Cheating in Europe: underreporting of self-employment income in comparative perspective

Merike Kukk[1,2] · **Alari Paulus**[3,4] · **Karsten Staehr**[1,2]

Abstract

Various national studies have used the expenditure method (Pissarides and Weber in J Public Econ 39(1):17–32, 1989) to estimate income underreporting by the self-employed relative to the wage earners. Within Europe, the studies mostly consider the UK or individual Nordic countries, while no data are available for most Southern European and Eastern European countries. This paper is the first to apply the expenditure method to a large number of EU countries using harmonised microdata and a common model specification to enhance cross-country comparability. We extend the number of countries studied using the expenditure method and contribute to the scarce comparative literature on tax non-compliance in general. Our estimates show substantial variation in income underreporting across countries, from under 10% to more than 40% of self-employed household income on average. The shares of underreporting do not appear to be related to the development level of the countries.

Keywords Income underreporting · Self-employment · Expenditure method · Tax compliance · Cross-country comparison

JEL Classification H26 · E21 · E26 · H24

1 Introduction

Estimates of the size of the unreported economy are important for tax policy, tax administration and auditing, and also for economic and social policy-making in general. This

✉ Alari Paulus
apaulus@essex.ac.uk

1 Tallinn University of Technology, Tallinn, Estonia

2 Bank of Estonia, Tallinn, Estonia

3 Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

4 Praxis Centre for Policy Studies, Tallinn, Estonia

is particularly the case for underreporting of income by the self-employed, since it is likely to be relatively prevalent due to the lack of third-party reporting and the substantial discretion the self-employed have about the information they make available to the tax and statistics authorities. Slemrod (2007) reports data for the USA in 2001 showing that the tax gap, the share of the taxable base not reported to the tax authorities, is likely marginal for wage and salary income, while it is large for capital gains and very substantial for business-related income. The underreporting of income is likely to stem from both underreporting of gross income and over-reporting of deductions and other offsets. A study of randomised tax audits in Denmark documents that income underreporting was much more prevalent among the self-employed than among the employed, and it is concluded that effective auditing of the self-employed is crucial for the compliance of the self-employed (Kleven et al. 2011).

The concerns about income underreporting of self-employed have only strengthened after the global financial crisis in the 2000s, since the subsequent recession led to notable changes in employment opportunities in many countries. In the UK, the jobs recovery after the recession has in large part been driven by growth in the number of self-employed (Hatfield 2015). Moreover, policy discussions often refer to a perceived need for more emphasis on entrepreneurship and self-employment (European Commission 2016). Increased emphasis on self-employment and the absence of third-party reporting underscore the importance of assessing how much the self-employed underreport income and evade taxation.

Even though tax and auditing schemes are national, it is valuable for policy makers to compare the extent of compliance across countries for benchmarking. Cross-country comparison may show the "seriousness" of non-compliance in a country and allow comparison with a peer group of countries. Such comparisons could be particularly valuable for European countries, as they share many economic and institutional features. It may also provide insights into the effectiveness of various tax and auditing schemes and open avenues for detailed country-specific investigation. The usefulness of cross-country studies of non-compliance and underreporting for policy-making explain the prevalence of such studies.

A number of studies produce cross-country estimates of various forms of non-compliance and underreporting using aggregate or macroeconomic data. Buehn and Schneider (2012) and Alm and Embaye (2013) provide estimates of the size of the shadow economy for countries around the world, while Tafenau et al. (2010) provide similar estimates for European countries. Poniatowski et al. (2016) compare the VAT gap in all EU countries. Studies that use aggregate data for the estimations typically find that the size of underreporting is inversely related to the development level of the countries.

There are relatively few studies using microdata to produce cross-country estimates. Eurobarometer (2014) presents survey results on the prevalence of undeclared work and the purchase of goods or services produced using undeclared work across the EU countries. Hazans (2011) uses data from the European Social Survey for 2004–2009 and concludes that the share of informal employment varies substantially across countries and exhibits no clear regional patterns. It is notable that studies using microeconomic data typically find no clear relationship between the extent of unreported economic activities and the economic development level of the countries and

also exhibit very limited geographical clustering. The same applies to surveys of tax morality.[1]

At the time of writing in 2019, no studies have provided cross-country comparison of income underreporting by the self-employed, a group of taxpayers seen as particularly prone to underreporting given the lack of third-party reporting. The most widely used method to estimate income underreporting by the self-employed, the so-called expenditure method, originates from Pissarides and Weber (1989) and aims to infer income gaps from household expenditure patterns. While the method has been used for a number of countries in Europe and elsewhere (see references in Sect. 2), the applications have focused on individual countries and within Europe mainly the UK and the Nordic countries. Moreover, the studies typically feature substantial differences in model specifications and collection of underlying household-level microdata and these differences in the individual country studies may impede cross-country comparison.

The paper addresses the gaps in the literature on non-compliance of the self-employed discussed above. We use the 2010 European Union Household Budget Survey to estimate the extent of income underreporting by the self-employed in a large number of EU countries employing a common modelling framework. We extend the relatively short list of European countries covered previously and provide estimates of income underreporting by the self-employed using the expenditure method for mainly Southern European and Eastern European countries for which such estimates have not been available. Overall, we contribute to the scarce comparative literature on underreporting and tax compliance using microeconomic data sources.

Our estimations show that the share of income not reported by the self-employed is relatively large in many European countries and there is substantial variation across the countries. The baseline estimates of the average share of unreported self-employment income range from under 10% to more than 40% of true income across the sample countries. The results are consistent with previous studies using the expenditure method for two individual countries, Estonia and Spain, which overlap with our country coverage, and this broadly corroborates the results found in this study using data from the EU-HBS. The shares of underreporting do not appear to be related to the development level of the countries, though suggest some broad country grouping based on structural, institutional and cultural commonalities. We also find that the estimates exhibit some sensitivity to the criterion used to define self-employed households so comparisons of results across countries must take into account these modelling choices.

The rest of this paper is organised as follows. Section 2 reviews previous studies of income underreporting by the self-employed using the expenditure method. Section 3 presents briefly the empirical model and estimation methodology used in this paper. Section 4 discusses the dataset and features of the data. Section 5 presents the baseline results, and Sect. 6 contains robustness checks. The results and their broader context are discussed in Sect. 7. Finally, Sect. 8 concludes.

---

[1] Lago-Peñas and Lago-Peñas (2010) use data for 2005 from the European Social Survey to map public attitudes to tax morality across countries in Europe. A much larger share of respondents in Estonia and Portugal than in Belgium and Germany support the statement that citizens should not cheat on their taxes.

2 Studies using the expenditure method

Pissarides and Weber (1989), or P&W, devised an inventive method to estimate the extent of income underreporting by the self-employed relative to a reference group perceived to be less prone to underreporting. The main idea is that households are likely to report consumption expenditures correctly to the statistics authorities collecting data for the household budget survey, while some households may choose to underreport their income. Households, including self-employed households, have little reason or incentive to underreport their income to the statistics authorities unless they have not truthfully reported their full income to the tax authorities (Pissarides and Weber 1989). This suggests that the estimated underreporting of income to the statistics authorities will be positively correlated with underreporting to the tax authorities and the estimates are thus of substantial academic and policy relevance.[2]

Household budget surveys typically require households to collect detailed data on their consumption spending using diaries or scanners, so consumption data are likely to be quite precise. However, household budget surveys usually do not require documentation of income data and this means that households can underreport their true income, an option which may be tempting if they have underreported income to the tax authorities and are concerned that income data reported to the statistics authorities may somehow be shared with the tax authorities. P&W argue that the self-employed are a group that may be particularly prone to this form of underreporting. If the self-employed and the employed have similar preferences for consumption, then it becomes possible to back out the share of income not reported by the self-employed.

The methodology has been used in a number of studies for individual countries, but has not been applied consistently to study and compare underreporting in a wide set of countries. Table 1 provides an overview of the key studies in the field, which mainly focus on European and North American countries. The share of unreported income in true income is typically reported as an interval because the variances of fluctuations in permanent income and the underreporting factors cannot be disentangled; see also the detailed discussion of the methodology in Sect. 3 and Appendix A.

All studies focus on one particular country with the exception of Kim et al. (2017), who produce estimates for two countries, Russia and South Korea, though at different time periods and using nation-specific data sources. Another feature is the geographical concentration of the countries studied. Within Europe, most studies consider the Nordic countries and the UK. For Southern Europe there are only studies for Spain and Northern Cyprus, and for Eastern Europe there are only studies for Estonia. There are no published studies for the majority of countries in Southern and Eastern Europe.

The majority of the studies use national household budget or expenditure surveys in which food consumption data are collected in detail. Paulus (2015) uses data from the Estonian version of the EU-SILC and register data from the Estonian tax authorities. Most of the published studies use data from before the mid-2000s although Cabral

---

[2] Hurst et al. (2014) apply the expenditure method to US data and get very similar results with two different surveys, the Consumer Expenditure Survey and the Panel Study of Income Dynamics, suggesting that the households report data in a consistent manner in the survey and potentially in other contexts. Paulus (2015) uses the Estonian SILC data linked with tax records and finds that there is a certain overlap between the reporting of income to the two entities.

