_s - \widetilde{\mu}_s\right) - \left(1+\rho_s\right)\sigma_m^2\right] = 0 \\
6\text{: IV} & E\left[z_m\left(m_s - \left(1+\rho_s\right) m_r - \alpha\right)\right] = 0
\end{array} \tag{8}$$

as a system of moment restrictions. With one instrument for the sixth moment, the model is just identified. However, adding more exclusion restrictions by using more than one instrument for the identification of ρ_s is straightforward.

Our method rests only on the assumptions that third-party measures of z_m are independent of first-party measurement error in the survey except through income, and that measurement error by third parties in our validation study is independent of the variable of interest. While untestable, our first assumption rests on the nature of third-party reports of our instruments, which we assume not to have any direct influence on the first-party measurement error process except through income. Our second assumption is supported by auxiliary tax audit data. Moreover, non-classical measurement error in our validation data would not undermine our estimation strategy but would merely change the interpretation of our estimates for ρ_s, which in this case would converge to $(\rho_s - \rho_r) / (1 + \rho_s)$. This quantity can be interpreted as a measure of relative non-classicality of our survey compared to administrative data.

In comparison to maximum likelihood estimation (Kapteyn and Ypma, 2007), we use only second moments m_s and m_r for identification of the parameters of interest, allowing for flexibility in the measurement error structure and robustness to different distributions of unobservables. Moreover, this method does not require a panel data structure (Abowd and Stinson, 2013).

B. Length of schooling

For years of schooling, mean-reverting errors may arise not only from response error but also from the bounded nature of the variable itself. Moreover, we draw a large portion of our validation data from census self-reports. Thus we cannot argue that measurement error in our validation source for schooling is classical or that an IV regression of m_s on m_r identifies ρ_s.

Since 1970, educational institutions have reported qualifications directly to the Ministry of Education. Because of this third-party reporting and because schooling seldom changes for seniors, we argue that institution reports of schooling are measured without error. We then use this information set to identify the parameters of interest.

Indicating with c whether the source of our validation data is the 1970 census ($c = 1$) or institution reports ($c = 0$), we can write the measurement error model for schooling in the survey and validation data as

$$\begin{aligned} m_s &= (1-c)\left(\mu_{m0} + \kappa_s + (1+\rho_s)(m - \mu_{m0}) + \varepsilon_s\right) \\ &\quad + c\left(\mu_{m1} + \kappa_s + (1+\rho_s)(m - \mu_{m1}) + \varepsilon_s\right) \\ m_r &= (1-c)(m) \\ &\quad + c\left(\mu_{m1} + \kappa_r + (1+\rho_r)(m - \mu_{m1}) + \varepsilon_r\right) \end{aligned} \tag{9}$$

where $\mu_{mj} = E(m \mid c = j)$ for $j \in \{0,1\}$. We allow for dependence between c and m, and we indicate as σ^2_{m1} and σ^2_{m0} the variance of m in the census and institution report samples, respectively. The first and second moments of m_s and m_r can be rearranged as linear functions of c as

$$\begin{aligned} E\left[m_r \mid c\right] &= \mu_{m0} + c\left(\mu_{m1} - \mu_{m0} + \kappa_r\right) = \widetilde{\mu}_{r|c} \\ E\left[m_s \mid c\right] &= \mu_{m0} + \kappa_s + c\left(\mu_{m1} - \mu_{m0}\right) = \widetilde{\mu}_{s|c} \\ E\left[\left(m_r - \widetilde{\mu}_{r|c}\right)^2 \mid c\right] &= \sigma^2_{m0} + c\left((1+\rho_r)^2\sigma^2_{m1} - \sigma^2_{m0} + \sigma^2_r\right) \\ E\left[\left(m_s - \widetilde{\mu}_{s|c}\right)^2 \mid c\right] &= (1+\rho_s)^2\sigma^2_{m0} + \sigma^2_{s0} \\ &\quad + c\left((1+\rho_s)^2\left(\sigma^2_{m1} - \sigma^2_{m0}\right) + \sigma^2_{s1} - \sigma^2_{s0}\right) \\ E\left[\left(m_r - \widetilde{\mu}_{r|c}\right)\left(m_s - \widetilde{\mu}_{s|c}\right) \mid c\right] &= (1+\rho_s)\sigma^2_{m0} \\ &\quad + c(1+\rho_s)\left((1+\rho_r)\sigma^2_{m1} - \sigma^2_{m0}\right) \end{aligned} \tag{10}$$

where $\sigma^2_{sj} = E\left[\varepsilon^2_s \mid c = j\right]$ for $j \in \{0,1\}$.

