Central European Journal of Economic Modelling and Econometrics

Estimate of the Underground Economy in Poland Based on Household Expenditures and Incomes

Mehmet Burak Turgut*, Tomasz Tratkiewicz†

Submitted: 11.03.2022, Accepted: 18.11.2022

Abstract

We estimate the size of income underreporting in Poland by following and extending the consumption method of Pissarides and Weber (1989). Our study shows that underreporting of income occurs among households with income from self-employment. We do not find any significant underreporting activity by the employees working in the private sector. The main findings indicate that roughly one-fourth of the total income of self-employed households is not reported in Poland. This share varies between 20 to 30 percent from 2005 to 2017 with a decreasing trend.

Keywords: underground economy, underreporting of income, income elasticity, food consumption, self-employed households

JEL Classification: E21, E26, H26

*Department of Macroeconomics and International Trade Theory, Faculty of Economic Sciences, University of Warsaw, Poland; e-mail: m.turgut3@uw.edu.pl; ORCID: 0000-0002-6231-2187

†Faculty of Economics and Sociology, University of Łódź, Poland; CASE - Center for Social and Economic Research; e-mail: tomasz.tratkiewicz@uni.lodz.pl; ORCID: 0000-0001-7772-1779

1 Introduction

The problem of classifying and measuring issues in tax non-compliance and income underreporting, or broadly understood underground economy, has long been the subject of research (Feige, 1990). While the concept of the underground economy itself has been quite accurately defined (Schneider, 2005), the best possible methods for measuring it are still being searched for. Measuring the underground economy is particularly difficult because the individuals who operate in it simply do not disclose it. Analysis of tax evasion, or the underground economy in general must proceed even in the absence of the direct observability of key variables, and theory should guide the construction and interpretation of evidence of the "invisible" (Slemrod and Weber, 2012).

In this paper, we estimate the size of income underreporting in Poland following the consumption method developed by Pissarides and Weber (1989) (hereinafter P&W) by using data from the Polish Household Budget Survey (hereinafter HBS). P&W followed Smith's (1986) research that underground economy activity is concentrated amongst the self-employed. This observation means that the underground economy does not include informal or even criminal activities other than self-employment, from which income is unobservable. Our estimates, therefore, do not include envelope wages obtained by private employees and informal income by public and private employees who might work informally without any contract.

P&W measured the size of the underground economy in the UK based on consumption propensities between two groups: "non-risky" (correctly reporting their income) and "risky" (underreporting their income). In particular, they used data from the Family Expenditure Survey (the British equivalent of HBS), estimated Engel curves of food expenditures for households of employees (non-risky) and self-employed (risky), and measured the extent of income under-reporting among the households of self-employed. Their key assumptions were that both types of households report their food expenditures correctly in the survey, the marginal propensity to consume with respect to income does not change between two types of households after controlling for household characteristics, and the survey income corresponds to the declared taxable income.

We improve on the original P&W method in several ways. First, we use public employees as our reference (non-risky) group in the spirit of Besim and Jenkins (2005) and Paulus (2015) and examine the possibility of underreporting of income by private sector employees in addition to the self-employed. Second, we use regular income in our analysis as a proxy of permanent income akin to Kukk and Staehr (2014, 2017) and obtain point estimates of under-reporting factors. Third, we identify a household as self-employed if the main source of income is self-employment, rent, and unemployment benefits. Lyssiotou et al. (2004) and Kukk et al. (2020) applied a similar identification by restricting a household to be self-employed only if the main source of income is from self-employment.

We estimate that roughly one-fourth of the total income of self-employed households

is not reported in Poland. This share varies between 20 and 30 percent over the years 2005 to 2017 with a decreasing trend. Our results are in the range of the estimates from previous studies that are discussed in detail in the next section. However, we find larger extent of underreporting in Poland compared to the results of Kukk et al. (2020) where authors calculated 13.8 percent of unreported income using 2010 European Union HBS. Our definition of risky group that includes a greater number of households prone to income underreporting can drive the difference in the results.

We do not find any evidence of income underreporting for the households working in the private sector relative to the public employees. This recalls the results of Paulus (2015) who did not find significant unreported income share for the private employees in Estonia based on survey data after using public employees as a reference group.

We contribute to the literature in the following ways. First, the P&W methodology asks for permanent income data that is generally not given in the survey data. This requires additional assumptions to have a proxy for permanent income resulting in lower and upper bound for the estimates of underreporting. The Polish HBS allows to disentangle household's income into regular and transitory income based on self-reported information. We use regular income as a proxy of permanent income to obtain point estimate of the income underreporting in Poland. This approach has been applied to Estonian data by Kukk and Staehr (2014, 2017). Moreover, we use temporary income as a control variable to account for the excess sensitivity of consumption to avoid omitted variable bias and obtain reliable estimates of income elasticity of consumption.

Second, we classify the households using the main source of income of the household's head. Previous literature typically classifies the "risky" and the "non-risky" households using three different methods. First method calculates the reported share of business income in reported total income and identifies the household as risky if this share exceeds a stated threshold. P&W, Schuetze (2002), and Kukk and Staehr (2014) used this method in their work. The second method uses the employment status of the household and identifies the household as risky if the employment status is reported as self-employment. Johansson (2000), Engström and Holmlund (2009), Hurst et al. (2014), Paulus (2015), and Kim et al. (2017) applied second method in their research. Kukk and Staehr (2017), Schmutz (2018), Nygård et al. (2019), and Cabral et al. (2019) utilized both methods in their papers. The third method uses main source of income of the household's head and identifies the household as risky if income from self-employment is the main source of income. This method is used in Lyssiotou et al. (2004) and Kukk et al. (2020). We followed the third method and defined a household as risky if main source of income of the household's head is self-employment, rent, or unemployment benefits. We believe our identification better captures the risky group since we include income categories such as rent income that give rise to the risk of not declaring income (see Albarea et al., 2020). We support this argument with empirical exercise and find that the share of unreported income

shrinks when unemployment benefits are excluded in the identification of the risky households.

Third, our data covers the post-EU accession period of Poland which constitutes numerous reforms and legislations. This allows us to analyse whether the EU membership had a beneficial impact on the extent of income underreporting in Poland. Although PIT is not harmonised in the EU, it seems that exposure to competition with entrepreneurs from other EU member states provided an incentive for policymakers in Poland to reform the PIT system. Among the changes that seem to have been able to foster a decrease in underreporting of income in the period under review, one can mention, firstly, the reduction in the amount of income taxation according to the tax scale in 2009 (elimination of the third threshold, increase of the limit for the first threshold and reduction of the tax rates). Between 2001 and 2008, average PIT taxation after deducting health insurance premiums oscillated around 9% of income, while in 2009 it dropped to 7.55% of income (Owsiak, 2016). Secondly, the possibility to use simplified forms of PIT taxation (19% flat rate and registered lump sum) was also "popularised" in the period under study due to legislative changes, which resulted in a significant increase in the number of persons using these forms of taxation.

Fourth, the literature maintained assumption of linear relationship between income and food expenditures. Nevertheless, the increase in food consumption can slow down with increasing income (Banks et al., 1997). In extension of our model, we allow for this possibility by including quadratic income term in the regression equation. Moreover, we show how to obtain share of unreported income using estimates from a non-linear consumption model. Our results show that non-linear effects do not have significant impact on the size of the underreporting shares estimated via linear model.

The paper is organized as follows. The next section reviews the literature. Section 3 explains the methodology. Section 4 provides definitions and summarizes data along with sources. Section 5 describes the empirical methodology and discusses the results. Section 6 conducts sensitivity analyses and presents the results of the extended empirical model. Section 7 concludes and discusses possible avenues for the future research.