Table 1 Overview of key studies using the expenditure method. *Sources*: Updated and revised from Kukk and Staehr (2014) and Paulus (2015); see also Schmutz (2018)

	Country	Database and time sample	Definition of self-employment	Unreported income as share of true income[a]
Pissarides and Weber (1989)	UK	Family Expenditure Survey, 1982	Share of business income ≥ 25%	Blue collar: 34–39% White collar: 22–35%
Schuetze (2002)	Canada	Family Expenditure Survey, 6 years within 1969–1992	Share of business income ≥ 30%	6–22%
Lyssiotou et al. (2004)	UK	Family Expenditure Survey, 1993	Main source of income from self-employment	Blue collar[b]: 27–29% White collar[b]: 8–21%
Johansson (2005)	Finland	Household Expenditure Survey, 1994–1996	Status as self-employed	One person self-employed: 9–19% Two people self-employed: 27–32%
Besim and Jenkins (2005)	Northern Cyprus	Households Consumption Expenditure Survey, 1998–1999	Status as self-employed; privately employed	Self-employed[c]: 11–13% Privately employed[c]: 10–14%
Engström and Holmlund (2009)	Sweden	Household Budget Survey, 1999–2004	Status as self-employed	Incorporated business: 14–15% Unincorporated business: 30–33%
Martinez-Lopez (2013)	Spain	Household Budget Surveys, 2006–2009	Status as self-employed	15–25%, sensitive to definition of the self-employed
Hurst et al. (2014)	USA	Consumer Expenditure Survey 1980–2003, Panel Study of Income Dynamics, 1980–1997	Status as self-employed	CEX: 25% PSID: 28–29%
Kukk and Staehr (2014)	Estonia	Household Budget Survey, 2002–2007	Business income ≥ 20%	62% when households with 0–20% business income excluded; lower if the cut-off is reduced
Paulus (2015)	Estonia	Estonian Social Survey and tax register, 2008	Status as self-employed	Survey: 20–44% Tax register: 48–71%
Ekici and Besim (2016)	Northern Cyprus	Households Consumption Expenditure Survey, 2008	Status as self-employed; privately employed	Self-employed[c]: 21% Privately employed[c]: 14%

Table 1 continued

	Country	Database and time sample	Definition of self-employment	Unreported income as share of true income[a]
Kim et al. (2017)	Russia, South Korea	Longitudinal Monitoring Survey 1994–2001 (Russia), Labour Income Panel Survey, 2000–2005 (Korea)	Status as self-employed	Russia: 28% Korea: 29%
Engström and Hagen (2017)	Sweden	Household Budget Survey, longitudinal tax database LINDA, 2003–2009	Self-employment income>0	15–24%
Kukk and Staehr (2017)	Estonia	Household Budget Survey, 2002–2007	Share of business income ≥ 0%; status as self-employed	Share of business income: 43% Status as self-employed: 28% Self-employed with business income: 53–56%
Schmutz (2018)	Switzerland	Household Budget Survey, 2006–2011	Share of business income ≥ 25%; status as self-employed	Share of business income: 16–29% Status of self-employment: 13–25%
Nygård et al. (2019)	Norway	Survey of Consumer Expenditure, 2003–2009 and 2012	Status as self-employed; wage earners with possibility for tax evasion	Status as self-employed[d]: 13% Wage earners with possibility for tax evasion[d]: 3%
Cabral et al. (2019)	UK	Living Costs and Food Survey, 2010–2012	Share of business income ≥ 25%, status as self-employed	Share of business income: 20% Status of self-employment: 22%

[a]Main estimates with instruments where available, reported as the share of true income

[b]Lyssiotou et al. (2004) find larger underreporting using a nonparametric method and a demand system approach

[c]The reference group consists of households where the household head is publicly employed

[d]The reference group consists of wage earners deemed to be without possibility for evasion based on their educational background

et al. (2019), Schmutz (2018) and Nygård et al. (2019) use datasets with some or all observations from after the global financial crisis and the subsequent recession. The data are typically cross-sectional and different waves are pooled so that the average rate of income underreporting over the sample period is computed.

The studies differ in the way they have defined or identified self-employed households. One approach is the one introduced by P&W who define households as self-employed if the reported share of business income in total reported income exceeds a given threshold. Schuetze (2002), Kukk and Staehr (2014, 2017) and Engström and Hagen (2017) use the same definition with different cut-off points. Lyssiotou et al. (2004) define households as self-employed if their main source of income is from self-employment. The second approach relies on the self-reported employment status of the household head and has been used by Besim and Jenkins (2005), Johansson (2005), Engström and Holmlund (2009), Martinez-Lopez (2013), Hurst et al. (2014), Paulus (2015) and Ekici and Besim (2016).

Many studies use a reference group made up of households where the household head is employed. Some studies have imposed more restrictions on the reference group. Hurst et al. (2014) exclude households where the reference person is a wage earner, but the spouse (if present) is self-employed. Besim and Jenkins (2005), Paulus (2015) and Ekici and Besim (2016) produce estimates where the reference group is restricted to public sector employees. Nygård et al. (2019) argue that wage earners that are trained within construction and repair work may even as wage earners be able to underreport their income and therefore leave out these workers from the reference group.

The expenditure method relies on the permanent income hypothesis, which posits that consumption is a function of the permanent income of households. Many studies, including the one by Pissarides and Weber (1989), do not have access to a measure of permanent income and this leads to a knotty identification issue since the variances in underreporting and in temporary income across households cannot be disentangled. The expenditure method addresses this issue by imposing some assumptions about these variances and this makes it is possible to calculate upper and lower bounds for the share of true income that is not reported.

Most studies cited in Table 1 only have data available on current income and therefore provide intervals for the share of income underreporting. Exceptions include Kim et al. (2017) and Engström and Hagen (2017) who have available panel data and therefore are able to compute proxy measures of permanent income, which makes it possible to produce point estimates of the extent of underreporting. Kukk and Staehr (2014, 2017) use a self-reported measure of regular income as an alternative. Other studies produce simplified point estimates by using various additional assumptions; see Besim and Jenkins (2005), Engström and Holmlund (2009), Martinez-Lopez (2013), Hurst et al. (2014), Ekici and Besim (2016) and Nygård et al. (2019).

When the results of earlier studies are considered, four features stand out. First, the studies generally find the shares of income unreported by the self-employed to be quite large, averaging around 20–30% of true income across the studies. It should be noted however that estimates of the overall size of the shadow economy are also substantial (Schneider 2007; Alm and Embaye 2013) and a large share of the shadow economy is likely to be linked to business or self-employment income that is left unreported.

The relatively large share of income underreporting reported in the published studies that use the expenditure method may, however, also reflect a publication bias, where studies with statistically and economically significant results end up being published, while others may remain unpublished, emphasising further the importance of our comparative study.

Second, there is substantial variation in the estimates within and across different studies. The estimates vary for different groups of interest such as white-collar and blue-collar workers, households with one person self-employed or two persons self-employed, and incorporated or unincorporated businesses. The differences in the results across different groups suggest that the estimates of income underreporting vary substantially depending on the method used to define the self-employed and we will explore it further in Sect. 6.

Third, the variety of sample criteria and model specifications makes it very challenging to compare the results across countries; see also Schmutz (2018). The diversity in the results may stem from the use of different datasets and methodological choices or they may reflect genuine differences across the countries in the underreporting of income by the self-employed. This uncertainty may be eliminated in our study given that we use data that are harmonised across countries and run the estimations with the same model specification and with the same sample selection criteria.

Fourth, the results across the different countries do not suggest clear patterns across levels of development or regions. Table 1 reveals that the estimates of the underreporting shares are substantial for countries with high per capita GDP such as the USA and the UK and smaller for a country such as Spain with substantially lower per capita GDP. Indeed, very small underreporting shares are found by Besim and Jenkins (2005) and Ekici and Besim (2016) for Northern Cyprus, one of the poorest regions in Europe. The diversity in the results across the countries is in line with the discussion in Sect. 1, which concluded that the estimates of underreporting based on microeconomic data often do not bear any direct relation to the development level of the countries considered.

3 Empirical approach

We follow closely the framework in Pissarides and Weber (1989); full details of the structural model and the derivation of the empirical specification are provided in Appendix A. We estimate a reduced-form consumption equation:

$$\log c_i = \alpha + \beta \log \hat{y}'_i + \gamma D_i + X'_i \phi + \eta_i \tag{1}$$

where c_i is the spending on food, $\hat{y}'_i$ is the instrumented income variable, D_i is a dummy variable taking the value $D_i = 1$ for households that are self-employed and $D_i = 0$ for households in the reference group, X_i is a vector of control variables, and η_i is an error term. The term α is a constant, β is the marginal propensity to consume coefficient, γ is a shift parameter for the self-employed relative to the wage earners, and ϕ is a vector of the coefficients of the control variables. Note that the final empirical specification in Eq. (1) corresponds to Eq. (A.5) in Appendix A where

$\beta(\mu_k - \mu_p)$ is replaced with γD_i and the three error terms collected into one labelled η_i.

