Escuela de Gobierno
Documento de Trabajo 2024/03

Money and lies: proxy respondents and the mismeasurement of income in surveys

Autoría: Estrada Ricardo; Goyheix, Daniela; Lombardi, María
Fecha: Julio 2024

UNIVERSIDAD TORCUATO DI TELLA
Escuela de Gobierno

ESCUELA DE GOBIERNO

DOCUMENTOS DE TRABAJO 2024/03

Money and lies

Proxy respondents and the mismeasurement of income in surveys

Ricardo Estrada

Daniela Goyheix

María Lombardi

Julio 2024

Documentos de trabajo: https://bit.ly/2REorES

UTDT: Av. Figueroa Alcorta 7350, C1428BCW Buenos Aires, Argentina

Money and lies: proxy respondents and the mismeasurement of income in surveys*

Ricardo Estrada † Daniela Goyheix ‡ María Lombardi §

July 2024

Abstract

When sampled individuals are not found at home, many surveys rely on a proxy respondent: another knowledgeable household member. We study the difference between self- and proxy-reported labor income in Mexico. We use the panel structure of the Mexican labor force survey and compare workers' income when they report it themselves to their income when another household member does the reporting. We find that the monthly wage of male workers is 6.1% lower when reported by a proxy. For female workers, the reporting gap is minute. We provide evidence that the gap in the reported income of male workers is due to asymmetry of information within the household, in part due to men hiding income from their relatives. Finally, we study the implications of using proxy respondents and find that it can lead to an underestimation of the gender wage gap by 60%.

JEL Classifications: C8, D13, J16, J31

Keywords: survey design, proxy reporting, labor income, household economics, gender wage gap

*We are grateful for comments from Martin Rossi, Pauline Rossi, Adan Silverio Murillo, Joe Vecci, and participants at seminars at CAF, Universidad Torcuato Di Tella, and the Max Weber Conference at the EUI.

†CAF–development bank of Latin America and the Caribbean, e-mail: restrada@caf.com

‡CAF–development bank of Latin America and the Caribbean, e-mail: daniela.goyheix_externo@caf.com

§Universidad Torcuato Di Tella, e-mail: mlombardi@utdt.edu

1 Introduction

When sampled individuals are not found at home, many surveys rely on proxy respondents (other knowledgeable household members). For example, nearly half of interviews in the Current Population Survey (CPS) in the United States are done by proxies (U.S. Census Bureau, 2019). The use of proxy respondents helps to lower costs, but may introduce measurement error in reported variables, such as labor income. If classic, measurement error could lead to less precise estimates. Of greater concern is that measurement error could introduce bias if proxy respondents systematically under(over)-estimate the income of sampled individuals. This could happen, for example, if there is asymmetric information in the household due to individuals' desire to have greater discretionary power over their income (Deschênes et al., 2020). By contrast, the use of proxy respondents could reduce bias if proxies are more likely to report the true income of workers. This could be the case, for example, if individuals are less subject to social desirability bias when reporting other household members' income—as opposed to their own income.

Despite its relevance, there is limited understanding of how the use of proxy respondents affects reported incomes, and the related mechanisms and implications. This research gap is probably due to a methodological challenge, namely that for a given worker, we typically observe either self- or proxy-reported income. This is problematic because workers who report their own income and those who use a proxy respondent likely differ in unobservable characteristics correlated with income, thereby compromising the causal interpretation of a naive comparison between the reported incomes of these two types of workers. In this paper, we overcome this difficulty by using the panel structure of the Mexican labor force survey and the fact that some individuals (around half of our sample) respond to their own questionnaires in some survey waves, but not in others. This variation enables us to compare these workers' income when they report it themselves to their income when another household member does the reporting. Our identification assumption is that for a given worker, having a proxy respondent in a survey wave does not

correlate with changes in actual (unobserved) income. We provide evidence in support of this assumption.

For the analysis, we use microdata from the Mexican labor force survey (ENOE as its Spanish acronym) for the 2005–2019 period. The ENOE has a rotating panel design in which sampled households are followed over five consecutive survey waves. Enumerators apply an employment questionnaire to all household members 15 years and older. If an individual is not at home at the time of the visit, another knowledgeable member of the household acts as a proxy. Crucially, the ENOE dataset identifies the informant.

We perform the analysis separately for male and female workers. We find significant differences between self- and proxy-reporting of income. Proxy respondents are less likely than workers themselves to report the worker's income. Proxies are 12.4 and 16.9 percentage points less likely to report the wages of male and female workers, respectively, with approximately 90% of workers reporting their wages when giving their own interview. When informants report wages, we observe substantial differences between genders in the effect of proxy respondents on reported wages. The monthly wage of men is approximately 6.1% lower when reported by a proxy; for married workers, this difference is larger when the proxy is the worker's spouse as opposed to another household member. By contrast, the difference between the self- and proxy-reported wages of female workers is minute (self-reported wages are 0.4% lower). Importantly, our results are robust to limiting our sample to workers who are less likely to have experienced income fluctuations throughout survey waves (e.g., workers who are only paid a salary or those who remain at the same firm), indicating that our findings stem from reporting discrepancies rather than from changes in actual income.

In terms of the mechanisms, our results indicate that the most likely explanation for the gap in reported income of male workers is an asymmetry of information within the household, in part due to men hiding income from their relatives. We find no support for an overreporting channel in which men overstate their income to enumerators because of

social desirability.

