ISER Working Paper Series

Income underreporting based on income-expenditure gaps: survey vs tax records

Alari Paulus
Institute for Social and Economic Research
University of Essex

No. 2015-15
June 2015

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

www.iser.essex.ac.uk

Non-technical summary

The paper provides evidence on income tax compliance at the household level. It estimates underreporting of employment income on the basis of household expenditure and income patterns for Estonia, following and extending the Pissarides and Weber (1989) approach. The method assumes that expenditures are accurately reported and have a robust relationship with household incomes. Unusual income-expenditure gaps therefore imply income underreporting.

While the method has been widely applied, there are two important limitations characterising previous applications. First, most of earlier studies have focussed on income underreporting by the self-employed, relying on employees as a reference group, who are assumed to report their incomes correctly. We relax this assumption by allowing for income misreporting among employees as well. Specifically, we argue that public sector employees can be assumed to be restricted in their choice to non-comply with taxes, while the same does not hold for private sector employees. Second, the method has been previously applied mainly to household survey data, implicitly assuming that household reporting behaviour is similar for tax and survey purposes. We are able to directly assess the validity of the second assumption by utilising a dataset, where household income survey information has been linked with tax reports.

Applying the method in a standard way – using all employees as the reference group and relying on survey income information – leads to the detection of sizeable income underreporting by the self-employed (20% on average). Distinguishing between public and private sector employees, yields a slight higher estimate for the self-employed (25%) but no significant underreporting for private employees. Our key findings, however, indicate that the extent of underreporting is much higher for register income (56%) and also substantial among private employees (23%). Estimating the model on different samples and adjusting key variable definitions, confirms the robustness of our results.

Our study indicates that people are much more truthful reporting their incomes in the survey. Applying this method on survey data alone is unlikely to reveal the full scale of tax non-compliance and previous studies may have underestimated the extent of non-compliance by a substantial margin. While lower income underreporting among private employees compared with the self-employed could be attributed to third-party reporting mechanism, the fact that non-compliance for private employees is found to be significant highlights the limitations of such mechanism. As self-employed income accounts for a relatively small share of total earnings, it calls for more attention to wages and salaries as their underreporting even by a modest proportion could result in a greater loss of tax revenue. Overall, our study contributes to extending empirical evidence on tax compliance outside the US and especially among the post-socialist countries.

Income underreporting based on income-expenditure gaps: survey vs tax records *

Alari Paulus

Institute for Social & Economic Research (ISER)
University of Essex

June 10, 2015

Abstract

We estimate the extent of income underreporting among working households, using data from an income survey linked with individual tax records for Estonia. Income underreporting is inferred from consumption propensities, following and extending the method by Pissarides and Weber (1989). Our dataset allows us to assess the validity of the key assumption in related studies that survey income corresponds to income reported to the tax authority. Our results show large underreporting of earnings by the self-employed and also substantial underreporting of earnings by private sector employees on the basis of register income, while a much smaller scale of non-compliance is detected for self-employed and no underreporting for private employees using survey incomes. This suggests that previous studies applying this methodology to survey data have underestimated the extent of non-compliance.

Keywords: tax compliance, tax reports, income survey, expenditure, Estonia
JEL codes: H26, H31

*Acknowledgements: The paper uses the 2007 and 2008 wave of the Estonian Social Survey (Estonian SILC), linked with administrative tax records and made available by Statistics Estonia. I gratefully acknowledge comments and suggestions from Mike Brewer, Holly Sutherland, Steve Pudney and financial support by the Economic and Social Research Council (ESRC) through the Research Centre on Micro-Social Change (MiSoC), award no. RES-518-28-001. Any errors and the views expressed in this paper are the author's sole responsibility.

1 Introduction

Reliable empirical evidence on tax non-compliance is difficult to obtain due to the very nature of the phenomenon. Apart from costly tax audits, various statistical methods have been developed to estimate income underreporting using micro-data from diverse sources. In this paper, we estimate the extent of income underreporting for Estonia following the approach in a well-known study by Pissarides and Weber (1989), PW for short. They seek to detect unreported income on the basis of consumption propensities, contrasting a particular population group with another for which incomes are assumed to be accurately measured. More specifically, using survey data PW estimate the extent of income underreporting among the self-employed in the UK, taking employees as a reference group and comparing their food expenditure. Their key assumptions are that both groups report their food expenditure correctly in the survey, survey income corresponds to the income reported to the tax authority, employees report their income accurately in the survey and that the marginal propensity to consume with respect to (permanent) income does not differ between the two groups (after controlling for household characteristics).

We improve on the Pissarides and Weber (1989) method in two ways. First, as demonstrated in Paulus (2015), a substantial part of salaries and wages can also be underreported, and hence, the original PW approach should be interpreted as estimating underreporting of income by the self-employed *relative* to employees, rather than in absolute terms. Instead of relying on all wage earners, as is commonly done in previous studies, we base our reference group on public sector employees, allowing us to estimate income underreporting not only for self-employed but also for employees working in the private sector. Besim and Jenkins (2005) were the first to try this for North Cyprus in a simplified approach (with survey data), while we introduce this extension in the full PW framework. Second, to interpret income underreporting in a survey more broadly as tax non-compliance, it needs to be established that there is no systematic variation between different population groups in the way their income in the survey compares to incomes in the tax reports. We explore the validity of the latter assumption by estimating income underreporting both with survey incomes and incomes declared to the tax authority (or register income), using a dataset for Estonia which links these two sources for the same individuals. While the PW method has been applied before to register incomes[1], to the author's knowledge, this is the first analysis carried out with a dataset containing

[1]Feldman and Slemrod (2007) use only information from tax returns for the US. Johansson (2005) and Engström and Holmlund (2009) use household surveys for Finland and Sweden, respectively, where income information has been added from administrative sources.

both types of income and, hence, is able to offer a direct comparison of results.[2] We further extend the empirical literature on tax non-compliance geographically with evidence for an Eastern European country, complementing a recent study by Kukk and Staehr (2014) who assess underreporting of self-employment income using the Estonian Household Budget Survey. Whereas previous studies have relied mainly on food expenditure, due to data limitations, we use instead information on housing related consumption expenditure (mostly utilities). We believe our results are not critically affected by this, for reasons discussed below.

Our results show large underreporting of earnings by the self-employed and also substantial underreporting of earnings by private sector employees on the basis of housing expenditures and register income, while a much smaller scale of non-compliance is detected for self-employed and no underreporting for private employees using survey incomes. This suggests that previous studies applying the PW methodology to survey data have underestimated the extent of non-compliance.

The paper proceeds as follows. The next section explains the methodology, starting with the Pissarides and Weber (1989) approach and modifications in later studies, before presenting our approach. Section 3 provides an overview of the data sources, sample selection, expenditure and income information, and descriptive statistics. Section 4 presents findings and Section 5 concludes.

2 Methodology

We first provide a methodological overview, outlining the original approach by Pissarides and Weber (1989) (Section 2.1) and discussing alternative specifications used in later applications (Section 2.2). Different approaches have yielded similar or even identical measures of income underreporting, but it is important to understand differences in their underlying assumptions when comparing results. We base our approach (Section 2.3) on the PW model, but estimate a slightly different measure of income underreporting following Hurst et al. (2014). We show that this particular form (applied to the standard PW model) is preferable to the one PW used as its estimation requires fewer assumptions. We then explore some of the main assumptions in more detail (Section 2.4).

2.1 The Pissarides-Weber approach

The general idea behind the Pissarides and Weber (1989) method is to infer income underreporting from contrasting expenditure and income patterns for different pop-

[2]Linking survey data with tax records also helps to overcome a serious limitation of the latter arising from generally very limited socio-demographic information.

ulation subgroups, assuming that marginal consumption propensities are identical and the reference group reports income correctly. Intuitively, one population group is used to estimate an expenditure function, which is then inverted to predict incomes for another group and compared with their reported incomes. PW studied income underreporting by the self-employed in the UK using food expenditure.[3]

The starting point is an expenditure function (Engel curve), relating log consumption expenditure on particular goods or services (c_i) by household i to household log *permanent* income (y_i^P) and a vector of household characteristics (z_i):

$$\ln c_i = z_i\alpha + \beta \ln y_i^P + \epsilon_i \tag{1}$$

where α is a vector of parameters, β the elasticity of consumption with respect to permanent income and ϵ_i a random term with zero mean and constant variance. Household *current* (true) income y_i^T fluctuates around the permanent income and is usually not (directly) observed:

$$y_i^T = p_i y_i^P \tag{2}$$

Instead, households report a measure of their current income (in the survey), which can differ from its actual value:

$$k_i y_i = y_i^T \tag{3}$$

with k_i denoting the adjustment (or scaling) factor needed to obtain the true income from the reported income. It is further assumed that p_i and k_i are stochastic terms distributed log-normally, that is

$$\ln p_i = \mu_p + u_i \tag{4}$$

$$\ln k_i = \mu_k + v_i \tag{5}$$

where μ_p and μ_k are mean log values, and u_i and v_i have zero means and constant variances σ_u^2 and σ_v^2. Combining equations (2) to (5), we obtain

$$\ln y_i = (\mu_p - \mu_k) + (u_i - v_i) + \ln y_i^P \tag{6}$$

Substituting this into (1) leads to

$$\ln c_i = z_i\alpha + \beta \ln y_i - \beta(\mu_p - \mu_k) - \beta(u_i - v_i) + \epsilon_i \tag{7}$$

[3]Similarly, (food) Engel curves have been used in other contexts, for example, to measure biases in consumer price indices (Hamilton, 2001) and purchasing power parities (Almås, 2012), impute total expenditure into income surveys (Blundell et al., 2008) and estimate household material living standards (Larsen, 2009).