The income equation contains the same explanatory variables as the consumption equation and additional variables or instruments. Using the instrumental variable approach serves three purposes. First, it provides the measures of the variances of the income shocks for the two groups that are needed for computing the underreporting share. Second, it addresses the possible endogeneity of the income variable, as emphasised by P&W. Third, the disposable income variable may be subject to measurement errors in survey data, as stressed by Lamarche (2017) in his study comparing income data in EU-SILC and EU-HBS.

The estimated coefficients of Eq. (1) allow us to calculate the lower and upper bound on the mean underreporting factor k (as explained in Appendix A):

$$\bar{k} \in \left[\bar{k}_{\mathrm{L}}, \bar{k}_{\mathrm{U}}\right] = \left[\exp\left(\frac{\gamma}{\beta} - \frac{1}{2}(\sigma_{\xi|S}^2 - \sigma_{\xi|W}^2)\right), \ \exp\left(\frac{\gamma}{\beta} + \frac{1}{2}(\sigma_{\xi|S}^2 - \sigma_{\xi|W}^2)\right)\right] \quad (2)$$

where $\sigma_{\xi|S}^2$ and $\sigma_{\xi|W}^2$ are the variances of the residuals ξ_i in the income equation for the self-employed and the wage earners, respectively. The lower bound $\bar{k}_{\mathrm{L}}$ may be considered a quite extreme or unlikely outcome given that it requires the underreporting factor to be identical for all self-employed households. The lower bound is thus a very conservative estimate with little practical relevance. A special case emerges when the variances of the residuals in the instrumentation step are identical for the two groups, narrowing the bounds into a simple point estimate calculated as[3]:

$$\bar{k}_0 = \exp\left(\frac{\gamma}{\beta}\right) \quad (3)$$

In studies following the P&W approach, it is customary to report not the underreporting factor but the underreporting share, which is the share of true income not reported. The underreporting factors in Eqs. (2) and (3) can be converted into the shares of true income not reported using the formulas:

$$\text{"Lower share"} = (\bar{k}_{\mathrm{L}} - 1)/\bar{k}_{\mathrm{L}}$$
$$\text{"Upper share"} = (\bar{k}_{\mathrm{U}} - 1)/\bar{k}_{\mathrm{U}}$$
$$\text{"Simple share"} = (\bar{k}_0 - 1)/\bar{k}_0 \quad (4)$$

When interpreting the different measures of the share of income underreporting in Eq. (4), it is important to bear in mind that they all pertain to the same underlying estimation results and merely are derived under different assumptions about the variances of the underreporting and income processes.

It is also important to recall that the estimates of underreporting by the self-employed are *additional to* the possible underreporting by the chosen reference group. Households in the reference group may also underreport their income in the survey. The

---

[3] This special case is also considered by e.g. Besim and Jenkins (2005), Engström and Holmlund (2009), Martinez-Lopez (2013), Hurst et al. (2014) and Ekici and Besim (2016).

way the reference group is defined does not rule out that the households in this group also underreport their income. The expenditure method cannot be used to estimate the possible underreporting by the reference group but only the *additional* underreporting of the self-employed on top of any underreporting by the reference group of wage earners.

The consumption and income equations are estimated simultaneously using the maximum likelihood (ML) method. As in Paulus (2015), this allows us to calculate standard errors for the lower and upper bounds of the share of income underreporting, which have been lacking in other studies. Technically, we specify a structural equation model (SEM) with two equations (consumption and income) and allow certain parameters to differ between the employed and the self-employed (intercepts, variance and covariance of the error term). One advantage of this is that it allows us to obtain the residual variance of income for the two groups in a straightforward way. The estimated elasticity of consumption and the coefficient of the self-employment dummy are used together with the estimated variances of the error term of the income regression to compute the various underreporting factors as expressed in Eqs. (2) and (3), and subsequently the shares of income underreporting using Eq. (4).

4 Dataset and the properties of the data

The underreporting by the self-employed is estimated using the 2010 wave of the EU Household Budget Survey. As of 2019, this is the only available wave of EU-HBS which is harmonised across countries. The choice of dataset is dictated by it having sufficiently detailed data on consumption, income, employment status and background characteristics for a large number of EU countries.[4]

The EU-HBS for 2010 provides harmonised data for 25 EU countries collected by national statistics offices.[5] In addition to detailed information on each consumption category, the survey contains information on household net income, the employment status of the household members, and demographic and socio-economic characteristics. However, the data available vary across the countries and this constrains our cross-country analysis to those countries for which the core variables needed to apply the expenditure method are available. We follow the standard approach taken in other studies and define food consumption as spending on food eaten at home (COICOP category 01) and outside the home (called catering services in the EU-HBS). National statistics offices use detailed diaries to collect data on expenditures though in some cases this is limited to certain categories such as food consumption (Eurostat 2015).

The method of collecting data on income varies somewhat across countries, as data are collected at a more detailed level in some countries than they are in other

---

[4] We also considered using other databases but found the EU-HBS to be the most appropriate. There are two potential datasets containing harmonised survey data across European countries: the EU Statistics on Income and Living Conditions (EU-SILC) for the EU countries and the Household Finance and Consumption Survey (HFCS) for the euro area countries. These two datasets have, however, very limited information on food consumption, and the HFCS has quite small sample sizes in several of the countries.

[5] The 2016 October release of EU-HBS we are using provides no data for Austria, Germany or the Netherlands.

countries. We had to exclude several countries from the sample because of the issues with the data availability or the low response rate. That leaves us with 14 countries for the analysis: Bulgaria (BG), Cyprus (CY), the Czech Republic (CZ), Estonia (EE), Greece (EL), Spain (ES), Croatia (HR), Hungary (HU), Ireland (IE), Lithuania (LT), Latvia (LV), Poland (PL), Portugal (PT) and Romania (RO).[6] Our sample covers a range of Southern and Eastern European countries for which no estimates have been provided so far. The samples overlap only for Estonia and Spain, cf. Table 1 in Sect. 2.

The standard approach in the literature is to restrict the sample to households with two adults where at least the household head is working. This ensures that the households exhibit a fairly similar family structure, which may be important if the family structure affects the composition of expenditures. Our sample consists of two-adult households where the household head is working and earnings are the main source of income of the household. In common with several previous studies, we further restrict the sample to households where the household head is of working age, 24–59, to ensure more homogenous preferences for consumption.[7]

As discussed in Sect. 2, two different methods may be used to define or categorise the group of households deemed particularly prone to income underreporting, either the self-reported employment status of the household head or the share of business income in total household income. The EU-HBS does not contain information on business income, and it is not possible to calculate the share of reported business income from total income, but the households report their main source of income as being either wage/salary, income from self-employment, property income, pensions/retirement benefits, unemployment benefit or other current benefits/other income. We define a household to be self-employed, and hence possibly prone to underreporting, if the following two criteria are both satisfied: (i) the household reports income from self-employment to be the main source of income, and (ii) the household reports at least one of the spouses to be self-employed.

The definition criteria for the reference group are equally noteworthy. Households which are not deemed prone to income underreporting are expectedly those where the main source of income is from wages or salaries and neither spouse is reported to be self-employed. Households where the main source of income and the self-reported type of employment appear to contradict each other are excluded from the sample altogether.[8]

Figure 4 in Appendix B shows the size of the samples for the countries used in the baseline estimations in Sect. 5. The size of the full estimation sample varies substantially across the countries, with more than 5000 households for Spain, Poland and Romania but fewer than 1000 households in half of the countries. The number of self-

---

[6] Data in the EU-HBS 2010 are for 2010 for all countries in our sample except Cyprus, Ireland and Lithuania, where the survey was collected in 2008–2009, 2009–2010 and 2008 instead.

[7] We also exclude outliers from the sample using the Chauvenet's criterion, identifying observations for which the data on food consumption or income of the household deviate substantially from the respective log-normal distributions. This reduces the number of observations in the sample by only 0.62% as 0.38% are considered outliers using income data and 0.25% using food consumption data.

[8] Kukk and Staehr (2014) find underreporting of income even by households which have only relatively small shares of business income.

employed is above 1000 for Spain, Poland and Romania, but below 100 for countries such as Bulgaria, the Czech Republic, Estonia and Latvia.

The relatively small estimation samples for several countries make it difficult to obtain precise estimates, and this complicates the interpretation of the results. We decided not to restrict the sample further. One common restriction is to exclude households that work in agriculture as they can grow and consume food which may not be reflected in the income and spending data. We considered excluding the households where the head is reported to be a farmer or an agricultural worker, but this would further restrict the sample of the countries in the analysis. Moreover, the data on detailed employment status are missing for some households in several countries, and this would restrict the sample further. We prefer to retain as many households and as many countries as possible in the sample and so we do not use the self-reported information about farmers in the baseline estimations.

