Examples of combining administrative records and survey data in validation studies

Claus Thustrup Kreiner†
David Dreyer Lassen‡
Søren Leth-Petersen§

Version 161111

Abstract

This paper shows how Danish administrative register data can be combined with survey data at the person level and be used to validate information collected in the survey. Register data are collected by automatic third party reporting and the potential errors associated with the two data sources are therefore plausibly orthogonal. Two examples are given to illustrate the potential of combining survey and register data. In the first example expenditure survey records with information about total expenditure are merged with income tax records holding information about income and wealth. Income and wealth data are used to impute total expenditure which is then compared to the survey measure. Results suggest that the two measures match each other well on average. In the second example we compare responses to a one-shot recall question about total gross personal income (collected in another survey) with tax records. Tax records hold detailed information about different types of income and this makes it possible to test if errors in the survey question are related to the reporting of particular types of income. Results show bias in the mean and that the survey error has substantial variance. Results also show that the error is correlated with conventional covariates suggesting that the error is not of the classical type. The latter example illustrates how Denmark can be used as a "laboratory" for future validation studies. Tax records with detailed information about different types of income are available for the entire Danish population and can be readily merged to survey data. This makes it possible to test the ability of respondents to accurately report different types of income using different interviewing techniques and questions. The examples presented in this paper are based on cross section data. However, the possibility to issue surveys repeatedly to the same persons and linking up to longitudinal tax records provides an opportunity to learn more about the time series properties of measurement errors, a subject about which little evidence exist, in the future.

Acknowledgements: The Danish Social Science Research Council provided support for the work presented in this paper. We are also thankful to Amalie Sofie Jensen and Gregers Nytoft Rasmussen for competent research assistance.

† Department of Economics, University of Copenhagen, Øster Farimagsgade 5, building 26, DK-1353 Copenhagen. Email: claus.thustrup.kreiner@econ.ku.dk.
‡ Department of Economics, University of Copenhagen, Øster Farimagsgade 5, building 26, DK-1353 Copenhagen. Email: david.dreyer.lassen@econ.ku.dk.
§ Department of Economics, University of Copenhagen, Øster Farimagsgade 5, building 26, DK-1353 Copenhagen, Denmark and The Danish National Centre for Social Research, Herluf Trolles Gade 11, DK-1052 Copenhagen, and AKF, Danish Institute of Governmental Research, Købmagergade 22, DK-1150 Copenhagen. Email: soren.leth-petersen@econ.ku.dk.

1. Introduction

Danish administrative register data can readily be combined at the person level with survey data. This makes it possible to compare survey based measures directly with corresponding measures based on information from administrative registers. Because register information is collected by third-party automatic reporting and completely independently from the survey collection we believe this provides cheap and powerful way to validate survey measures.

The objective of this paper is to illustrate how Danish register and survey data may be combined at the person or household level and used for validating measures collected by survey, and we illustrate the potential of this methodology by two examples. In the first example we use administrative records about disposable income and wealth to validate the total expenditure measure collected in the Danish family expenditure survey and in the second example we use third party reported information about gross personal income to validate a survey measure of gross personal income. Information about gross personal income relies entirely on third-party automatically reported information and is thought to be very close to the "truth" and the validation exercise therefore relies on few assumptions. Validating total expenditure requires making more assumptions as the register measure of total expenditure is it self ridden with error. The most important assumption is that the errors of the two measures are uncorrelated. This is not likely to be a restrictive assumption since the data are collected from completely independent sources.

The next section will outline the Danish institutional setup facilitating the collection of administrative register data and the merging of register and survey records. Section 3 will outline the analytical framework that we use to asses the importance of measurement error in the survey data. Section 4 will show how income-tax records with information about income, tax payments and wealth have been used to impute a measure of total household expenditure that is then matched at the household level to data from the Danish expenditure survey in order to check how well the total expenditure measure in the survey matches the register based imputation. The analysis presented in that section complements the analysis presented in Browning and Leth-Petersen (2003) and is based on the same data. In section 5 we will combine income-tax records with new survey data containing a measure of total gross personal income to directly validate the survey measure of gross income. Section 5 sums up and gives perspectives for future validation studies based on combining Danish register and survey data.

2. Administrative register data and matching with survey data

All persons in Denmark are assigned a unique personal identification number (CPR). This number is used by all government institutions to store person specific information including information relevant for taxation, for example the information on the tax return, but also information about contacts to the health system, the educational system, and about family composition and place of residence, so that household units can be constructed. Many administrative registers, including population registers and income tax registers, are collected by Statistics Denmark who merge them and give access to researchers working at authorized Danish research institutions. The data are confidential and are kept on servers at Statistics Denmark, and they are accessed under comprehensive security precautions. The data must be kept at the servers and only aggregated numbers such as regression coefficients can be extracted.