The identification strategy is based on the assumption that one can distinguish between two population sub-groups: individuals in group A (e.g. employees) report all their income correctly, that is $k_i = 1\ \forall i \in A$ (and hence $\mu_{k_A} = 0$ and $\sigma^2_{v_A} = 0$), while individuals in group B (e.g. self-employed) may underreport (or overreport) their income. It is also assumed that parameters α and β in the expenditure function and the mean of p_i (i.e. $\bar{p}_A = \bar{p}_B$) are the same for two groups. Given the properties of the log-normal distribution, $\ln \bar{p} = \mu_p + \frac{1}{2}\sigma^2_u$, this yields

$$\mu_{p_A} - \mu_{p_B} = \frac{1}{2}(\sigma^2_{u_B} - \sigma^2_{u_A}) \tag{8}$$

Note that we can expect the current income of the self-employed to be more volatile than that of employees, that is $\sigma^2_{u_B} > \sigma^2_{u_A}$. Using an indicator variable D_i, which takes a value of 1 for individuals in group B and 0 otherwise, these assumptions can be incorporated in (7):

$$\begin{aligned} \ln c_i &= z_i\alpha + \beta \ln y_i + \beta(\mu_{k_A} - \mu_{p_A}) + \beta D_i\left[(\mu_{k_B} - \mu_{p_B}) - (\mu_{k_A} - \mu_{p_A})\right] + \eta^{PW}_i \\ &= -\beta\mu_{p_A} + z_i\alpha + \beta \ln y_i + \gamma^{PW} D_i + \eta^{PW}_i \end{aligned} \tag{9}$$

where $\gamma^{PW} = \beta\left[\mu_{k_B} + \frac{1}{2}(\sigma^2_{u_B} - \sigma^2_{u_A})\right]$ and $\eta^{PW}_i = \epsilon_i - \beta(u_i - v_i)$. The error term η^{PW}_i is heteroskedastic due to the assumed differences in the variance of u_i and v_i between group A and B. Furthermore, as $\ln y_i$ and η^{PW}_i are correlated[4], income is instrumented with a set of x (we discuss the choice of instruments in Section 4):

$$\ln y_i = z_i\delta_z + \delta_d D_i + x_i\delta_x + \xi_i \tag{10}$$

It follows that the average adjustment factor for group B is

$$\bar{k}^{PW}_B = \exp\left[\mu_{k_B} + \frac{1}{2}\sigma^2_{v_B}\right] = \exp\left[\frac{\gamma^{PW}}{\beta} + \frac{1}{2}(\sigma^2_{v_B} + \sigma^2_{u_A} - \sigma^2_{u_B})\right] \tag{11}$$

To obtain variance estimates in (11) note that ξ_i in (10) absorbs u_i and v_i as well as any unexplained variation in y^P_i (cf. equation 6). By its nature, permanent income is not correlated with shocks in current and reported income. Assuming also that the unexplained variation in permanent income is the same for two groups of individuals, the difference in residual variation (σ^2_ξ) between the two groups can be expressed as

$$\sigma^2_{\xi_B} - \sigma^2_{\xi_A} = \sigma^2_{u_B} + \sigma^2_{v_B} - 2\,\mathrm{Cov}(u_B, v_B) - \sigma^2_{u_A} \tag{12}$$

[4]It follows from equation (6) that $\mathrm{E}[\ln y_i \eta^{PW}_i] \neq 0$.

While combining (11) with (12) is not sufficient to obtain an identifiable point estimate of $\bar{k}_B^{PW}$, PW discuss its plausible range by making the following arguments. Assuming that u and v are uncorrelated, the lower bound is obtained with the lowest σ_{v_B} value (zero, i.e. everyone in group B misreport their income by the same proportion) and the upper bound with the lowest σ_{u_B} value (equal to σ_{u_A}, i.e. current incomes in group B are no more volatile than current incomes in group A). With these additional assumptions, the range of $\bar{k}$ can be expressed as

$$\bar{k}_B^{PW} = \exp\left[\frac{\gamma^{PW}}{\beta} \pm \frac{1}{2}\left(\sigma_{\xi_B}^2 - \sigma_{\xi_A}^2\right)\right] \tag{13}$$

PW further show that allowing for small positive correlation between u and v, does not have a large effect on the estimated range of $\bar{k}$ in the UK context.

2.2 Alternative specifications

The PW approach has been followed in its original form by Schuetze (2002) and Johansson (2005). There have been also several attempts to obtain a point estimate instead of bounds by utilising various proxies for permanent income or relying on different assumptions, which we summarise in this section.

Kim, Gibson, and Chung (2009) use *average* log income over time for the same household, i.e. $\overline{\ln y_{it}} = (1/T)\sum_{t=1}^{T} \ln y_{it}$, constructed from panel data. They argue that this eliminates variation in p_i, hence, yielding $\gamma^{KGC} = \beta\mu_{k_B}$ in an equivalent expression to (9). Taking into account that (12) is now reduced to $\sigma_{\xi_B}^2 - \sigma_{\xi_A}^2 = \sigma_{v_B}^2$, the average adjustment factor becomes

$$\bar{k}_B^{KGC} = \exp\left[\frac{\gamma^{KGC}}{\beta} + \frac{1}{2}\sigma_{v_B}^2\right] = \exp\left[\frac{\gamma^{KGC}}{\beta} + \frac{1}{2}\left(\sigma_{\xi_B}^2 - \sigma_{\xi_A}^2\right)\right] \tag{14}$$

This is numerically identical to the upper bound in the PW approach (cf. equation 13), which involved slightly weaker assumptions of $\sigma_{u_A}^2 = \sigma_{u_B}^2$ and $\mathrm{Cov}(u_B, v_B) = 0$. The approach by Kim et al. (2009) raises however a question whether such 'between estimates' should also balance out variation in the reported income (σ_v^2), which the authors do not address.

Kukk and Staehr (2014) draw on data where people report both their current and regular income, and use the latter as a direct measure of permanent income. This allows them to rely explicitly on $y_i^P = k_i y_i$ instead of (2) and (3) above, and leads to

$$\ln c_i = z_i\alpha + \beta \ln y_i + \gamma^{KS} D_i + \eta_i^{KS} \tag{15}$$

where $\gamma^{KS} = \beta\mu_{k_B}$ and $\eta_i^{KS} = \epsilon_i + \beta v_i$. As with Kim et al. (2009), the average adjustment factor is

$$\bar{k}_B^{KS} = \exp\left[\frac{\gamma^{KS}}{\beta} + \frac{1}{2}(\sigma_{\xi_B}^2 - \sigma_{\xi_A}^2)\right] \tag{16}$$

Hurst, Li, and Pugsley (2014) assume instead of $\bar{p}_A = \bar{p}_B$ that the transitory income component is the same for the two groups after controlling for their characteristics, i.e. $\ln p_i = z_i\psi + \nu_i$. Unlike other studies, Hurst et al. (2014) focus on the proportion of true income which is reported, $\kappa_i = 1/k_i$, rather than k_i, and assume it is constant for group B (self-employed). Due to this assumption, κ and k are entirely equivalent in their application. However, the choice between the two indicators does matter for the standard PW approach as we will show in the next subsection.[5] Noting that z_i does not include the group indicator D_i, this leads to

$$\ln c_i = z_i(\alpha - \beta\psi) + \beta \ln y_i + \gamma^{HLP} D_i + \eta_i^{HLP} \tag{17}$$

where $\gamma^{HLP} = -\beta \ln \kappa_B$ and $\eta_i^{HLP} = \epsilon_i - \beta\nu_i$. Their equivalent to (11) is the average proportion of reported income for group B, which is straightforward to estimate:

$$\bar{\kappa}_B^{HLP} = \exp\left(-\frac{\gamma^{HLP}}{\beta}\right) \tag{18}$$

The Hurst et al. (2014) approach is based on the set of strongest assumptions, effectively combining the assumptions behind the PW lower and upper bound. To see this, substitute both $\sigma_{v_B}^2 = 0$ (lower bound) and $\sigma_{u_A}^2 = \sigma_{u_B}^2$ (upper bound) into (11), yielding an equivalent expression to (18).

Besim and Jenkins (2005) and Engström and Holmlund (2009) estimate $\bar{k}_B = \exp(\gamma/\beta)$, which equals $1/\bar{\kappa}_B^{HLP}$ in equation (18), but they only discuss reduced-form estimation without elaborating on the underlying structural model. Similarly, Feldman and Slemrod (2007) focus directly on current income rather than permanent income and assume that a given income source is underreported by the same proportion.[6] On the one hand, this simplifies the model as in Hurst et al. (2014). On the other hand, they distinguish between multiple income sources, allowing each to have a separate adjustment factor k, which results in a non-linear system and is estimated with non-linear least squares.

Lyssiotou, Pashardes, and Stengos (2004) estimate a complete demand system instead of a single expenditure function. They argue that this avoids mistaking

[5] Hurst et al. (2014) do not elaborate on this and their study is the only one among those reviewed, which estimates the average proportion of reported income ($\bar{\kappa}$) instead of average adjustment factor ($\bar{k}$).

[6] Using current income could make more sense in their case as they relate charitable contributions to taxable income using income tax returns.

preference heterogeneity for income effects and classifying households according to their main source of income, which can be rather arbitrary. On the other hand, their demand system makes simplifications in other dimensions as they also ignore the transitory component of current income and assume that self-reported income is underreported by a constant fraction. Furthermore, the demand system is potentially more sensitive to the measurement error in consumption data and they additionally include income in quadratic terms. They also provide a non-parametric (single equation) estimate, which seems to suggest that a linear functional form for food expenditure may cause a downward bias. A non-parametric method is also used in Tedds (2010) to avoid imposing the functional form a priori. Her analysis for Canada suggests, however, that the reporting function is indeed linear although it also includes a constant.

For further details on previous studies estimating income underreporting on the basis of expenditure and income micro-data, see Table A1 in Appendix. Besides summarising methodological aspects, the table also covers key estimates obtained in these studies. To offer a better comparison with earlier studies and demonstrate the sensitivity of results to the model specification, in the empirical part we estimate the Pissarides and Weber (1989) and Hurst et al. (2014) type of measures alongside with our preferred specification, which is explained next.

2.3 Current approach

Our approach follows the PW model but seeks to estimate the average proportion of true income which is reported, $\bar{\kappa}$, as in Hurst et al. (2014). It does not matter whether the model is specified in terms of κ_i or k_i, as one can be substituted with the other, but as we see below calculating $\bar{\kappa}$ requires fewer assumptions than $\bar{k}$. While $\kappa_i = 1/k_i$, in general, $\bar{k} \neq 1/\bar{\kappa}$. In the case of Hurst et al. (2014), $\bar{k} = 1/\bar{\kappa}$ as they assumed the fraction of underreporting to be constant (i.e. $\sigma_v^2 = 0$), which we do not impose here by following the original PW framework.

Instead of (3), we now have $y_i = \kappa_i y_i^T$ and if k_i is log-normally distributed, so is κ_i. For convenience, we re-define equation (5) as

$$\ln \kappa_i = \mu_\kappa + v_i \tag{19}$$

Equation (6) and (7) then become

$$\ln y_i = (\mu_p + \mu_\kappa) + (u_i + v_i) + \ln y_i^P \tag{20}$$

$$\ln c_i = z_i\alpha + \beta \ln y_i - \beta(\mu_p + \mu_\kappa) - \beta(u_i + v_i) + \epsilon_i \tag{21}$$

Substituting (8) into (21) and using again the indicator D_i, we obtain

$$\ln c_i = -\beta\mu_{p_A} + z_i\alpha + \beta \ln y_i + \gamma D_i + \eta_i \tag{22}$$

where $\gamma = -\beta\left[\mu_{\kappa_B} + \frac{1}{2}(\sigma^2_{u_A} - \sigma^2_{u_B})\right]$ and $\eta_i = \epsilon_i - \beta(u_i + v_i)$. We can express the average proportion of true income reported by group B as

$$\bar{\kappa}_B = \exp\left[\mu_{\kappa_B} + \frac{1}{2}\sigma^2_{v_B}\right] = \exp\left[-\frac{\gamma}{\beta} + \frac{1}{2}(\sigma^2_{v_B} + \sigma^2_{u_B} - \sigma^2_{u_A})\right] \tag{23}$$

where the σ^2_u terms appear with opposite signs compared to (11). Combining (23) with (12), which remains the same (apart from the sign for the covariation term), and assuming as PW that u_B and v_B are uncorrelated, allows us to write $\bar{\kappa}_B$ in a form, which can be estimated without further assumptions:

$$\bar{\kappa}_B = \exp\left[-\frac{\gamma}{\beta} + \frac{1}{2}(\sigma^2_{\xi_B} - \sigma^2_{\xi_A})\right] \tag{24}$$

The lower bound for the adjustment factor in equation (13) is numerically equal to $1/\bar{\kappa}_B$ in (24), but unlike $\bar{\kappa}_B$ is obtained with a strong assumption that everyone in group B misreports their income in the same proportion ($\sigma^2_{v_B} = 0$). We expect $\bar{\kappa}_B < 1$, meaning that individuals in group B on average underreport their income.