2 Literature review

P&W used income and consumption data drawn from the 1982 Family Expenditure Survey. They calculated that the share of true income not reported by self-employed individuals as 35 percent which corresponds to 5.5 percent of the UK's GDP. They also found higher underreporting shares for the blue-collar self-employed relative to the white-collar self-employed.

The following works applied P&W's methodology for different countries. Johansson (2000) estimated that 10 to 47 percent of income is left unreported in Finland depending on the number of self-employed in the household. Schuetze (2002) obtained unreported income share between 11 to 23 percent for the self-employed in Canada

using six years of data from 1969 to 1992. Engström and Holmlund (2009) found that 15 to 50 percent of income is left unreported by the self-employed in Sweden. Hurst et al. (2014) concluded self-employed in the US did not report 30 percent of their income to the tax authorities.
Kukk and Staehr (2014) estimated that 62 percent of true income is not reported in Estonia by the self-employed whose share of business income exceeds 20 percent of total reported income. Paulus (2015) obtained a larger underreporting share for the self-employed when registered income from the Estonian tax records is used instead of reported income from the Estonian Social Survey.
In more recent papers, Kukk and Staehr (2017) found the share of unreported income in Estonia is higher when self-employed is identified using the share of business income rather than the status of employment. Kim et al. (2017) used the status of employment to identify the risky group in Russia and Korea and estimated the share of unreported income as 28 percent for the former and 29 percent for the latter. Schmutz (2018) obtained an unreported income share for Switzerland between 13 and 25 percent and Nygård et al. (2019) found the same share as 13 percent for Norway using the status of self-employment in the identification of self-employed. In a very recent attempt, Kukk et al. (2020) studied the extent of income underreporting for the 14 European countries using the 2010 EU HBS, and their estimates for the share of unreported income ranged from 10 to 40 percent. There were attempts in the literature that used models as an alternative to the P&W method for estimating income underreporting. Lyssiotou et al. (2004) applied non-parametric estimation to a consumer demand system approach using 1993 UK FES and found a slightly larger share of unreported income relative to the P&W. Lichard et al. (2021) developed an endogenous switching model with unknown sample separation and estimated both the probability of hiding income and the expected amount of unreported income for each household in Czechia and Slovakia.

3 Methodology

The starting point of the P&W approach is the following expenditure function (Engel curve):

$$\ln C_i = \alpha + \beta \ln Y_i^P + X_i\delta + \eta_i, \tag{1}$$

where C_i is the food consumption, Y_i^P is the permanent income of the household i, and X_i is a vector of household characteristics aims to control for household heterogeneity that can affect food consumption. The term β measures the elasticity of consumption with respect to permanent income and η_i is the residual with zero mean and constant variance.
The Polish HBS separates the reported current income of the household into regular and transitory components. We assume that reported regular income by the household excludes short-term income fluctuations and captures the variations in the permanent

income of the same household. However, it can differ from the permanent income due to misreporting in the survey. These assumptions allow us to express permanent income as following:

$$Y_i^P = k_i Y_i^R, \tag{2}$$

where Y_i^R denotes the reported regular income in the survey and k_i captures the income underreporting factor. The reported regular income should be multiplied with the underreporting factor to derive the permanent income of the household. We assume that each household belongs one of the three household type j that are public employees (GE), self-employed (SE), and private employees (PE). We follow P&W and specify the underreporting factor k_i as log-normally distributed:

$$\ln k_i = \mu_j + v_i, \tag{3}$$

where μ_j is the mean log-value within household type j and v_i is normally distributed error term with zero mean and constant variance $\sigma^2_{v,j}$. Combining Equations (2)–(3) and plugging into Equation (1) gives:

$$\ln C_i = \alpha + \beta \ln Y_i^R + X_i\delta + \beta\mu_j + \beta v_i + \eta_i. \tag{4}$$

We assume that among household types GE is not subject to under-reporting (non-risky group), and SE and PE are potentially subject to under-reporting (risky group). This classification implies:

$$\begin{aligned} &k_i = 1, && \mu_j = 0, && \sigma^2_{v,j} = 0 && \text{if } i \in j = GE \\ &k_i > 1, && \mu_j > 0, && \sigma^2_{v,j} > 0 && \text{if } i \in j = SE, PE. \end{aligned} \tag{5}$$

Following P&W, we assume that the parameters β and δ are same across the three types of households. Using Equation (5) and an indicator variable $D_{i,j}$, which takes the value of 1 for individuals in group SE or PE and 0 otherwise, and maintaining previous assumptions, Equation (4) can be rewritten as following:

$$\ln C_i = \alpha + \beta \ln Y_i^R + X_i\delta + \beta \sum_{j=SE,PE} D_{i,j}\mu_j + \beta v_i + \eta_i, \tag{6}$$

or

$$\ln C_i = \alpha + \beta \ln Y_i^R + X_i\delta + \sum_{j=SE,PE} D_{i,j}\gamma_j + \varepsilon_i. \tag{7}$$

A comparison of Equations (4) and (7) reveals the following relationships:

$$\gamma_j = \beta\mu_j \Rightarrow \mu_j = \frac{\gamma_j}{\beta}; \quad j = SE, PE, \tag{8}$$

$$\varepsilon_i = \beta v_i + \eta_i. \tag{9}$$

Using the log-normal assumption, the mean of the under-reporting factor across the households in a specific group can then be derived as following:

$$k_j = \exp\left(\mu_j + \frac{1}{2}\sigma^2_{v,j}\right) = \exp\left(\frac{\gamma_j}{\beta} + \frac{1}{2}\sigma^2_{v,j}\right); \quad j = SE, PE. \tag{10}$$

By using Equations (2) and (3), one can show the link between reported and permanent income becomes:

$$\ln Y_i^R = \ln Y_i^P - \mu - v_i. \tag{11}$$

The variance of the reported income, Y_i^R, is given by σ_j^2 and varies according to the household types. It is, however, assumed that the variance of the permanent income, Y_i^P, is the same across all household types and given by $\sigma^2_{Y^P}$. This assumption can be incorporated into Equation (11) to get:

$$\sigma_j^2 = \sigma^2_{Y^P} + \sigma^2_{v,j}; \quad j = GE, SE, PE.$$

Since $\sigma^2_{v,GE} = 0$ for public employees, it holds that $\sigma^2_{GE} = \sigma^2_{Y^P}$. When we subtract the variance of the reported income of the public employee (GE) from the variance of the self-employed (SE) and private employees (PE), we obtain:

$$\sigma_j^2 - \sigma^2_{GE} = \sigma^2_{v,j}; \quad j = SE, PE, \tag{12}$$

which gives the second term in Equation (10).
The permanent income of the average household in the risky groups can be found by multiplying the reported regular income by the mean underreporting factor k_j. By using Equation (12) in Equation (10), the mean underreporting factor for the risky groups can be found as:

$$k_j = \exp\left(\frac{\gamma_j}{\beta} + \frac{1}{2}\left(\sigma_j^2 - \sigma^2_{GE}\right)\right); \quad j = SE, PE. \tag{13}$$

There are a couple of issues worth mentioning related to the main assumptions of the methodology. First, we assumed the relationship between regular and permanent income is independent of household characteristics. However, this relationship can potentially be affected, for instance, by age-related patterns, i.e., smaller the young-aged relative to the old-aged group. We remove pensioners from our analysis to reduce the risk of this possibility affecting the result unduly. Moreover, Hurst et al. (2014) modelled p_i to depend on household characteristics, including age. Their analysis showed that the vector of household characteristics may capture the relationships between regular and permanent income other than underreporting, and so allowing μ to differ across household groups is enough for establishing the link between regular and permanent income.

Second, by restricting α and β to be the same across different household types we exclude preference heterogeneity from the analysis. This might cause downward bias of the estimates of the under-reporting factors (see Lyssiotou et al. 2004). Since our data only allows us to estimate a single demand equation, we restrict our sample to a minimum of two-adult households to obtain more homogenous groups of households.