The register data have many outstanding features, but some of the features that are most important in this context is that it covers the entire population and contains tax records with third party reported information about income and wealth. In this study we shall rely on register data from the income tax registers to validate survey information about spending and income. The income-tax register is collected by the tax authorities in order to calculate the amount of taxes to be paid by all persons in Denmark by the end of each calendar year. The tax authorities collect information from many sources. Most important for this study are earnings and employers pension contributions collected directly from employers, information about transfer income from government institutions, and information about interest payments/income and the value assets and liabilities by the end of the year which is collected directly from banks. In a new study Kleven et al. (2011) have conducted a large scale randomized tax auditing experiment in collaboration with the Danish tax authorities and documented that the level of tax evasion is small, in particular among wage earners, in Denmark. This means that the third party reported income information collected by the tax authorities is of very high quality.

The tax authorities use the information for different purposes. Information about earnings and capital income is preprinted on the tax return whereas wealth information is used to cross check if the reported level of income is consistent with the level of asset accumulation from one year to the next. While the tax authorities collects this information at a high level of detail corresponding to individual entries at the tax return for income and at the account level for wealth, this information is in some cases only transferred to Statistics Denmark's research data base as summary variables, for example as the sum of earnings from different employers, and for some capital income sub-components only net income is available. Apart from covering the entire population and being based on third party reported information the income-tax registers also has the attractive feature that income and wealth

information is not top coded and longitudinal information can be retrieved back to 1980 for some variables.

Another crucial feature, for the present purpose, is that it is possible to link to survey data via the CPR number. Matching surveys with register data is done at relatively low cost. As an example the survey used in example 2 consists of 40 questions, it was collected by telephone interview and includes 6,000 completed interviews. The sample was randomized from the population based on register data covering the entire population, and the survey data was merged on to register data after collection. The total costs were about 200,000 USD.[1]

3. Analytical framework

There are several ways of summarizing the accuracy of the survey data. In this paper we focus on the magnitude of the attenuation bias in OLS regressions of the register measure on the survey measure. The analytical setup is a generalization of the setup presented by Bound and Krueger (1991).

Consider

$$z^S = z^* + u^S \qquad (1)$$

$$z^R = z^* + u^R \qquad (2)$$

Where z^S is the observed survey based measure, z^* is the true but unobserved measure, and u^S is the survey measurement error. Correspondingly, z^R is the observed register based measure, and u^R is the register measurement error. All variables are measured in natural logarithms.[2] This amounts to assuming that the measurement error is multiplicative in levels. Subscripts identifying that each observation of $\left(z^*, z^R, z^S, u^R, u^S\right)$ pertains to an individual are suppressed. While in one of the examples, the case of gross income, we believe that the register based measure is very close to the "truth", this is obviously not the case in the other example where we compare total expenditure from survey and imputed from the register data. The framework presented lays out the assumptions necessary to validate the survey measure.

1 A number of survey agencies are specialized in conducting surveys and linking to administrative register data. Two of those are SFI survey (http://www.sfi.dk/Default.aspx?ID=2832) and Epinion (www.epinion.dk) who have collected the survey used in example 2. Also Statistics Denmark (www.dst.dk) conduct surveys that can subsequently me merged on to register data.

2 The analytical framework, of course, does not require that the variables are measured in logarithms.

Assume

$$\operatorname{cov}\left(z^{*},u^{R}\right)=0 \qquad (A.1)$$

$$\operatorname{cov}\left(u^{S},u^{R}\right)=0 \qquad (A.2)$$

(A.1) assumes that the error of the register measure is uncorrelated with the true level. This assumption is not testable and may in some cases be a reasonable assumption while in others it may not be. (A.2) assumes that the error of the survey measure is uncorrelated with the error of the register measure. This seems to be a reasonable assumption in both of the examples as will be discussed in connection with each example.

Consider a regression of z^{R} on the true but unobserved measure z^{*}.

$$z^{R}=\delta_{0}+\delta_{1}z^{*}+u^{R} \qquad (3)$$

Now substitute in the survey measure for the true measure

$$z^{R}=\delta_{0}+\delta_{1}z^{S}+u^{R}-\delta_{1}u^{S} \qquad (4)$$

Using A.1 and A.2 the probability limit of the OLS estimator of δ_{1} can then be written

$$p\lim\widehat{\delta_{1}}=\delta_{1}\lambda \qquad (5)$$

Where $\lambda=\operatorname{cov}\left(z^{R},z^{S}\right)\Big/\operatorname{var}\left(z^{S}\right)$ is just the OLS regression of the register measure on the survey measure. The bias due to the measurement error in the survey measure is then $\left(1-\lambda\right)$.

This expression covers the case with classical measurement error where u^{S} are *iid* and $\lambda=\sigma_{u^{S}}^{2}\Big/\left(\sigma_{z^{*}}^{2}+\sigma_{u^{S}}^{2}\right)$ but is not limited to this special case. In particular, the present framework is more general because it allows for cases where errors are not *iid*.

4. Example 1: Total Expenditure

Total expenditure is one of the most important variables collected in expenditure surveys and this variable is central in numerous studies of demand and intertemporal consumption allocation. However, there is little evidence on the quality of the information collected in expenditure surveys. In Denmark it is possible to link the household level information from the Danish Family expenditure survey to administrative income tax records including third party reported information about income and wealth that can be used to impute total expenditure.