One of the central aims of the paper is to establish whether income differences between population subgroups in the survey indeed correspond to how they declare their incomes to the tax authority. It is not obvious that the PW assumption about survey reporting is correct and to assess this, we estimate equation (22) in turn with the survey income ($y_i = y^s_i$) and the register income ($y_i = y^r_i$), available for each individual in the dataset. If people from type B households report consistently to the tax authority and in the survey, i.e. $\kappa^B_i(y^r_i) \simeq \kappa^B_i(y^s_i)$, then we would expect to find a similar extent of underreporting with either income concept. If people are (more) truthful in the survey, i.e. $\kappa^B_i(y^s_i) \simeq 1$, then estimation with the survey income should yield no substantial underreporting even if the estimation with the register income does.

Another extension relevant in our context concerns the composition of the reference group, which typically comprises all employees. Depending on opportunities for employees to collude with their employers to circumvent the third party reporting requirements, there can also be underreporting of wages and salaries. Therefore, it makes sense to distinguish between public sector and private sector employees as, in principle, there should be little (if any) possibility for the former to engage successfully in tax evasion activities due to the lack of incentives on the side of employers

in the public sector. This requires extending the model as follows.

Starting from equation (21), define three subgroups: group A (public sector employees), group B (private sector employees) and group C (self-employed). As before, we assume that group A reports correctly ($\kappa_i = 1 \; \forall i \in A$) and the expected current income, conditional on permanent income, is the same for all groups (i.e. $\bar{p}_A = \bar{p}_B = \bar{p}_C$). Hence, we can rewrite equation (8) as

$$\mu_{p_A} - \mu_{p_j} = \frac{1}{2}(\sigma^2_{u_j} - \sigma^2_{u_A}) \quad \text{where} \quad j = B, C \tag{25}$$

and equation (22) becomes

$$\ln c_i = -\beta \mu_{p_A} + z_i \alpha + \beta \ln y_i + \sum_{j=B,C} \gamma^j D_i^j + \eta_i \tag{26}$$

where $\gamma^j = -\beta \left[\mu_{\kappa_j} + \frac{1}{2}(\sigma^2_{u_A} - \sigma^2_{u_j}) \right]$ and $\eta_i = \epsilon_i - \beta(u_i + v_i)$. Given the differences in the variance of the residual term ξ_i between the groups (if u and v are uncorrelated):

$$\sigma^2_{\xi_j} - \sigma^2_{\xi_A} = \sigma^2_{u_j} + \sigma^2_{v_j} - \sigma^2_{u_A} \tag{27}$$

the average proportion of true income reported by group j is now

$$\begin{aligned} \bar{\kappa}_j &= \exp\left[\mu_{\kappa_j} + \frac{1}{2}\sigma^2_{v_j} \right] = \exp\left[-\frac{\gamma^j}{\beta} + \frac{1}{2}(\sigma^2_{v_j} + \sigma^2_{u_j} - \sigma^2_{u_A}) \right] \\ &= \exp\left[-\frac{\gamma^j}{\beta} + \frac{1}{2}(\sigma^2_{\xi_j} - \sigma^2_{\xi_A}) \right] \quad \text{where} \quad j = B, C \end{aligned} \tag{28}$$

Following the same logic, the framework can be easily extended to N type of households, for example, allowing for different types of private employees and self-employed. In the empirical part, we additionally estimate equation (26), both with survey and register income, to see how much such a breakdown affects results.

2.4 Main assumptions

Having laid out the PW framework and various modifications, we now consider the main (parametric) assumptions in more detail and discuss their implications.

First, what happens if the assumption about the reference group (i.e. employees) reporting correctly is not valid? The original PW framework has been extended to such a case by Martinez-Lopez (2012), showing that this affects primarily the interpretation of estimates, which then indicate the scale of underreporting relative to the reference group. We demonstrate it for our main specification. The term μ_{κ_A} is now retained in (22) with $\gamma = -\beta \left[(\mu_{\kappa_B} - \mu_{\kappa_A}) + \frac{1}{2}(\sigma^2_{u_A} - \sigma^2_{u_B}) \right]$ and equation

(23) becomes:

$$\bar{\kappa}_B = \exp\left[-\frac{\gamma}{\beta} + \frac{1}{2}(\sigma^2_{v_B} + \sigma^2_{u_B} - \sigma^2_{u_A}) + \mu_{\kappa_A}\right] \tag{29}$$

Equation (12) includes additional terms as well:

$$\sigma^2_{\xi_B} - \sigma^2_{\xi_A} = \sigma^2_{u_B} + \sigma^2_{v_B} + 2\,\mathrm{Cov}(u_B, v_B) - \left[\sigma^2_{u_A} + \sigma^2_{v_A} + 2\,\mathrm{Cov}(u_A, v_A)\right] \tag{30}$$

which combined with (29) (and assuming zero covariance terms) yields:

$$\begin{aligned}\bar{\kappa}_B &= \exp\left[-\frac{\gamma}{\beta} + \frac{1}{2}(\sigma^2_{\xi_B} - \sigma^2_{\xi_A}) + \mu_{\kappa_A} + \frac{1}{2}\sigma^2_{v_A}\right] \\ &= \exp\left[-\frac{\gamma}{\beta} + \frac{1}{2}(\sigma^2_{\xi_B} - \sigma^2_{\xi_A})\right]\bar{\kappa}_A \end{aligned} \tag{31}$$

This is equal to (24) adjusted with the average proportion of reported income for group A. In other words, if (24) is estimated when group A also misreports on average ($\bar{\kappa}_A \neq 1$) then the result for group B cannot be interpreted in absolute terms but *relative* to the level of misreporting by group A (and vice versa). It is not possible in this case to estimate misreporting for any group in absolute terms. This underlines the need to find a reference group characterised by minimal misreporting and ideally with $\bar{\kappa} \simeq 1$, which we hope to have achieved by focussing on public employees.

Second, PW and most of later studies have relied on the traditional demand function relating log expenditure to log income (see eq. 1). This functional form implies quite restrictive assumptions on consumer preferences (see e.g. Blundell, 1988): it links substitution effects strictly to income effects and demand is characterised by constant income elasticities. More flexible forms with a budget share as the dependent variable have been used instead by Lyssiotou et al. (2004) and Kim et al. (2009) in the same framework. On the other hand, the Working-Leser/AIDS type of functional form does not restrict the budget share to increase monotonically in overall budget and hence it may not be possible to invert the Engel curve for all values (see e.g. Tedds, 2010). We are not able to construct budget shares in our case as the dataset at our disposal does not contain information on total expenditure (more in Section 3). In some instances, if savings can be ignored, one might use total income as a proxy for total expenditure to derive budget shares but this would appear highly problematic in our context, where incomes are thought to be misreported (and expenditure not). In fact, this seems to be an overlooked aspect in Lyssiotou et al. (2004) approach when they set up a household expenditure function with the dependent variable (budget share) defined in terms of total expenditure,

while using true income as the budget constraint on the right hand side.[7] This potential inconsistency is avoided with the usual log-log specification of expenditure function, which we consider as a sufficient approximation for modelling demand.

Another assumption concerns the variable p_i, which determines how current income is related to permanent income. This is modelled independently of household characteristics (equation 4), while there could be for example age-related patterns with permanent income exceeding current income for young people ($p_i < 1$) and the opposite for the middle-aged group. The approach is less restrictive, however, than it initially appears. Hurst et al. (2014) allow $\ln p_i$ to explicitly depend on household characteristics ($z_i\psi$) but consequently the additional parameters are absorbed in the general vector of household characteristics, $z_i(\alpha - \beta\psi)$, see equation (17). What matters is the potential difference in the intercept for the two groups, which is captured by allowing μ_p to differ between the groups.

The assumption that all α-s and β-s are the same for the two groups could be more restrictive. Lyssiotou et al. (2004) point out that one of the key advantages of their complete demand system approach over the PW single-equation method is that it avoids confusing preference heterogeneity with income underreporting. They show that not accounting for preference heterogeneity can bias the estimate of underreporting downwards. While data constraints allow us only to estimate a single demand equation, we limit our sample to more homogenous households (couples with the head working full-time) similar to other studies. As part of the sensitivity analysis, we also test some additional sample restrictions.

3 Data

3.1 Data sources and linkage

We use the Estonian Social Survey (*Eesti Sotsiaaluuring*, ESU), linked with individual tax records. ESU is an annual household income survey, which also provides the Estonian component in the European Union Statistics on Income and Living Conditions (EU-SILC) database. It has a rotating sample design where households are followed in four consecutive waves and a quarter of the sample is replaced in every wave. The survey collects basic demographic information for all household

[7] Starting from a household cost function, Lyssiotou et al. (2004)[p. 625] derive a household expenditure function where the dependent variable, $w_i(p, U)$, is the share of total expenditure allocated for good i (p and U denote prices and utility). When they subsequently substitute U for the indirect utility function V and use true income as the budget constraint, they keep budget shares as they are. This would be strictly correct only if true income equals total expenditure (which they observe in their dataset), in which case it would be very straightforward to calculate misreported income.

members and detailed information for persons aged 16 or over, with a particular focus on their incomes. Interviews are carried out in the first half of year and the overlap with the end period for submitting annual tax reports (i.e. end of March) is thought to reduce recall errors.[8]

ESU has been linked with individual tax records allowing us to apply the method presented in the previous section both on survey and register income data for the same sample. The data linkage is based on the unique personal identification code, which is assigned to every person[9], and was legally carried out by Statistics Estonia without being required to inform sample members and obtain their consent. This is an important feature as it avoids the potential problem where those who are less compliant might be more likely to refuse data linkage, therefore, leading to a biased sample. Tax records refer either to a personal tax declaration or an (employer) tax withholding report, if the former was not submitted, and match the income reference period in ESU (i.e. the previous calendar year). Note that registered self-employed people are required to file a tax report.[10] Despite the different structure of personal and employer declarations, the informational content is broadly similar, and the tax withholding reports are also used to pre-populate individual tax reports. Both types of report show income by type and provider – employer or government institution administrating a given benefit. While tax records exclude non-taxable income sources (such as private transfers between households, the child benefit and the subsistence benefit), the share of such income components in aggregate disposable income is very small (about 2% according to ESU) and we are anyway mainly focussing on household earnings, which are not affected by this.[11]

We use the pooled 2007 and 2008 waves to increase the number of observations and reduce sensitivity to outliers in a given year.[12] The combined waves contain nearly 10,000 household observations (see Table 1). 98.6% of these had all household members identified in the tax register (no matter whether they had taxable

[8]For detailed information, see the national quality reports for the Estonian SILC at http://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/national-quality-reports (or starting from Eurostat's main page, follow Topic: Population and social conditions – Income, Social Inclusion and Living Conditions – Quality – National quality reports).