Given that we observe c, this model has 10 exclusion restrictions and 11 parameters, and is not identified if we allow such heterogeneity in the measurement error structure.[10] Therefore, to estimate this model, we impose the additional assumption $\rho_r = \rho_s$. This assumption derives from the common nature of the survey and the census data, because for the census subsample we observe two responses to similar questions by the same individual at different times. As both are first-party responses, we assume that the nonclassical measurement error component in the two surveys is the same.

Under this additional assumption, the different error structure across the two samples in the two measures provides identification. We interpret our identification strategy as an external validation study within an internal validation study. Specifically, the intercept of the variance of m_r point identifies σ^2_{m0}. Once σ^2_{m0} is known, the intercept of the covariance between m_r and m_s identifies $\rho_s = \rho_r$. Similarly, each of the coefficients in the five linear expressions of the first and second moments of m_r and m_s and their covariance identifies a parameter of the general model.

We estimate the parameters using only this system of moment equations, thus maintaining a high degree of generality in that we do not impose additional distributional assumptions on the stochastic components of our model. A more structured model such as that developed by Kane et al. (1999), adapted to the specific data available, may provide more precise estimates and allow correction for both biases and efficiency losses. However, our framework is more generally applicable, and provides sufficient information for us to apply the general results presented in Section II.

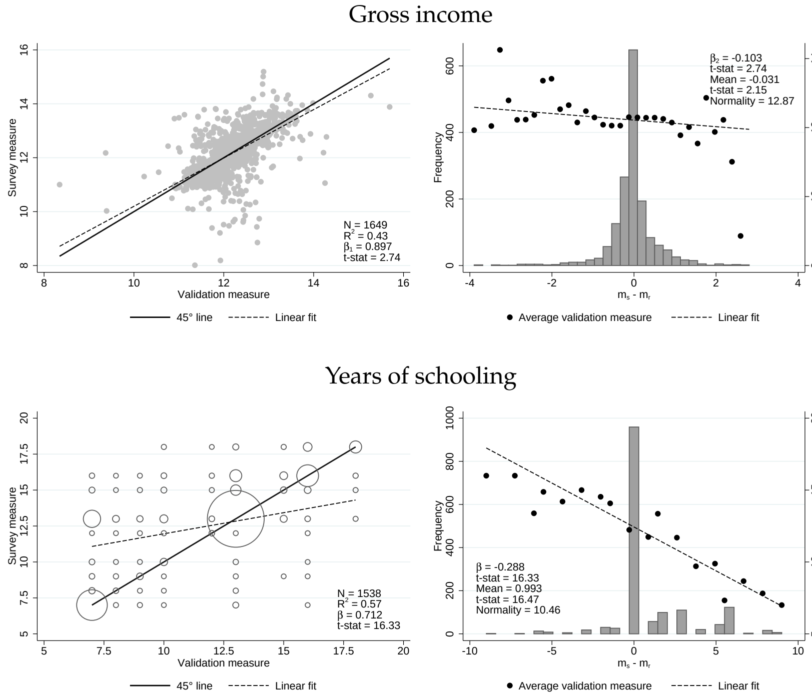
V. Results

To establish a benchmark, we begin by assuming that the validation data always represents the true value of our variable of interest, and we construct our first measure of the error as the difference between the survey and the register values, $m_s - m_r$, as in the studies presented in Table 1. We then compare the estimates of measurement error parameters and bias factors under this restrictive assumption with those obtained from our more general models.

Figure 1 shows the relationship between the two measures as a first analysis of the measurement error in our validated variables, assuming that our validation dataset is always error-free ($m_r = m$). The left panes of Figure 1 show scatterplots of the income and schooling data, where the vertical and horizontal axes represent the survey and the register measures, respectively. We compare the 45° line, shown in solid black, with the OLS regression of m_r on m_s, reporting the corresponding coefficient β_1 in the graph.

The right panes show histograms of the difference in the two measures, where for each bin we report the conditional average of the validation measure in black dots (scale on the right axis). We use these values, frequency

[10] More specifically, we could identify only parameters relative to the comparison within the institution report subsample (ρ_s, σ^2_{s0}, σ^2_{m0} and conditional means).