4 Data

We use data from the Polish HBS between 2005 to 2017 for empirical analysis. The Polish HBS is conducted monthly among a representative cross-section of households in Poland. The household is defined as people who live in the same residence and share their expenses. The member of the household with the highest income is the household head. The number of households surveyed was around 37,000 per year over the study period. The subject of the survey is primarily the household budget, i.e., the amount of income and outgoings (monetary and non-monetary) of all members of the surveyed household and the quantitative consumption of selected items and services (Myck and Najsztub, 2015). Each month, a different household participates in the survey. The household records their outgoings and incomes during this period in special notebooks, called budget books. Respondents also have the option of collecting receipts, which they do not have to rewrite in the paper books. Since 2016, the option has been made available to respondents to record income and outgoings in an electronic application – on the respondent's equipment. An additional interview is conducted with households surveyed in each month of the calendar quarter at the end of the quarter (CSO, 2018). There were no significant methodological changes to the survey during the period under study that could be relevant to the objectives of our study. It can only be noted that in 2018, the sampling was slightly altered in order to obtain better precision indices for the estimators, the variance of which is highly dependent on the information obtained from wealthier households.

The current income (total income) of households in the Polish HBS is the sum of the following components: the income from employment, the income from self-employment outside the individual farm in agriculture, the income from the individual farm in agriculture, the income from the rental of property, the pension contributions, remaining cash from the previous month, gifts, and other income. Besides, the households report the regular monthly incomes from employment, self-employment, rent, and pensions that do not include temporary or extraordinary income components.

We exclude regular income from the pension contributions and use regular income from employment, self-employment, and rent as a proxy of permanent income in the empirical analysis. We treat all temporary income reported in the Polish HBS as separate item from the regular income. The income from temporary activities includes temporary employment and self-employment income earned in the country and abroad, as well as cash, gifts, and other income. The exclusion of the temporary

income aims to improve the correctness of estimates since it could affect the amount of expenditure on food only in the period in question. However, our empirical analysis is based on the Permanent Income Hypothesis (PIH) which manifests that consumption depends on the permanent income (regular income in our case) and not on current income. Therefore, we separate the temporary income from the total income and use it as a control variable to account for excess sensitivity of consumption. The PIH also rationalizes the exclusion of pension contributions from the definition of regular income and pensioners from the empirical analysis since their consumption-saving behaviour is likely to be different from the rest of the households in the sample due to lower permanent and current income ratio.

The regular income reported by the households in the Polish HBS includes income from regular employment and regular self-employment at home and abroad. We include income from abroad in because we assume that if the survey covers people with a regular income from abroad but with a life interest centred in Poland, these people would have to pay PIT in Poland even though they stay 183 days or more outside the country.

P&W identifies the household as self-employed (subject to under-reporting) if the head of the household in the Family Expenditure Survey declares at least 25 percent of his/her total income comes from self-employment. In our analysis, rather than relying on this identification, we exploit the information from the Polish HBS to identify the households subject to under-reporting. In the survey head of the household reports the main source of income (D2G column in HBS) under one of the following categories: white-collar job, blue-collar job, self-employment in agriculture, self-employment other than the agriculture, self-employment (freelancing), rental incomes, pensions, benefits, and other. Also, the households declare whether their income is from private or public sources. We combine these two types of information and distinguish three types of households using the following definition:

Definition 1. *The household is defined as*

1. **self-employed (SE)** if the head of the household's main source of income is from self-employment, ownership, rent, benefits, and other;
2. **private employee (PE)** if the head of the household's main source of income is from a white-collar job or a blue-collar job and the ratio of private income to public income is more than a half;
3. **public employee (GE):** if the head of the household's main source of income is from a white-collar job or a blue-collar job and the ratio of public income to private income is more than half.

In the above definition, self-employed and private employees constitute the risky groups (potentially subject to under-reporting), and public employees constitute the non-risky group (not subject to under-reporting). Hence, the public employees are our reference group, and their under-reporting factor is equal to one by definition. This

way of selecting a reference group from public employees is in line with the previous literature (Paulus, 2015; Ekici and Besim, 2016).

It is important to mention that Definition 1 excludes households with the main source of income from agricultural activities because they are likely to have different expenditure patterns on food than those in other occupations. Since one of the main assumptions of the underlying methodology is the same income elasticity of consumption between risky and non-risky households, we opt to exclude the households with the main source of income from agriculture. Also, we restrict our sample to households with at least two-adults similar to the procedure in previous studies (see P&W and Paulus, 2015) to maintain homogenous consumption patterns across household groups.

Another important feature of this paper is that it uses both at-home and away-from-home expenditures on food. This approach is justified by a greater tendency of non-retired households with relatively high income to replace home meals with away-from-home meals, i.e., eating lunch in a company canteen or eating out in a restaurant. Hence, excluding the food expenditure consumed away from home may affect the results unduly.

The literature to date shows that most studies only included food bought to be prepared at home in expenditure. We identified four studies stating explicitly that food expenditures away-from-home were also included in the total food expenditure (Schuetze, 2002; Kukk and Staehr, 2014; Cabral et al., 2019; Lichard et al., 2021). The main concern of including the latter into food expenditures is that since food consumed out of the house tends to be more expensive, differences in consumption patterns between the household groups (risky and non-risky) could explain the estimation results. We tested whether this concern is observed in data and there was a difference in the preference for eating at home and away in the different household groups during the study period. Figure 1 shows the evolution of the share of food expenditures away-from-home in the total food expenditures for each household group between the years 2005 and 2017. An evident trend from Figure 1 is that the share of expenditure on food away-from-home increased in all groups, which seems reasonable when society is "getting richer". This finding implies that the share of food expenditure at home in total food expenditure does not show significant variations across household groups in Poland over the period studied. Therefore, we came to a similar conclusion as researchers working for the Tax Administration Research Centre in the UK, that the higher level of food expenditure observed for the self-employed cannot be justified by higher food expenditure away from home (Cabral et al., 2019).

Table 1 presents the summary statistics of the dataset obtained from Polish HBS according to the regular income used in this study and household types based on Definition 1. The nominal food expenditures and income variables are deflated by HICP (Harmonised Index of Consumer Prices) to express the expenditures and incomes at constant 2005 prices.

Figure 1: The share of food expenditures away from home in total food expenditure

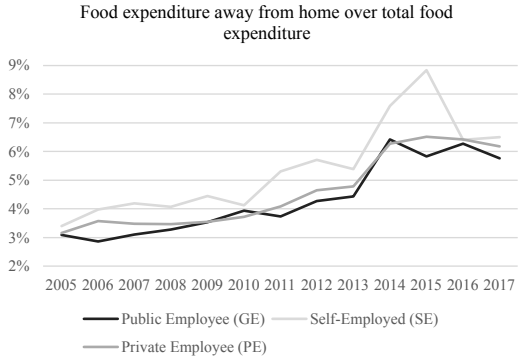

Table 1: Summary statistics

	Full Sample	SE	PE	GE
Sample size	220,232	29,579	129,557	61,096
Real Food Consumption (in PLN)	916.72	1,014.29	892.59	920.64
Real Regular Income (in PLN)	3,059.29	3,640.86	2,874.40	3,169.80
Real Total income (in PLN)	4,708.36	5,471.62	4,428.09	4,932.40
Age	43.23	44.08	42.09	45.25
Female Ratio	28.27%	23.95%	24.40%	38.56%
Number of child	1.36	1.41	1.38	1.29
Number of cars	0.84	0.98	0.79	0.88
Size of house (m^2)	84.06	102.07	79.74	84.49

Notes: SE, PE, and GE stand for self-employed, employee in private sector, and employee in public sector, respectively.