Figure 5 in Appendix B shows the share of self-employed households as a percentage of the total number of households in the estimation sample for each country. The prevalence of the self-employed among working households varies greatly across the sample countries. In Greece, the share of the self-employed is over 30%, while in Hungary, Estonia and Ireland it is 10% or less. The ranking of the countries by the share of the self-employed is similar to the rankings using employment data from the Labour Force Survey (Eurostat) and from the National Accounts (AMECO).[9]

Table 3 in Appendix B shows the mean of the log food consumption and of the log income for employed and self-employed households and for the difference between the two groups. The statistics reveal substantial differences in the reporting of food consumption and disposable income across the sample countries. In many countries, the self-employed report higher food consumption than the employed, while the self-employed report lower income than the employed.

Figure 1 shows the fitted linear relationships between reported disposable income and food consumption for self-employed households and for employed households. The lines depict the predictions of a linear regression of disposable income on food consumption for the households in each of the groups. For all countries, the self-employed report higher food consumption than the wage earners at all or most income levels. In most countries, the slope depicting the relationship between consumption and income is slightly smaller for the self-employed than for the wage earners. Hurst et al. (2014) and Engström and Hagen (2017) find similar patterns for households in the USA and Sweden, respectively. At high-income levels, the differences in the food consumption of the self-employed and wage earners are small.[10]

The unconditional results in Fig. 1 (and Table 3) point to notable differences in the patterns of food consumption between the self-employed and wage earners, but also suggest that there is notable heterogeneity across the sample countries. The expenditure method makes it possible to investigate these differences and reach estimates of the extent of income underreporting by the self-employed relative to the underreporting by wage earners.

---

[9] Data from Eurostat LFS (detailed annual survey results, code: *lfsa*) and AMECO (code: *NSDT*).

[10] This may be seen using the LOWESS method; the results are not reported here but available upon request.
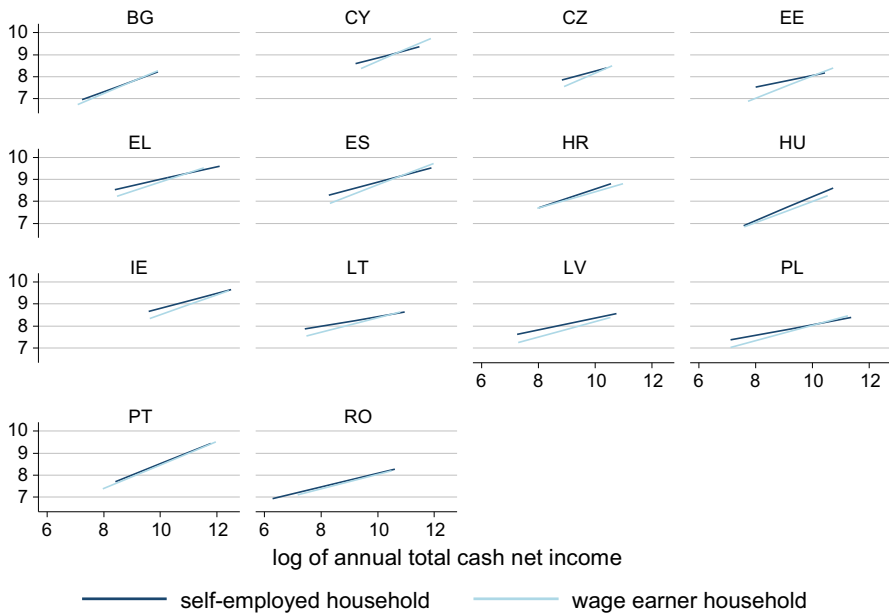
Fig. 1 Linear Engel curves for household food consumption. *Notes* Linear fit of log household food expenditures (at home and outside) on log household total disposable income using sample weights and excluding outliers in log expenditures and incomes (Chauvenet's criterion). The sample consists of couples with earnings as the main source of income where the household head is employed and 24–59 years old. Self-employed households are based on self-reported employment status and the main source of income. *Source*: Authors' calculations using the EU-HBS

5 Baseline estimations

For the baseline, we estimate Eq. (1) using the IV-ML procedure discussed in Sect. 3. Reported current income is instrumented, and the food regression is then estimated using the instrumented income variable, the dummy variable for self-employment and various control variables. The different underreporting factors and their associated shares are computed using the estimated coefficients from the food regression with instrumented income and the variances from the income regressions.

For the instrumentation of the reported income, we use variables for the educational attainment and gender of the household head and their interactions with a dummy for the self-employed household, together with all the explanatory variables from Eq. (1). Estimations of Mincer regressions generally find that the level of education is a suitable instrument for income. We considered using more instruments such as part time work, temporary work contract, occupation, industry, region or being an immigrant, but these variables are missing for several countries and would result in a loss of countries.

The control variables in the food regression are the five-year age groups of the household head and the spouse, the number of children within the age groups 0–4, 5–9, 10–14 and 15–24 excluding those aged 20–24 who are employed. The data availability in the EU-HBS conditions the specification of the empirical model to some extent, and

we do not add more control variables to the baseline model since this would reduce our sample unduly.

As discussed in Sect. 3, we consider three different underreporting factors and their corresponding shares of unreported income. A special measure of the underreporting factor can be computed solely from the estimated elasticity of consumption, β, and the coefficient of the dummy variable, γ, using Eq. (3). The lower and upper bounds of the underreporting factor can be found from estimates of β and γ and the variances of the residuals from the instrumentation of income, $\sigma^2_{\xi|W}$ and $\sigma^2_{\xi|S}$, using Eq. (2). In all cases, the factors are converted to shares of unreported income in true income using Eq. (4).

Table 2 presents the results of the baseline estimations using IV-ML estimation. The endogeneity test reported in column (8) suggests that income is an endogenous variable except for Bulgaria, Cyprus, Croatia, Ireland and Lithuania. We instrument, however, the income variable even in the cases where the test does not reveal endogeneity because the income may be subject to measurement error and because we aim to use the same model specification for all countries.

The diagnostics for the instrumentation are shown in columns (9)–(11). The instruments are generally statistically significant (not shown) and help explain 30–50% of the variation in the income variable. The F-statistic indicates that the instruments are suitably correlated with the instrumented variable. The results of the Hansen j test confirm the validity of the over-identifying instruments for all of the sample countries except Poland and Romania and with Spain as a borderline case. For these countries, we could use additional instruments, but we prefer to retain the same model specification for all countries so as to maintain the comparability of the results.

Columns (1) and (2) in Table 2 show the estimated values for the elasticity of food consumption to current income, β, and the coefficient of the self-employment dummy, γ. The elasticity of food consumption to current income varies between 0.3 and 0.7 across the countries. This range is in line with the findings in the studies in Table 1 where typical estimates are 0.2–0.3 for high-income countries and 0.4–0.6 for emerging economies. The estimated coefficient of the self-employment dummy varies between 0 and 0.2, but it is not statistically significant for four of the sample countries.

The simple share is shown in column (3). The estimated share is statistically insignificant for the same countries for which the self-employment dummy is insignificant, i.e. Bulgaria, Cyprus, Croatia and Portugal. The largest simple shares of above 30% are found for Latvia and Ireland, while shares of between 20 and 30% are found for the Czech Republic, Estonia, Hungary and Lithuania. Lower but still statistically significant simple shares of between 10 and 20% are obtained for Greece, Spain, Poland and Romania. We will discuss these results in greater detail in Sect. 7.

Columns (4)–(5) in Table 2 show the estimates of the lower and upper underreporting shares. We find, as expected, that the variance of the income residuals for wage earners is smaller than the variance for the self-employed ($\sigma^2_{\xi|W} < \sigma^2_{\xi|S}$) for all countries except Bulgaria and Cyprus, for which the differences are generally very small (not reported).

The lower bound denotes, under various assumptions, the lowest value of the share of unreported income that can be obtained, and the highest bound denotes the highest possible value. As discussed in Sect. 3, the assumptions required to attain the lowest

Table 2 IV-ML estimations of Engle curves and the computed shares of unreported income in true income. *Source*: Author's calculations using the EU-HBS

| | Estimated coefficients | | Computed shares of unreported income | | | Sample sizes | | Diagnostics | | | |
	(1) β	(2) γ	(3) Simple share	(4) Lower share	(5) Higher share	(6) Obs. total	(7) Obs. self-employed	(8) Endogeneity	(9) Partial R^2	(10) F-statistic	(11) Hansen j test
BG	0.524*** (0.103)	0.047 (0.055)	0.086 (0.103)	0.074 (0.104)	0.098 (0.110)	463	54	0.939	0.151	14.2	0.715
CY	0.398*** (0.093)	0.038 (0.054)	0.091 (0.118)	0.080 (0.122)	0.102 (0.114)	788	104	0.342	0.207	25.6	0.958
CZ	0.563*** (0.108)	0.139*** (0.037)	0.219*** (0.051)	0.193*** (0.054)	0.244*** (0.049)	638	77	0.054	0.143	26.0	0.348
EE	0.690*** (0.113)	0.171** (0.080)	0.220** (0.095)	0.186* (0.105)	0.251** (0.096)	911	89	0.063	0.171	21.4	0.873
EL	0.502*** (0.074)	0.075** (0.034)	0.139** (0.067)	0.021 (0.077)	0.242*** (0.065)	725	248	0.014	0.191	20.6	0.426
ES	0.515*** (0.031)	0.075*** (0.021)	0.135*** (0.036)	0.107*** (0.037)	0.163*** (0.036)	7071	1402	0.017	0.214	148.1	0.088
HR	0.368*** (0.079)	0.067 (0.051)	0.168 (0.113)	0.140 (0.117)	0.195* (0.111)	514	97	0.931	0.256	29.6	0.505
HU	0.555*** (0.043)	0.152*** (0.026)	0.240*** (0.039)	0.222*** (0.040)	0.257*** (0.038)	2296	192	0.000	0.325	77.9	0.729
IE	0.468*** (0.063)	0.168*** (0.037)	0.301*** (0.053)	0.251*** (0.058)	0.348*** (0.050)	1350	137	0.109	0.254	30.2	0.116