[9]Personal identification code is known for all sampled individuals from the Population Register and asked for other household members during the interview. The remaining individuals were matched with the Population Register using their characteristics (e.g. gender and the date of birth determine 7 out of 11 digits of the code) and the address. Nearly all people were matched and while the matching may have involved some error, this is likely to be insignificant. The final dataset used here is anonymised, without names, addresses etc.

[10]A personal tax declaration is also required to claim additional tax allowances, if applicable, and to benefit from optional joint assessment for married couples.

[11]On the other hand, benefit receipt tends to be generally underreported in survey datasets (Bound et al., 2001; Meyer et al., 2009).

[12]For example, Kim et al. (2009) demonstrate substantial year-to-year variation in their estimate of income underreporting for Korea in 2000-2005.

income or not). Excluding households with zero housing expenditure (see the next subsection) has also a negligible effect on the sample size, while excluding households with no earnings (employment and self-employment income[13]) leaves about 7,400 households. We further focus on households whose *head* has positive earnings (95% of all households with positive earnings) so that we can include head characteristics as covariates in the regression analysis. We select the head of household among the persons who state to be responsible for accommodation (or their partners), prioritising the one with the highest earnings[14], as the income position of that person is likely to have the largest influence on total housing related expenses.

[TABLE 1 HERE]

Among household heads with earnings we distinguish between public employees, private employees and self-employed. All household heads reporting (either full-time or part-time) self-employment as their main activity in any month of the income reference period (previous year) or working specifically as a (registered) sole proprietor[15] are considered as self-employed in the main analysis. In the sensitivity analysis we also consider alternative definitions for self-employed where household-level information or earnings related to activities as a sole proprietor are additionally taken into account.[16] There are 643 households whose head is self-employed and reported positive earnings in ESU and 568 households whose head is self-employed and reported positive earnings in the tax records. Other household heads with earnings are classified as employees, distinguishing further heads employed in the public sector and in the private sector. In ESU, people are asked about the proprietor of the enterprise where they work (state/municipality vs private individual/entity). As this refers to their current status rather than the income reference period, we consider those household heads who have changed jobs in between or have multiple jobs (or unspecified affiliation) as private employees. On the other hand, tax records allow us to distinguish between earnings received from private and public entities. Here, we consider those household heads as public employees who have received payments

[13]Throughout we exclude net losses from (registered) self-employment from our survey income measure to be consistent with income information in the tax reports. The number of affected households is less than 50 in each wave.

[14]To ensure a unique match, additional criteria include being the oldest and, finally, being male. By default, the head of household is defined as the person with the highest income in ESU.

[15]Respectively, variables G35* and H22.

[16]Sole proprietors (FIE) pay both employer contributions and personal taxes, but can deduct related business expenses from their taxable income. (Employer contributions are further deducted from their tax base for the income tax purposes.) People not registered as sole proprietors but engaged in individual work activities (e.g. private consultancy) are liable to declare and pay taxes on such income similar to salaries and wages. ESU also distinguishes earnings related to non-FIE self-employment, which we consider for total earnings but not for determining the self-employment status to be consistent with the tax records.

only from public entities, non-profit organisations or foundations. Household heads classified as private employees account for 73% of all heads with positive earnings, public employees 19-20% and self-employed 7-8% (the share varying only slightly depending on whether ESU or MTA information is used). These proportions are also very similar across the two waves.

There is no sufficiently detailed information on work intensity to account for its impact on the variation in households' earnings. ESU indicates a person's *main* activity in every month of the income reference period, distinguishing between working full-time and part-time but not in greater detail (weeks, hours), and the tax records do not contain any information about work intensity. We therefore limit our sample further to the household heads who have indicated in ESU working the whole year full-time as an employee or a self-employed. While the resulting sample is not representative of the whole working population, it still accounts for about 80% of all households with earnings. We relax this criterion as part of the sensitivity analysis.

Finally, as consumption patterns are strongly influenced by household structure, we focus on couples (both with and without children) who account for about 60% of the remaining sample. This is a similar approach to most previous studies (see Table A1 in Appendix). Another selection criterion commonly used is limiting sample to working age people, which we test as part of the sensitivity analysis. The final effective sample contains just over 3,400 households.

3.2 Expenditure and income information

The main disadvantage of ESU for this analysis is very limited expenditure data. While previous studies applying the PW method have relied primarily on food expenditure, this is not available in our dataset. Instead, this paper uses household costs of running the home. Our measure of housing related costs includes heating and power consumption (central heating, electricity, gas, other fuels); water, sewerage and other services; home insurance; housing maintenance and regular repairs – all collected separately in ESU. We exclude payments for housing per se in the main analysis as this is observed in the form of rent and mortgage interest payments for relatively few households: only 10% of households rent their accommodation and just 19% of homeowners report mortgage interest payments. (See Table A2 in Appendix for descriptive statistics for all non-monetary variables used.) Most households own their house without mortgage, in which case the cost of housing is implicit and can only be estimated indirectly (see e.g. Frick et al., 2010). The high proportion of owner-occupation is largely a result of housing privatisation in the early 1990s (see e.g. Pichler-Milanovich, 2001). In the sensitivity analysis, we also consider our measure of 'housing costs' with rent and mortgage interests pay-

ments.[17]

It is not clear without a direct comparison, how modelling based on housing costs (utilities) rather than food expenditure might affect results. Importantly, both consumption items are necessities and represent a substantial part of the total household budget. For example, Blundell et al. (1993, 1998) provide evidence for the UK on food and domestic fuel having similar relationship with household total expenditure. While food expenditure might offer more variability and hence potentially better identification, housing costs could have a more stable relationship with permanent income. Expenditures on utilities depend largely on the choice of dwelling, which is made for a longer period ahead – typically for a year at least – compared with choices related to food consumption and therefore should better reflect income potential in the medium term. Blundell et al. (1998) provide also evidence for domestic fuel being less sensitive to household composition than food and hence our approach could be more robust to potential specification errors.

There is also no particular reason for information on housing expenditure to be more problematic in terms of potential measurement errors. If both housing and food expenses are collected without a diary then housing costs may even have smaller recall errors due to involving less transactions (in a given reference period) and transactions being made on a more regular basis. The survey is also carried out when heating costs – the key component of our housing cost variable – are seasonally high and therefore amounts spent are likely to gain more attention by households.[18] While systematic measurement errors in expenditure would bias the coefficient for permanent income (for example, tendency to underreport expenditure would result in a downward bias), what would be more critical for our estimation strategy are differences in systematic measurement errors in expenditure between employees and self-employed. It is not obvious why this should be the case. A potential scenario could involve the self-employed reporting some of housing costs under business expenses rather than personal consumption. Fortunately, ESU collects information on non-cash income from self-employment and the share of self-employed who report that their business expenses include utilities is very marginal (about 2%).[19] What is perhaps the most reassuring evidence supporting our expenditure measure is that Hurst et al. (2014) obtain very similar results of income underreporting by

[17]In comparison, the COICOP category for housing expenditure includes electricity, gas and other fuels; water, sewage and other services; maintenance and repairs as well as actual and imputed rent but not housing-related insurance and mortgage interest payments.

[18]For example, recommendations made by Browning et al. (2003) for improving the measurement of total household expenditure in general purpose surveys include asking specifically questions about food and utilities, followed by housing costs (rent and interest payments).

[19]In comparison, the most common items (motor fuel and mobile phone services) were reported by about 10-15% of the self-employed.

the self-employed in the US both with food and utilities.

Table 2 provides the first look at how household expenditure and income compare across different types of households. It shows the (unconditional) mean of log expenditure (housing costs) and earnings, separately for the 2007 and 2008 wave. As part of the sensitivity analysis, we also use net total household income and housing expenses together with rent and mortgage interests.

[TABLE 2 HERE]

We can see that mean (log) consumption is highest for the self-employed in both waves and the differences with other two groups are statistically significant. Mean consumption is also somewhat higher for private employees compared to public employees but not significantly so at the 95% level. The same ranking emerges for survey earnings (y^s) in the 2008 wave (with a very marginal difference between private employees and self-employed), while the 2007 wave exhibits a different pattern: mean (log) survey earnings of private employees still exceed that of public employees, but it is the self-employed who have the lowest mean earnings. However, earnings in the tax records (y^r) show exactly the opposite ranking to consumption levels, with mean income being the highest for public employees and the lowest for self-employed – consistently across two waves (the differences between public and private employees are again not statistically significant though).[20]

As a consequence, the ratio of mean consumption to mean income (reflected in the difference in mean logs, Δ) varies across household types, being notably higher for self-employed household heads. This is robust to both waves and data sources, though the difference with other household types is larger with earnings in the tax records. The latter also indicate a higher consumption-to-income ratio for private employees compared to public employees, while it is the opposite with survey earnings. As such it provides preliminary evidence for income underreporting by self-employed and possibly by private employees.

4 Estimation and results

We now proceed with the econometric analysis. For a comparison with previous studies, we first estimate equation (22), distinguishing between wage earners and

[20]We use gross earnings from the tax records and survey earnings in net terms, which is how most survey respondents have stated them. This is to minimise the share of sample for which we have to rely on incomes derived from corresponding gross or net values. We do not expect it to have much impact on our comparison of income underreporting in the survey and in the tax records due to the flat income tax with a constant marginal rate above a relatively low income threshold, resulting in a fairly proportional tax system in Estonia.

self-employed. In the second stage, we distinguish also between public and private employees, estimating equation (26) with three groups of households.

It follows from the structural model that current observed income ($\ln y_i$) is endogenous and needs exclusion restrictions (i.e. instruments) to be properly identified in the consumption equation. Suitable instruments in this context are variables relevant for the income generation process but with no direct effect on housing expenditure, with various proxies for human capital or work effort being natural candidates.[21] We use dummies for the education level, occupation and industry of the household head as instruments in our case, on the basis that these are strong income predictors and there is no obvious reason why these should affect our measure of housing related costs except through income. With a single endogenous regressor and multiple instruments, the model is over-identified. In the sensitivity analysis, we also test a reduced set of instruments containing only information on head's education, which is the variable most often used in earlier studies. Other covariates (z_i), used in both consumption and income equation, include household head characteristics (gender, age, age squared, nationality, marital status), household characteristics (number of children and other adults, region, rural area, survey wave) and housing characteristics (type, year of construction, number of rooms, size in m^2, type of ownership).

Model estimates are obtained with the maximum likelihood method using survey weights and robust standard errors with clustering at the household level to account for multiple observations when pooling two waves.[22] On this basis we calculate the average proportion of reported income $\bar{\kappa}$ (equation 24), which is our main measure of income misreporting. In addition, we calculate $\bar{\kappa}$ according to Hurst et al. (2014) (equation 18) as well as the lower and the upper bound for the average adjustment factor $\bar{k}$ according to Pissarides and Weber (1989) (equation 13). This allows us to assess the sensitivity of results to additional assumptions underlying these approaches (see Section 2.2) and compare our results with previous estimates in the literature. In all cases, we present results in terms of the proportion of *underreported income* for an easier comparison.[23] The standard errors for all statistics are calculated using the delta method. We also include estimates of the income elasticity

[21] Instrumental variables used in the previous studies vary markedly. For example, Pissarides and Weber (1989), Schuetze (2002), Lyssiotou et al. (2004) and Johansson (2005) use a rather extensive set of instruments including head's education and/or work intensity for one or both spouses, physical assets and interactions of the self-employment status with other characteristics. Kukk and Staehr (2014) employ education level, gender and nationality of the household head as well as regional dummies. On the other hand, Engström and Holmlund (2009) rely only on income from capital and property taxes and Hurst et al. (2014) on dummies for educational attainment.