Figure 2: Distribution of income sources by household groups

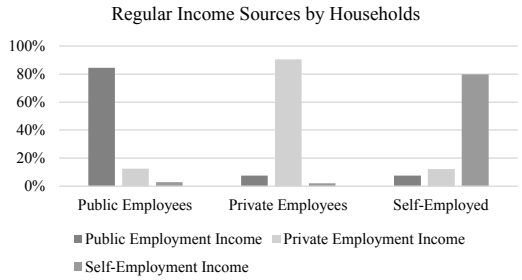

Figure 2 shows the distribution of the households' regular income sources by household types and suggests that the identification based on Definition 1 well matches household types with their regular income types. Even though our identification is based on the main income source of the head of the household, this does not result in the inclusion of different regular income sources from the other members of the household. For example, self-employment and rent income comprise approximately 80 percent of the regular income of the self-employed households, and private employment income constitutes more than 90 percent of the regular income of the private employees. These findings reinforce the sufficiency of Definition 1 in distinguishing the household types.

5 Estimation and results

We estimate the food consumption model by running the following regression:

$$\ln C_i = \alpha + \beta \ln Y_i^R + \gamma_{SE} D_{i,SE} + \gamma_{PE} D_{i,PE} + X_i \delta + \varepsilon_i, \tag{14}$$

where C_i is the real food expenditure at 2005 prices, Y_i^R is the real reported regular income at 2005 prices, $D_{i,SE}$ and $D_{i,PE}$ are the dummy variables taking value 1 if the household belongs to self-employed or private employee types (one of the risky groups), respectively, and 0 otherwise (see Definition 1). The vector X_i consists of log temporary income, quarterly and yearly dummies and the control variables to account for household characteristics. These include demeaned age, demeaned age square, size of the household, size of the house (m^2), and dummies indicating the region, the size of the city in which the household resides, education level, civil status, parental status (whether the household lives with children or not), occupation sector of the household head and finally, the homeownership status to proxy for the wealth of the household. The summary statistics of the variables of which vector X_i consists of are given in Appendix.

We estimate Equation (14) on monthly Polish HBS data from 2005 to 2017. Since most of the households change every year, it is not possible to use a panel model in our analysis. Therefore, we employed a pooled cross-sectional model by combining observations from all years into one sample. This structure allows us to capture the seasonality by including quarterly dummies and to capture the aggregate year effects by including yearly dummies. We cluster the standard errors at the household level to account for multiple observations when pooling across different years. The reported income is likely to be correlated with error term because we cannot control for cross-sectional unobserved heterogeneity. Moreover, the reported income is likely to be measured with errors. Therefore, we estimate Equation (14) using the method of Two-Stage Least Squares (2SLS).

In the first stage, we estimate the following income regression for the three types of households separately:

$$\ln Y_i^R = \theta + X_i \varphi_1 + Z_i \varphi_2 + \xi_i, \tag{15}$$

where Z_i is a set of identifying instruments and ξ_i is the error term. We include the number of private cars and the gender of the household's head as identifying instruments. We use number of private cars in spirit of P&W that is expected to be a proxy for human capital or work effort which can be a relevant indicator for income generation process but not for food expenditures. We include gender to control for the income gaps in Poland that are observed between genders (Machrowska et al., 2014). We also tried other instruments that are used in the previous studies such as education level, occupational sector of the household's head, nationality, and household ownership dummy but none of these instruments passed the validity tests. The variance of the error term in Equation (15), ξ_i, is denoted as σ_j^2 for the risky households where $j = SE, PE$ and σ_{GE}^2 for the non-risky households. It is assumed that any differences in the reported income variances between the risky and non-risky households stem from the variance in the log under-reporting factor. We also assume that the regular income reported in HBS is a direct measure of permanent income since the temporary change in the income is eliminated. Under these two assumptions, the mean under-reporting factor can be found as in Equation (13):

$$\overline{k}_j = \exp\left(\frac{\gamma_j}{\beta} + \frac{1}{2}\left(\sigma_j^2 - \sigma_{GE}^2\right)\right), \quad j = SE, PE,$$

where $\overline{k}_j$ is the average under-reporting factor that indicates the number by which the regular incomes reported by the risky households should be multiplied to arrive at their total (true) incomes. In the above equation, γ_j and β are the estimates from Equation (14), and σ_j^2 for $j = SE, PE$, and σ_{GE}^2 are the estimates from Equation (15) over the three sub-samples.

Another way to express the extent of under-reporting is the average under-reporting share $\overline{s}_j$:

$$\overline{s}_j = \frac{\overline{k}_j - 1}{\overline{k}_j}, \quad j = SE, PE. \tag{16}$$

The under-reporting share $\overline{s}_j$ is the mean share of total (reported and unreported) income not reported by self-employed and private employees. We use the delta method to compute the standard errors of the under-reporting factors and shares for each risky household group.

Table 2 presents the estimated coefficients, mean under-reporting factors, $\overline{k}_j$, and shares, $\overline{s}_j$, by household groups, and first-stage tests from the estimation of Equations (14) and (15) using 2SLS with errors clustered at the household level. We also present the Ordinary Least Squares (OLS) estimates of Equation (14) in Table 2.

Column (1) in Table 2 presents the results from the instrumented model where the estimated elasticity of food consumption to the regular income, β, is 0.2780, the coefficient of the self-employment dummy, γ_{SE}, is 0.0559 and the coefficient of the private employee dummy, γ_{PE}, is -0.0110. All these variables are statistically

Table 2: Estimation by household groups

	2SLS (1)	OLS (2)
Log Regular Income	0.2780***	0.2142***
	(0.0061)	(0.0020)
Self-Employed (SE)	0.0559***	0.0540**
	(0.0032)	(0.0032)
Private Employees (PE)	-0.0110***	-0.0130***
	(0.0023)	(0.0023)
$\bar{k}_{SE}$	1.360***	1.285***
	(0.017)	(0.019)
$\bar{s}_{SE}$	0.265***	0.223***
	(0.009)	(0.012)
$\bar{k}_{PE}$	0.983***	0.941***
	(0.008)	(0.010)
$\bar{s}_{PE}$	-0018**	-0.062***
	(0.009)	(0.011)
Constant	3.2255***	3.6671***
	(0.0426)	(0.0202)
Centered R^2	0.3377	0.3432
No. of Observations	220,232	220,232
Endogenity test (p-value)	0.0000	
Kleibergen-Papp rk Wald F stat	5,813	
Hansen J-test (p-value)	0.3373	

Notes: Standard errors are in parenthesis. For all estimates in 2SLS, $\sigma^2_{SE} = 0.411$, $\sigma^2_{PE} = 0.241$ and $\sigma^2_{GE} = 0.197$. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

significant at the 1 percent level. The Kleibergen-Paap rk Wald F statistic and the p-value of the Sargan-Hansen test cannot reject that our identifying instruments for the first-stage regression in Equation (15) are valid.

We calculate the mean under-reporting factors and shares using Equations (13) and (16). The mean under-reporting factor and share for the self-employed are equal to 1.360 and 0.265, respectively, and both are significant at the 1 percent level. This means that self-employed individuals on average conceal 26.5 percent of their total income. The estimated under-reporting factor of 1.360 for self-employed is in the same ballpark as the previous estimates in the literature (see Kukk et al., 2020 for an extensive survey). The under-reporting factor for private employees is less than one, indicating no under-reporting activity by private employees in Poland, a result similar to the one obtained in the report on Undeclared Work in the European Union

(Special Eurobarometer 498, 2020). Our results are also similar to previous work in the literature, i.e., Paulus (2015) separates public and private employees and does not find the under-reporting factor to be larger than one for private employees using Estonian survey data. The OLS estimates shown in Column (2) of Table 2 produce results similar to the 2SLS estimates although the under-reporting factors and shares for both household groups calculated through OLS estimates are lower than in the case of the 2SLS estimates. The reason is that the variance term in Equation (3) declines as there is no first-stage regression in the OLS estimation.