3.1 data

The sample used consists of the households entering the Danish Family expenditure survey (DES) 1994-1996. The households in this survey have been contacted at different times of the year so that observations are distributed across the calendar year. Each household has participated in a comprehensive interview, where they have answered questions about holdings of durables and about purchase of durables within the past 12 months from the interview date. Furthermore, each household has kept a diary for two weeks, where they have kept a detailed account of all expenditures in the household. This information is scaled to obtain an expression of annual consumption.

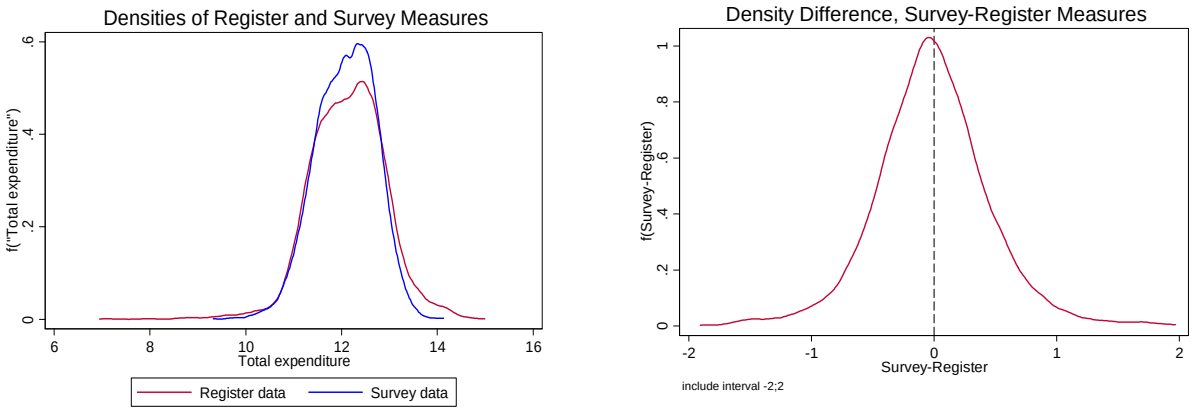
For the households entering the DES administrative register, data are collected on income, tax payments and wealth at the end of the year (corresponding to the survey year) together with wealth information for the previous year, and this is merged with the DES data. Total expenditure is then imputed from the income and wealth information by simply calculating $\widehat{c_t} = y_t - \Delta W_t$, where y_t is disposable income and W_t is net wealth measured at the end of period t. While simple in theory there are many details involved in implementing this and we refer to Browning and Leth-Petersen (2003) for details. For the analysis we use the same sample selection criteria as Browning and Leth-Petersen (2003). This leaves us with a sample of 3,352 observations

3.2 Results

We start out by presenting in figure 1 two graphs of the distribution of the two measures of total expenditure and their individual level difference. The left panel show that the distributions have modal points very close to each other and the right panel show that at the individual level the differences are centered at zero. It is, however, also evident that there are important differences in the spread of the distributions of the two measures with the register based measure exhibiting larger dispersion. The way the data are constructed implies that a fair amount of noise is expected. First, the interviews are distributed across the calendar year, and this means that recall questions about durable

purchases, for example, do not necessarily pertain to the calendar year. Moreover, Browning and Leth-Petersen (2003) show that the measurement error in the imputed measure is related to capital gains on wealth components used in the imputation.

Figure 1. Densities of the survey and register based measures of total expenditure and of the individual differences.



Note: the right panel includes only data in the interval -2;2. 32 observations are selected away

Figure 2 plots the data together with the diagonal and a nonparametric regression line. If the survey and the register measure coincided all points would be located on the diagonal. The blue line is a nonparamtric regression through the data cloud and comparing its' slope to the diagonal shows the attenuation bias. One thing to notice is that the blue line is almost linear and it is also noticeable that the bias is apparent.

Figure 2 presents a non parametric regression of the register based measure on the survey measure.