NOTE.—In the left pane, which shows the scatterplot of survey versus register measure and their linear relationship, we report from top to bottom the sample size, the R^2 and β_1 for the linear regression and the t statistic from a t-test with H_0: $\beta_1 = 1$. The right pane shows the histogram of the difference between the two measures, the average register value for each histogram bin, and the linear regression line between $m_s - m_r$ and m_r. We report from top to bottom the associated β_2, the t statistic from a t-test with H_0: $\beta_2 = 0$, the average difference between measures, the t statistic from a t-test with H_0: $E[m_r] = 0$ and the z-statistic from a Shapiro-Wilk normality test.

FIG. 1.—Measurement error, assuming $m_r = m$

weighted, to show with a dashed line the negative relationship between survey measurement error and register measure. This regression mirrors the OLS regression fitted in the left pane. The β_2 coefficient estimated in the right pane is equal to $\beta_1 - 1$.

Figure 1 highlights both similarities and differences between the measurement error distributions in gross income and years of schooling. In the gross income pane, despite a considerable amount of noise, the data are scattered around the 45° line, shown in solid black. However, the dashed linear prediction line shows that β_1, the estimated OLS coefficient for m_r, is equal to 0.897 and statistically different from unity.

This scatterplot representation further clarifies the intuition behind the consequences of measurement error in the validation variable. As Section II shows, OLS estimators suffer from attenuation bias if the independent variable is measured with error. Thus a coefficient of 0.897 might result from a negative correlation between measurement error in the survey and gross income, measurement error in the validation data—with variance equal to 11.5% of the variance of gross income or a combination of the two.

The top right pane of Figure 1 suggests that the negative correlation is largely due to outliers in the difference in measures distribution. Such a finding is common in the measurement error literature (Bollinger, 1998; Bound et al., 2001; Kristensen and Westergaard-Nielsen, 2007). However, because these observations could not be detected as outliers without a validation source, such results are not useful outside a validation study context.

While the panes for years of schooling tell a similar story, all magnitudes are amplified. In the bottom left pane, the area of each circle is proportional to the number of observations sharing a particular combination of survey and validation measure. The largest cell, corresponding to 13 years of schooling (vocational education) in both measures, has 452 observations (29% of the sample). The scatterplot shows that large deviations from the validation variable are more common above the 45° line, suggesting a strong negative correlation between the difference in measurements and the validation measure.

This negative correlation is especially evident in the bottom right pane of the figure, because schooling is a bounded variable and, people with few years of schooling can only err upwards. Therefore, people cannot have random response errors if their true level of education is at the boundaries of the distribution. This simple analysis, which graphically replicates a standard validation study and maintains the assumption of perfect validation data, suggests that both schooling and income are subject to various degrees of non-classical measurement error.

We now drop the assumption of observing the true value in the validation dataset and turn to the models described in equations (6) and (9). For gross income, we allow for non-classical measurement error in the survey data and classical measurement error in the validation data. Although for years of schooling, both survey and validation measures can be contaminated with measurement error, but exploit the knowledge that part of our validation dataset comes from third-party reports, which we assume to be error-free because of the static, stock nature of length of schooling for seniors. [11]

Table 2 presents the results from the GMM estimations and highlights the differences in estimates with the models in which we assume $m_r = m$. Columns 1 and 3 of Table 2 show the results for the models in which we impose a restrictive error structure as a benchmark, by assuming that the validation measure represents the true value of the income variable. Columns

[11] While it is possible that the institution-reported information does not record education achievements obtained abroad unless converted into a Danish degree, only three individuals say, when interviewed in SHARE Denmark, that they obtained an education abroad. Of those three, only one has a much lower value in the register measure (13 years) than in the survey measure (18 years). Excluding those three observations from the analysis does not change our results (see appendix).

Table 2
GMM estimates of measurement error characteristics

	Gross income		Years of schooling	
σ^2_m	0.318**	0.280**		
	(0.0175)	(0.0220)		
σ^2_{m1}			11.39**	15.02**
			(0.286)	(1.792)
σ^2_{m0}			4.686**	4.647**
			(0.291)	(0.296)
σ^2_s	0.340**	0.306**		
	(0.0262)	(0.0315)		
σ^2_{s1}			5.477**	3.089**
			(0.232)	(0.212)
σ^2_{s0}			2.012**	1.950**
			(0.221)	(0.240)
σ^2_r		0.0375**		3.507**
		(0.0181)		(0.249)
ρ_s	-0.103**	0.0194	-0.301**	-0.276**
	(0.0375)	(0.0710)	(0.0186)	(0.0406)
Observations	1649	1649	1538	1538
F. stat		129.3		
Hansen's J p-val.		0.606		
OLS bias	0.479	0.477	0.726	0.964
IV bias	1.115	0.981	1.431	1.381
LHS bias	0.897	1.019	0.699	0.724

NOTE—* $p < 0.1$, ** $p < 0.05$. Robust standard errors in parenthesis. The table reports estimated parameters of the measurement error models described in section IV for both gross income and years of schooling. For each variable we compare results from the restricted model (assuming perfect validation data) with those from the general model in which we allow for measurement error in the validation dataset. Given the point estimates of the measurement error parameters, the bottom pane of the table summarizes the expected bias factors.