The gender differences in the impact of proxy respondents on reported wages could lead to the mismeasurement of common metrics. To illustrate this, we analyze the case of the gender wage gap, a widely used measure in both economics and other social sciences. To do so, we focus on the sample of workers for whom we observe both self- and proxy-reporting and examine how the gender wage gap varies depending on whether or not we consider income reported by proxies. We find that if proxy responses are excluded, the estimated gender wage gap is 60% larger than the business-as-usual estimate, which relies on both self- and proxy-reported responses. These results indicate that the use of proxy respondents can lead to a substantial underestimation of the gender wage gap.

Our paper is connected to various strands of literature. Our main contribution is to the literature on survey design (Bound et al., 2001; De Weerdt et al., 2020; Dillon et al., 2020) and, more specifically, on the use of proxy respondents and the measurement of income.[1] In developed countries, Mellow and Sider (1983) and Bound and Krueger (1991) compare earnings reported in the CPS to tax records and find that proxy respondents do not significantly contribute to the mismeasurement of labor income in the United States.[2] It is unclear whether these findings generalize to contexts characterized by different gender and social norms, as such norms influence communication both within and outside the household. In developing countries, several papers find that proxy respondents have effects on the measurement of different indicators (labor force participation, child labor, returns to education, assets, bargaining power, agricultural productivity, etc.), using ad hoc surveys with relatively small sample sizes and sparse geographical coverage (Bardasi et al., 2011; Dillon et al., 2012; Serneels et al., 2017; Silverio-Murillo, 2018; Ambler et al.,

[1]Income misreporting is an old concern in survey design and there are multiple papers devoted to estimating, for example, the extent to which income is underreported in surveys compared to in administrative sources such as tax records. Rather than estimating overall misreporting, the purpose of our paper is to understand the marginal effect of using proxy reporting on the measurement of income. In other words, our results can be interpreted as an effect on top of any underlying tendency of, for example, rich individuals to underreport their income.

[2]Tamborini and Kim (2013) performs a similar exercise using a different survey (the 2004 Survey of Income and Program Participation) and finds similar results.

2021; Dervisevic and Goldstein, 2023). Within this literature, Fisher et al. (2010) and Masselus and Fiala (2024) study household income in rural contexts (Malawi in the first case and Paraguay and Uganda in the second). Masselus and Fiala (2024) use a survey experiment to study the effect of proxy respondents on reported income.[3] They find that in Paraguay (Uganda), the income of males is 12% (between 3% and 19%) lower when it is reported by the applicant's spouse compared to when it is self-reported, although this difference is only statistically significant in Paraguay. Unlike in our study, they find that proxy-reporting leads to lower reported income for women in Uganda (no comparable data in Paraguay). Our contribution here is threefold. First, we provide evidence using a nationally representative survey. Second, we study the context of an upper-middle-income country with substantial gender imbalances (Matulevich et al., 2021). In this context, measurement error has important implications, as this survey is the main source of labor market statistics. Third, we provide evidence on the mechanisms driving the results, and their implications.

We also contribute to the literature on gender norms and the generation and sharing of resources in the household. Our paper is closely connected to a line of studies on income hiding between relatives, which is mainly focused on Africa and, to a lesser extent, South Asia. A series of lab-in-the-field studies find that individuals are willing to pay to hide income from their spouses (Ashraf, 2009; Castilla, 2019; Pouliquen, 2023) and relatives (Jakiela and Ozier, 2016; Boltz et al., 2020). In addition, Chen and Collins (2014) and Ziparo (2020) use ad hoc surveys in which spouses are asked about each other's incomes. They find that both men and women tend to underestimate the income reported by their spouses. The authors interpret this as evidence of income hiding. We contribute to this literature by providing evidence of income hiding, using a large, national survey from an upper-middle-income country. Furthermore, we link this body of literature to the

[3]The main outcomes of the paper are household income and food consumption, and the experiment randomizes whether applicants to a loan program answer the survey alone or in the presence of their spouse, or whether both of them answer the survey separately (this last treatment arm is only implemented in Uganda).

literature on survey design by providing evidence on the effects of asymmetric information in the household on the measurement of income in household surveys.

Finally, our paper is connected to the literature on the gender wage gap, which is the subject of enormous attention in the social sciences (Blau and Kahn, 2017; Goldin, 2024). We contribute to this literature by showing how the use of proxy respondents in household surveys can lead to a significant underestimation of the gender wage gap. To the best of our knowledge, we are the first to study this issue.

The rest of the paper is organized as follows. Section 2 presents the data, Section 3 the estimation strategy, and Section 4 the main results and validity checks. Section 5 analyzes the mechanisms and Section 6 presents evidence on the implications for measuring the gender wage gap. Section 7 concludes.

2 Data and descriptive statistics

We use microdata from the National Survey of Occupation and Employment ENOE, which is the main source of information on the Mexican labor market. This survey is conducted by the national statistical office and is representative of the national population. It provides quarterly data on occupation, net labor income, and number of hours worked per week, as well as sociodemographic information on all household members.[4] The survey has a rotating panel design in which sampled households are followed over five consecutive quarters, maintaining 80% of the sample each quarter. While the intention is for individuals to answer their own questionnaire, if an individual is not at home, another knowledgeable household member acts as a proxy. The ENOE dataset enables the identification and characterization of the worker (individual being surveyed), the informant, and the relationship between the two. This feature is crucial for our empirical

[4]The survey is designed to capture net labor income. For salaried workers, enumerators are asked to record the disposable pay after taxes, social security contributions, and any other automatic deductions. For self-employed workers, enumerators must record the net income (revenue minus expenditures) (INEGI, 2009).

strategy because it enables us to leverage the naturally occurring variation in the survey respondent to assess differences in reported income by type of informant across waves.