5.1 Under-reporting by years

To assess the evolution of under-reporting activity in Poland over the years, we interact risky household dummies with yearly dummies and run the following regression:

$$\ln C_i = \alpha + \beta \ln Y_i^R + \gamma_{SE} D_{i,SE} + \gamma_{PE} D_{i,PE} + \psi_{SE} D_{i,SE} * Year_i + \\ + \gamma_{PE} D_{i,PE} + \psi_{PE} D_{i,PE} * Year_i + X_i \delta + \varepsilon_i, \quad (17)$$

where $Year_i$ shows the participation year of the household i to the survey. Table 3 presents the estimated coefficients, mean under-reporting shares, $\overline{s}_j$, by household groups and years, and first-stage tests. The yearly estimates of $\overline{k}_j$ in Equation (13) used to derive $\overline{s}_j$ in Equation (16) are calculated by dividing the sum of γ_j and ψ_j to β. The estimates of the coefficient on log reported regular income and income variances from the first-stage regression are restricted to be unchanged over the sample. The former is due to the fact that the share of food consumption over income between 2005 and 2017 in the Polish HBS did not show significant variation.

The estimated elasticity of food consumption to the permanent income, β, is 0.2752 and it is statistically significant at 1 percent level. The estimated coefficient on the self-employment dummy, γ_{SE}, is 0.0772, and private employee dummy, γ_{PE}, is -0.0019 in the base year 2005, but only the former is significant at 1 percent level. The Kleibergen-Paap rk Wald F statistic and p-value of the Sargan-Hansen test show that our identifying instruments are valid.

The self-employment coefficients are statistically significant at the 1 percent level for all of the years and the largest under-reporting occurred in years 2005, 2006, 2014, and 2015, and the lowest occurred in the years 2011 and 2016. The under-reporting share of the private employees fluctuates around 0 across all years in line with the results from Table 2. Also, the under-reporting share is significant at the 1 percent level in all years for the self-employed but only in years 2008, 2012, 2013, and 2016 for the private employees and only at 5 percent significance levels.

These results are again in line with the results from the baseline estimation and suggest that only self-employed households are prone to underreporting in Poland since the under-reporting share for private employees generally moves around zero without any clear trend during the period of study.

Table 3: Estimation by household groups and years

	Self-Employed		Private Employee	
	2SLS	$\bar{s}_{SE}$	2SLS	$\bar{s}_{PE}$
Log Regular Income	0.2752***		0.2752***	
	(0.0059)		(0.0059)	
Year 2005	0.0772***	0.321***	-0.0019	0.015
	(0.0089)	(0.022)	(0.0057)	(0.021)
Year 2006	0.0765***	0.319***	-0.0023	0.014
	(0.0088)	(0.022)	(0.0057)	(0.020)
Year 2007	0.0594***	0.276***	-0.0118**	-0.021
	(0.0091)	(0.024)	(0.0057)	(0.021)
Year 2008	0.0537***	0.261***	-0.0187***	-0.047**
	(0.0093)	(0.025)	(0.0059)	(0.023)
Year 2009	0.0564***	0.268***	-0.0104*	-0.016
	(0.0092)	(0.025)	(0.0060)	(0.022)
Year 2010	0.0463***	0.240***	-0.0150**	-0.033
	(0.0094)	(0.026)	(0.0059)	(0.022)
Year 2011	0.0356***	0.210***	-0.0141**	-0.030
	(0.0092)	(0.027)	(0.0061)	(0.023)
Year 2012	0.0538***	0.261***	-0.0190***	-0.048**
	(0.0092)	(0.025)	(0.0061)	(0.023)
Year 2013	0.0438***	0.234***	-0.0198***	-0.051**
	(0.0095)	(0.027)	(0.0064)	(0.025)
Year 2014	0.0740***	0.313***	-0.0055	0.002
	(0.0094)	(0.024)	(0.0063)	(0.023)
Year 2015	0.0734***	0.312***	0.0022	0.030
	(0.0099)	(0.025)	(0.0064)	(0.023)
Year 2016	0.0305***	0.195***	-0.0197***	-0.051**
	(0.0097)	(0.028)	(0.0067)	(0.026)
Year 2017	0.0426***	0.230***	-0.0078	-0.006
	(0.0100)	(0.028)	(0.0067)	(0.024)
Constant	3.2465***			
	(0.0404)			
R^2	0.3341			

Notes: The table continues on the next page.

Table 3: Estimation by household groups and years cont.

No. of Obs. 220,232	
Endogenity test (p-value)	0.0000
Kleibergen-Papp rk Wald F stat	6,227.58
Hansen J-test (p-value)	0.3964

Notes: Standard errors are in parenthesis. The year coefficients other than the base year are the sum of estimated coefficients on the risky household dummies and the interactions between the risky household dummy and year dummy. For all estimates, $\sigma^2_{SE} = 0.410$, $\sigma^2_{PE} = 0.241$ and $\sigma^2_{GE} = 0.197$. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Comparing the results across years, we see that the under-reporting share for self-employed moves between 20 to 30 percent with a slightly decreasing trend. A decreasing trend in terms of the underground economy in general over a similar period to the one we studied has also been reported in other studies using alternative methods of examining it (Dybka et al. 2019). It should be noted that the highest underreporting of income among self-employed occurred in 2005 and 2006 and 2014 and 2015. These results are close to the phenomena observed in the study of the Polish Central Statistical Office (CSO) on the underground economy, according to which, in the year of Poland's accession to the EU, the number of people working in the underground economy was almost twice as high as in subsequent years, although no increase in the number of people working in the underground economy was observed in the CSO's study in 2014, such an increase was, however, recorded similarly to the presented consumption method study in 2017. (CSO, 2019). This is not surprising, however, if we look at the VAT gap, which recorded some of the highest values in 2014 and 2015 after Poland's accession to the EU (Mazur et al. 2019). It should also be noted that it is difficult to conduct a reliable comparative study with the CSO study of those working in the underground economy, as the presented study 'uses' permanent income in the model, omitting temporary income as a rule, while the CSO uses the criterion of main and additional work, not necessarily coinciding with the approach presented here.

6 Sensitivity analysis and extensions

6.1 Sensitivity analysis

We check the sensitivity of our baseline results through various robustness tests. Table 4 summarizes the results for the baseline estimation and the robustness checks in columns specified as following: (i) Column (1) for Baseline 2SLS estimation; (ii) Column (2) for the estimation with alternative definition of the households where the head of the household is defined as self-employed if regular self-employment income is more than 75% of the total regular income, as a private employee if regular self-employment income is less than 25% and private to public employment income

ratio is larger than half, and as public employee otherwise; (iii) Column (3) for the estimation with pensioners in the sample and placed in the non-risky group along with public employees; (iv) Column (4) for the estimation where single households (only one adult) are included in the sample; (v) Column (5) for the estimation using only food expenditures at home as the dependent variable; (vi) Column (6) for the estimation using the total income (i.e. the sum of regular and temporary incomes) as an independent variable instead of regular income.