2 and 4 show the point estimates from the more general models presented in Section IV. Then, at the bottom of the table, we use the point estimates of the measurement error parameters to compute the expected bias factors according to equations (2), (3), and (4).

In the first column, consistent with the evidence presented in Figure 1, we estimate a negative correlation, significantly different from zero, between measurement error in gross income and its true value. This apparent non-classicality is not trivial, as it suggests that all results estimating returns on gross income are biased and underestimated by about 10%.[12] However, we find that this negative correlation is a construct in our data due to the presence of mild measurement error in the validation dataset.

The second column shows how the results change if we allow the validation data to be measured with error. The model underlying the estimates in Column 1 is a special case of this general model. As discussed in Section IV, to estimate this model we need instruments that, while correlated with income, are otherwise independent of the survey measurement error component ε_s. Our instruments are thus third-party reported variables we draw from register data–gender and the logarithm of assets at the end of 2003. These variables are likely to be strongly correlated with income, and we assume that they are independent of survey measurement error except through income. The instruments have a high F-statistic, and they pass an overidentification test for exogeneity.

Using this identification strategy, we find no evidence of a correlation between measurement error in the survey and the true value of income. We provide evidence that the negative correlation between $m_s - m_r$ and m_r is due to moderate measurement error in m_r. When we allow for non-zero ρ_s and σ_r^2, our estimate of the expected bias in OLS estimators in SHARE Denmark data decreases to less than 50%. However, we do not expect any bias while using m_s as a dependent variable or as an independent variable in an IV estimation.

We estimate the measurement error properties in years of schooling by splitting our sample according to the source of validation data by the strategy outlined in Section IV. While our application is specific to the nature of our validation variable, the strategy itself is generally applicable. For example, validation studies are often only able to match a potentially non-random portion of the survey data to a precise validation source, while for the rest of the sample another imperfect information source is available (e.g., a subsequent wave of a survey). In these cases, our strategy can be appropriate for assessing the measurement error properties of variables that can be assumed to be measured correctly in the validation dataset, not varying across survey waves (e.g., completed schooling for seniors). For estimating measurement error characteristics in both samples, assuming that at least one of these characteristics is constant across the subsamples is sufficient.

Table 2 compares the measurement error properties estimated under the assumption that the validation dataset is correct (column 3) with those estimated by treating only institutional reports as precise (column 4). That is, we allow for non-classical measurement error in the 1970 census portion of our validation data, and we identify measurement error parameters by performing an external-within-internal validation. The two subsamples do not need to be randomly selected. In the census subsample, which is disproportionately representative of older cohorts, both mean and variance of years of schooling are lower than in the reports from educational institutions.

While measurement error in our validation data leads to overestimation

[12]Using labor market earnings only, instead of gross income, we estimate an even stronger bias factor equal to 0.73. (See table A.5 in the appendix.)

of the variance of the classical measurement error component in the survey, we find limited evidence that it confounds the estimate of ρ_s. Due to the bounded nature of years of schooling as a variable, measurement error is strongly non-classical, even allowing for imperfect validation data, with severe consequences for the estimation of linear models using survey data. According to our estimates (column 4), while we expect a very small bias from OLS estimators due to the composite effect of non-classical and classical measurement error, IV estimators overestimate the true returns to education by 38% in our sample.

Meaurement error alone explains approximately a 42% difference between the OLS and IV estimators. However, due to non-classical measurement error, this difference is almost completely explained not by, as is commonly thought, attenuation bias in the OLS estimates but by a strong amplification bias in the IV estimator.

VI. Conclusions

Measurement error is pervasive in both surveys and administrative data. Yet most validation studies assume no measurement error in their validation source. A standard treatment of measurement error assumes it to be classical for an explanatory variable, leading to attenuation bias in OLS estimates that can be corrected by IV. However, if measurement error in an independent variable is non-classical and is correlated with the true value of the quantity of interest, an IV estimator will be biased: A positive correlation leads to IV attenuation bias, and a negative correlation leads to IV amplification bias. This result implies that because measurement error in discrete or bounded variables tends to be negatively correlated with the true value of the quantity of interest, program evaluation techniques relying on IV estimators on average overestimate treatment effects if the treatment dummy is measured with error.