[22] Using `sem` command in Stata 12 and restricting coefficients in both equations to be the same for different groups, apart from intercepts, variances and covariances which are allowed to differ.

[23] That is $1 - \bar{\kappa}$ and $1 - 1/\bar{k}$, bearing in mind that generally $\bar{k} \neq 1/\bar{\kappa}$.

(β), dummies for self-employed and private employees (γ), and variances of the first stage error terms (σ^2_ξ).

4.1 Employees vs self-employed

Table 3 shows the estimated results for misreporting among self-employed compared to employees. As explained in Section 2.4, the results should be interpreted relative to the level of reporting by employees (which could be also incomplete). Consider first results with the survey income (first two columns). The IV estimate of 0.308 for the income elasticity of housing expenditure (β) (column 2) is statistically highly significant and notably higher than the estimate without instrumenting income (0.121, column 1), showing the extent of bias when income endogeneity is ignored. Our estimate is also consistent with those in the previous studies cited here which are mostly in the range of 0.2-0.4.[24]

The residual variance from the (reduced-form) income regression (σ^2_ξ) is also substantially higher for self-employed as expected, and the shift parameter for self-employed households (γ) is positive and highly significant. The estimate for income underreporting ($1-\bar{\kappa}_B$) suggests that, on average, 20% of household (net) earnings are underreported by households whose head is self-employed. Our estimate of the standard PW range is 20-43% and overlaps with those obtained in the earlier studies (see Table A1 in Appendix), all shown in terms of the average proportion of underreported income. There is substantial variation, however, among earlier studies and about half of them do not provide a measure of statistical precision for their estimates of income underreporting. The point estimate of 62% underreporting for the self-employed by Kukk and Staehr (2014), using the Estonian Household Budget Survey, is the main exception which is difficult to reconcile not only with our estimates but also with other studies. Their approach is unique for relying on a self-reported measure of regular income as a proxy for permanent income, though without a direct comparison with estimates based on the usual measure of current income it is not possible to ascertain whether this is indeed the primary source of differences.

[TABLE 3 HERE]

The last two columns in Table 3 report equivalent estimates using register income. Not only are all estimates highly statistically significant but they also reveal

[24]The exceptions are Besim and Jenkins (2005), Feldman and Slemrod (2007) and Kukk and Staehr (2014) whose estimate of β is higher, about 0.5-0.6. Besim and Jenkins (2005) do not use instrumental variables and have the smallest sample, among else. Feldman and Slemrod (2007) use very different income and expenditure data (declared incomes and charitable contributions in the tax records). Kukk and Staehr (2014) use a measure of regular income, arguably less affected by transitory movements in income.

much larger income underreporting on average. Our main estimate ($\bar{\kappa}_B$) indicates that 48% of household (gross) earnings are underreported by households with a self-employed head, the PW upper bound 71% and the HLP measure 61% (column 4). This is due to the estimate of income elasticity (β) being smaller and the estimated shift parameter (γ) being larger compared to the IV estimates with survey income (column 2), though this is partly counterbalanced by larger differences in variance estimates between the two groups (cf. equation 24). The variance estimates themselves are almost twice as large compared to estimates from survey income. The fact that register data allows us to detect substantially larger income underreporting suggests that self-employed are more truthful in reporting their income in the survey compared to the tax declarations.

The bottom section of Table 3 shows typical diagnostic tests for our instruments.[25] For both data sources, the endogeneity test rejects the null hypothesis that household earnings are exogenous.[26] Furthermore, partial R^2 and the F-test of excluded instruments confirm that instruments are reasonably strong in all models. Finally, the Hansen J-statistic fails to reject the null hypothesis that these are valid instruments in the case of survey incomes, while it raises some doubt for register incomes. It appears to be of limited importance though as in subsequent specifications, the test is passed for both income sources. We also get very similar estimates for β and γ when using only dummies for educational attainment as instruments, in which case the p-value for the Hansen statistic is about 0.4. Hence, without instrumenting earnings, we would obtain biased estimates of income underreporting, indicating much larger income underreporting than is actually the case.

4.2 Public employees vs private employees and self-employed

Until now we have estimated income underreporting among self-employed using (all) employees as the reference group as in previous studies using the same method, apart from Besim and Jenkins (2005), but, as demonstrated in Paulus (2015), there can be substantial non-compliance also among employees. In the next step, we further distinguish between households whose head works in the public sector and the private sector, and assume that only the latter have opportunities to underreport their income. The results are shown in Table 4.

[TABLE 4 HERE]

[25] These are estimated with the help of `ivreg2` package in Stata. While this also supports (limited-information) maximum likelihood estimation method with cluster-robust variance estimates, it does not allow specifying the model structure in the same detail as `sem` (e.g. different σ^2_ξ by subgroups). Nevertheless, we consider these test diagnostics to represent our main model sufficiently accurately.

[26] The test statistic is defined as the difference of two Sargan-Hansen statistics.

Similar to Table 3, estimates with survey income (column 2) detect income underreporting for households with self-employed heads, now to a slightly larger extent (25% on average). The estimates also show a modest underreporting for households whose head is a private sector employee (7%), though these are not statistically significant and hence do not suggest substantial differences between public and private sector employees when it comes to income reporting. Estimates from register income (column 4) on the other hand, yield strong evidence for substantial income underreporting among households with privately employed heads (23%), while the estimate for households with self-employed heads is now 56%. The PW upper bound implies average underreporting of 34% and 78% among the two groups. There are also notable differences between estimates from survey income and register income for the residual income variance of households whose head is a private employee: while survey data do not suggest much difference with households whose head is a public employee, estimates with register data show a much higher variance for private employees which exceeds that of public employees by almost two-fold.[27] The income elasticity (β) estimates are essentially not affected by distinguishing between the three groups of households rather than two, and the model fit (according to the AIC and BIC statistics) improves for estimates from either data source. As before, the IV estimation leads to lower underreporting compared to estimation without instruments. The results and conclusions of diagnostic tests for instrumental variables in Table 4 are also very similar to those discussed above for Table 3.

The estimate of underreporting for private sector employees from register income (23%) is of similar magnitude to the one estimated in Paulus (2015) (16%), noting that the first estimate is the average scale of underreporting while the other is the share of undeclared earnings in total earnings (i.e. the aggregate scale). Besides this, the two approaches also differ for the overall method, the unit of analysis (household vs individual) and sample (households of couples whose head is working full-time vs employees working full-time).

On the other hand, our estimates differ substantially from Besim and Jenkins (2005) which is the only other PW-type of study that distinguishes between public sector and private sector employees as well as self-employed. Their estimates of underreporting for the self-employed (10-11%) are the lowest among all the studies considered here and, surprisingly, very similar to that for private employees (13%) – in fact, the latter are even slightly higher. Taken together with unusually high β estimates, their results warrant extra caution. There are several different method-

[27] As another sensitivity check, we excluded all households with self-employed heads and estimated the model for employees only. The results with both survey and register income changed only marginally for the private sector employees (relative to the public sector employees). See Table A3 in Appendix.

ological choices, which could undermine the comparability of their results with other studies. First, they estimate the extent of underreporting at the average income level (rather than the average rate of misreporting). Second, they employ the OLS method. Third, they have the smallest sample among such studies and impose very few sample restrictions. For example, studies relying on food expenditure typically exclude households engaged in agricultural production as their food purchases might be strongly affected. Based on our own sensitivity analysis (see the next subsection), there can be also variation in consumption patterns due to differences in household structure (e.g. singles vs couples) and a very heterogenous sample might render the estimates of income underreporting unstable.

4.3 Sensitivity analysis

We further explore the sensitivity of our results to alternative sample and variable definitions. Table A4 in Appendix summarises estimates of β, γ and our principal measure of income underreporting, along with diagnostic indicators, for the base scenario (column 1) and for alternative configurations: an alternative set of instruments (column 2), expanded samples (columns 3-5); narrower samples (columns 6-9); alternative expenditure and income measures (columns 10-11); and alternative self-employment definitions (columns 12-13).

Model (2) shows that very similar results to the baseline are obtained when the set of instrumental variables is limited to the dummies for head's education (though inevitably this reduces the explanatory power of instruments). Model (3) expands the sample to include those households whose head worked part-time or part-year (about an 8% increase) and the results are also affected very little. Model (4) and (5) are based on samples combining couple households with single households and other type of households, respectively. Here, we see slightly more variation in results with some estimates becoming less precise, and more so when single households are included (model 4), though the estimates are broadly still stable.

Model (6) and (7) focus on more homogenous samples by restricting the age range of the household head to 25-55, and excluding those with earnings reported only in one data source, respectively. The sample for model (6) is about 80% of the main sample and, while estimates of underreporting with survey income change very little, estimates with register income become significantly larger: 33% and 64% on average for households whose head is, respectively, a private employee and a self-employed. The sample for model (7) is only slightly smaller compared to the main sample, indicating that the latter contains relatively few people with positive earnings in one data source and not in the other, and results are just marginally different from the baseline. Model (8) and (9) provide estimates for the 2007 and 2008 wave

separately. Splitting the sample obviously increases standard errors (which is the very reason for using pooled waves in the main analysis) though point estimates are generally quite similar, apart from the dummy for self-employed heads (γ_C) for survey income and the dummy for head in the private sector (γ_B) for register income.

Model (10) tests a measure of housing costs, which includes rent and mortgage interest payments and model (11) uses a broader income measure (total household income) instead of earnings.[28] The alternative expenditure measure is limited, however, in the sense of considering actual expenses only and not implicit rent for homeowners. In both cases, the estimate of the income elasticity of housing expenditure (β) is higher than in the baseline, which is expected as rent and mortgage interests ought to be more elastic than utilities, and total income potentially more relevant for household expenses than earnings alone. However, the model fit to the data (based on AIC and BIC) becomes poorer with (10) for both survey and register income, and $\bar{\kappa}_B$ becomes even higher than one, implying that households whose heads are private employees *overreport* their income on average, though this estimate is not statistically reliable. Estimates of income underreporting with model (11) are slightly lower compared to the baseline, which is expected as reporting accuracy for other income components, which are now included (e.g. public pensions and other social transfers), should not be affected by the type of household.[29] However, the estimates of underreporting do not decrease much as earnings are the dominant source of income for this sample of households. Although the model fit is improved when using total household income, the reason for not choosing model (11) as the baseline is because of our explicit focus on underreporting of earnings.

Finally, model (12) and (13) test alternative definitions for self-employed households, the former considers a household self-employed if *any* of its members is working as a self-employed (not specifically the head of household) and the latter extends the number of self-employed households by including heads who have not indicated self-employment status (in ESU) but reported income related to registered self-employment income. This expands their numbers when using register income as the reported self-employment status and income are already aligned for survey income based sample. The estimates for model (12) and (13) with survey income differ only marginally from the baseline; the same applies to model (13) with register income, while model (12) estimates of underreporting are slightly smaller.

