Economics Letters 237 (2024) 111626

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Misreporting in household income and expenditure: Evidence from the Chinese Household Income Project

Feng Li [a,*], Xintao Wang [b]

[a] School of Economics, Hangzhou Normal University, No.2318 Yuhangtang Rd., Yuhang District, Hangzhou, Zhejiang, China
[b] Institute of Economics, Chinese Academy of Social Sciences, China

ARTICLE INFO

JEL classification codes:
C81
O15
D31

Keywords:
Misreporting
Income
Expenditure
Mean version
Subjective well-being

ABSTRACT

This study provides new evidence on misreporting in household income and expenditure using recall and diary data from the 2013 Chinese Household Income Project. Using more accurate diary records as benchmarks, we observe substantial and systematic misreporting in income and expenditure from recall data. Two main patterns of misreporting are identified: mean reversion and correlation with subjective well-being (i.e., happier respondents tend to overreport).

1. Introduction

Household income and expenditure data are crucial for a wide range of both basic research and policy analysis. Meanwhile, misreporting in income or expenditure data in surveys has been extensively recognized through comparisons with recall and administrative data. Early studies conducted by Bound and Krueger (1991) and Bound et al. (1994), using data from the U.S., were the first to find that errors in self-reported earnings data in surveys exhibit a negative correlation with true earnings obtained from administrative tax records. This negative correlation suggests a pattern of mean reversion in misreporting earnings. The mean reversion pattern is supported by more recent studies, such as Kreiner et al. (2013) using data from Denmark, Kim and Tamborini (2014) using data from the U.S. (2014), Valet et al. (2019) using data from Germany, and Angel et al. (2019) using data from Austria. Furthermore, Brzozowski et al. (2017) also identified this pattern of mean reversion in misreporting in food expenditure data from Canada. In contrast to the mean version pattern, Hariri and Lassen (2017) found that individuals with higher income tend to overreport their income, based on data from the Netherlands. All these studies are conducted using data from developed countries.

In developing countries like China, administrative data on income and expenditure are generally unavailable for research purposes, resulting in a lack of studies on misreporting in income and expenditure. Nevertheless, in the absence of administrative data as benchmarks, a strand of literature attempts to infer true income in China, particularly among high-income households: Wang and Woo (2011) utilized the relationship between Engel's coefficient and income level to deduce the true level of household income; Gao et al. (2015) inferred residents' hidden income from their housing affordability; and Zhang and Zhao (2019) estimated the size and structure of unreported income using cash flow statements. However, to the best of our knowledge, no studies have explored the pattern of income or expenditure misreporting in China by comparing survey data with administrative data or diary data. This study aims to address this gap by leveraging recall and diary data from the 2013 Chinese Household Income Project (CHIP) survey.

Our new empirical evidence reveals substantial misreporting in both household income and expenditure in China, identifying two primary patterns of misreporting. For one, we find strong evidence of mean reversion in both self-reported household income and expenditure data. Specifically, using more precise diary records as benchmarks, we observe that in survey data, respondents from high-income (expenditure) households tend to underreport their household income (expenditure), while respondents from low-income (expenditure) households tend to overreport their household income (expenditure). For another, considering that self-reported income and expenditure are to some

* Corresponding author.
E-mail address: fengli@hznu.edu.cn (F. Li).

https://doi.org/10.1016/j.econlet.2024.111626
Received 11 January 2024; Received in revised form 28 February 2024; Accepted 29 February 2024
Available online 17 March 2024

degree subjective data, we examine whether they are influenced by respondents' subjective factors during the survey. Particularly, we explore the correlation between respondents' subjective well-being and their misreporting. The regression results show that respondents who report "very happy" tend to overreport their household income and expenditure by over 10 log points compared to respondents who report "not happy at all".

Our contribution to the existing literature is twofold. Firstly, our empirical evidence confirms that in the Chinese context, mean reversion also significantly contributes to misreporting in income and expenditure, aligning with prior literature (e.g., Brzozowski et al. 2017, Angel et al. 2019). On the one hand, income and consumption inequality are pronounced in China, leading lower-income (expenditure) households to potentially overreport to avoid embarrassment, while higher-income (expenditure) households may feel compelled to underreport to avoid hurting others' feelings. On the other hand, conspicuous consumption and flaunting wealth are prevalent in China (Jinkins, 2016), potentially influencing all households, including those with higher income (expenditure), to overreport their income and expenditure in order to show off wealth. Our empirical results support the dominance of the former argument.