The data we use spans the 2005–2019 period, comprising a total of 60 survey waves and approximately 120,000 households per quarter. We restrict the analysis to salaried and self-employed workers who are aged 25 to 64 at the time of the survey. We exclude individuals for whom information on the informant is lacking, which accounts for a minimal fraction of cases. In addition, we exclude workers who do not have the same type of position (self-employed or salaried) in every survey wave, observations with zero working hours, and workers who appear in only one period.[5] We winsorize income at the 3rd and 97th percentiles to limit the influence of outliers.

Our sample has a total of 3,880,117 individual interviews conducted on 946,983 workers. Approximately 35% of those interviews are self-reported, while 65% are proxy-reported, as shown in Table 1. Men are more likely to have a proxy informant than women. Slightly over 50% of the workers in our sample have both self- and proxy-reported responses across survey waves.

Table 2 reports the descriptive statistics of our main sample, separated by gender. Female workers are less likely to be head of household or be cohabiting with a partner, are more educated, have a lower monthly wage, and work fewer hours per week compared to their male counterparts. These findings underscore the existence of gender-based disparities in both socioeconomic and labor market dimensions within our study sample.

[5]There are 2,801,855 individuals aged 25–64 in the 60 survey waves of our sample. We exclude 647,713 individuals in this age bracket who are unemployed, employers, or unpaid family workers. We then drop 945,664 individuals who do not have the same occupation status in all survey waves, 183,560 observations with non-positive hours of work, and 248,613 individuals who only appear in one survey wave. Finally, we drop 18 individuals for whom the informant is missing. Our final sample is composed of 946,983 workers, for whom we have 3,880,117 interviews.

3 Estimation strategy

For a given worker, we usually observe either self- or proxy-reported income in national surveys. Since workers who report their own income and those who use a proxy respondent may differ in unobservable characteristics correlated with income, identifying the effect of using a proxy respondent is challenging. We overcome this difficulty by taking advantage of the panel structure of the ENOE survey. As shown in Table 1, there is variation across survey waves in the type of respondent for 51% of the workers in our sample.[6] This variation enables us to observe the income of the same person when they report it themselves compared to when another household member does the reporting. We estimate the following equation:

$$Y_{it} = \beta_1 Proxy\ informant_{it} + \gamma_i + \alpha_t + U_{it} \tag{1}$$

Y_{it} is an outcome for worker i in period t. Our main outcomes are a dummy for whether worker i's income is reported in the survey conducted in period t, the worker's monthly and hourly wages (in MEX$ of 2019, and in ln), and the weekly hours worked. As the effect of proxy-reporting could vary by the gender of the worker, we estimate this equation separately for male and female workers. Our main regressor, $Proxy\ informant_{it}$, is a dummy variable that takes a value of 1 if the informant for worker i in survey period t is another household member, and 0 if the worker self-reports. Finally, γ_i and α_t are individual and survey-wave fixed effects, and U_{it} is the error term. We cluster our standard errors at the individual level.

For our coefficient of interest ($\hat{\beta}_1$) to capture differences in *reported* income by respondent type, having a proxy respondent in a given wave should not correlate with differences in *actual* (unobserved) income. Our individual fixed effects control for any time-invariant differences in employment conditions between workers who are present at the time of the

[6] Appendix Table A.1 compares the characteristics of these workers with those of workers who always self-report or always rely on proxy-reporting.

survey and workers who are not. However, in principle, there could be individual-specific determinants of income that vary across time and correlate with the type of respondent. This concern is partially mitigated by the fact that the workers in our sample are observed over a relatively short period (five consecutive quarters) and our sample is restricted to individuals who are employed in the same type of position (self-employed or salaried) in every period. Nevertheless, we provide evidence in support of this assumption in Section 4.1.

4 Results

Overall, proxy respondents are less likely to report the worker's income than workers themselves, as shown in columns 1 and 5 of Table 3. Male and female workers report their own wages in 89% and 91% of interviews, respectively, whereas proxies are 12.4 and 16.9 percentage points less likely to provide a response. Both coefficients are statistically significant at the 1% level.

When the (net) monthly income of workers is reported (in 80% of cases), there are statistically significant differences between the self- and proxy-reported incomes of male workers. The monthly wage reported for male workers is 6.1% lower when reported by a proxy. While proxies report higher wages for female workers than what they self-report, this difference is very small (0.4%). Importantly, our results are very similar if we use nominal instead of real wages (Appendix Table A.2).

We find discrepancies in the reported weekly hours worked for both genders, and these differences are statistically significant at the 1% level. For male and female workers, respectively, proxies report 0.82 and 0.70 more hours than the workers themselves. These differences in reported hours worked imply that the gap in reported hourly wages for male workers is even larger than the gap in monthly wages (8.4% vs. 6.1%). Furthermore, there is a smaller but statistically significant gap in reported hourly wages for female workers,

with proxies reporting hourly wages that are 2.4% lower than those reported by the worker herself.

4.1 Validity checks

As discussed in Section 3, the causal interpretation of our estimates requires that the type of respondent in a given survey wave not correlate with other time-varying determinants of income. The fact that we only find differences in reported monthly wages for male workers mitigates this concern, as it is unclear why only men should experience time-varying shocks in income that correlate with being absent at the time of the survey. Nonetheless, we perform several robustness checks to examine the validity of our identifying assumption. In particular, we show that our results are robust to limiting our sample to workers who are less likely to have experienced fluctuations in income throughout the survey waves. The main results from these validity checks are summarized in Figure 1.