[28]The total income in the tax records is limited to taxable incomes only. See Section 3.

[29]In the case of survey data, other income sources (even if non-taxable) are still likely to be measured imperfectly due to, for example, recall errors, social stigmas related to the receipt of welfare benefits etc, but it is not obvious why this should vary systematically between public employees, private employees and self-employed.

5 Conclusions

In this paper we extend the method of Pissarides and Weber (1989) for estimating income underreporting and apply this to a household income survey linked with individual tax records for Estonia. This allows us to test the validity of the two main assumptions underlying this method: that employees are fully compliant and patterns of survey income reporting correspond to the way incomes are declared to the tax authority (i.e. the actual tax compliance behaviour). As a further methodological contribution, we identify a way to obtain a point estimate of underreporting with fewer assumptions. We also review other studies applying this kind of method to provide an overview of their methodological differences. Similar to Besim and Jenkins (2005), but in a more rigourous framework, we distinguish between public sector and private sector employees, relaxing the assumption of full income reporting by the latter.

Our key findings are the following. We detect large underreporting of earnings by households (of couples) whose head is a self-employed (56% on average) and also substantial underreporting of earnings by households whose head is a private sector employee (23%) on the basis of (housing related) expenditures and incomes in the tax records. However, the scale of underreporting by the self-employed and private employees is estimated to be much smaller with survey incomes (respectively, 25% and 7%) and the latter estimate is also statistically non-significant. Importantly, this indicates that people are more truthful in the surveys than often assumed and previous studies using this method may have underestimated the extent of non-compliance by a substantial margin. Moreover, an obvious advantage of using tax records is that this allows us to attribute income underreporting to non-compliance with much greater certainty compared to survey data where misreporting may also occur due to recall errors, stigma effects etc.

There are several policy implications. Higher reporting of wages and salaries compared to self-employment income is an indication that the third-party reporting reduces non-compliance substantially. However, what is equally important to emphasise is that it does not rule out tax evasion altogether as the employee and the employer can still collude. Furthermore, in absolute terms, much more tax revenue is lost through the underreporting of employment income compared to the underreporting of self-employment income as the latter accounts for only a marginal share of total earnings (less than 2-4% according to ESU, without corrections for underreporting). Hence, the underreporting of wages and salaries by a small proportion can in monetary terms easily exceed underreported self-employment income even if the latter was entirely concealed.

Despite utilising a rich and novel data source in the field of tax compliance and among the PW-type of studies in particular, the paper was limited to a cross-sectional analysis and to a single type of expenditure. More waves and larger samples are required to take the analysis further by utilising the panel data element and studying specific subgroups in more detail. Richer consumption data would also allow us to estimate more complex demand systems.

References

Almås, I. (2012). "International income inequality: Measuring PPP bias by estimating Engel curves for food." *The American Economic Review*, *102*(2), 1093–1117.

Besim, M., and Jenkins, G. P. (2005). "Tax compliance: when do employees behave like the self-employed?" *Applied Economics*, *37*(10), 1201–1208.

Blundell, R. (1988). "Consumer behaviour: Theory and empirical evidence – A survey." *The Economic Journal*, *98*(389), 16–65.

Blundell, R., Duncan, A., and Pendakur, K. (1998). "Semiparametric estimation and consumer demand." *Journal of Applied Econometrics*, *13*(5), 435–461.

Blundell, R., Pashardes, P., and Weber, G. (1993). "What do we learn about consumer demand patterns from micro data?" *The American Economic Review*, *83*(3), 570–597.

Blundell, R., Pistaferri, L., and Preston, I. (2008). "Consumption inequality and partial insurance." *The American Economic Review*, *98*(5), 1887–1921.

Bound, J., Brown, C., and Mathiowetz, N. (2001). "Measurement error in survey data." In J. J. Heckman, and E. Leamer (Eds.), *Handbook of Econometrics*, vol. 5, chap. 59, 3705–3843, Amsterdam: Elsevier.

Browning, M., Crossley, T. F., and Weber, G. (2003). "Asking consumption questions in general purpose surveys." *The Economic Journal*, *113*(491), F540–F567.

Engström, P., and Holmlund, B. (2009). "Tax evasion and self-employment in a high-tax country: evidence from Sweden." *Applied Economics*, *41*(19), 2419–2430.

Feldman, N. E., and Slemrod, J. (2007). "Estimating tax noncompliance with evidence from unaudited tax returns." *The Economic Journal*, *117*, 327–352.

Frick, J. R., Grabka, M. M., Smeeding, T. M., and Tsakloglou, P. (2010). "Distributional effects of imputed rents in five European countries." *Journal of Housing Economics*, *19*(3), 167–179.

Hamilton, B. W. (2001). "Using Engel's Law to estimate CPI bias." *The American Economic Review*, *91*(3), 619–630.

Hurst, E., Li, G., and Pugsley, B. (2014). "Are household surveys like tax forms: evidence from income underreporting of the self-employed." *Review of Economics & Statistics*, *96*(1), 19–33.

Johansson, E. (2005). "An estimate of self-employment income underreporting in Finland." *Nordic Journal of Political Economy*, *31*, 99–109.

Kim, B., Gibson, J., and Chung, C. (2009). "Using panel data to exactly estimate income under-reporting by the self-employed." Working Paper 09-02, Korea Institute for International Economic Policy.

Kukk, M., and Staehr, K. (2014). "Income underreporting by households with business income: evidence from Estonia." *Post-Communist Economies*, *26*(2), 257–276.

Larsen, E. (2009). "Using inverted Engel curves to estimate material standards of living in a household." *Empirical Economics*, *36*(1), 109–132.

Lyssiotou, P., Pashardes, P., and Stengos, T. (2004). "Estimates of the black economy based on consumer demand approaches." *The Economic Journal*, *114*(497), 622–640.

Martinez-Lopez, D. (2012). "The underreporting of income by self-employed workers in Spain." *SERIEs*, 1–19.

Meyer, B. D., Mok, W. K. C., and Sullivan, J. X. (2009). "The under-reporting of transfers in household surveys: Its nature and consequences." NBER Working Paper 15181.

Paulus, A. (2015). "Tax evasion and measurement error: an econometric analysis of income survey data linked with tax records." ISER Working Paper 2015-10, University of Essex, Colchester.

Pichler-Milanovich, N. (2001). "Urban housing markets in Central and Eastern Europe: convergence, divergence or policy 'collapse'." *European Journal of Housing Policy*, *1*(2), 145–187.

Pissarides, C. A., and Weber, G. (1989). "An expenditure-based estimate of Britain's black economy." *Journal of Public Economics*, *39*, 17–32.

Schuetze, H. J. (2002). "Profiles of tax non-compliance among the self-employed in Canada: 1969 to 1992." *Canadian Public Policy / Analyse de Politiques*, *28*(2), 219–238.

Tedds, L. M. (2010). "Estimating the income reporting function for the self-employed." *Empirical Economics*, *38*(3), 669 – 687.

Table 1: Sample size

	Number of households			
	Total	Public	Private	Self-empl.
All survey respondents	9,890			
Matched with tax records	9,754			
Household with non-zero consumption	9,728			
Household with positive earnings				
in the survey	7,457			
in the tax records	7,426			
Head with positive earnings				
in the survey	7,064	1,289	5,132	643
in the tax records	7,038	1,449	5,021	568
Head worked full-time (for whole year)				
in the survey	5,954	1,175	4,251	528
in the tax records	5,789	1,101	4,219	469
Couples (with or without children)				
in the survey	3,498	595	2,549	354
in the tax records	3,413	598	2,502	313

Notes: ESU 2007 and 2008 waves pooled; *public/private* = household head is employed in the public/private sector and not considered a self-employed; (employees in ESU who have multiple jobs or unspecified affiliation or switched jobs are included among private employees, public employees in the tax records include those with earnings only from a public entity, an NGO or a foundation); *self-employed* = household head worked as a self-employed in the income reference period (based on ESU).

Table 2: Mean household expenditure and income

	Survey income			Register income		
	$\ln c$	$\ln y^s$	Δ	$\ln c$	$\ln y^r$	Δ
			2007 wave			
Public	7.074	9.295	-2.221	7.082	9.527	-2.444
	(0.026)	(0.042)	(0.040)	(0.027)	(0.049)	(0.047)
	324	324		317	317	
Private	7.121	9.426	-2.305	7.122	9.475	-2.352
	(0.015)	(0.022)	(0.022)	(0.015)	(0.027)	(0.026)
	1307	1307		1280	1280	
Self-employed	7.228	9.223	-1.996	7.232	8.995	-1.763
	(0.039)	(0.083)	(0.082)	(0.042)	(0.098)	(0.093)
	178	178		160	160	
N total	1,809			1,757		
			2008 wave			
Public	7.257	9.472	-2.215	7.208	9.750	-2.543
	(0.024)	(0.045)	(0.046)	(0.029)	(0.046)	(0.052)
	271	271		281	281	
Private	7.307	9.586	-2.279	7.324	9.690	-2.366
	(0.013)	(0.020)	(0.020)	(0.012)	(0.026)	(0.025)
	1242	1242		1222	1222	
Self-employed	7.440	9.594	-2.154	7.440	9.357	-1.917
	(0.039)	(0.077)	(0.077)	(0.042)	(0.113)	(0.106)
	176	176		153	153	
N total	1,689			1,656		

Notes: estimated using survey weights on a sample of couple households whose head has positive earnings and worked full-time (2007-08 waves pooled); standard errors (shown in parantheses) are clustered at the household level; c = hh monthly housing costs (excl. rent and mortgage interests) in EEK; y^s = hh net earnings in the survey in EEK (annual amount divided by 12); y^r = hh gross earnings in the tax records in EEK (annual amount divided by 12); Δ = difference in mean log values; *public/private* = household head is employed in the public/private sector and not considered a self-employed; *self-employed* = household head worked as a self-employed in the income reference period (based on ESU).

Table 3: Estimates with employees (A) and self-employed (B)

	Survey income		Register income	
	(1) ML	(2) ML-IV	(3) ML	(4) ML-IV
β	0.121***	0.308***	0.078***	0.187***
	(0.018)	(0.054)	(0.014)	(0.036)
γ_B	0.087***	0.123***	0.121***	0.177***
	(0.032)	(0.036)	(0.036)	(0.040)
$\sigma^2_{\xi_A}$	0.190***	0.190***	0.326***	0.326***
	(0.008)	(0.008)	(0.024)	(0.024)
$\sigma^2_{\xi_B}$	0.536***	0.536***	0.904***	0.908***
	(0.168)	(0.169)	(0.183)	(0.184)
$1-\bar{\kappa}_B$	0.421**	0.202**	0.717***	0.481***
	(0.182)	(0.101)	(0.142)	(0.113)
$1-1/\bar{k}_B^{PW_l}$	0.421**	0.202**	0.717***	0.481***
	(0.182)	(0.101)	(0.142)	(0.113)
$1-1/\bar{k}_B^{PW_u}$	0.590***	0.436***	0.841***	0.710***
	(0.127)	(0.095)	(0.083)	(0.073)
$1-\bar{\kappa}_B^{HLP}$	0.513***	0.329***	0.788***	0.612***
	(0.146)	(0.082)	(0.107)	(0.084)
# of employees (A)	3,017	3,017	2,975	2,975
# of self-employed (B)	345	345	306	306
Total obs	3,362	3,362	3,281	3,281
AIC	13,840,023	13,836,263	13,666,398	13,663,828
BIC	13,840,439	13,836,692	13,666,813	13,664,255
Partial R^2		0.1323		0.1572
F-test for excluded instr.-s		19.14		26.35
Endogeneity test (p-value)		0.0000		0.0004
Hansen J-test (p-value)		0.4467		0.0459

Notes: estimated using survey weights on a sample of couple households whose head has positive earnings and worked full-time (2007-08 waves pooled); standard errors (shown in parantheses) are clustered at the household level; *dependent variable* = ln housing costs; *income* = ln earnings; *instruments* = head education level, occupation and industry; *covariates* = head gender, age (centered), age squared, nationality, marital status; no of children and (other) adults in the hh, region, rural area, wave and housing characteristics (type, year of construction, no of rooms, size in m2, ownership); *self-employed* = household head worked as a self-employed in the income reference period (based on ESU); * p<0.1, ** p<0.05, *** p<0.01.