Table 2 continued

	Estimated coefficients		Computed shares of unreported income			Sample sizes		Diagnostics			
	(1) β	(2) γ	(3) Simple share	(4) Lower share	(5) Higher share	(6) Obs. total	(7) Obs. self-employed	(8) Endogeneity	(9) Partial R^2	(10) F-statistic	(11) Hansen j test
LT	0.414*** (0.079)	0.122** (0.048)	0.255* (0.097)	0.200* (0.109)	0.306*** (0.092)	1674	220	0.401	0.156	23.0	0.240
LV	0.565*** (0.112)	0.243*** (0.080)	0.350*** (0.105)	0.258** (0.123)	0.430*** (0.094)	628	78	0.020	0.176	12.5	0.107
PL	0.416*** (0.015)	0.062*** (0.010)	0.138*** (0.022)	0.068*** (0.025)	0.203*** (0.022)	11,208	2294	0.000	0.213	282.9	0.002
PT	0.700*** (0.066)	0.080 (0.068)	0.108 (0.084)	0.051 (0.092)	0.162** (0.081)	2065	303	0.000	0.364	67.4	0.300
RO	0.374*** (0.016)	0.067*** (0.017)	0.165*** (0.035)	0.068 (0.040)	0.252*** (0.033)	5923	1027	0.000	0.285	138.6	0.017

IV-ML estimations where education levels and gender of the reference person (and interactions) are used as instruments for the income variable. Robust standard errors are reported in brackets under the coefficient estimates

Superscripts ***, ** and * indicate that the coefficient is statistically different from 0 at the 1%, 5% and 10% levels, respectively. p values are reported for the endogeneity test in column (8) and the Hansen j test in column (11)

bound are quite improbable in practice and arguably mostly of theoretical interest. The upper bound is reached when the variance of the transitory income for the wage earners is the same as the variance for the self-employed.

For many of the countries, the difference between the upper and the lower share estimates is comparatively small. The main exceptions are Greece, Lithuania, Poland, Portugal and Romania where the difference is above 10 percentage points. For Greece, Portugal and Romania, the estimate of the lower share is neither statistically nor economically significant.

6 Robustness checks

We run various robustness checks using alternative sets of covariates, instruments and definitions of the self-employed. The results are most sensitive to the way the group of self-employed households and the reference group is defined, but robust to different covariates and instruments beyond the main ones. Below, we present the estimates of income underreporting using different definitions of the self-employed and the reference group, while the full details of other robustness checks can be found in the earlier version of the paper (Kukk et al. 2018).

Figure 2 shows the results when different definitions are used, comparing the baseline estimates of income underreporting using the simple share defined in Eq. (4). In the baseline specification, the self-employment group is selected using two criteria concurrently: (i) the household reports income from self-employment to be the main source of income and (ii) the household reports that at least one of the household members is self-employed. The reference sample includes only households where the household reports wages as the main source of income and none of the household members is self-employed. As the first alternative definition, we use only criterion (i) and *all* other households are in the reference group (circle), meaning that the reference group and the sample overall is slightly broader than the reference group used in the baseline estimations. The results when only the main income source is used to identify the self-employed are typically close to the baseline results, with Bulgaria, Croatia and Romania as partial exceptions.

Next, we define the self-employed households as those where the household head reports to be self-employed regardless of the main source of household income, while the rest of the sample is in the reference group (diamond). The estimates are in most cases lower than those of the baseline, which is an expected result since in this case the self-employment group may also include households with low self-employment income. The difference between the baseline and the self-reported status of the household head is particularly large for the Czech Republic, Hungary and Latvia.

Finally, we define households as self-employed if either of the household members is self-employed (cross), with the rest of the sample again considered as the reference group. This definition implies that the self-employment income in the household can be very low and we indeed find that the share of unreported income differs substantially in many cases from the baseline, with Latvia as the most extreme example.

The way in which the group of self-employed and the reference group of wage earners are defined was also examined in Kukk and Staehr (2017) for Estonia, and
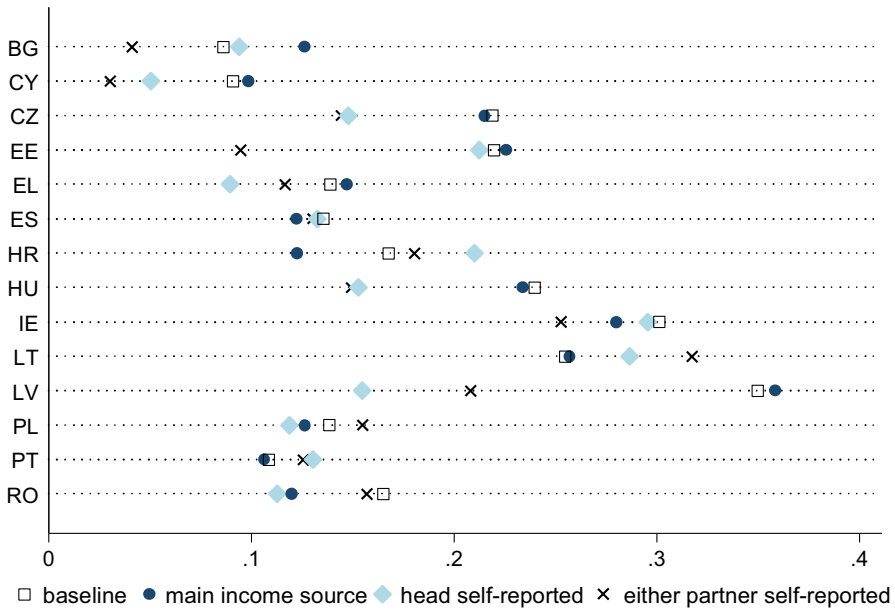
Fig. 2 Estimates of the simple share of unreported income with alternative definitions of self-employed households. *Notes* The share of unreported income is calculated from Eq. (4). *Source*: Authors' calculations using the EU-HBS

they obtained very different results depending on whether income or the reported self-employment status is used. Similarly, Paulus (2015) and Ekici and Besim (2016) show that the results are substantially affected by the way the underreporting and the reference groups are defined. Contrary to these results, Pissarides and Weber (1989) posit that their results are broadly similar regardless of how the self-employed are defined. Engström and Hagen (2017) reach very similar results for Sweden regardless of the criteria used for sample selection.

The upshot of our robustness checks is that for a cross-country comparison, the model specification in terms of covariates and instruments is not materially important, as long as it is sufficiently broad, whereas this is not the case for the way the reference group and the self-employed are defined. This suggests that the definitions need to be harmonised across countries in order to attain comparability of the results across countries.

7 Discussion of the results

The baseline estimations in Sect. 5 use a common specification for all the 14 EU countries for which sufficient data are available in the EU-HBS. The results show that the share of underreported income exhibits substantial heterogeneity across the countries. The analysis in Sect. 6 indicates that the results can be somewhat sensitive to the way the self-employment and reference groups are defined so these modelling choices must be taken into account when studies are compared.

Previous studies are only available for two of the 14 EU countries in the sample, i.e. Spain and Estonia. Our underreporting results for Spain are very close to those reported in Martinez-Lopez (2013), and indeed the estimated income elasticity and the coefficient of the dummy variable reported in Table 2 are also very similar to his estimates. This congruence is reassuring given that our study and Martinez-Lopez (2013) use household budget data from time periods close to each other and define the self-employed and the reference group in broadly similar ways.

Several studies provide underreporting estimates for Estonia using the expenditure method. The arguably most comparable result appears in Kukk and Staehr (2017), who use data from the household budget survey for 2002–2007 and estimate the share of unreported income to be around 28% of true income when the self-employed are identified using employment status.[11] Paulus (2015) uses the Estonian SILC for 2008 and runs estimations in which the self-employed are identified by their reported employment status. The share of underreported income is found to be 20–44% of true income. The upshot is that the results for Estonia in this study are within the intervals found in previous studies although in the lower end of these intervals.