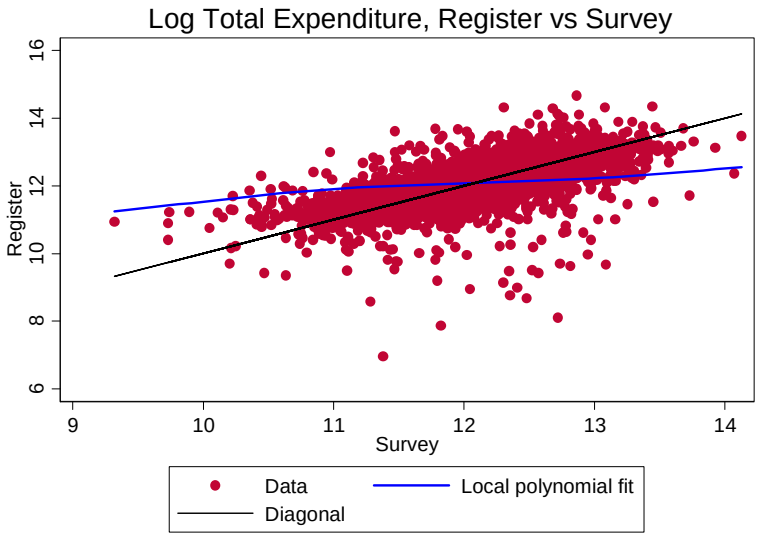


Table 2 presents the results from estimating the regression line by OLS. The estimate in column (1) shows that the bias is 0.21, which suggests a fair amount of noise in the survey measure. Restricting the size of the errors does not change the estimate much indicating that the bias is not caused by outliers. Of course, concluding that the survey measure is noisy relies on assumptions 1 and 2 being correct, in particular that the measurement error of the two measures is uncorrelated. Since errors in the survey are related to the accurateness of the survey response and the register error is related to capital gains on the portfolio this assumption does not appear restrictive. The assumption that the register error be uncorrelated with the true (but unobserved level) is not testable, however.

Table 2. Estimates of λ

	(1) $z^S - z^R$ unrestricted		(2) $-2 < z^S - z^R < 2$	
λ	0.791	***	0.816	***
	(0.0148)		(0.0128)	
Constant	2.519	***	2.237	***
	(0.1792)		(0.1546)	
N	3,352		3,320	

* p<0.05, ** p<0.01, *** p<0.001

3.3 Summary, example 1

In this example the possibility to construct a register based measure of total expenditure that can be compared with the survey measure is illustrated. While the validity of this exercise hinges on two important assumptions, we believe that the register approach provides a cheap way to get some insight about the precision of the survey measure that can not otherwise be obtained easily.

5. Validating survey questions about gross income using third party reported information from the income tax registers

A variable included in almost any survey collected by social scientists is income, and surveying is often the only way to collect income jointly with other variables of interest. Danish register data on income are of very high quality because they are automatically third party reported and are reported separately for different types of income. In this section we compare the responses to a one-shot recall question about gross personal income collected by telephone interview in January 2010 to the tax records of the respondents in order to assess the quality of the survey measure. As opposed to the previous example the register information is now perceived to be close to the "truth". We therefore expect to be relying much less on assumption (A.1), and (A.2)

4.1 data

In January 2010 the authors of this paper organized a telephone survey including 6004 completed interviews. The purpose of the survey was to obtain information about their response to a stimulus policy implemented in 2009. The sample is drawn randomly from the population of persons in employment at some point in the period 1998-2003 including about 3.9 mill persons or about 75% of the Danish population. As part of the survey respondents were asked a one-shot recall question about their gross annual income in 2009. The question was:

"*We are also interested in knowing about the development in your income before taxes. We are thinking about income such as earnings (incl. employers pension contribution), pension payments, payments from unemployment insurances, cash benefits or other forms of transfer income. What was approximately your income before taxes in 2009?*"

5.394 persons answered the question. Self-employed persons effectively self-report income to the tax authorities and we therefore do not have as much faith in the register information for this group as we have for wage earners and persons receiving transfer income. We therefore deselect persons with own-business income. Finally, we deselect two observations with negative gross income[3] and are left with 4,793 observations. The survey data were subsequently merged at the person level with administrative register data about income from the income tax register covering the tax year 2009, i.e. exactly the same period that the survey question was intended to cover.

4.2 Main results

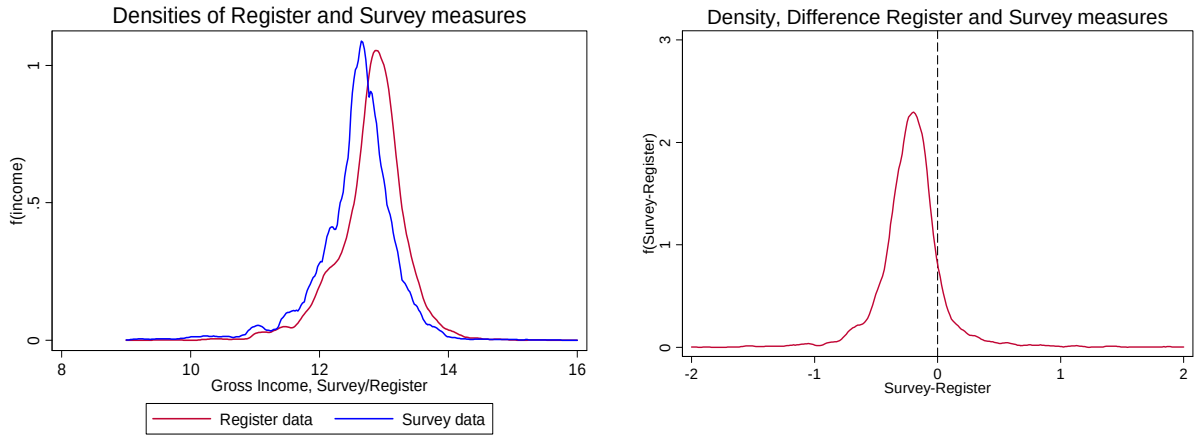
Figure 4.1 presents the densities of the survey and the register measure (left panel) and the density of the individual level differences between the two measures (right panel). The left panel clearly reveals that the means of the two measures are not equal. It also suggests that the spread of the survey measure is larger as would be expected if the survey measure carries an error and the register measure is accurate. The right panel confirms that the means are different when individual level differences are considered and also that individual level errors have considerable spread, i.e. that the survey measure is noisy.