We show that ignoring errors in the validation data leads to incorrectly inferring non-classical measurement error in a validated variable. We build a framework that allows us to estimate the sufficient statistics determining measurement error bias in IV and OLS estimators of linear models through imperfectly measured validation data for length of schooling and gross income. In contrast to most validation studies, once we allow for imperfect validation measures, we find evidence of classical measurement error in gross income when comparing SHARE with Danish administrative registers. The substantial noise in the survey measure of gross income does not cause bias when income is the dependent variable or when using IV estimators. When income is the outcome of interest, we find that measurement error only affects the efficiency of linear estimators.

We acknowledge that because years of schooling is a bounded variable, its measurement error is likely to be non-classical, even if it is independent of the response error generation process. As expected, we find that measurement error in years of schooling is strongly negatively correlated with the true value of length of schooling, even when we allow for imperfect validation data. We show that in our sample, while measurement error alone accounts for a 42% increase between an OLS and an IV estimate on the same sample in the absence of omitted variable bias, the vast majority of this bias is due to the amplification bias affecting the IV estimates and not, as commonly thought, to attenuation bias in OLS estimates. In the absence of heterogeneous treatment effects, this result changes the interpretation of the phenomenon for which most empirical studies report higher IV estimates of returns to schooling than OLS estimate, even though omitted variable bias typically runs in the opposite direction.

The general and flexible approach that we develop can be tailored according to the type of validation data available for assessing measurement error in other variables and in other contexts. While our approach does not provide a sufficiently detailed description of the measurement error generation process for correcting for efficiency losses, it allows identification of those characteristics that are sufficient for determining the bias in OLS and IV estimation of linear models. Our approach extends the classical validation study techniques used in the labor economics literature, because the traditional approach–assuming that the validation data are measured without error–is a special case of our model. Moreover, our framework can test this assumption. Further research should consider the consequences of other types of non-classical measurement error, for example, correlation of measurement error with other explanatory variables in the model.

References

Abowd, John M, and Martha H Stinson (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics 95*, no. 5:1451–1467.

Aigner, Dennis J (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics 1*, no. 1:49–59.

Barron, John M., Mark C. Berger, and Dan A. Black (1997). How well do we measure training? *Journal of Labor Economics 15*, no. 3:507–528.

Biemer, Paul P., Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman (Eds.) (2004). *Measurement errors in Surveys*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Bollinger, Christopher R. (1998). Measurement error in the current population survey: A nonparametric look. *Journal of Labor Economics 16*, no. 3:576–594.

Börsch-Supan, Axel, Martina Brandt, Christian Hunkler, Thorsten Kneip, Julie Korbmacher, Frederic Malter, Barbara Schaan, Stephanie Stuck, and Sabrina Zuber (2013). Data resource profile: The survey of health, ageing and retirement in europe (share). *International Journal of Epidemiology*.

Bound, John, Charles Brown, Greg J. Duncan, and Willard L. Rodgers (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics 12*, no. 3:345–368.

Bound, John, Charles Brown, and Nancy Mathiowetz (2001). Measurement error in survey data. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5, Chapter Chapter 59, 3705 – 3843. Elsevier.

Bound, John, and Alan B. Krueger (1991). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics 9*, no. 1:1–24.

Bricker, Jesse, and Gary V. Engelhardt (2008). Measurement error in earnings data in the health and retirement study. *Journal of Economic & Social Measurement 33*, no. 1:39 – 61.

Browning, Martin, and Søren Leth-Petersen (2003). Imputing consumption from income and wealth information*. *The Economic Journal 113*, no. 488:F282–F301.

Card, David (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*, no. 5:1127–1160.

Christelis, Dimitris (2011). Imputation of missing data in waves 1 and 2 of share. Working Paper Series 01-2011, SHARE - Survey of Health, Ageing and Retirement in Europe.

Duncan, Greg J., and Daniel H. Hill (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics 3*, no. 4:508–532.

Fuller, Wayne A. (1987). *Measurement Error Models*. Wiley series in probability and mathematical statistics. New York: John Wiley and Sons.

Hausman, Jerry (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *The Journal of Economic Perspectives 15*, no. 4:57–67.

Hyslop, Dean R., and Guido W. Imbens (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics 19*, no. 4:475–481.

Jensen, Vibeke M., and Astrid W. Rasmussen (2011). Danish education registers. *Scandinavian Journal of Public Health 39*, no. 7 suppl:91–94.