The expenditure method was first used on the UK by Pissarides and Weber (1989) and variations of the method have subsequently been used on UK data by Lyssiotou et al. (2004) and Cabral et al. (2019). The studies find shares of underreporting from 8 to 39% of true income, but most of the estimates are around 20–30%. The UK is not included in our main sample of EU countries since data on education or other variables suitable for instrumentation of the income variable are not available in the EU-HBS. It is therefore impossible to estimate the lower and upper underreporting shares. It is, however, possible to estimate the Engle curve using OLS and compute the simple underreporting share. We obtain an estimate of 28.9% of true income with the standard error of 9.49 percentage points (available on request). To the extent the results can be compared, our result for the simple underreporting share is thus in line with the results in the three published studies for the UK.

There are substantial overlaps between our results and those for Spain, Estonia and—arguably also—the UK for which estimates are available in the academic literature. We take this as a confirmation that our approach of using the expenditure method on data from the EU-HBS for a large number of EU countries produces plausible results.

A striking feature of our results is that some of the lowest shares of income underreporting are found for Southern European countries such as Bulgaria, Greece, Portugal and Romania. We have not found any studies applying the expenditure method for these countries, making direct comparison impossible. Besim and Jenkins (2005) and Ekici and Besim (2016) consider the case of Northern Cyprus. In both papers, the reference group consists of households for which the household head is *publicly* employed, making comparison with other studies difficult. However, the differences between the underreporting by the privately employed and the self-employed, both relative to the reference group of publicly employed, are small. This suggests that the share of underreporting by the self-employed would be marginal if *all* employed

---

[11] Kukk and Staehr (2014, 2017) find substantially higher underreporting shares if other definitions of the self-employed and the reference group are used.

households were used as the reference group. The relatively low estimates we obtain for Bulgaria, Greece, Portugal and Romania would thus seem in line with the results for Northern Cyprus.

The estimates of income underreporting in this study are generally quite low. There may be various reasons. First, the published results may be subject to various selection biases, both across and within countries. Researchers may not complete empirical analyses for an individual country if they find very low or statistically insignificant shares of underreporting in early stages of their research. Likewise, a publication bias may have led to publication of the studies with relatively high underreporting estimates. Second, the estimates in this study also refer to a later time period than most of the studies of individual countries. We use a snapshot from a year after the outbreak of the global financial crisis, and the crisis may have had very different consequences for various households within a country and across countries (Pissarides 2013).

The estimates of the share of underreporting are often quite different for countries with broadly similar income levels. Figure 3 presents a cross-plot of GNP per capita and the simple share of unreported income from Table 2. The income levels of Bulgaria and Latvia are among the lowest in the sample, but the estimates of the share of unreported income vary markedly. Likewise, Cyprus and Ireland are among the richest in the sample, but the estimates vary greatly. In the sample of countries used in this study, there is no clear correlation between the economic development level of the countries and the underreporting by the self-employed.[12] This finding is in line with the results from other studies that compare non-compliance across countries using microdata (Eurobarometer 2014; Hazans 2011).

It is important to recall that the income underreporting estimated using the expenditure method is in addition to or on top of the possible underreporting by the reference group. If income underreporting among the wage earners is prevalent, overall income underreporting in a country may still be substantial, even if the difference between the underreporting by wage earners and the self-employed is limited. It is also worth noting that the underlying data are derived from household budget surveys where individuals are interviewed on their consumption, income, etc. There may be differences in the way individuals in various countries trust the statistics authorities and consequently in the truthfulness in which they provide answers to the surveys.

There are, however, some hints of broad country groupings in Fig. 3 on the basis of structural, institutional and cultural commonalities with the Baltic States located in the upper left part of the plot, Southern European countries (Portugal, Greece, Spain, Cyprus) in the lower right part and the post-communist Central and East European countries in the middle. We have examined further how far it is possible to relate our underreporting results to various measures of institutional development, governance

---

[12] The lack of a pattern is arguably in large part due to Ireland which has high income and a very large share of income underreporting. Without the results for Ireland, there is a negative correlation between income and underreporting, a result found in many studies using macroeconomic data but not in studies using microeconomic data, cf. discussion in Sect. 1. It is clear, however, that the evidence for such a negative correlation is at best indicative and would require further investigation once more data become available.
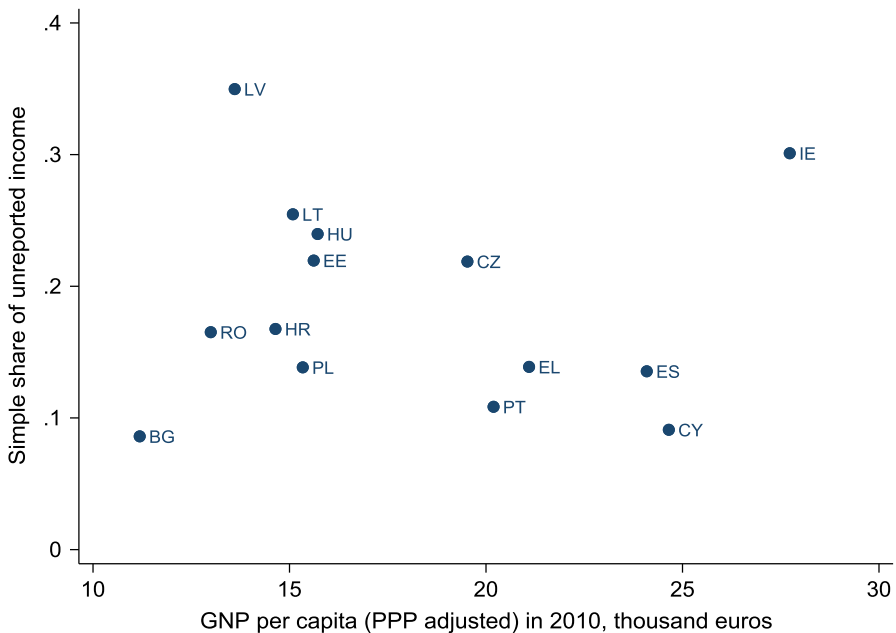
Fig. 3 Estimates of the simple share of unreported income and GNP per capita (PPP). *Source*: Authors' calculations using the EU-HBS and AMECO (code: *HVGNP*)

quality and tax burdens, but found no conclusive evidence, possibly due to the limited regional representation in our resulting sample of countries.[13]

8 Concluding remarks

Various studies have sought to provide estimates of income underreporting by the self-employed for individual countries using the expenditure method (Pissarides and Weber 1989). This paper contributes to the literature on income underreporting and non-compliance by being the first to provide comparative estimates with the expenditure method across 14 EU member countries, using the 2010 wave of the EU-HBS. The comparison is aided by the countries sharing many institutional features, and data in all EU countries being collected using a fairly standardised methodology. To obtain a high degree of comparability, we opted for the same specification of the estimation models for all 14 countries, as generic as data allowed.

Our estimates show substantial variation of income underreporting across countries with the highest country-level baseline estimates exceeding 40% of self-employed household income on average. The results are robust to changes in the model specification and the choice of instruments, whereas the way in which the self-employed

---

[13] Data on tax rates are from Eurostat (2013), and data on government effectiveness, regulatory quality, the rule of law and corruption control are from the Worldwide Governance Indicators published by the World Bank (http://info.worldbank.org/governance/wgi/index.aspx#home). The results were reported in the earlier version of this paper (Kukk et al. 2018).

households are defined are important for the results. We find no clear patterns across economic development levels, in line with the results in earlier studies for individual countries, though point out some suggestive evidence for country groupings based on their structural, institutional and cultural commonalities. In a broader context, the large variability of the underreporting estimates in our study is consistent with the findings in most other studies that use microeconomic data to estimate the prevalence of undeclared wages, tax evasion or other unreported activities.

Arguably the most surprising result is that the estimates for some Southern European countries are relatively low. This result may, however, be in accordance with a few earlier studies using the expenditure method to estimate the extent of under-reporting in individual countries (where these overlap with our country selection). Moreover, it is important to recall that the measure of underreporting estimated using the expenditure method is a relative measure and should be interpreted as being in addition to the possible underreporting by the reference group. There may also be differences in the way individuals in different countries provide answers to the statistics authorities. Future analyses could seek to include more high-income countries from western Europe, as additional data become available for instance in the next round of the EU-HBS. Comparative estimates over several periods will undoubtedly provide additional insights, including insights into the factors driving income underreporting and tax evasion by the self-employed.

Appendix A: Pissarides–Weber (1989) methodology

The starting point for Pissarides and Weber (1989) is the estimation of a food consumption function or Engel curve using household data. Households are assumed to smooth their food consumption over time so that it depends on the permanent disposable income of the household. The Engel curve for household i can be expressed as follows:

$$\log c_i = \alpha + \beta \log y_i^{\mathrm{P}} + X_i \phi + \varepsilon_i \tag{A.1}$$

The variable c_i is the spending on food, y_i^{P} is the true permanent income of the household, X_i is a vector of control variables, and ε_i is an error term. The control variables account for household characteristics that are likely to affect consumption,

such as household composition and the age of household members. The term α is a constant, β is the marginal propensity to consume, and ϕ is a vector of the coefficients of the control variables.