Table 4: Sensitivity Analysis

	(1)	(2)	(3)	(4)	(5)	(6)
Log Regular Income (β)	0.2780***	0.2734***	0.3092***	0.2862***	0.2292***	0.3777***
	(0.0061)	(0.0061)	(0.0054)	(0.0057)	(0.0059)	(0.0073)
Self-Employed (SE)	0.0559***	0.0457***	0.0544***	0.0631***	0.0144***	0.0313***
	(0.0032)	(0.0039)	(0.0029)	(0.0030)	(0.0031)	(0.0030)
Private Employees (PE)	-0.0110***	-0.0096***	-0.0151***	-0.0073***	-0.0268***	-0.0126***
	(0.0023)	(0.0023)	(0.0020)	(0.0021)	(0.0023)	(0.0022)
$\bar{s}_{SE}$	0.265***	0.217***	0.251***	0.294***	0.156***	0.102***
	(0.009)	(0.012)	(0.007)	(0.008)	(0.016)	(0.007)
$\bar{s}_{PE}$	-0.018**	-0.005	-0.028***	-0.002	-0.099***	-0.036***
	(0.009)	(0.008)	(0.007)	(0.008)	(0.012)	(0.006)
Constant	3.2255***	3.3434***	3.0729***	3.2929***	3.6700***	2.9506***
	(0.0426)	(0.0407)	(0.0364)	(0.0374)	(0.0400)	(0.0505)
R^2	0.3377	0.3280	0.3481	0.4109	0.3052	0.3467
Total Obs.	220,232	208,536	308,888	250,589	220,232	224,507
Endogeneity test (p-value)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Weak iden. test (F stat)	5,813.00	5,874.09	7,850.52	6,859.02	6,248.52	5,276.30
Hansen J-test (p-value)	0.3373	0.4479	0.6376	0.0000	0.0000	0.7494

Notes: Standard errors are in parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The under-reporting shares in columns (2)-(4) are very close to those in the baseline estimation and the share for the self-employed ranges between 0.22 to 0.29. The picture is slightly different in the columns (5)-(6) where the under-reporting shares are smaller relative to the shares in column (1). However, these findings are not surprising; for example, by including temporary income, we allow temporary variations in the income to affect the food consumption which results in a higher consumption elasticity and lower under-reporting shares as shown in column (6). The findings from the sensitivity analysis support the validity of the results from the baseline estimation.
We conduct additional robustness checks and present the results in Table A2 of Appendix with the columns specified as following: (i) Column (1) for the Baseline

2SLS estimation; (ii) Column (2) where we include private employees into the non-risky household group; (iii) Column (3) where we use alternative definition of the self-employed that is the household if the main source of income of the head of the household is from self-employment, ownership, and rent; (iv) Column (4) where we limit the sample only to the working-age population in Poland, i.e. people aged between 18-65; (v) Column (5) where we limit the sample to the survey data in 2017; (vi) Column (6) where we only use explanatory variables without missing observations in the estimation.
This additional battery of robustness checks also confirms the soundness of the baseline estimates and under-reporting shares.

6.2 Extension: Under-reporting with non-linear income elasticity of consumption

This section presents an extension to the original P&W framework to account for the impact of possible non-linear effects of income on consumption. One of the main assumptions of the baseline model given in Equation (14) is that income elasticity of consumption, β, is log-linear in income. However, this elasticity can be a decreasing function of the income. We test for the non-linear effects of income on food consumption by including quadratic income term into the baseline model as following:

$$\ln C_i = \alpha + \beta \ln Y_i^R + \varphi \left(\ln Y_i^R\right)^2 + \gamma_{SE} D_{i,SE} + \gamma_{PE} D_{i,PE} + X_i \delta + \varepsilon_i. \tag{18}$$

We label the model given in Equation (18) as non-linear model and estimate it using OLS because we did not find suitable instruments for log income squared. Table 5 presents the summary of the results together with baseline 2SLS and OLS estimates. We see that the coefficient of the squared log income, φ, is statistically significant. We also see the sign of the β changes. The immediate question that arises is whether the non-linear effect of income on consumption changes the under-reporting factors and shares estimated in the baseline specification. We develop a new framework to answer this question because Equation (13) is not valid anymore to calculate the under-reporting factor due to the non-linear term in Equation (18).
The framework to account for the non-linear income effects starts with adding the quadratic income term to the P&W's expenditure function given in Equation (1):

$$\ln C_i = \alpha + \beta \ln Y_i^P + \theta \left(\ln Y_i^P\right)^2 + X_i \delta + \eta_i. \tag{19}$$

By using the same assumptions given in Equations (2)–(3), we can rewrite Equation (19) as follows:

$$\ln C_i = \alpha + \beta \ln Y_i^R + \beta\mu + \beta v_i + \theta \left(\ln Y_i^R\right)^2 + \theta(\mu)^2 + \theta \left(v_i\right)^2 + X_i \delta + \eta_i, \tag{20}$$

and applying Equation (6) in Equation (20) gives:

$$\ln C_i = \alpha + \beta \ln Y_i^R + \theta \left(\ln Y_i^R\right)^2 + \sum_{j=SE,PE} D_{i,j} \left[\beta \mu_j + \theta \left(\mu_j\right)^2\right] + X_i \delta + \varepsilon_i. \quad (21)$$

The following can be deduced from Equations (18) and (21):

$$\theta \left(\mu_j\right)^2 + \beta \mu_j - \gamma_j = 0. \quad (22)$$

The Equation (22) is a quadratic equation, and its roots are given by:

$$\mu_j = \frac{-\beta \pm \sqrt{\beta^2 - 4\theta\gamma_j}}{2\theta}. \quad (23)$$

The mean of the log under-reporting factor, μ_j, can be obtained from Equation (23) since μ_j is the mean of a log variable, the solution root should be strictly larger than zero. In our estimations, we only selected the root that satisfies this requirement. Finally, the variance of the under-reporting factor is equal to zero for the all household types since we estimate Equation (18) through OLS. This allows us to derive the under-reporting factors and shares from the non-linear model:

$$\overline{k}_j = \exp\left(\mu_j\right); \qquad j = SE, PE, \quad (24)$$

$$\overline{s}_j = \frac{\overline{k}_j - 1}{\overline{k}_j}; \qquad j = SE, PE. \quad (25)$$

Column (1) in Table 5 presents the under-reporting factors and shares for self-employed and private employees given in Equations (24) and (25) along with the estimated parameters obtained from the regression of the non-linear model specified in Equation (18). Columns (2) and (3) in Table 5 present the results for the same set of parameters obtained from the 2SLS and OLS estimations of baseline empirical model given in Equation (14), respectively.

We see that allowing for non-linear effects of the income on food expenditures does not have a large effect on the under-reporting shares for the self-employed and the private employees. For both groups, the under-reporting shares are slightly and moderately lower relative to the shares obtained from baseline OLS and 2SLS estimations, respectively. The results in Table 5 suggest that exclusion of the non-linear effects in the baseline model does not lead to a downward bias in the under-reporting shares for the risky groups.

Table 5: Estimation with quadratic income variable

	Non-Linear OLS (1)	Baseline 2SLS (2)	Baseline OLS (3)
Log Regular Income	-0.2134***	0.2780***	0.2142***
	(0.0204)	(0.0061)	(0.0020)
Squared Log Regular Income	0.0276***		
	(0.0013)		
Self-Employed (SE)	0.0399***	0.0559***	0.0544***
	(0.0031)	(0.0032)	(0.0031)
Private Employees (PE)	-0.0164***	-0.0110***	-0.0127***
	(0.0022)	(0.0023)	(0.0022)
$\bar{k}_{SE}$	1.211***	1.360***	1.289***
	(0.031)	(0.017)	(0.019)
$\bar{s}_{SE}$	0.174***	0.265***	0.224***
	(0.021)	(0.009)	(0.011)
$\bar{k}_{PE}$	0.927***	0.983***	0.943***
	(0.011)	(0.008)	(0.010)
$\bar{s}_{PE}$	-0.079***	-0.018**	-0.061***
	(0.013)	(0.009)	(0.011)
Constant	5.4159***	3.2255***	3.6808***
	(0.0797)	(0.0426)	(0.0190)
Centered R2	0.3358	0.3377	0.3392
No. of Observations	220,232	220,232	220,232
Endogeneity test (p-value)		0.0000	
Kleibergen-Papp rk Wald F stat		5,813	
Hansen J-test (p-value)		0.3373	

Notes: Standard errors are in parenthesis. For 2SLS estimates, $\sigma^2_{SE} = 0.410$, $\sigma^2_{PE} = 0.241$ and $\sigma^2_{GE} = 0.197$. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

7 Conclusions

In this paper, we estimate the income underreporting and size of the underground economy in Poland by using the consumption method developed by Pissarides and Weber (1989) on Polish Household Budget Survey between the years 2005 and 2017. In contrast to the original method of Pissarides and Weber (1989), we differentiate between the regular and temporary income of the household and use the former as a proxy for permanent income in estimating the income elasticity of food consumption. We distinguish between public and private employees and relax the assumption of truthful income reporting of the latter. Moreover, we extend theoriginal method to account for the non-linear effects of income on food consumption.