Secondly, our identification of a new source of reporting bias, namely subjective well-being, might deepen our understanding of the impact of income and consumption on individuals' subjective well-being. The current body of literature consistently highlights positive effects of income and consumption on subjective well-being (e.g., Carver and Grimes 2019). However, when utilizing income and consumption data from household surveys, caution is warranted as these effects may be overestimated. This is due to the tendency of respondents with higher subjective well-being to systematically overreport their income and consumption. Therefore, careful interpretation is needed when assessing the causal effects of self-reported income and consumption on subjective well-being.

2. Data

The 2013 Chinese Household Income Project (CHIP) survey provides an opportunity to explore misreporting in household income and expenditure in the Chinese context. The 2013 CHIP datasets contain household income and expenditure from two sources, real-time diaries and end-of-year recall.[1] The 2013 CHIP questionnaire collecting recall data was administered to the sample households in mid-2014.[2] The collection of recall data occurs more than half a year after the recording of real-time diaries, potentially resulting in significant misreporting in recall data due to a substantial time gap. By employing more accurate diary records as benchmarks, we can investigate the extent and patterns of misreporting in recalled household income and expenditure, in line with the approach taken by Brzozowski et al. (2017).

In the 2013 CHIP datasets, n3701 and n4202 serve as variable names for household income and expenditure derived from diary data, while F01_1 and F02_1 represent household income and expenditure obtained through recall data. The definitions of household income and expenditure remain consistent for both diary and recall data. Household income is gauged through household disposable income, comprising wage income, net business income, net property income, and net transfer income. Household expenditure consists of spending on eight categories: (1) food and tobacco, (2) clothing, (3) housing, (4) facility and services, (5) communication and transportation, (6) education, entertainment, and cultural activities, (7) healthcare, and (8) miscellaneous goods and services. A total of 16,908 households have data on both recall and diary-based income and expenditure, comprising 6262 from the urban household subsample, 9973 from the rural household subsample, and 673 from the rural-to-urban migrant household subsample.

Table 1
Summary statistics.

Variables	Obs	Mean	Std. Dev.	Min	Max
Household income (recall) in 10,000 RMB	16,492	5.5193	3.8448	0.45	23.59
Household income (diary) in 10,000 RMB	16,577	5.2834	3.9633	0.3003	23.2642
Household expenditure (recall) in 10,000 RMB	16,455	3.2832	2.3737	0.318	15
Household expenditure (diary) in 10,000 RMB	16,572	3.8049	2.779	0.5098	17.8943
Urban household (1 = Yes)	16,908	0.3704	0.4829	0	1
Rural household (1 = Yes)	16,908	0.5898	0.4919	0	1
Rural-to-urban migrant household (1 = Yes)	16,908	0.0398	0.1955	0	1
Family size	16,908	3.4197	1.3736	1	13
Education years of the respondent	16,011	8.6097	3.5545	0	22
Age of the respondent	16,324	49.3406	13.031	18	97
Gender of the respondent (1 = male; 0 = female)	16,405	0.5885	0.4921	0	1

In Table 1, we provide basic descriptive statistics. To eliminate extreme outliers, we exclude the top and bottom 1% of household income and expenditure from both recall and diary data. Since typically only adults are acquainted with household income and expenditure conditions, respondents younger than 18 are excluded. On average, recalled household income (55,193 RMB) are higher than diary-based household income (52,834 RMB), while recalled household expenditure (32,832 RMB) are lower than diary-based household expenditure (38,049 RMB). If we use diary data as benchmarks, Chinese households, on average, tend to overreport income by 4.46% and underreport expenditure by 13.71%.