Table 4: Estimates with public employees (A), private employees (B) and self-employed (C)

	Survey income		Register income	
	(1) ML	(2) ML-IV	(3) ML	(4) ML-IV
β	0.121***	0.302***	0.080***	0.215***
	(0.020)	(0.054)	(0.012)	(0.034)
γ_B	0.018	0.023	0.049**	0.073***
	(0.022)	(0.022)	(0.023)	(0.024)
γ_C	0.102***	0.141***	0.163***	0.253***
	(0.038)	(0.041)	(0.040)	(0.046)
$\sigma^2_{\xi_A}$	0.181***	0.181***	0.200***	0.199***
	(0.013)	(0.013)	(0.015)	(0.015)
$\sigma^2_{\xi_B}$	0.192***	0.192***	0.354***	0.355***
	(0.007)	(0.007)	(0.017)	(0.017)
$\sigma^2_{\xi_C}$	0.535***	0.536***	0.906***	0.909***
	(0.168)	(0.169)	(0.184)	(0.184)
$1-\bar{\kappa}_B$	0.132	0.068	0.414**	0.232***
	(0.161)	(0.069)	(0.177)	(0.088)
$1-1/\bar{k}_B^{PW_l}$	0.132	0.068	0.414**	0.232***
	(0.161)	(0.069)	(0.177)	(0.088)
$1-1/\bar{k}_B^{PW_u}$	0.141	0.078	0.498***	0.343***
	(0.160)	(0.069)	(0.152)	(0.076)
$1-\bar{\kappa}_B^{HLP}$	0.137	0.073	0.458***	0.290***
	(0.160)	(0.068)	(0.164)	(0.081)
$1-\bar{\kappa}_C$	0.485***	0.250**	0.812***	0.561***
	(0.186)	(0.110)	(0.106)	(0.095)
$1-1/\bar{k}_C^{PW_l}$	0.485***	0.250**	0.812***	0.561***
	(0.186)	(0.110)	(0.106)	(0.095)
$1-1/\bar{k}_C^{PW_u}$	0.638***	0.474***	0.907***	0.784***
	(0.130)	(0.097)	(0.054)	(0.054)
$1-\bar{\kappa}_C^{HLP}$	0.568***	0.372***	0.868***	0.692***
	(0.151)	(0.090)	(0.075)	(0.066)
# of public employees (A)	580	580	579	579
# of private employees (B)	2,437	2,437	2,396	2,396
# of self-employed (C)	345	345	306	306
Total obs	3,362	3,362	3,281	3,281
AIC	13,241,775	13,238,113	12,662,605	12,658,765
BIC	13,242,216	13,238,572	12,663,044	12,659,222
Partial R^2		0.1325		0.1502
F-test for excluded instr.-s		19.33		23.71
Endogeneity test (p-value)		0.0000		0.0000
Hansen J-test (p-value)		0.5083		0.1582

Notes: estimated using survey weights on a sample of couple households whose head has positive earnings and worked full-time (2007-08 waves pooled); standard errors (shown in parantheses) are clustered at the household level; *dependent variable* = ln housing costs; *income* = ln earnings; *instruments* = head education level, occupation and industry; *covariates* = head gender, age (centered), age squared, nationality, marital status; no of children and (other) adults in the hh, region, rural area, wave and housing characteristics (type, year of construction, no of rooms, size in m2, ownership); *public/private* = household head is employed in the public/private sector and not considered a self-employed; *self-employed* = household head worked as a self-employed in the income reference period (based on ESU); * p<0.1, ** p<0.05, *** p<0.01.

Table A1: Studies on income underreporting based on income-expenditure gap

Study	Country	Data source	Data type	Sample selection	Sample size
Pissarides and Weber (1989)	UK	FES 1982	cross	couples; head working male	2,208
Schuetze (2002)	Canada	FES 1969, 1974, 1984, 1986, 1990, 1992	pooled	urban couples; head male, aged 25-64, worked full-time/full-year, non-farmer	8,463
Lyssiotou et al. (2004)	UK	FES 1993	cross	married couples; head employed	1,750
Johansson (2005)	Finland	HBS 1994-1996 (income from register)	pooled	couples; head aged $<$ 64, worked full year, non-farmer	2,053
Besim and Jenkins (2005)	North-Cyprus	HCES 1998/99	cross	head/spouse: 1 type of empl. income	907
Feldman and Slemrod (2007)	US	unaudited income tax returns 1999	cross	taxpayers with itemised deductions	76,647
Engström and Holmlund (2009)	Sweden	HBS 1999-2004 (except 2002)	pooled	couples (one/both working), non-farmers	4,600-6,000
Kim et al. (2009)	Korea	KLIPS 2000-2005	panel (average)	urban couples aged 20-65,	6,593
	Russia	RLMS 1994-2000	panel (average)	male head	5,243
Tedds (2010)	Canada	FES 1982, 1986, 1992, 1996	pooled	married couples (no children) aged 25-64, head employed and non-farmer	3,880
Martinez-Lopez (2012)	Spain	HBS 2006-2009	pooled	non-farmers	13,158-16,451
Hurst et al. (2014)	US	CES 1980-2003	pooled	head male, aged 25-55, 30h+, 40 weeks, non-farmer	27,219
		PSID 1980-1997 (except 1988-89)	pooled, panel (3y average)		36,434/18,233
Kukk and Staehr (2014)	Estonia	HBS 2002-2007	pooled	couples: head active	6,016

Notes: FES=Family Expenditure Survey(s), HBS=Household Budget Survey, HCES=Household Consumption Expenditure Survey, KLIPS=Korea Labor Income Panel, RLMS=Russian Longitudinal Monitoring Survey, CES=Consumer Expenditure Survey, PSID=Panel Study of Income Dynamics.

Table continues on the next page.

Table A1 continues

Study	Group def	Expenditure	Income	Method	Estimation	Results
Pissarides and Weber (1989)	SE ($\geq$ 25% THI)	food	net HE	k bands (EC)	IV-2SLS	white collar 22-35%, blue collar 34-39%
Schuetze (2002)	SE ($\geq$ 30% THI)	food	net THI	k bands (EC)	IV-2SLS	6-22% (across years)
Lyssiotou et al. (2004)	SE (main income), SE income	food, DS	HE	k bands (EC), NP	IV-2SLS, GMM	white collar 8-20%, NP 32%*, DS 39%*; blue collar 27-29%, NP 43%*, DS 54%*
Johansson (2005)	SE ($\geq$ 6 months)	food	net THI (R)	k bands (EC)	(OLS), IV-2SLS	9-19% (head SE), 27-32% (couple SE)
Besim and Jenkins (2005)	SE (income), private EE (reported)	food	net THI	k point	OLS	10-11% (self-empl.), 13% (priv. empl.)
Feldman and Slemrod (2007)	tax schedules (C-F)	charitable contributions	taxable income (R) by schedules	k point	NLS	15-65%* (across schedules)
Engström and Holmlund (2009)	SE (reported)	food	net THI (R)	k point	OLS, IV	14-15% (incorp.), 33% (unincorp.)
Kim et al. (2009)	SE (main job), SE (reported)	food share	THI	k bands (EC), k point (EC)	between OLS, (FE)	38%* (Korea) 47%* (Russia)
Tedds (2010)	SE ($>$ 0% THI)	food	net THI + savings	NP	LOWESS	20%*
Martinez-Lopez (2012)	SE (reported)	food	net THI/THHI	k bands (EC)	IV-2SLS	15-25%
Hurst et al. (2014)	SE (reported)	food, total, non-durables	gross/net THI	κ point	OLS, IV	19-32%* (CES) 28-32%* (PSID)
Kukk and Staehr (2014)	SE ($\geq$ 20% THI)	food	regular net THI	k point (EC)	IV-GMM	62%*

Notes: all results shown in terms of underreporting ($1 - \bar{\kappa}$ and $1 - 1/\bar{k}$), * estimates with t-values or standard errors; SE=self-employed, TH(H)I=total household (head) income, HE=household earnings, R=register-based; EC=(with) error correction, DS=demand system, NP=non-parametric.

Table A2: Descriptive statistics for non-monetary variables

	mean	st.dev.	N
Education=basic or less	0.07	0.26	4,014
Education=secondary	0.56	0.50	4,014
Education=tertiary	0.37	0.48	4,014
Occupation=senior managers, legislators	0.20	0.40	3,983
Occupation=professionals	0.15	0.35	3,983
Occupation=technicians, associate professionals	0.11	0.32	3,983
Occupation=service/sales workers	0.07	0.25	3,983
Occupation=craft/related trade workers	0.20	0.40	3,983
Occupation=clerks, plant/machine operators	0.20	0.40	3,983
Occupation=agricultural workers, elementary occupations	0.06	0.24	3,983
Industry=agriculture, forestry, fishing	0.05	0.21	3,938
Industry=manufacturing, mining, electricity, gas, water supply	0.25	0.43	3,938
Industry=construction	0.15	0.36	3,938
Industry=trade, hotels, restaurants, transport, communication	0.26	0.44	3,938
Industry=finance, real estate, renting, business activities	0.10	0.29	3,938
Industry=public admin, education, health; own production	0.20	0.40	3,938
Age (centered)	-0.00	1.13	4,014
Age (centered) squared	1.28	1.50	4,014
Gender=male	0.72	0.45	4,014
Nationality=Estonian	0.71	0.45	4,014
Marital status=married	0.73	0.44	4,014
Region=north	0.40	0.49	4,014
Region=central	0.11	0.31	4,014
Region=north-east	0.11	0.32	4,014
Region=west	0.13	0.33	4,014
Region=south	0.24	0.43	4,014
Area=rural	0.29	0.45	4,014
No of persons aged 15+ in the hh (other than couple)	0.45	0.75	4,014
No of children aged 14 or younger in the hh	0.69	0.88	4,014
Housing type=house	0.33	0.47	4,013
Housing type=flat	0.67	0.47	4,013
Construction period=before 1946	0.14	0.35	3,957
Construction period=1946-1960	0.09	0.28	3,957
Construction period=1961-1970	0.17	0.38	3,957
Construction period=1971-1980	0.25	0.43	3,957
Construction period=1981-1990	0.22	0.41	3,957
Construction period=1991-1999	0.05	0.23	3,957
Construction period=2000 or later	0.08	0.27	3,957
Housing size (m2, capped at 450)	75.45	42.57	3,988
Housing ownership=owned	0.90	0.30	4,014
Housing ownership=rented	0.10	0.30	4,014
No of rooms (capped at 6)	3.10	1.18	4,014
2008 wave	0.51	0.50	4,014

Notes: ESU 2007 and 2008 waves pooled; estimated using survey weights on a sample of couple households whose head has positive earnings (in either data source) and worked full-time; person characteristics refer to the head of household; age variable is centered at sample mean (and divided by 10).