Figure 4.2 graphs the register measure against the survey measure together with a smooth line through the data and a diagonal. The picture shows some very large outliers and also that the regression line has a smaller slope than the diagonal indicating that the attenuation appears to be considerable.[4]

[3] Negative gross income can occur because some components of capital income are available in our data set only as net measures and therefore adds negatively to gross income if the net-value is negative. This seems to be a small problem in the data set. For most people the major capital expenditure components are constituted by interest payments on bank debt and mortgages. Interest payments on bank and mortgage debt are observed, and when we take these components out 62 cases are observed with negative capital income and half of these observations have negative capital income within 1000USD. We therefore conclude that this is a minor problem.

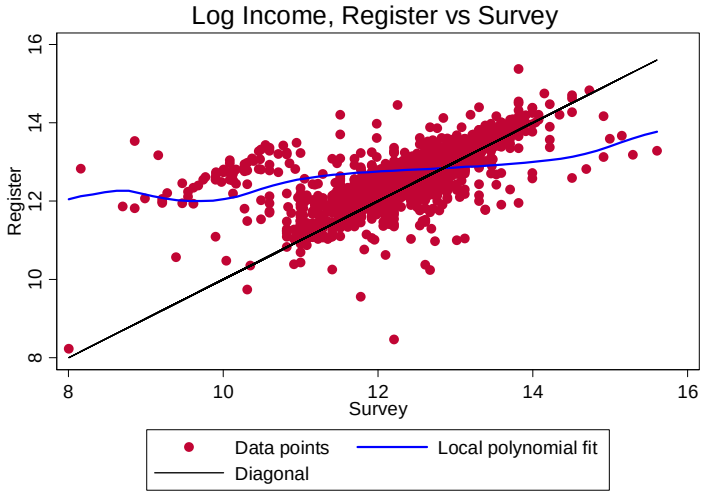
[4] There is a graphically striking cluster of data points in the north eastern corner of the graph appearing to fall along a fairly tight regression line that is different from the mass of the data points. The apparent importance of this cluster is a visual deception because the cloud consists of only 64 observations. We have not been able to identify any significant differences between these observations and the rest of the data set apart from finding that they are on average 4 years younger than the rest of the sample.

Figure 4.1. Densities of the survey and register based measures of total expenditure and of the individual differences.



Note: the right panel includes only data in the interval -2;2. 60 observations are selected away.

Figure 4.2 presents a non parametric regression of the register based measure on the survey measure.



Note: the graph includes only observations in the interval 8;16.

This is confirmed by a parametric regression reported in table 4.1. Regressing the register measure on the survey measure on the unrestricted sample yields an estimate of λ of 0.57 indicating important individual level deviations between the survey and the register measure. In column (2) the errors are restricted to be within the -2;2 interval and this increases the estimate of λ to 0.84 suggesting that a limited number of outliers are responsible for a large part of the attenuation bias.

Table 4.1. Estimates of λ

	(1) $z^S - z^R$ unrestricted	(2) $-2 < z^S - z^R < 2$
λ	0.570 ***	0.835 ***
	(0.0081)	(0.0072)
Constant	5.651 ***	2.283 ***
	(0.1024)	(0.0912)
N	4,793	4,707

* p<0.05, ** p<0.01, *** p<0.001

In table 4.2 the two measures of gross income and the individual level errors are regressed on a set of "external" covariates. The idea is to see how the noise influence the covariance with other variables often used in empirical analyses. Comparing the numbers in column (1) and (2) in table 4.2 suggests that the register and the survey measure have similar covariance with the set of external variables, but the parameter estimates obtained using the survey measure do differ significantly from the parameter estimates obtained using the register measure for age, for woman, for number of children, for single, and for owner. Regressing the individual level error on the same set of covariates suggests differences for the same variables. If one take the register measure to be the "truth" then the results of column (3) suggests that the measurement error associated with the survey measure is not classical.