As shown in Table A.3, we find very similar results if we limit our sample to full-time workers (30 or more weekly hours of work), a group for which wages should be relatively stable over short periods of time. In particular, we find that the reported monthly wage of full-time male workers is 6.5% lower when a proxy is reporting than when the worker himself is (vs. 6.1% in the original sample).[7] Our full sample includes self-employed workers, and one could be concerned that the income of self-employed workers fluctuates in ways that correlate with the worker's presence in the home. Importantly, as shown in Appendix Tables A.4 and A.5, we find a very similar gap in male workers' reported income if we limit our sample to salaried workers and, more restrictively, salaried workers who are paid only a salary (as opposed to commissions, performance bonuses, etc.). We also obtain similar results if we eliminate interviews conducted in December, the month in which salaried workers receive their 13^{th} salary (Appendix Table A.6), and if we eliminate

[7]The differences in hours worked by type of respondent for female workers found in Table 3 are mostly driven by part-time workers, as the differences are much smaller once we focus on full-time workers.

observations in which individuals are not working their usual hours (Appendix Table A.7).

The basic questionnaire, applied in all survey waves, does not have information on job tenure. However, interviews conducted in the first quarter of the year and in some other periods use an extended questionnaire, which asks workers about their tenure in the company in which they are currently employed.[8] We conduct a robustness check restricting our sample to workers who remain employed at the same firm from the quarter in which their household is initially surveyed to their final interview with the extended questionnaire. We drop interviews conducted after the last interview with the extended questionnaire, as we do not know whether workers changed firms after this point. Our main estimates are almost unchanged if we focus on this sample (see Table A.8 in the Appendix).

A potential concern is that proxies systematically act as respondents in earlier interviews (each worker is interviewed in five survey waves), and our coefficient of interest thus captures the natural increase in wages that occurs over time for workers who progress in their careers. Mitigating this concern, Appendix Figure A.1 shows that the estimated gap between proxy- and self-reported wages is unaffected by the interview number in which the proxy acts as an informant.

Overall, our results indicate that the reported income of male workers is substantially lower when using a proxy respondent, and that these differences stem from reporting discrepancies rather than from changes in actual income.

[8]Besides the first quarter of every year, the extended questionnaire was also applied in all survey waves of 2005, and in the second quarters of 2006, 2007 and 2008.

4.2 Heterogeneous effects by civil status and relationship with the informant

The effect of proxy respondents could vary with the relationship between the worker and the informant.[9] Appendix Table A.9 shows the relationship between the worker and his/her proxy, distinguishing between married and single workers.[10] As shown in Panel A, the proxy informant for 76% of the married workers in our sample is their spouse; their children serve as proxies in 17% of cases, and other household members in the remaining 7% of cases. For married male workers, the likelihood of having a spouse as their proxy is higher than for female workers (80% vs. 61%). For single workers, proxies are most likely to be their children (42%), followed by their parents and their siblings (23% and 18%, respectively); in the remaining 17% of cases, the proxy is another household member.

We estimate the effect of proxy respondents separately for married and single workers.[11] For married workers, we distinguish between proxy respondents who are the worker's spouse or another household member, motivated by the literature on household economics that documents asymmetric information between spouses with respect to each other's income (Deschênes et al., 2020). Therefore, we expand equation (1) and estimate the following equation in the case of married workers:

$$Y_{it} = \beta_1 Spouse\ \ informant_{it} + \beta_2 Other\ informant_{it} + \gamma_i + \alpha_t + U_{it} \tag{2}$$

, where $Spouse\ informant_{it}$ and $Other\ informant_{it}$ are dummy variables for whether the informant for worker i in survey period t is his/her spouse or another household mem-

[9] We identify cohabiting couples using data on the relationship with the household head from individuals who report being married or in a union. We could not unambiguously identify the spouse of only 1% of the married/in union workers in our sample.

[10] We are able to identify the relationship between the worker and his/her proxy in 98% of cases by looking at the relationship with the head of household of both the worker and the proxy.

[11] For ease of exposition, we use the term married to refer to individuals who are married or in a union and cohabit with their spouse. Similarly, we use the term single to refer to workers who are single or who report being married or in a union but not cohabiting.

ber, respectively (the omitted category comprises interviews in which the worker is the informant).

Our results for married workers are presented in Panel A of Table 4, and those for single workers in Panel B.[12] As shown in columns 1 and 5 of Panel A, the gap in the likelihood of reporting the worker's income is much smaller when the proxy is the worker's partner as opposed to another household member. In particular, the partners of male and female workers are 7.7 and 9.2 percentage points less likely than the worker to provide a response, respectively. For other household members, the difference in the likelihood of reporting is larger (more than 20 percentage points for both genders).

As can be seen in column 2 of Panel A, we find a larger gap between the proxy- and self-reported income of married male workers when the informant is the spouse (6.4%) as opposed to another household member (2.2%). All of these coefficients are statistically significant at the 1% level. This gap between self- and partner-reported income is smaller than in Masselus and Fiala (2024) (12% in Uganda and between 3% and 19% in Paraguay), possibly because we study this question in a context with less gender imbalance—see Section 5 below. The monthly wage self-reported by female workers is 0.4% lower than that reported by their spouse, and 3.4% lower than that reported by other family members.

For single male workers (Panel B), we also find a large and statistically significant gap between proxy- and self-reported monthly income (7.9%). In the case of female workers who are single, the gap between self- and proxy-reported income is statistically significant, but very small (0.9%).[13]

[12]We drop 29,760 observations (0.77% of the sample) corresponding to married/in union workers for whom we could not ascertain whether or not they cohabit with their spouse.

[13]In Appendix Table A.10, we show the results of regressions in which we allow the effect of proxy respondents to differ according to the relationship between the worker and the proxy.

5 Why is there a gap between self- and proxy-reported income for male workers?

The observed discrepancies in reported income by respondent type could arise from asymmetric information within the household (Deschênes et al., 2020). This asymmetry can arise if workers conceal their income to have more discretionary power over their spending, or if proxies infer other household members' wages based on consumption that is imperfectly observed. An alternative explanation for our results is that male workers overstate their income to enumerators. In this section, we examine which of these mechanisms is at play.