Some studies have an income variable available that may be a good proxy for permanent income, but in the EU-HBS only current income is available so we follow P&W and define the variable p_i as the fraction of true current income relative to true permanent income. The logarithm of true current income y_i can then be expressed as:

$$\log y_i = \log p_i + \log y_i^{\mathrm{P}} \qquad (A.2)$$

The variable p_i is assumed to follow a log-normal distribution over the households in the sample. If the mean of $\log p_i$ is labelled μ_p, $\log p_i$ can be written as $\log p_i = \mu_p + u_i$, where the disturbance u_i satisfies the conditions that $E[u_i] = 0$ and $\sigma_u^2 = \mathrm{Var}(u_i)$. The variance of the disturbance for the self-employed group ($\sigma_{u|S}^2$) is expectedly larger than the variance for the group of wage earners ($\sigma_{u|W}^2$), so $\sigma_{u|S}^2 > \sigma_{u|W}^2$, a feature we return to when we consider the possible range of the mean underreporting factor.

Given the way consumption data are reported to household budget surveys, it is conceivable that the households report their consumption correctly, while they could hide part of their true disposable income, especially if they fear that data may be shared with the tax authorities. The factor by which the reported current income y_i' has to be multiplied to obtain the true current income y_i is labelled k_i so the relationship between true and reported current income can be written as:

$$\log y_i = \log k_i + \log y_i' \qquad (A.3)$$

Household i underreports its true income if $k_i > 1$ and reports accurately if $k_i = 1$. It is assumed that the factor follows a log-normal distribution so that $\log k_i = \mu_k + v_i$ with $E[v_i] = 0$ and $\sigma_v^2 = \mathrm{Var}(v_i)$. All households in the reference group of wage earners are assumed to have $k_i = 0$ so by definition $\sigma_{v|W}^2 = 0$ for this group, while underreporting can occur in the group of self-employed, which means that $\sigma_{v|S}^2 > 0$.

It follows from Eqs. (A.2) and (A.3) that the true permanent income y_i^{P} can be expressed in logarithmic form as:

$$\log y_i^{\mathrm{P}} = \log y_i' + \log k_i - \log p_i = \log y_i' + (\mu_k + v_i) - (\mu_p + u_i) \qquad (A.4)$$

When the expression for the true permanent income in Eq. (A.4) is inserted into Eq. (A.1), the Engel curve becomes:

$$\log c_i = \alpha + \beta \log y_i' + \beta(\mu_k - \mu_p) + X_i \phi + \beta(v_i - u_i) + \varepsilon_i \qquad (A.5)$$

We are interested in the average factor of income underreporting for the self-employed, which we denote $\bar{k} = E[k_i|S]$. The mean μ_p cannot be identified, but since p_i follows a log-normal distribution, we know that $\mu_{p|S} - \mu_{p|W} = -\frac{1}{2}(\sigma_{u|S}^2 - \sigma_{u|W}^2) \leq 0$ if we follow P&W and assume that the means of p_i are the same for the two groups. The inequality comes from the assumption about the different variances of u_i for the two groups.

The term $\beta(\mu_k - \mu_p)$ in Eq. (A.5) is replaced by γD_i, where D_i is a dummy variable taking the value $D_i = 1$ for households that are self-employed and $D_i = 0$ for households in the reference group. The coefficient γ shifts the intercept for the self-employed relative to the wage earners. The substitution implies that $\gamma = \beta(\mu_k + \frac{1}{2}(\sigma_{u|S}^2 - \sigma_{u|W}^2))$, where we have used $\mu_{p|S} - \mu_{p|W} = -\frac{1}{2}(\sigma_{u|S}^2 - \sigma_{u|W}^2)$. Rearranging the terms, the mean underreporting factor by the self-employed, $\bar{k}$, can then be expressed as:

$$\bar{k} = E[k_i|S] = \exp\left(\frac{\gamma}{\beta} + \frac{1}{2}(\sigma_{v|S}^2 + \sigma_{u|W}^2 - \sigma_{u|S}^2)\right) \tag{A.6}$$

The mean underreporting factor for the self-employed is the exponential of the coefficient of the self-employment dummy over the marginal propensity to consume plus an adjustment term comprising the variance of the underreporting of the self-employed and the variances of the transitory income for both groups. The variance terms are not known, so it is generally not possible to compute a point estimate of the mean underreporting factor $\bar{k}$. However, P&W show that it is possible to establish a range within which $\bar{k}$ must lie under various assumptions.

It follows from Eq. (A.5) that the reported income is endogenous as the error term is $\beta(v_i - u_i) + \varepsilon_i$. This suggests that the reported current income should be instrumented in the food regression. The instrumentation may be carried out for the two groups separately, in which case there would be residuals ξ_i with variances related to the variance $\sigma_{u|S}^2$ and $\sigma_{u|W}^2$.

The residual variance from the reduced-form regressions for income, ξ, is the sum of three error components: (i) (unexplained) variations in permanent income; (ii) deviations of current income from permanent income u_i; and (iii) deviations of reported income from current income v_i. Under the assumption that the variance of permanent income is the same for the self-employed and for wage earners, the difference between the variance of the residuals for the self-employed $\sigma_{\xi|S}^2$ and the variance of the residuals for the wage earners $\sigma_{\xi|W}^2$ can be written as:

$$\sigma_{\xi|S}^2 - \sigma_{\xi|W}^2 = \sigma_{u|S}^2 + \sigma_{v|S}^2 - 2\text{Cov}(u, v \mid S) - \sigma_{u|W}^2 \tag{A.7}$$

We follow P&W and derive lower and upper bounds for $\bar{k}$ under the assumption that $\text{Cov}(u, v \mid S) = 0$, which implies that for the self-employed there is no relationship between the deviations of current income from permanent income u_i and the deviations of reported income from current income v_i. The lower bound for underreporting factor can then be found as $\bar{k}_L = \exp(\gamma/\beta - \frac{1}{2}(\sigma_{\xi|S}^2 - \sigma_{\xi|W}^2))$, which follows when we consider that the lowest possible value of the variance $\sigma_{v|S}^2$ is 0 and insert this into Eq. (A.7), and then insert the resulting relationship into Eq. (A.6). The upper bound is $\bar{k}_U = \exp(\gamma/\beta + \frac{1}{2}(\sigma_{\xi|S}^2 - \sigma_{\xi|W}^2))$ which emerges from Eqs. (A.7) and (A.6) when we use that $\sigma_{u|W}^2$ can at most take the value $\sigma_{u|S}^2$.

Appendix B: Additional figures and tables
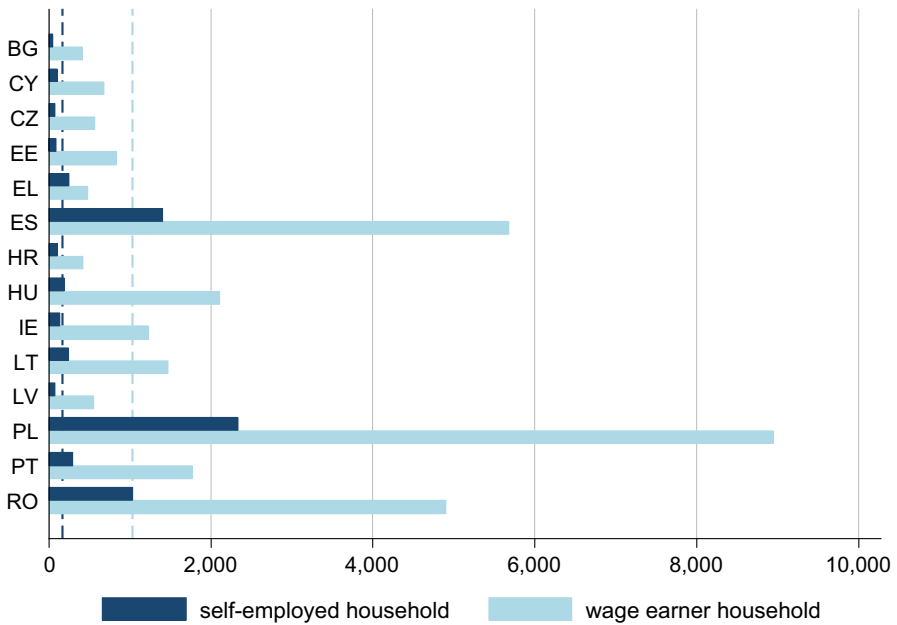
See Figs. 4, 5 and Table 3.