3. The patterns of misreporting in household income and expenditure

We investigate the patterns of misreporting in household income and expenditure using a linear regression model following the approach of Hariri and Lassen (2017):

$$Misreporting_i(income\ /\ expenditure) = \alpha + X_i\beta + \varepsilon_i \quad (1)$$

Eq. (1) represents the linear regression model for misreporting in income or expenditure. The dependent variable $Misreporting_i$ ($income$) denotes the misreporting in household income of respondent i, as defined in Eq. (2) by the difference between the log of recalled income ($income\_recall_i$) and the log of diary-based income ($income\_diary_i$):

$$Misreporting_i(income) = ln(income\_recall_i) - ln(income\_diary_i) \quad (2)$$

Similarly, the dependent variable $Misreporting_i$ ($expenditure$) is defined in Eq. (3) as the difference between the log of recalled expenditure ($expenditure\_recall_i$) and the log of diary-based expenditure ($expenditure\_diary_i$):

$$Misreporting_i(expenditure) = ln(expenditure\_recall_i) - ln(expenditure\_diary_i) \quad (3)$$

[1] The 2018 CHIP datasets, the latest wave of CHIP, contain full sets of diary data on household income and expenditure. However, in the recall data, respondents only reported their individual total income, comprising wage income and net business income, rather than reporting household income and expenditure. As a result, we can only partially replicate the results from the 2013 datasets with the 2018 datasets. Nevertheless, the two main patterns of misreporting are also identified in the 2018 datasets. For further details, please refer to the appendix.

[2] For additional details regarding the 2013 CHIP datasets, refer to Sicular et al. (2017).