Table 4.2. Regressing on external covariates

	(1) z^R	(2) z^S	(3) $z^S - z^R$
Age	0.095***	0.114***	0.019***
	(0.0049)	(0.0066)	(0.0052)
Age^2	-0.001***	-0.001***	-0.000***
	(0.0001)	(0.0001)	(0.0001)
Woman	-0.145***	-0.232***	-0.087***
	(0.0127)	(0.0172)	(0.0136)
Single	-0.020	-0.067**	-0.047**
	(0.0162)	(0.0219)	(0.0173)
Number of children	0.012	-0.009	-0.021**
	(0.0072)	(0.0098)	(0.0077)
Education, short	0.113***	0.147***	0.034*
	(0.0160)	(0.0216)	(0.0171)
Education medium	0.257***	0.288***	0.031
	(0.0188)	(0.0255)	(0.0202)
Education long	0.362***	0.362***	-0.000
	(0.0241)	(0.0327)	(0.0258)
House owner	0.356***	0.245***	-0.111***
	(0.0147)	(0.0198)	(0.0157)
Constant	10.466***	9.954***	-0.512***
	(0.0980)	(0.1327)	(0.1049)
N	4,793	4,793	4,793

* p<0.05, ** p<0.01, *** p<0.001. Standard errors in parentheses.

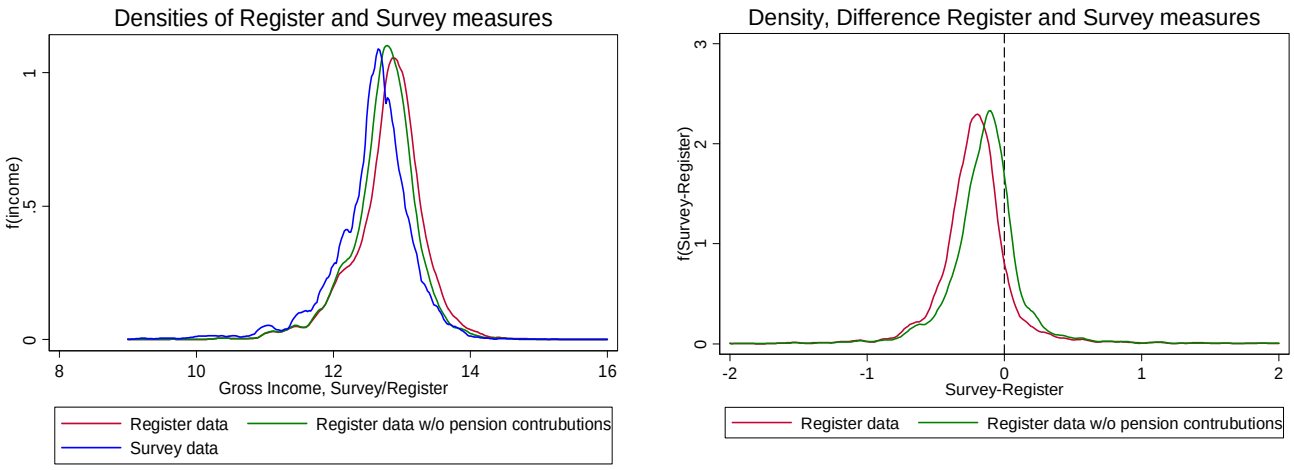
4.3 Robustness

The survey question ask people to recall gross income including earnings, employers pension contributions, transfer income, and capital income. A couple of these are not salient to the respondent, for example employers pension contributions. This number does not appear separately on the pay check nor on the tax return or the annual statement from the tax authorities since it is not liable to taxation before it is paid out. In a robustness check we subtract employers pension contribution from the register measure and repeat the analysis from the previous section to check if the survey measure perform better when compared to the adjusted register measure. To do this we define an alternative gross income measure constructed from the registers where employers pension contributions are deducted from the register measure used in the previous section. The idea is to investigate if

respondents are more likely to have stated their income without employers pension contributions even though it is clearly stated in the survey question that it should be included.

Figure 4.3 shows density graphs for the survey measure, of the original register measure, and of the register measure where employers pension contributions are subtracted. The right panel shows densities of differences between the register measures and the survey measure.

Figure 4.3. Densities of the survey measure, of the original register measure and of the register measure where employers pension contributions are subtracted and of the individual differences between the register measures and the survey measure.



The figure shows that subtracting employers pension contributions reduces the mean bias, but also that the spread is almost unaffected. Estimating λ by OLS reveals that this has not improved on the precision at the individual level. In fact, the estimates of λ in table 4.3 suggest that the attenuation bias has become more serious, if anything.

Further, examining the correlation with the external covariates reveals even stronger correlations between the external covariates and the differences between the register and the survey measure. These results are presented in table 4.4. In the regression of individual level differences between the two measures, column (3), on external covariates, the parameters are now more significant. In particular, education dummies are now more significant. This suggests that the ability of the respondens to include employers pension contributions when reporting their income vary across educational levels.