Kane, Thomas J., Cecilia Elena Rouse, and Douglas Staiger (1999). Estimating returns to schooling when schooling is misreported. Working Paper 7235, National Bureau of Economic Research.

Kapteyn, Arie, and Jelmer Y. Ypma (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics 25*, no. 3:513–551.

Kleven, Henrik Jacobsen, Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica 79*, no. 3:651–692.

Kreiner, Claus Thustrup, David Dreyer Lassen, and Søren Leth-Petersen (2015). Measuring the accuracy of survey responses using administrative register data: Evidence from Denmark. In J. S. Christopher Carroll, Thomas Crossley (Ed.), *Improving the Measurement of Household Consumption Expenditures*, NBER Book Series Studies in Income and Wealth. University of Chicago Press.

Kristensen, Nicolai, and Niels Westergaard-Nielsen (2007). A large-scale validation study of measurement errors in longitudinal survey data. *Journal of Economic and Social Measurement 32*, no. 2:65–92.

Mellow, Wesley, and Hal Sider (1983). Accuracy of response in labor market surveys: Evidence and implications. *Journal of Labor Economics 1*, no. 4:331–344.

Pischke, Jörn-Steffen (1995). Measurement error and earnings dynamics: Some estimates from the psid validation study. *Journal of Business & Economic Statistics 13*, no. 3:305–314.

Statistics Denmark (1977). Folke- og boligtællingen 9. november 1970: C.4. uddannelse (housing and population census 9th november 1970: Section 4: Education). Technical report, Statistics Denmark.

Stefanski, Leonard A. (1985). The effects of measurement error on parameter estimation. *Biometrika 72*, no. 3:583–592.

Stefanski, L. A. (2000). Measurement error models. *Journal of the American Statistical Association 95*, no. 452:1353–1358.

Appendix

A. Survey and census questions

In this section we report the questions originally asked in the first wave of SHARE and how the variables for education and gross household income were constructed. For further information and the exact Danish wording, we refer to the SHARE guideline and country-specific questionnaires available at `www.share-project.org`.

a. Education questions in the 1970 Census

The central registration of education in Denmark began with the general population and housing census of November 9th, 1970, when all residents of Denmark had to respond using their CPR numbers. The census asked 13 housing questions and 13 people questions, three of which were about schooling. These were under the heading "Education and vocational training status"[13]. Five pages of instructions were followed for the later coding of the education responses, with the objective of placing the written responses to each of the three education questions into a 3-digit coding frame (Statistics Denmark, 1977).

Educational qualifications received abroad are not recorded in the administrative registers unless converted into an equivalent Danish degree. Only three of the SHARE Denmark sample state they have a foreign qualification. For immigrants, Statistics Denmark conducted a schooling census in 1999, and has since surveyed new immigrants at two year intervals. There are 42 immigrants in the SHARE Denmark sample and we consider this source of information as self-reports similar to the 1970 census.

We hereby report the official English translation of the census questions regarding education level:

Section B. Education and vocational training status To be filled in for all persons who have turned 14, but not 70 years (i.e. born between November 9th, 1900 and November 8th, 1956)

6 Education or vocational training in progress
Persons who are **not** in process of education or vocational training, write: none For **school pupils** (i.e. up to and including secondary level) the class is to be listed, eg. 7th class, "2nd real", "1.g" **apprentices and trainees** should list this and the trade, eg, bricklayer's apprentice, cabinet maker's apprentice, traffic trainee, bank trainee For **students and others receiving an education**, the kind of education is to be listed as accurately as possible, eg.university student with language major or the like, correspondent - 3 languages, laboratory technician training,

[13]The first question was about education or vocational training in progress; the second about completed schooling and the third, about completed education or vocational training. See appendix.

teacher's training, specialist teacher's training, agricultural school student.

7 Completed schooling
For persons who have left school, the highest examination passed is to be listed, e.g. "mellemskoleeks" (i.e. exam after 9 years of schooling), "realeks" (i.e. exam after 10 years of school), "nyspr. student" (i.e. exam after 12 years of school with language major), "HF" (i.e. exam after 11 years of school) or highest class in school which has been completed, e.g. 7th school year, 9th class, "2. real" (i.e. 10 years of school). For persons who have attended school abroad, the corresponding information is to be listed, the total number of years in school, and name of the country