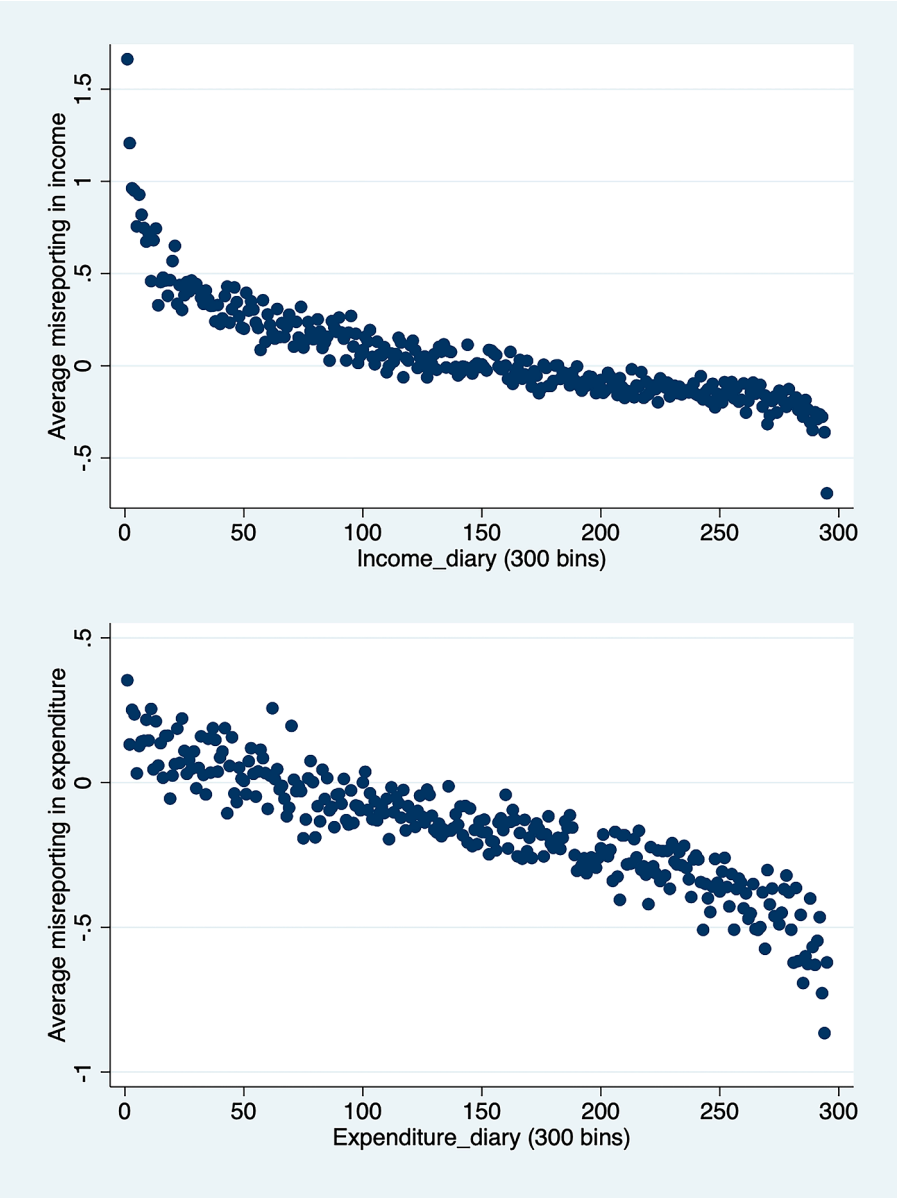

Fig. 1. Average misreporting in household income (expenditure) by income (expenditure) level.
Note: The horizontal axis displays household income (expenditure) from diary data in 300 equally sized bins ranked from the lowest household income (expenditure) to the highest. The vertical axis presents the average misreporting in income (expenditure) for households within each bin.

X_i is a vector of variables which are plausibly correlated with misreporting behavior. County fixed effects are included, and standard errors are clustered at the county level.

3.1. The pattern of mean reversion

Fig. 1 illustrates the observed pattern of mean reversion in both household income and expenditure. Following Hariri and Lassen (2017), we sort observations by household income (expenditure) from diary data, arranging them from the lowest to the highest. Subsequently, we divide them into 300 equally sized bins,[3] which are ranked from the lowest household income (expenditure) to the highest along the horizontal axis in Fig. 1. The average misreporting in household income (expenditure) for each bin is presented on the vertical axis. It is evident from the figure that households with lower income (expenditure) are inclined to overreport household income (expenditure), and vice versa. Furthermore, the lower the household income (expenditure), the more pronounced the overreporting, and conversely.

Table 2 reports regression results. Columns (1) and (2) demonstrate a negative correlation between misreporting in household income (expenditure) and diary-based household income (expenditure), confirming the presence of the mean reversion pattern. Specifically, an increase of 10,000 RMB in household income (expenditure) implies a 4.8 (6.92) log point underreporting in household income (expenditure).

3.2. Subjective well-being and misreporting

The CHIP questionnaire contains a question assessing respondents' subjective well-being, phrased as: "All things considered, do you feel happy?" The response options provided are: very happy, happy, so-so, not very happy, not happy at all, and unsure/ no answer. The few

[3] The pattern of mean version in misreporting can also be observed in graphs using raw data, 200 equally sized bins, or 400 equally sized bins. For more details, please refer to the appendix.

Table 2
Regression results on the impact of factors on misreporting behavior.

Dependent variable:	Mis-reporting income (1)	Mis-reporting expenditure (2)	Mis-reporting income (3)	Mis-reporting expenditure (4)	Mis-reporting income (5)	Mis-reporting expenditure (6)
Household income (diary)	−0.0480***		−0.0493***		−0.0563***	
	(0.0031)		(0.0032)		(0.0035)	
Household expenditure (diary)		−0.0692***		−0.0697***		−0.0865***
		(0.0044)		(0.0045)		(0.0052)
Very happy			0.0741***	0.0496***	0.0788***	0.0460***
(1 = Yes)			(0.0164)	(0.0163)	(0.0159)	(0.0160)
Happy			0.0557***	0.0334***	0.0510***	0.0252**
(1 = Yes)			(0.0100)	(0.0107)	(0.0100)	(0.0109)
Not very happy			−0.0657***	−0.0162	−0.0559**	0.0019
(1 = Yes)			(0.0230)	(0.0216)	(0.0218)	(0.0225)
Not happy at all			−0.0492	−0.0681*	−0.0297	−0.0618
(1 = Yes)			(0.0341)	(0.0386)	(0.0352)	(0.0393)
Rural household					0.0562***	−0.1005***
(1 = Yes)					(0.0183)	(0.0203)
Rural-to-urban migrant household					0.0137	0.0084
(1 = Yes)					(0.0232)	(0.0228)
Family size					0.0748***	0.0664***
					(0.0064)	(0.0055)
Education years of the respondent					0.0142***	0.0131***
					(0.0018)	(0.0020)
Age of the respondent					−0.0000	0.0004
					(0.0004)	(0.0005)
Gender of the respondent					−0.0119	−0.0153
(1 = male; 0 = female)					(0.0091)	(0.0095)
Observations	16,272	16,219	15,683	15,651	15,259	15,225
R-squared	0.2738	0.2399	0.2817	0.2434	0.3171	0.2729

Note:
*** $p < 0.01$,
** $p < 0.05$,
* $p < 0.1$.
County fixed effects are included, and standard errors are clustered at the county level.