We test whether workers intentionally conceal their income from family members by analyzing if self-reported income varies with the presence of other family members at the time of the survey. For single workers, we add to our estimating equation a dummy variable for whether the worker self-reports his/her income and at least one other household member is present at the time of the survey; the comparison group is workers who self-report when nobody else is around. We consider other household members to be absent if a proxy gives the interview; accordingly, we have an upper bound of actual presence in the interview, as others might be present in the dwelling, but not necessarily in the room where the interview takes place. For married workers, we distinguish between the presence of their spouse and other household members, as several studies in other contexts have documented income hiding between spouses (Ashraf, 2009; Castilla, 2019; Pouliquen, 2023). We thus add to our estimating equation a dummy variable for whether the worker self-reports and his/her spouse is present at the time of the survey, and a dummy for whether the worker self-reports and another household member (but not his/her spouse) is present at the time of the survey. The results of these estimations are presented in Table 5.

As shown in the estimations for married workers in Panel A, the self-reported income

of male workers is 3.0% lower when their spouse is present than when they self-report with nobody around, and this difference is statistically significant at the 1% level.[14] When other household members are present, there is no difference in the self-reported income of married men. In the case of single male workers, their self-reported income is 3.3% lower when others are present compared to when the worker answers the questionnaire with nobody else around. More than half of the single male workers in our sample live with their parents and could possibly have incentive to hide their income so as to retain discretion over their spending. For female workers, income self-reported in solitude is only 0.5% lower than income self-reported in the presence of their spouse; this difference is statistically significant only at the 10% level.[15] These results are consistent with male workers hiding income from other household members. However, the income reported by proxies is lower than the income self-reported by workers in the presence of others. This might be due to our measure of the presence of others, which probably overestimates actual presence. Furthermore, income hiding might not be the only explanation for our findings. Proxy informants may also underreport the income of male workers if they systematically underestimate it, regardless of whether or not the worker has the intention of hiding his/her income.

Differences in bargaining power may lead to higher income hiding and asymmetric information within couples (Doss, 2013). We examine whether the difference between the self- and proxy-reported income of married workers varies with bargaining power in the couple, which we proxy using differences in educational attainment. Figure 2 reports the main coefficients of regressions for the sample of married workers in which we fully interact equation (2) with dummies for whether the respondent has less, equal, or more education than his/her spouse, and use real monthly wage (in ln) as the dependent vari-

[14] Masselus and Fiala (2024) find that the reported income of male workers is 11–14% higher when their spouse is not present (although these differences are not statistically significant); this could explain why the gap in self- and partner-reported income for male workers is higher in their study.

[15] We find that male and female workers are less likely to report their income when their spouse or other household members are present, as shown in columns 1 and 5, another manifestation of income hiding in the household.

able. We report the coefficients of the interaction effects for the dummy of the spouse informant in the top graph, and the interactions with the dummy for whether the informant is another household member in the bottom graph. The estimates of regressions in which the worker is male (female) are presented in blue (orange). We find that the gap between self- and spouse-reported income for male workers is significantly larger when the worker has more education than his spouse. These results are compatible with income hiding in couples with power asymmetry.

Research shows that measurement errors in earnings tend to be mean-reverting (Bound and Krueger, 1991; Bound et al., 1994; Angel et al., 2019; Flachaire et al., 2023), with low-income individuals overstating their wages and high-income individuals understating them. If our findings were driven by overreporting of income by male workers, the discrepancy in reported income would only arise in the case of low-income workers. We explore this hypothesis by testing for heterogeneous effects across workers' income and present our results in Figure 3.[16]

We find that the gap in reported income increases with male workers' income, which is incompatible with the explanation that our findings are due to male workers overstating their income. We also find a gap in reported income among high-earning female workers (those in the top two quintiles), although it is smaller than among their male counterparts. These findings are consistent with the idea that asymmetry of information might increase with income level, because there is more room for discretionary spending, which is difficult for other household members to observe (Dynan et al., 2004). Furthermore, these findings suggest that the average gap in reported income among male workers might be due to differences in both income level with respect to female workers and preferences for higher discretionary spending.

Another hypothesis is that men overstate their wages to conform to societal expectations of being the primary breadwinner (Bertrand et al., 2015; Slotwinski and Roth, 2020).

[16]We use workers' average income throughout all survey waves to compute their position in the income distribution.

If this were the case, the gap in reported income would be larger for men whose spouse is employed. We explore this hypothesis by testing for heterogeneous effects across the employment status of spouses in the subsample of married workers. Figure 4 reports the main coefficients of regressions in which we fully interact equation (2) with dummies for whether the worker's partner is employed or unemployed/inactive, and use real monthly wage (in ln) as the dependent variable. As shown in the top graph, the difference between the self-reported income of male workers and their income reported by their spouse is larger when the worker's spouse is not employed. In particular, the gap in reported income is 6.9% for male workers with an unemployed/inactive spouse and 6.1% with a spouse who is employed; the difference between both coefficients is statistically significant at the 1% level. These results indicate that our findings are not due to men overstating their income to comply with breadwinner norms; men with unemployed or inactive spouses already conform to these norms and thus would have no incentive to lie about their income.

In sum, our results indicate that the most likely explanation for the gap in reported income of male workers is an asymmetry of information within the household, in part due to men hiding their income from other household members.