Table 4.3. Estimates of λ for register measure without employers pension contributions

	(1) $z^S - z^R$ unrestricted	(2) $-2 < z^S - z^R < 2$
λ	0.528***	0.773***
	(0.0077)	(0.0069)
Constant	6.093***	2.995***
	(0.0971)	(0.0872)
N	4,793	4,707

* p<0.05, ** p<0.01, *** p<0.001

Table 4.4. Regressing on external covariates

	(1) Register	(2) Survey	(3) u_reg2
Age	0.086***	0.114***	0.028***
	(0.0046)	(0.0066)	(0.0053)
Age^2	-0.001***	-0.001***	-0.000***
	(0.0001)	(0.0001)	(0.0001)
Woman	-0.133***	-0.232***	-0.098***
	(0.0119)	(0.0172)	(0.0137)
Single	-0.007	-0.067**	-0.059***
	(0.0152)	(0.0219)	(0.0174)
Number of children	0.015*	-0.009	-0.024**
	(0.0068)	(0.0098)	(0.0078)
Education short	0.102***	0.147***	0.045**
	(0.0150)	(0.0216)	(0.0172)
Education medium	0.237***	0.288***	0.051*
	(0.0176)	(0.0255)	(0.0202)
Education long	0.323***	0.362***	0.039
	(0.0226)	(0.0327)	(0.0259)
Owner	0.347***	0.245***	-0.102***
	(0.0137)	(0.0198)	(0.0157)
Constant	10.563***	9.954***	-0.608***
	(0.0918)	(0.1327)	(0.1053)
N	4793	4793	4793

* p<0.05, ** p<0.01, *** p<0.001. Standard errors in parentheses.

It is, of course, possible to construct many other concepts where other income components are subtracted. The most salient feature of income is arguably earnings and transfer income which is regular and where the recipient receives a letter notifying about the amount paid out. Capital income is less regular and may therefore also be difficult to give an account for. We have experimented with also subtracting capital income from the register measure. This made little difference to the results and we therefore do not report the results from that check. The calculations given in this paper are merely examples intended to illustrate the possibilities to identify different subcomponents of income and how this may be used to identify what components of income respondents find difficult answering.

4.4 Comparison to US findings

CEX and PSID use annual recall questions about income but ask about different income components separately, for example earned income, transfer income, capital income. Bound et al. (1994) have performed a validation study of the earnings question in the PSID by comparing answers to the PSID questions about earnings with company records for 418 workers from a single manufacturing company in 1983. This sample is called the PSID Validation Study. They find that the mean difference between the survey and the register measure is small but that the standard deviation of the difference is substantial amounting to 0.67 of the standard deviation of the company records. The corresponding measure in our data is 0.89. The slope coefficient from a regression of record on interview measure are similar between the two studies, 0.76 in the Bound et al study and 0.84 in the trimmed version in our study.[5]

Both studies suggest substantial measurement error. One shortcoming of the Bound et al. study is that it is confined to validate the survey responses for a homogenous and small group. This could explain the smaller error. Our results suggest that the survey error is correlated with standard covariates and that the error is therefore not of the classical type. This leaves the possibility that validation studies based on narrowly defined samples such as the PSID Validation Study do not give a complete picture of the size of the error in the main sample.

Another difference is that our study focus on gross income including transfer income and capital income. This leaves the possibility that our income measure is more noisy only because we include

[5] Bound et al. (2001) survey nine studies validating survey based earnings measures from different US surveys against administrative records. Four of these studies report regression coeffients from a regression of the administrative record measure on the survey measure and three of these report regression coefficients in the vicinity of 0.75 using different data sets.

non-earned income. Gottschalk and Moffitt (2011) show evidence about the development of transitory family non labor income from the PSID, but to our knowledge the measures of transfer income and capital income in the PSID have not been validated.

Overall the results from the present study and the study by Bound et al. have implications for studies in many areas, but perhaps in particular for the interpretation of estimates from studies decomposing income variances in to temporary and permanent components. The validation results suggest that there is considerable noise in survey measures. This may explain why studies estimating income processes on US data collected in different ways find different results. Specifically, in a series of papers Gottschalk and Moffitt (1994) and Moffitt and Gottschalk (2002, 2011) use the PSID to decompose income in to permanent and transitory variations and find that the transitory component is relatively big and increasing in the 1980s. For example, Moffitt and Gottschalk (2002) find that the variance of transitory log earnings for males is around 0.15-0.3. Kopczuk, Saez and Song (2010) use Social Security Administration longitudinal earnings data for the period 1937-2004 and find that the transitory component is almost constant across time and relatively small, about 0.06-0.08 for the whole period and about 0.06 for the period 1980-, and that it cannot explain the increase in the variance of log earnings in the US during the 1980s. While there are many other differences between these studies than the data collection mode this does open for the possibility that the size of the measurement error is important and not constant across time.

4.5 Summary, example 2

The analysis of the quality of the recall question about annual gross income revealed that a one-shot recall question is inaccurate. Respondents tend to underreport their income level and the survey measure is noisy. Changing the definition of the register measure by excluding employers pension contributions corrected for some of the mean bias, but did not reduce the spread much, and in particular did not reduce the attenuation bias in a regression of the register measure on the survey measure. The analysis also suggested that the individual level differences between the survey and the register measure were correlated with observed characteristics of the respondents suggesting that the errors associated with the survey measure are not of the classical type.

Summary and suggestions for future work

This paper has provided two examples illustrating how Danish third-party reported register data can be matched at the individual or household level to survey records and used to validate the accuracy of responses to survey questions. The first example suggests that expenditure survey evidence on total expenditure is mean unbiased but noisy, and the second example suggests that a one-shot recall question about annual gross income is both mean biased and noisy.