Table A3: Estimates with public employees (A) vs private employees (B)

	Survey income		Register income	
	(1) ML	(2) ML-IV	(3) ML	(4) ML-IV
β	0.132***	0.306***	0.081***	0.214***
	(0.018)	(0.058)	(0.015)	(0.038)
γ_B	0.017	0.022	0.050**	0.074***
	(0.021)	(0.021)	(0.023)	(0.023)
$\sigma^2_{\xi_A}$	0.181***	0.181***	0.199***	0.198***
	(0.014)	(0.014)	(0.019)	(0.019)
$\sigma^2_{\xi_B}$	0.192***	0.192***	0.354***	0.355***
	(0.008)	(0.008)	(0.021)	(0.021)
$1-\bar{\kappa}_B$	0.119	0.064	0.417**	0.234***
	(0.139)	(0.064)	(0.173)	(0.085)
$1-1/\bar{k}_B^{PW_l}$	0.119	0.064	0.417**	0.234***
	(0.139)	(0.064)	(0.173)	(0.085)
$1-1/\bar{k}_B^{PW_u}$	0.128	0.074	0.501***	0.345***
	(0.138)	(0.064)	(0.148)	(0.074)
$1-\bar{\kappa}_B^{HLP}$	0.123	0.069	0.461***	0.292***
	(0.138)	(0.064)	(0.160)	(0.079)
# of public employees (A)	580	580	579	579
# of private employees (B)	2,437	2,437	2,396	2,396
Total obs	3,017	3,017	2,975	2,975
AIC	11,935,962	11,933,260	11,519,963	11,516,532
BIC	11,936,371	11,933,681	11,520,371	11,516,951
Partial R^2		0.1319		0.1563
F-test for excluded instr.-s		18.53		21.13
Endogeneity test (p-value)		0.0000		0.0001
Hansen J-test (p-value)		0.6236		0.1744

Notes: estimated using survey weights on a sample of couple households whose head has positive earnings and worked full-time (2007-08 waves pooled); standard errors (shown in parantheses) are clustered at the household level; *dependent variable* = ln housing costs; *income* = ln earnings; *instruments* = head education level, occupation and industry; *covariates* = head gender, age (centered), age squared, nationality, marital status; no of children and (other) adults in the hh, region, rural area, wave and housing characteristics (type, year of construction, no of rooms, size in m2, ownership); *public/private* = household head is employed in the public/private sector and not considered a self-employed; * p<0.1, ** p<0.05, *** p<0.01.

Table A4: Sensitivity analysis

	Estimates with survey income												
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
β	0.302***	0.329***	0.281***	0.313***	0.301***	0.293***	0.294***	0.273***	0.297***	0.452***	0.362***	0.295***	0.302***
	(0.054)	(0.064)	(0.046)	(0.042)	(0.051)	(0.060)	(0.052)	(0.064)	(0.077)	(0.064)	(0.061)	(0.048)	(0.052)
γ_B	0.023	0.025	0.030	0.035*	0.026	0.020	0.027	0.018	0.026	-0.057*	0.018	0.022	0.024
	(0.022)	(0.021)	(0.020)	(0.020)	(0.018)	(0.022)	(0.021)	(0.034)	(0.032)	(0.030)	(0.021)	(0.020)	(0.025)
γ_C	0.141***	0.145***	0.133***	0.152***	0.140***	0.136***	0.132***	0.179***	0.100**	0.102*	0.119***	0.124***	0.142***
	(0.041)	(0.041)	(0.037)	(0.040)	(0.038)	(0.037)	(0.038)	(0.054)	(0.050)	(0.053)	(0.038)	(0.033)	(0.044)
$1-\bar{\kappa}_B$	0.068	0.072	0.096	0.091	0.078	0.063	0.084	0.039	0.098	-0.140*	0.034	0.068	0.066
	(0.069)	(0.058)	(0.067)	(0.057)	(0.059)	(0.070)	(0.065)	(0.120)	(0.092)	(0.077)	(0.058)	(0.064)	(0.078)
$1-\bar{\kappa}_C$	0.250**	0.214**	0.256**	0.237**	0.235**	0.251**	0.242**	0.350**	0.189	0.048	0.203**	0.258***	0.250**
	(0.110)	(0.095)	(0.101)	(0.093)	(0.107)	(0.102)	(0.097)	(0.140)	(0.129)	(0.114)	(0.090)	(0.084)	(0.110)
# of public employees (A)	580	582	634	843	882	414	577	315	265	580	580	514	556
# of private employees (B)	2,437	2,498	2,625	3,087	3,410	2,020	2,341	1,258	1,179	2,437	2,437	2,263	2,461
# of self-employed (C)	345	347	390	399	462	295	316	174	171	345	345	585	345
Total obs	3,362	3,427	3,649	4,329	4,754	2,729	3,234	1,747	1,615	3,362	3,362	3,362	3,362
AIC	13,238,113	13,030,728	14,609,185	20,963,726	18,091,817	9,976,054	12,717,809	6,261,371	6,253,333	13,439,434	13,135,990	13,248,323	13,268,314
BIC	13,238,572	13,031,121	14,609,651	20,964,229	18,092,328	9,976,498	12,718,266	6,261,770	6,253,727	13,439,893	13,136,449	13,248,782	13,268,773
Partial R^2	0.1325	0.0562	0.1293	0.1666	0.1148	0.1224	0.1405	0.1413	0.1315	0.1325	0.1253	0.1341	0.1333
F-test for excluded instr.-s	19.33	47.07	20.13	33.97	22.67	13.84	19.67	12.85	13.34	19.33	18.82	19.63	19.42
Endogeneity test (p-value)	0.0000	0.0003	0.0000	0.0000	0.0000	0.0004	0.0000	0.0006	0.0019	0.0000	0.0000	0.0000	0.0000
Hansen J-test (p-value)	0.5083	0.4835	0.4685	0.5283	0.0991	0.2233	0.6322	0.4654	0.7607	0.6435	0.5391	0.4609	0.5069
	Estimates with register income												
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
β	0.215***	0.228***	0.190***	0.237***	0.209***	0.203***	0.216***	0.201***	0.222***	0.293***	0.228***	0.201***	0.217***
	(0.034)	(0.041)	(0.031)	(0.034)	(0.033)	(0.038)	(0.036)	(0.049)	(0.051)	(0.044)	(0.037)	(0.034)	(0.035)
γ_B	0.073***	0.073***	0.071***	0.051**	0.056***	0.096***	0.073***	0.043	0.098***	0.005	0.069***	0.066***	0.073***
	(0.024)	(0.024)	(0.022)	(0.021)	(0.021)	(0.025)	(0.022)	(0.036)	(0.036)	(0.032)	(0.023)	(0.022)	(0.023)
γ_C	0.253***	0.258***	0.243***	0.229***	0.226***	0.275***	0.245***	0.256***	0.243***	0.247***	0.244***	0.186***	0.245***
	(0.046)	(0.052)	(0.040)	(0.054)	(0.042)	(0.052)	(0.042)	(0.068)	(0.064)	(0.063)	(0.045)	(0.036)	(0.045)
$1-\bar{\kappa}_B$	0.232***	0.213***	0.251***	0.145*	0.171*	0.325***	0.231***	0.125	0.306***	-0.061	0.207**	0.224**	0.233***
	(0.088)	(0.081)	(0.089)	(0.074)	(0.091)	(0.093)	(0.085)	(0.165)	(0.112)	(0.114)	(0.085)	(0.088)	(0.086)
$1-\bar{\kappa}_C$	0.561***	0.519***	0.595***	0.423***	0.518***	0.639***	0.558***	0.593***	0.529***	0.386***	0.546***	0.498***	0.538***
	(0.095)	(0.097)	(0.099)	(0.124)	(0.100)	(0.122)	(0.105)	(0.132)	(0.127)	(0.118)	(0.093)	(0.098)	(0.098)
# of public employees (A)	579	589	648	803	837	440	579	307	272	579	579	503	577
# of private employees (B)	2,396	2,449	2,575	3,049	3,395	1,969	2,396	1,233	1,163	2,396	2,396	2,213	2,366
# of self-employed (C)	306	307	343	363	405	262	284	156	150	306	306	565	338
Total obs	3,281	3,345	3,566	4,215	4,637	2,671	3,259	1,696	1,585	3,281	3,281	3,281	3,281
AIC	12,658,765	12,853,090	13,998,692	20,473,284	17,464,964	9,535,421	12,588,820	6,114,329	6,024,277	12,846,529	12,613,991	12,731,497	12,659,990
BIC	12,659,222	12,853,482	13,999,155	20,473,785	17,465,473	9,535,863	12,589,276	6,114,726	6,024,669	12,846,986	12,614,448	12,731,954	12,660,447
Partial R^2	0.1502	0.0750	0.1414	0.1220	0.1232	0.1502	0.1474	0.1551	0.1504	0.1502	0.1512	0.1571	0.1520
F-test for excluded instr.-s	23.71	67.08	24.27	19.11	25.08	18.23	23.05	16.00	15.13	23.71	23.11	25.18	24.03
Endogeneity test (p-value)	0.0000	0.0001	0.0000	0.0000	0.0000	0.0007	0.0000	0.0007	0.0017	0.0000	0.0000	0.0001	0.0000
Hansen J-test (p-value)	0.1582	0.4859	0.2574	0.6819	0.0296	0.0578	0.1380	0.4199	0.2601	0.3902	0.1603	0.1050	0.1668

Notes: model (1) estimated using survey weights on a sample of couple households whose head has positive earnings and worked full-time (2007-08 waves pooled); standard errors (shown in parantheses) are clustered at the household level; *dependent variable* = ln housing costs; *income* = ln earnings; *instruments* = head education level, occupation and industry; *covariates* = head gender, age (centered), age squared, nationality, marital status; no of children and (other) adults in the hh, region, rural area, wave and housing characteristics (type, year of construction, no of rooms, size in m2, ownership); *public/private* = household head is employed in the public/private sector and not considered a self-employed; *self-employed* = household head worked as a self-employed in the income reference period (based on ESU); sensitivity tests: (2) only head's education level used as instruments, (3) sample also includes hh heads working part-time/part-year, (4) also single households, (5) also other (non-single) households, (6) only hh heads aged 25-55, (7) only hh heads with positive earnings in both data sources, (8) 2007 wave only, (9) 2008 wave only, (10) housing cost includes rent and mortgage interests payments, (11) total household income (instead of earnings), (12) also self-employed hh if non-head hh member working as a self-employed, (13) also self-employed hh if head has self-employment income; * p<0.1, ** p<0.05, *** p<0.01.