6 Implications for measuring the gender wage gap

The results presented in Section 4 show that the income of male workers is substantially lower when reported by a proxy, while the gap between the self- and proxy-reported income of female workers is small. Therefore, the widespread use of proxy respondents could lead to the mismeasurement of common metrics. In this section, we focus on the effect of proxy-reporting on the gender wage gap.

We estimate the gender wage gap using the following equation:

$$ln(Income_{it}) = \beta_1 Female_{it} + X_{it}\rho + \gamma_t + U_{it} \quad (3)$$

$Income_{it}$ is the hourly wage of worker i in quarter t and $Female_i$ is a dummy for whether worker i is female (0 if male). X_{it} is a vector of covariates that controls for age, educational attainment, urban residence, and state of residence, and γ_t captures survey-wave fixed effects. β_1 is the parameter of interest and captures the average difference between wages earned by female and male workers conditional on the characteristics included in X_{it}.

We estimate equation (3) in the sample of full-time workers (with at least 30 hours of work per week) for which we observe both self- and proxy-reported income and examine how the gender wage gap (β_1) varies depending on whether or not the income reported by proxies is considered. Standard errors are clustered at the individual level. Table 6 reports the results.

The estimated gender wage gap when using both self- and proxy-reporting is minus 6.8% (column 1), whereas using only self-reported income leads to an estimated gender wage gap of minus 10.9% (column 2). Not surprisingly, restricting the analysis to proxy respondents produces a smaller estimate of the gender wage gap—of minus 3.8% (all results significant at the 1% level). Therefore, if proxy respondents are excluded, the estimated gender wage gap is 60% larger than the business-as-usual estimate, which uses both self- and proxy-reported responses. These results indicate that the use of proxy respondents can lead to a substantial underestimation of the true gender wage gap.

7 Conclusions

Our study brings new evidence on the impact of proxy respondents on reported incomes, and related mechanisms and implications. Using the panel structure of the Mexican labor

force survey, we demonstrate substantial differences between self-reporting and proxy-reporting of income, with proxy respondents reporting lower wages for male workers, but not for female workers. Our analysis of mechanisms reveals that this gap is due to asymmetric information within households and a tendency for men to hide their income from other household members. Furthermore, we show that the gender differences between self- and proxy-reporting of labor income can lead to a significant underestimation of the gender wage gap.

The use of proxy respondents helps to lower costs, but, as we show, it can introduce measurement error in key variables such as labor income, particularly in contexts where social norms affect the flow of information within the household. Researchers can benefit from assessing how the use of proxy respondents affects their results, particularly if gender plays a role in their analysis. To this end, it is necessary for statistical offices to include in their survey datasets information about the identity of the respondents. For this article, we assessed 10 national surveys carried out by national statistical offices from Latin American countries. It was possible to know if a proxy gave the interview in only five out of 10 of the surveys (in Ecuador, Mexico, Paraguay, Peru, and Uruguay, but not in Argentina, Bolivia, Brazil, Chile, and Costa Rica). Overall, our research underscores the importance of understanding the role of proxy respondents in survey data collection to ensure the accuracy and reliability of socioeconomic indicators.

References

Ambler, Kate, Cheryl Doss, Caitlin Kieran, and Simone Passarelli (2021) "He says, she says: Spousal disagreement in survey measures of bargaining power," *Economic Development and Cultural Change*, 69 (2), 765–788.

Angel, Stefan, Franziska Disslbacher, Stefan Humer, and Matthias Schnetzer (2019) "What did you really earn last year?: explaining measurement error in survey income data," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182 (4), 1411–1437.

Ashraf, Nava (2009) "Spousal control and intra-household decision making: An experimental study in the Philippines," *American Economic Review*, 99 (4), 1245–1277.

Bardasi, Elena, Kathleen Beegle, Andrew Dillon, and Pieter Serneels (2011) "Do labor statistics depend on how and to whom the questions are asked? Results from a survey experiment in Tanzania," *The World Bank Economic Review*, 25 (3), 418–447.

Bertrand, Marianne, Emir Kamenica, and Jessica Pan (2015) "Gender identity and relative income within households," *Quarterly Journal of Economics*, 130 (2), 571–614.

Blau, Francine D and Lawrence M Kahn (2017) "The gender wage gap: Extent, trends, and explanations," *Journal of Economic Literature*, 55 (3), 789–865.

Boltz, Marie, Karine Marazyan, and Paola Villar (2020) "Is informal redistribution costly? Evidence from a lab-in-the-field experiment in Senegal," *The World Bank Economic Review*, 34 (Supplement_1), S72–S78.

Bound, John, Charles Brown, Greg J Duncan, and Willard L Rodgers (1994) "Evidence on the validity of cross-sectional and longitudinal labor market data," *Journal of Labor Economics*, 12 (3), 345–368.

Bound, John, Charles Brown, and Nancy Mathiowetz (2001) "Measurement error in survey data," in *Handbook of econometrics*, 5, 3705–3843: Elsevier.

Bound, John and Alan B Krueger (1991) "The extent of measurement error in longitudinal earnings data: Do two wrongs make a right?" *Journal of Labor Economics*, 9 (1), 1–24.

Castilla, Carolina (2019) "What's yours is mine, and what's mine is mine: Field experiment on income concealing between spouses in India," *Journal of Development Economics*, 137, 125–140.

Chen, Joyce J and LaPorchia A Collins (2014) "Let's talk about the money: Spousal communication, expenditures, and farm production," *American Journal of Agricultural Economics*, 96 (5), 1272–1290.

De Weerdt, Joachim, John Gibson, and Kathleen Beegle (2020) "What can we learn from experimenting with survey methods?" *Annual Review of Resource Economics*, 12, 431–447.

Dervisevic, Ervin and Markus Goldstein (2023) "He said, she said: The impact of gender and marriage perceptions on self and proxy reporting of labor," *Journal of Development Economics*, 161, 103028.