who report "unsure/ no answer" are excluded from the analysis. Columns (3) and (4) present the estimates regarding the relationship between respondents' subjective well-being and misreporting in household income and expenditure, using respondents who report "so-so" as the reference category. In comparison to those reporting "so-so", respondents who report "very happy" and "happy" significantly overreport household income (expenditure) by 7.41 (4.96) and 5.57 (3.34) log points, respectively. Conversely, those reporting "not very happy" and "not happy at all" underreport household income (expenditure) by 6.57 (1.62) and 4.92 (6.81) log points. This suggests that respondents' subjective well-being is a significant factor contributing to misreporting.

3.3. Sociodemographic characteristics and misreporting

Columns (5) and (6) illustrate the association between sociodemographic characteristics and misreporting. In contrast to urban households, rural households exhibit a 5.59 log point overreporting in household income and a 10.05 log point underreporting in household expenditure, indicating notable urban-rural disparities in misreporting. Rural-to-urban migrant households show no significant difference from urban households. Moreover, respondents with more years of education or from larger-sized families tend to overreport both household income and expenditure, whereas the age or gender of the respondent does not significantly impact misreporting.

4. Conclusion

Misreporting in recalled household income and expenditure in the 2013 CHIP datasets is substantial and systematic. Two patterns of misreporting are identified: mean reversion and correlation with subjective well-being. More precise measurement of income and expenditure, along with an exploration of patterns such as urban-rural disparities in misreporting, provides opportunities for future research.

Data availability

Data will be made available on request.

Acknowledgement

Feng Li expresses gratitude for the support from his wife, Caiyun Zhu. We acknowledge the financial support from the National Natural Science Foundation of China (Grant #72203056).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econlet.2024.111626.

References

Angel, S., Disslbacher, F., Humer, S., Schnetzer, M., 2019. What did you really earn last year?: explaining measurement error in survey income data. J. R. Stat. Soc. Ser. A Stat. Soc. 182 (4), 1411–1437.
Bound, J., Brown, C., Duncan, G.J., Rodgers, W.L., 1994. Evidence on the validity of cross-sectional and longitudinal labor market data. J. Labor Econ. 12 (3), 345–368.
Bound, J., Krueger, A.B., 1991. The extent of measurement error in longitudinal earnings data: do two wrongs make a right? J. Labor Econ. 9 (1), 1–24.
Brzozowski, M., Crossley, T.F., Winter, J.K., 2017. A comparison of recall and diary food expenditure data. Food Policy 72, 53–61.
Carver, T., Grimes, A., 2019. Income or consumption: which better predicts subjective well-being? Rev. Income Wealth 65, S256–S280.
Gao, Q., Ying, Q., Luo, D., 2015. Hidden income and occupational background: evidence from Guangzhou. J. Contemp. China 24 (94), 721–741.
Hariri, J.G., Lassen, D.D., 2017. Income and outcomes: social desirability bias distorts measurements of the relationship between income and political behavior. Public Opin. Q. 81 (2), 564–576.
Jinkins, D., 2016. Conspicuous consumption in the United States and China. J. Econ. Behav. Organ. 127, 115–132.
Kim, C., Tamborini, C.R., 2014. Response error in earnings: an analysis of the survey of income and program participation matched with administrative data. Sociol. Methods Res. 43 (1), 39–72.

Kreiner, C.T., Lassen, D.D., Leth-Petersen, S., 2013. Measuring the Accuracy of Survey Responses using Administrative Register Data: Evidence from Denmark (No. w19539). National Bureau of Economic Research.
Sicular, T., Li, S., Yue, X., & Sato, H. (2017). Changing trends in China's inequality: key issues and main findings (No. 2017-12). CHCP Working Paper.
Valet, P., Adriaans, J., Liebig, S., 2019. Comparing survey data and administrative records on gross earnings: nonreporting, misreporting, interviewer presence and earnings inequality. Qual. Quant. 53, 471–491.
Wang, X., Woo, W.T., 2011. The size and distribution of hidden household income in China. Asian Econ. Pap. 10 (1), 1–26.
Zhang, J., Zhao, W., 2019. The unreported income and its impact on Gini coefficient in China. J. Chin. Econ. Bus. Stud. 17 (3), 245–259.