8 Completed education or vocational training
This space is also to be filled in by persons who are economically inactive. The most important education or vocational training or further training is to be listed. For persons with an **exam or school leaving certificate** from university, higher school, or the like, the kind of education is to be listed as accurately as possible, e.g. university degree in languages or the like, university degree in engineering, degree from technical engineering school (college), chartered accountant, "HA" (i.e. degree from school of business and economics), school teacher, social worker. For persons with **apprentice's training or other vocational training**, the vocation is to be listed, e.g. electrician, trained office clerk, book seller's assistant, skilled baker, nurse, assistant nurse, technical assistant, laboratory worker, agricultural technician, catering officer. For persons whose vocational training is entirely practical this is to be listed and the nature of the work, e.g. practical office training, practical agricultural training. Persons without completed education or training including school pupils should write: none.

b. Education in SHARE and recoding into years of schooling

The questions from which we draw information about education are those in module DV of SHARE wave 1, named DN010_ and DN012_ in the questionnaires. Table A.1 shows the Danish wording of the options, the corresponding English translation, and the associated imputed standard years of schooling, which we also use to recode data from the census and institutional reports.

c. Gross income in SHARE

Gross income is the sum of a list of variables, each capturing a different portion of the income process of an individual, each asked separately to the financial respondent(s). Table A.2 shows the variables that form gross income and their source within the questionnaires. All questions refer to previous year income.

Table A.1
SHARE questions and years of schooling recoding

Danish	English	Years of schooling
Please look at card 2. What is the highest school leaving certificate or school degree that you have obtained?		
7. klasse	7^{th} grade or lower	7
8. klasse	8^{th} grade	8
9. klasse	9^{th} grade	9
10. klasse, realeksamen	10^{th} grade	10
Studentereksamen eller HF	Gymnasium	12
HH, HG, HHX, HTX	Technical secondary	12
Please look at card 3. Which degrees of higher education or vocational training do you have?		
Specialarbejderuddannelse	Vocational	13
Lærlinge eller EFG-uddannelse	Vocational	13
Anden faglig uddann. $>$ 12 mdr.	Vocational $>$ 12 months	13
Kort videregående uddannelse	Higher education ($<$3y)	15
Mellemlang videregående uddannelse	Higher education (3-4y)	16
Lang videregående uddannelse	Higher education ($>$4y)	18

Table A.2
SHARE gross income components

Variable	Question	Description
ydipv	ep205	Annual gross income from employment previous year
yindv	ep207	Annual gross income from self-employment previous year
ybaccv	as005	Interest income from bank accounts
ybondv	as009	Interest income from bonds
ystocv	as015	Dividends from stocks/shares
ymutfv	as058	Interest and dividend income from mutual funds
yrentv	ho030	Income from rent
yltcv	ep086	Monthly long-term care insurance previous year
pen1v	ep078_1	Monthly public old age pension
pen2v	ep078_3	Monthly public early or pre-retirement pension
pen3v	ep078_4	Monthly main public DI pension, or sickness benefits
pen4v	ep078_6	Monthly public unemployment benefit or insurance
pen5v	ep078_7	Monthly public survivor pension from partner
pen7v	ep078_9	Monthly war pension
pen8v	ep324_1	Monthly private (occupational) old age pension
pen9v	ep324_4	Monthly private (occupational) early retirement pension
pen10v	ep324_5	Monthly private (occupational) disability insurance
pen11v	ep324_6	Monthly private (occupational) survivor pension from partner's job
reg1v	ep094_1	Monthly life insurance payment received
reg2v	ep094_2	Monthly private annuity or private personal pension
reg4v	ep094_4	Monthly alimony received
reg5v	ep094_5	Monthly regular payments from charities received

See Christelis (2011) for more information

B. Relationship between difference in measurements and survey and validation variables

Table A.3
Income measurement error dependence, assuming $x_r = x$

	(1)	(2)	(3)	(4)
Register variables				
Income	-0.103**	-0.119**	-0.387**	-0.125**
	(0.0375)	(0.0397)	(0.0506)	(0.0180)
Assets		0.00619	0.0128**	0.00372
		(0.00531)	(0.00488)	(0.00266)
Female		-0.0700	-0.0377	-0.0642**
		(0.0559)	(0.0529)	(0.0242)
Couple		-0.125	-0.0169	-0.0375
		(0.112)	(0.107)	(0.0507)
Female×couple		0.0228	-0.101	0.00936
		(0.0647)	(0.0638)	(0.0317)
Survey variables				
Financial respondent			0.0527	0.0349*
			(0.0340)	(0.0203)
Age			-0.0104**	-0.00377**
			(0.00207)	(0.000911)
Labor income share			0.229**	0.118**
			(0.0593)	(0.0241)
Imputed income (dummy)			0.0960**	0.0282
			(0.0368)	(0.0187)
Imputed income share			-0.0394	-0.0165
			(0.0665)	(0.0337)
Years of schooling			0.0240**	0.0103**
			(0.00471)	(0.00234)
Observations	1649	1649	1638	1638
Adjusted R^2	0.009	0.015	0.092	

NOTE—* $p < 0.1$, ** $p < 0.05$. The table shows the estimated coefficients of a linear regression model of $x_s - x_r$ on different sets of covariates. In the first two columns we use only register measures as covariates. In the third and fourth column we include regressors drawn from survey data.