Deschênes, Sarah, Christelle Dumas, and Sylvie Lambert (2020) "Household resources and individual strategies," *World Development*, 135, 105075.

Dillon, Andrew, Elena Bardasi, Kathleen Beegle, and Pieter Serneels (2012) "Explaining variation in child labor statistics," *Journal of Development Economics*, 98 (1), 136–147.

Dillon, Andrew, Dean Karlan, Christopher Udry, and Jonathan Zinman (2020) "Good identification, meet good data," *World Development*, 127, 104796.

Doss, Cheryl (2013) "Intrahousehold Bargaining and Resource Allocation in Developing Countries," *The World Bank Research Observer*, 28 (1), 52–78, http://www.jstor.org/stable/24582372.

Dynan, Karen E, Jonathan Skinner, and Stephen P Zeldes (2004) "Do the rich save more?" *Journal of Political Economy*, 112 (2), 397–444.

Fisher, Monica, Jeffrey J Reimer, and Edward R Carr (2010) "Who should be interviewed in surveys of household income?" *World Development*, 38 (7), 966–973.

Flachaire, Emmanuel, Nora Lustig, and Andrea Vigorito (2023) "Underreporting of top incomes and inequality: A comparison of correction methods using simulations and linked survey and tax data," *Review of Income and Wealth*, 69 (4), 1033–1059.

Goldin, Claudia (2024) "Nobel Prize Lecture," https://www.nobelprize.org/prizes/economic-sciences/2023/goldin/lecture/.

INEGI (2009) *Manual del Entrevistador de la ENOE*: Instituto Nacional de Estadistica, Geografia e Informatica.

Jakiela, Pamela and Owen Ozier (2016) "Does Africa need a rotten kin theorem? Experimental evidence from village economies," *The Review of Economic Studies*, 83 (1), 231–268.

Masselus, Lise and Nathan Fiala (2024) "Whom to ask? Testing respondent effects in household surveys," *Journal of Development Economics*, 103265.

Matulevich, Rubiano, Eliana Carolina Galeano, and Diana Maria servian (2021) "Mexico - Country Gender Scorecard," Country Gender Scorecards FY22 Washington, D.C.: World Bank Group, available at http://documents.worldbank.org/curated/en/159831645690741028/Mexico-Country-Gender-Scorecard.

Mellow, Wesley and Hal Sider (1983) "Accuracy of response in labor market surveys: Evidence and implications," *Journal of Labor Economics*, 1 (4), 331–344.

Pouliquen, Victor (2023) "This Business is Mine! Individual Property Rights and Intrahousehold Relationships," Working paper.

Serneels, Pieter, Kathleen Beegle, and Andrew Dillon (2017) "Do returns to education depend on how and whom you ask?" *Economics of Education Review*, 60, 5–19.

Silverio-Murillo, Adan (2018) "Wife vs. Husband: Does It Matter Who Answers the Survey?", Working Paper.

Slotwinski, Michaela and Anja Roth (2020) "Gender norms and income misreporting within households," *ZEW-Centre for European Economic Research Discussion Paper* (20-001).

Tamborini, Christopher R and ChangHwan Kim (2013) "Are proxy interviews associated with biased earnings reports? Marital status and gender effects of proxy," *Social science research*, 42 (2), 499–512.

U.S. Census Bureau (2019) "Current Population Survey Design and Methodology," Technical Paper 77.

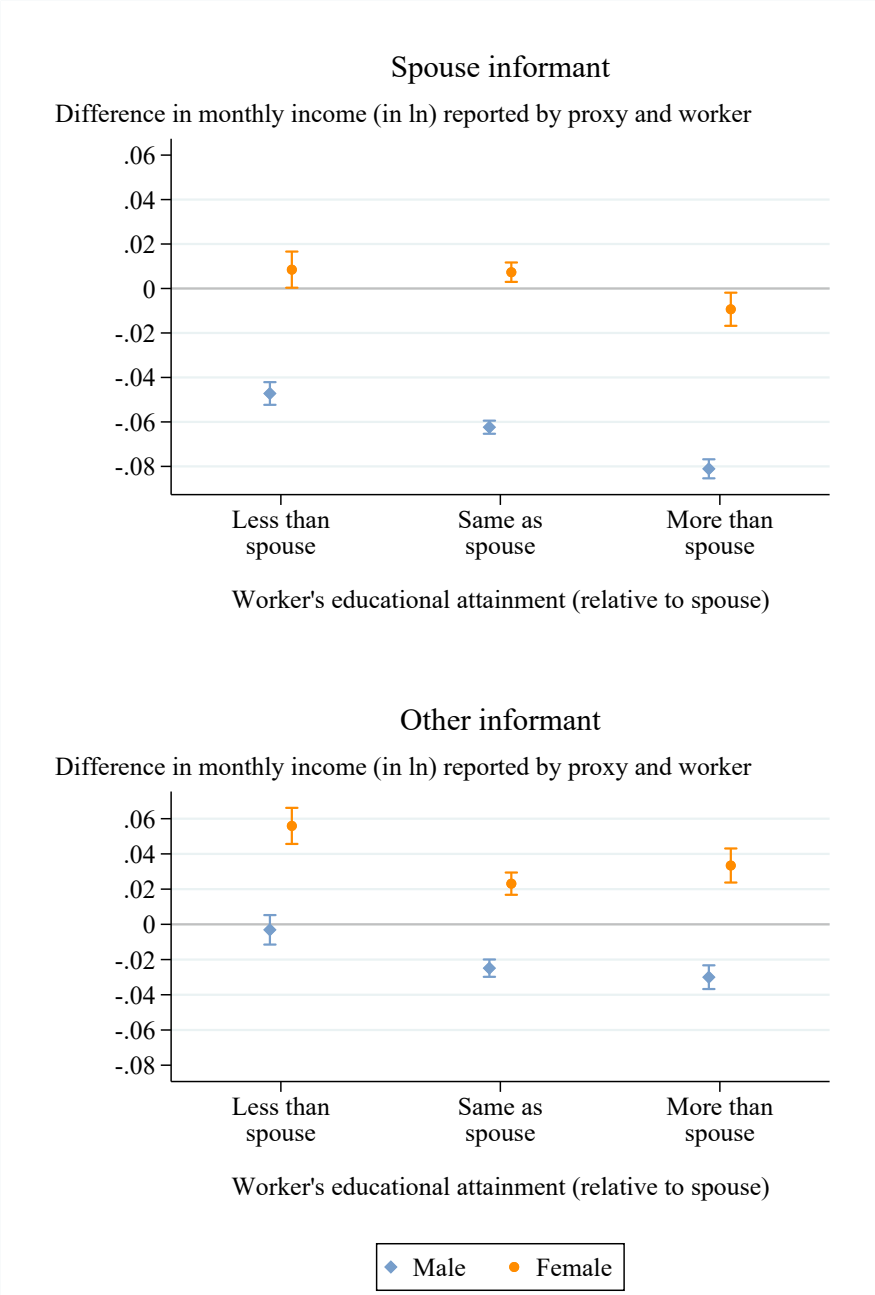
Ziparo, Roberta (2020) "Why do spouses communicate: love or interest? A model and some evidence from Cameroon," Working paper.