We estimated an under-reporting factor of 1.36 for the self-employed which means that self-employed individuals conceal 26.5 percent of their total income, on average. The under-reporting factor for private employees is close to one, indicating no under-reporting activity by private employees in Poland. These results can be interpreted in absolute terms as long as our assumption of truthful reporting by the public employees holds. When we decompose the underreporting activity by years, we find that the share of unreported income fluctuates between 20 and 30 percent across 2005 to 2017 but with a decreasing trend. These results are robust to various sensitivity checks, including non-linear expenditure function.

This paper is the first comprehensive study of the income underreporting in Poland using the consumption method over relatively long horizon. Our results suggest that income underreporting in Poland occurred only across self-employed households during the period of study. This result is not surprising since self-employed is more prone to income underreporting due to the limited third-party reporting (Soos, 1990; Bruce, 2000; Alm et al., 2009). We also found that the extent of underreporting by the self-employed decreased between 2005 and 2017 probably due to the implementation of the PIT tax reduction policy and allowing more self-employed people to benefit from simplified forms of taxation in this tax.

There are a few policy implications of our findings. First, third-party reporting substantially reduces the income underreporting (in our study, employees who receive their salaries through third-party did not underreport their income). Second, self-employed households should be in the radar of the tax authorities to detect non-compliance. Third, the policies implemented after the EU membership such as reducing the burden of taxation in income taxes, also in terms of compliance costs, and sealing up the VAT area (B2C and B2B reporting requirements e.g. cash registers and SAF-T) should be strengthened since they are effective tools to reduce the non-compliance.

Richer consumption data and matching of the surveyed households with their tax returns data would be useful in order to take the analysis further. As of now, the paper is limited to a single type of expenditure despite building on a rich data source. It would be possible to estimate demand system equations with richer consumption data. By utilizing tax returns data, it would be possible to measure the discrepancies

between survey and fiscal incomes and assess the impact of these discrepancies on the under-reporting activity.

Acknowledgments

We would like to thank two anonymous referees and the editor for helpful comments, Centrum Projektów Europejskich for financial support (Project. No. POWR.04.03.00-00-0047/17) and Adam Śmietanka (CASE) for arranging the database. We are solely responsible for the interpretation of the data and all errors.

References

[1] Albarea A., Bernasconi M., Marenzi A., Rizzi D., (2020), Income underreporting and tax evasion in Italy: Estimates and distributional effects, *Review of Income and Wealth* 66(4), 904–930.

[2] Banks J., Blundell R., Lewbel A., (1997), Quadratic Engel curves and consumer demand, *Review of Economics and Statistics* 79(4), 527–539.

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

[4] 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.

[5] Central Statistical Office (Główny Urząd Statystyczny), (2018), Zeszyt metodologiczny. Badanie budżetów gospodarstw domowych, GUS, Warszawa.

[6] Central Statistical Office (Główny Urząd Statystyczny), (2019), Praca nierejestrowana w Polsce w 2017 r., GUS, Warszawa.

[7] Dybka P., Kowalczuk M., Olesiński B., Torój A., Rozkrut M., (2019), Currency Demand and Mimic Models: Towards a Structured Hybrid Method of Measuring the Shadow Economy, *International Tax and Public Finance* 26(1), 4–40.

[8] 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.

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

[10] Feige E., (1990), Defining and Estimating Underground and Informal Economies: The New Institutional Economics Approach, *World Development* 18(7), 989–1002.

[11] Hurst E., Li G., Pugsley B., (2014), Are Household Surveys Like Tax Forms: Evidence from Income Underreporting of the Self Employed, *The Review of Economics and Statistics* 96(1), 19–33.

[12] Johansson E., (2000), An Expenditure-Based Estimation of Self-Employment Income Underreporting in Finland, Swedish School of Economics and Business Administration, Working papers 433.

[13] 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.

[14] Kukk M., Staehr K., (2014), Income Underreporting by Households with Business Income: Evidence from Estonia, *Post-Communist Economies* 26(2), 257–76.

[15] 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.

[16] Kukk M., Paulus A., Staehr K., (2020), Cheating in Europe: underreporting of self-employment income in comparative perspective, *International Tax and Public Finance* 27, 363–390.

[17] Lichard T., Hanousek J., Filer R.K., (2021), Hidden in plain sight: using household data to measure the shadow economy, *Empirical Econonomics* 60, 1449–1476.

[18] Lyssiotou P., Pashardes P., Stengos T., (2004), Estimates of the Black Economy Based on Consumer Demand Approaches, *The Economic Journal* 114(497), 622–640.

[19] Majchrowska A., Strawiński P., Konopczak K., Skierska A., (2014), Why are women paid less than men? An investigation into gender wage gap in Poland, University of Warsaw, Faculty of Economic Sciences Working Papers No: 31/2014(148).

[20] Mazur T., Bach D., Juźwik A., Czechowicz I., Bieńkowska J., (2019), Raport na temat wielkości luki podatkowej w podatku VAT w Polsce w latach 2004–2017. MF Opracowania i Analizy, No 3-2019.

[21] Myck M., Najsztub M., (2015), Data and Model Cross-Validation to Improve Accuracy of Microsimulation Results: Estimates for the Polish Household Budget Survey, *International Journal of Microsimulation* 8(1), 33–66.

[22] Nygård O. E., Slemrod J., Thoresen T. O., (2019), Distributional implications of joint tax evasion, *The Economic Journal* 129(620), 1894–1923.

[23] Owsiak S., (2016), System podatkowy Polski w okresie transformacji – próba oceny, *Annales Universitatis Mariae Curie-Skłodowska Lublin – Polonia* L(H), 15–27.

[24] Paulus A., (2015), Income underreporting based on income expenditure gaps: Survey vs tax records (No. 2015-15), ISER Working Paper Series.

[25] Pissarides A., Weber G., (1989), An Expenditure-Based Estimate of Britain's Black Economy, *Journal of Public Economics* 39(1), 17–32.

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

[27] Schneider F., (2005), Shadow economies around the world: What do we really know?, *European Journal of Political Economy* 21, 598–642.

[28] Schuetze H. J., (2002), Profiles of Tax Non-Compliance Among the Self-Employed in Canada: 1969 to 1992, *Canadian Public Policy/Analyse de Politiques*, 219–238.

[29] Slemrod J., Weber C., (2012), Evidence of the invisible: toward a credibility revolution in the empirical analysis of tax evasion and the informal economy, *Tax Public Finance* 19(1), 25–53.

[30] Smith S., (1986), *Britain's shadow economy*, Oxford University Press, USA.

[31] Special Eurobarometer 498, (2020), Undeclared Work in the European Union, European Union, available at: `https://europa.eu/eurobarometer/surveys/detail/2250`.