Fig. 4 Number of self-employed and wage earner households in the estimation sample. *Notes* Vertical dashed lines mark median values. The sample consists of households with couples that report earnings as their main source of income and the household head as being employed and 24–59 years old. Households are defined as self-employed if income from self-employment is reported to be the main source of income and at least one of the spouses is reported to be self-employed. *Source*: Authors' calculations using the EU-HBS
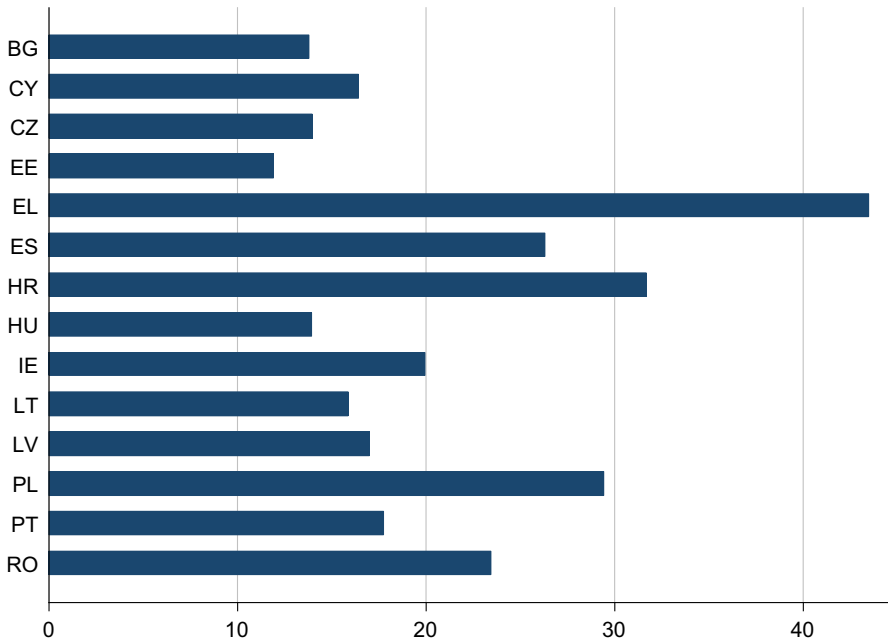
Fig. 5 Percentage of self-employed households in the estimation sample. *Notes*: The sample consists of households with couples that report earnings as their main source of income and the household head as being employed and 24–59 years old. Households are defined as self-employed if income from self-employment is reported to be the main source of income and at least one of the spouses is reported to be self-employed. *Source*: Authors' calculations using the EU-HBS

Table 3 Mean log values of consumption and income by household type and country. *Source*: Authors' calculations using the EU-HBS

	Mean log food consumption			Mean log reported income		
	Employed	Self-employed	Difference	Employed	Self-employed	Difference
BG	7.496 (0.022)	7.585 (0.055)	− 0.089 (0.060)	8.490 (0.024)	8.575 (0.078)	− 0.084 (0.082)
CY	9.009 (0.021)	8.976 (0.051)	0.033 (0.055)	10.577 (0.017)	10.336 (0.044)	0.241 (0.047)
CZ	8.033 (0.014)	8.108 (0.031)	− 0.075 (0.034)	9.765 (0.012)	9.600 (0.044)	0.164 (0.046)
EE	7.648 (0.029	7.886 (0.079)	− 0.238 (0.084)	9.309 (0.020)	9.348 (0.072)	− 0.039 (0.075)
EL	8.907 (0.022)	9.069 (0.029)	− 0.161 (0.036)	10.086 (0.026)	10.241 (0.050)	− 0.155 (0.057)
ES	8.813 (0.010)	8.907 (0.019)	− 0.094 (0.021)	10.140 (0.008)	10.113 (0.018)	0.027 (0.020)
HR	8.270 (0.024)	8.280 (0.052)	− 0.010 (0.058)	9.548 (0.024)	9.388 (0.058)	0.159 (0.063)
HU	7.598 (0.012)	7.744 (0.037)	− 0.147 (0.039)	9.169 (0.011)	9.143 (0.038)	0.026 (0.040)
IE	8.952 (0.014)	9.050 (0.058)	− 0.099 (0.060)	11.003 (0.012)	10.783 (0.066)	0.220 (0.067)
LT	8.109 (0.016)	8.263 (0.033)	− 0.154 (0.036)	9.196 (0.017)	9.104 (0.072)	0.092 (0.074)
LV	7.847 (0.023)	8.118 (0.098)	− 0.271 (0.101)	9.012 (0.028)	8.933 (0.130)	0.079 (0.133)
PL	7.767 (0.005)	7.911 (0.009)	− 0.144 (0.010)	9.231 (0.006)	9.346 (0.015)	− 0.115 (0.016)
PT	8.437 (0.025)	8.396 (0.054)	0.041 (0.060)	9.958 (0.019)	9.783 (0.053)	0.175 (0.056)
RO	7.574 (0.006)	7.428 (0.015)	0.146 (0.016)	8.585 (0.008)	7.903 (0.028)	0.681 (0.029)

Food consumption includes catering outside the home. The mean values are estimated with survey weights and refer to the samples used in the estimations in Sect. 5. The standard errors of the mean estimates are shown in brackets

References

Alm, J., & Embaye, A. (2013). Using dynamic panel methods to estimate shadow economies around the world, 1984–2006. *Public Finance Review, 41*(5), 510–543.

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

Buehn, A., & Schneider, F. (2012). Shadow economies around the world: Novel insights, accepted knowledge, and new estimates. *International Tax and Public Finance, 19*(1), 139–171.

Cabral, A. C. G., Kotsogiannis, C., & Myles, G. (2019). Self-employment income gap in Great Britain: How much and who?". *CESifo Economic Studies, 65*(1), 84–107.

Ekici, T., & Besim, M. (2016). A measure of the shadow economy in a small economy: Evidence from household-level expenditure patterns. *Review of Income and Wealth, 62*(1), 145–160.

Engström, P., & Hagen, J. (2017). Income underreporting among the self-employed: A permanent income approach. *European Economic Review, 92*(C), 92–109.

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

Eurobarometer. (2014). *Undeclared work in the European Union*. Special Eurobarometer, 402, European Commission.

European Commission. (2016). *Employment and social developments in Europe 2015*. Luxembourg: Publications Office of the European Union.

Eurostat. (2013). *Taxation trends in the European Union*. Luxembourg: Publications Office of the European Union.

Eurostat. (2015). *EU Household Budget Survey 2010 quality report*. European Commission, DOC HBS/2015/01/EN.

Hatfield, I. (2015). *Self-employment in Europe*. London: Institute for Public Policy Research.

Hazans, M. (2011). *Informal workers across Europe: Evidence from 30 European countries*. Policy Research Working Paper WPS 5912, World Bank.

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

Johansson, E. (2005). An estimate of self employment income underreporting in Finland. *Nordic Journal of Political Economy, 31*(1), 99–109.

Kim, B., Gibson, J., & Chung, C. (2017). Using panel data to estimate income under-reporting of the self-employed. *Manchester School, 85*(1), 41–64.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica, 79*(3), 651–692.

Kukk, M., Paulus, A., & Staehr, K. (2018). *Income underreporting by the self-employed in Europe: A cross-country comparative study*. Eesti Pank Working Paper 4/2018.

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

Kukk, M., & Staehr, K. (2017). Identification of households prone to income underreporting: Employment status or reported business income? *Public Finance Review, 45*(5), 599–627.

Lago-Peñas, I., & Lago-Peñas, S. (2010). The determinants of tax morale in comparative perspective: Evidence from European countries. *European Journal of Political Economy, 26*(4), 441–453.

Lamarche, P. (2017). *Measuring income, consumption and wealth jointly at the micro-level*. European Commission, Eurostat Methodological Note, Luxembourg. http://ec.europa.eu/eurostat/documents/7894008/8074103/income_methodological_note.pdf.

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

Martinez-Lopez, D. (2013). The underreporting of income by self-employed workers in Spain. *SERIEs, 4*(4), 353–371.

Nygård, O. E., Slemrod, J., & Thoresen, T. O. (2019). Distribution implications of joint tax evasion. *Economic Journal*, *129*(620), 1894–1923.

Paulus, A. (2015). *Income underreporting based on income-expenditure gaps: survey vs tax records*. ISER Working Paper Series, No. 2015-15, Institute for Social and Economic Research, University of Essex.

Pissarides, C. A. (2013). Unemployment in the Great Recession. *Economica, 80*(319), 385–403.

Pissarides, C. A., & Weber, G. (1989). An expenditure based estimate of Britain's black economy. *Journal of Public Economics, 39*(1), 17–32.

Poniatowski, G., Bonch-Osmolovskiy, M., & Belkindas, M. (2016). *Study and reports on the VAT gap in the EU-28 Member States: 2016 final report*. Case Research Paper, No. 483.

Schmutz, F. (2018). Income underreporting by the self-employed in Switzerland: An international comparison. *FinanzArchiv/Public Finance Analysis, 74*(4), 481–534.

Schneider, F. (2007). Shadow economies and corruption all over the world: New estimates for 145 countries. *Economics: The Open-Access Open-Assessment E-Journal, 1*(5), 1–29.

Schuetze, H. (2002). Profiles of tax noncompliance among the self employed in Canada: 1969–1992. *Canadian Public Policy, 28*(2), 219–237.

Slemrod, J. (2007). Cheating ourselves: The economics of tax evasion. *Journal of Economic Perspectives, 21*(1), 25–48.

Tafenau, E., Herwartz, H., & Schneider, F. (2010). Regional estimates of the shadow economy in Europe. *International Economic Journal, 24*(4), 629–636.