The analyses presented in this paper are possible because all persons living in Denmark are assigned a unique identification number to which all public authorities link up person specific information and because surveys can be collected using the same person identifier. The potential of this validation methodology is big. In the Danish context, it is possible to match survey and register data for (potentially) the entire population, and it is also possible to match in the longitudinal dimension. In this way Denmark can be thought of as a "laboratory" where much more detailed and focused validation studies can be organized and where the impact of survey methodology on the accuracy of the survey responses are investigated so as to optimize the survey methodology across different groups and balancing this with the survey costs. For example, in the context of validating income questions the Danish setup allows to merge survey records with tax records containing detailed information about different types of income and this provides a unique opportunity to test the ability of respondents to accurately report different types of income using different interviewing techniques and questions. Using the register data it is also possible to consider individual as well as household units and to assess to what extent is important to ask all household members or just one member in order to assess household income accurately.

This study focused on cross sections of Danish households and persons. The Danish setup also allows asking the same people repeatedly and to match with panel data on income and wealth. Very little is known about the time series properties of measurement error in recall data. Bound et al (1994) used panel data on earnings for the PSID validation study, but this is limited in size and only concerns a very narrowly defined group of people for two years. The Danish setup is much broader in scope since it potentially covers the entire Danish population with longitudinal information from the administrative registers. This provides a unique opportunity to learn about the time series properties of survey errors in the future. For example, it should be critical to understand if the size of the survey error is constant across time, if it always over-/undershoots at the individual level, if the error is mean reverting but persistent, etc.. The survey used in example 2 in this study has been repeated to cover questions concerning income in 2010 and will be repeated to cover 2011 through to 2013. When register data have been updated we will be able to examine the time series properties of the survey errors.

The Danish setup allows matching new survey data with register data relatively easy and at relatively low costs. Matched survey and register data are kept at Statistics Denmark's servers and only researchers working at authorized Danish research institutions can get access to work with the matched data. However, researchers or statistical agencies with good research questions and appropriate funding wishing to start new research projects using combined survey and register data can do that in collaboration with Danish based researchers. This can for example be done by contacting one of the authors of this paper.

References

John Bound and Alan B. Krueger (1991); "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?"; *Journal of Labor Economics*, Vol. 9, No. 1 (Jan., 1991), pp. 1-24.

John Bound, Charles Brown, Greg J. Duncan, Willard L. Rodgers (1994); "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data"; *Journal of Labor Economics*, Vol. 12, No. 3 (Jul., 1994), pp. 345-368

John Bound, Charles Brown, Nancy Mathiowetz (2001); "Measurement Error in Survey Data", Handbook of Econometrics, Volume 5, chapter59, Edited by J Heckman and E Leamer.

Martin Browning and Søren Leth-Petersen (2003); "Imputing Consumption from income and wealth information". *The Economic Journal*, Vol. 113: Issue 488, F282-301, June 2003.

Henrik Jacobsen Kleven, Martin 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, 2011, 651–692

Wojciech Kopczuk, Emmanuel Saez and Jae Song (2010); "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data since 1937" (with). *Quarterly Journal of Economics*, 2010, 125(1), 91-128

Peter Gottschalk, Robert Moffitt (1994); "The Growth of Earnings Instability in the U.S. Labor Market"; Brookings Papers on Economic Activity, Vol. 1994, No. 2 (1994), pp. 217-272

Peter Gottschalk and Robert A. Moffitt (2011); "The Rising Instability of U.S. Earnings: Addendum"; Manuscript. (http://www.econ2.jhu.edu/people/moffitt/jep%20addendum%20v1%202011-7-17.pdf)

Robert A. Moffitt and Peter Gottschalk (2011); "Trends in the Transitory Variance of Male Earnings in the U.S., 1970-2004"; forthcoming *Journal of Human Resources*.

Robert A. Moffitt and Peter Gottschalk (2002); "Trends in the Transitory Variance of Earnings in the United States"; *The Economic Journal*, Vol. 112, No. 478, Conference Papers (Mar., 2002), pp. C68-C73

Appendix: Sample statistics

Table A.1. Sample statistics, expenditure survey and expenditure imputation from register data from example 1

	Register	Survey	Survey-Register
N	3.352	3.352	3.352
mean	12.077	12.085	0.008
variance	0.518	0.381	0.296
min	6.951	9.3.02	-21.490
p1	10.106	10.550	-13.331
p50	12.105	12.121	-0.0210
p99	13.626	13.371	18.046
max	14.660	14.127	46.236
iqr	0.988	0.875	0.551

Table A.2. Sample statistics, Income survey and income register data from example 2

	Register	Survey	Survey-Register
N	4793	4793	4793
mean	12.804	12.561	-0.243
variance	0.282	0.439	0.221
min	8.236	2.485	-8.934
p1	11.180	10.275	-0.254
p50	12.861	12.612	-0.214
p99	13.988	13.816	0.880
max	15.375	17.148	3.739
iqr	0.547	0.575	0.238