Appendix

Table A1: Summary statistics of household characteristics

	Full-Sample	Self-Employed (SE)	Private Employee (PE)	Public Employee (GE)
Sample size	220,232	29,579	129,557	61,096
Real Food Consumption	916.72 zł	1014.29 zł	892.59 zł	920.64 zł
Real Regular Income	3,059.29 zł	3,640.86 zł	2,874.40 zł	3,169.80 zł
Real Total income	4,708.36 zł	5,471.62 zł	4,428.09 zł	4,932.40 zł
Age	43.23	44.08	42.09	45.25
Female Ratio	28.27 %	23.95%	24.40%	38.56%
Number of children	1.36	1.41	1.38	1.29
Number of cars	0.84	0.98	0.79	0.88
Size of house	84.06 m2	102.07 m2	79.74 m2	84.49 m2
Distribution of Households by Education Status				
Education Level 1	6.63 %	4.70 %	8.18 %	4.26 %
Education Level 2	7.08 %	8.51 %	6.95 %	6.67 %
Education Level 3	2.28 %	2.64 %	1.83 %	3.08 %
Education Level 4	59.54 %	57.57 %	65.64 %	47.57 %
Education Level 5	24.42 %	26.54 %	17.35 %	38.39 %
Education Level 6	0.04 %	0.04 %	0.05 %	0.03 %
Distribution of Households by Voivodeship				
Dolnośllaskie	7.83 %	7.65 %	8.16 %	7.23 %
Kujawsko-pomorskie	5.16 %	5.13 %	5.46 %	4.54 %
Lubelskie	4.91 %	4.70 %	4.24 %	6.41 %
Lubuskie	2.74 %	2.78 %	2.76 %	2.67 %
Łódzkie	6.82 %	6.48 %	6.98 %	6.64 %
Małopolskie	8.87 %	9.90 %	8.94 %	8.22 %
Mazowieckie	15.24 %	16.06 %	14.85 %	15.68 %
Opolskie	2.72 %	2.60 %	2.81 %	2.58 %
Podkarpackie	5.40 %	4.51 %	5.49 %	5.62 %
Podlaskie	2.47 %	2.61 %	2.25 %	2.87 %
Pomorskie	6.08 %	6.96 %	6.29 %	5.23 %
Śląskie	12.05 %	10.48 %	11.07 %	14.88 %
Świętokrzyskie	3.14 %	3.00 %	3.17 %	3.14 %
Warmińsko-mazurskie	3.53 %	3.10 %	3.55 %	3.70 %
Wielkopolskie	8.90 %	9.34 %	10.11 %	6.11 %
Zachodniopomorskie	4.16 %	4.70 %	3.88 %	4.48 %

Notes: The table continues on the next page.

Table A1: Summary statistics of household characteristics cont.

	Full-Sample	Self-Employed (SE)	Private Employee (PE)	Public Employee (GE)
Distribution of Households by Location				
Location 1	12.76 %	15.06 %	12.42 %	12.38 %
Location 2	9.22 %	10.32 %	8.70 %	9.78 %
Location 3	7.51 %	7.17 %	7.20 %	8.32 %
Location 4	18.36 %	17.47 %	17.55 %	20.52 %
Location 5	11.96 %	11.96 %	11.38 %	13.20 %
Location 6	40.19 %	38.01 %	42.75 %	35.81 %
Distribution of Households by Industry				
PKD 1	25.23 %	18.27 %	17.89 %	44.15 %
PKD 2	3.58 %	0.97 %	4.81 %	2.21 %
PKD 3	1.99 %	1.83 %	2.85 %	0.26 %
PKD 4	17.72 %	9.32 %	25.52 %	5.27 %
PKD 5	12.21 %	21.84 %	14.78 %	2.10 %
PKD 6	9.42 %	12.60 %	9.59 %	7.50 %
PKD 7	1.50 %	2.30 %	1.84 %	0.37 %
PKD 8	2.62 %	2.58 %	2.68 %	2.52 %
PKD 9	4.53 %	7.52 %	2.65 %	7.05 %
PKD 10	5.96 %	2.10 %	1.38 %	17.54 %
PKD 11	3.48 %	2.80 %	1.21 %	8.62 %
PKD 12	0.77 %	1.83 %	0.68 %	0.47 %
PKD 13	11.00 %	16.04 %	14.12 %	1.93 %

Education level

Code	Education level
1	without education
2	Primary, lower secondary
3	Secondary general, teacher training college, language college, for social service workers
4	post-secondary
5	basic vocational, secondary vocational
6	bachelor's or engineer's degree, master's degree or equivalent, higher with an academic degree

Notes: The table continues on the next page.

Table A1: Summary statistics of household characteristics cont.

Location

Code	Location
1	500,000 inhabitants or more
2	200,000 - 499,000 inhabitants
3	100,000 - 199,000 inhabitants
4	20,000 - 99,000 inhabitants
5	below 20,000 inhabitants
6	Rural areas

PKD

Scope of PKD codes (2005-2007) (1)	Scope of PKD codes (2008-2016) (2)	Category used in the model (3)	Description (4)
15	10-11	1	Section C - MANUFACTURING (manufacture of food products, manufacture of beverages)
36	31-32	2	Section C - MANUFACTURING (manufacture of furniture, other manufacturing)
16-35	12-30, 33	3	Section C - MANUFACTURING (other categories from Section C)
45	41-43	4	Section F - CONSTRUCTION
50-52	45-47	5	Section G - WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES
60-64	49-53	6	Section H - TRANSPORTATION AND STORAGE
55	55-56	7	Section I - ACCOMMODATION AND FOOD SERVICE ACTIVITIES
65-67	64-66	8	Section K - FINANCIAL AND INSURANCE ACTIVITIES
72-74	69-75	9	Section M - PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES
80	85	10	Section P - EDUCATION
85	86-88	11	Section Q - HUMAN HEALTH AND SOCIAL WORK ACTIVITIES
	Other	12	Other sections, divisions - other than those listed above
	Non-classified	13	No indication of PKD code

Notes: The table gives the mean value of the socio-economic variables and distribution of households by education level, location and size of the place of residence and industry the head of the household works. The following tables provide the description of education level, location, and PKD variables.

Table A2: Additional sensitivity analysis

	(1)	(2)	(3)	(4)	(5)	(6)
Log Regular Income (β)	0.2780*** (0.0061)	0.2760*** (0.0059)	0.3285*** (0.0062)	0.2743*** (0.0059)	0.3281*** (0.0196)	0.3086*** (0.0047)
Self-Employed (SE)	0.0559*** (0.0032)	0.0634*** (0.0027)	0.0390*** (0.0032)	0.0555*** (0.0031)	0.0443*** (0.0114)	0.0347*** (0.0024)
Private Employees (PE)	-0.0110*** (0.0023)		-0.0106*** (0.0022)	-0.0110*** (0.0022)	-0.0057 (0.0086)	-0.0334*** (0.0018)
$\bar{s}_{SE}$	0.265*** (0.009)	0.275*** (0.015)	0.169*** (0.010)	0.266*** (0.009)	0.190*** (0.029)	0.201*** (0.007)
$\bar{s}_{PE}$	-0.018** (0.009)		-0.016* (0.008)	-0.018** (0.008)	-0.002 (0.026)	-0.091*** (0.006)
Constant	3.2255*** (0.0426)	3.2423*** (0.0399)	3.1851*** (0.0429)	3.2447*** (0.0405)		2.8961*** (0.0318)
R^2	0.3377	0.3337	0.3348	0.3337	0.3065	0.5021
Total Obs.	220,232	220,232	217,915	217,803	16,836	437,803
Endogenity test (p-val.)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Kleibergen-Papp rk Wald F stat	5,813.00	6,284.25	5,970.35	6,210.81	763.36	9,863.14
Hansen J-test (p-val.)	0.3373	0.3545	0.1770	0.3184	0.4931	0.3308

Notes: Standard errors are in parenthesis. *$p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.