C. Robustness checks

Table A.4
GMM estimation of length of education measurement error model, robustness checks

	Baseline	**Add zero-outlier**	**No educ. abroad**	**No born abroad**
σ^2_{m1}	15.02**	15.23**	15.15**	14.30**
	(1.792)	(1.825)	(1.822)	(1.559)
σ^2_{m0}	4.647**	4.647**	4.635**	4.524**
	(0.296)	(0.296)	(0.297)	(0.287)
σ^2_{s1}	3.089**	3.099**	3.089**	3.133**
	(0.212)	(0.212)	(0.212)	(0.217)
σ^2_{s0}	1.950**	1.950**	1.954**	1.864**
	(0.240)	(0.240)	(0.241)	(0.232)
σ^2_{r}	3.507**	3.493**	3.507**	3.360**
	(0.249)	(0.248)	(0.249)	(0.242)
ρ_s	-0.276**	-0.276**	-0.279**	-0.259**
	(0.0406)	(0.0406)	(0.0408)	(0.0374)
Observations	1538	1539	1535	1489
OLS bias	0.964	0.967	0.968	0.941
IV bias	1.381	1.381	1.387	1.349
LHS bias	0.724	0.724	0.721	0.741

NOTE—* $p < 0.1$, ** $p < 0.05$. Robust standard errors in parenthesis. The table replicates the results of the tfourth column of Table 2 for different samples. The first column of Table A.4 reports the results from our general model as a term of comparison. The second column adds to the sample the single observation reporting zero years of education, which we exclude for graphical presentation. The third and fourth columns exclude respondents stating having received a qualification abroad and born abroad respectively. One of the respondents receiving a qualification abroad was born in Denmark.

Table A.5
GMM estimation of income measurement error model, robustness checks

	No proxy		No imputations		No mismatch		Only earnings	
σ_x^2	0.317**	0.273**	0.311**	0.290**	0.320**	0.281**	0.191**	0.136**
	(0.0185)	(0.0219)	(0.0193)	(0.0242)	(0.0177)	(0.0222)	(0.0225)	(0.0191)
σ_s^2	0.339**	0.298**	0.282**	0.261**	0.337**	0.303**	0.239**	0.199**
	(0.0273)	(0.0322)	(0.0313)	(0.0366)	(0.0264)	(0.0318)	(0.0309)	(0.0323)
σ_r^2		0.0443**		0.0204		0.0377**		0.0490**
		(0.0179)		(0.0188)		(0.0182)		(0.0199)
ρ_s	-0.0976**	0.0502	-0.0199	0.0491	-0.100**	0.0223	-0.271**	0.00149
	(0.0395)	(0.0724)	(0.0431)	(0.0759)	(0.0377)	(0.0714)	(0.0727)	(0.119)
Observations	1550	1550	1088	1088	1630	1630	745	745
F-statistic		124.7		85.09		127.0		60.91
Hansen's J p-val.		0.755		0.994		0.628		0.0367
OLS bias factor	0.480	0.479	0.525	0.524	0.483	0.481	0.408	0.406
IV bias factor	1.108	0.952	1.020	0.953	1.111	0.978	1.371	0.999
LHS bias factor	0.902	1.050	0.980	1.049	0.900	1.022	0.729	1.001

NOTE—* $p < 0.1$, ** $p < 0.05$. Robust standard errors in parenthesis. The table reports estimated parameters of the measurement error models described in section IV for gross income according to various sample selection or variable creation criteria. In the columns 1 and 2 we exclude proxy reports. In columns 3 and 4 we exclude incomes that have been at least partially imputed. In columns 5 and 6 we exclude observations whose year of birth differs in the survey and validation datasets. In column 7 and 8 we apply the model to gross earnings only, for those individuals that report a non-zero amount. For each sample we compare results from the restricted model (assuming perfect validation data) with those from the general model in which we allow for measurement error in the validation dataset. Given the point estimates of the measurement error parameters, the bottom pane of the table summarizes the expected bias factors.