Figures and Tables

Figure 1: Differences in reported income by type of respondent – summary of validity checks

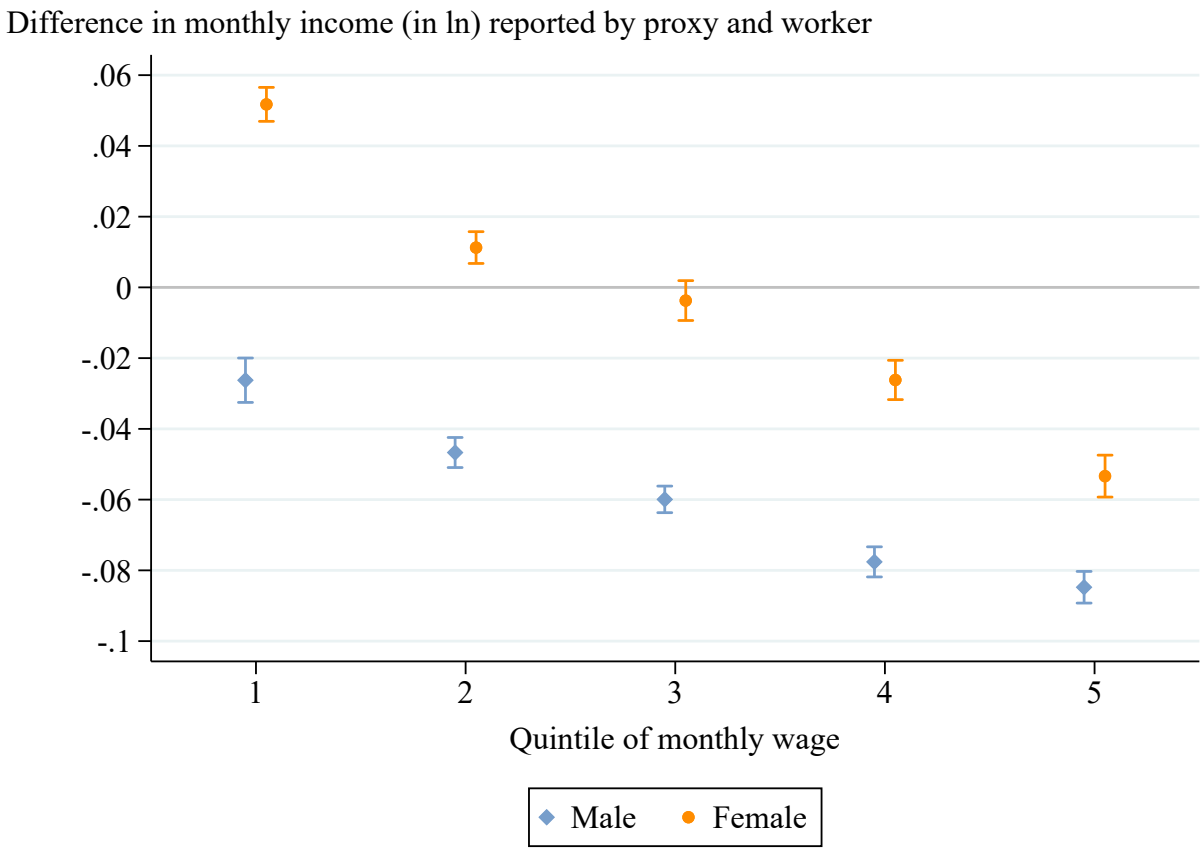
Note: This figure presents the main coefficients and 95% confidence intervals for the estimation of equation (1) for different subsamples. The dependent variable is the monthly wage (in MEX$ of 2019 and in ln). The estimates of regressions in which the worker is male (female) are presented in blue (orange). The original sample includes all salaried and self-employed workers aged 25–64; we exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. The sample of full-time workers is further limited to full-time workers (30 or more hours worked per week). The samples of salaried workers and salaried workers who only receive a salary are limited to workers who have a salaried position in all periods and workers who have a salaried position for which they only receive a salary in all periods, respectively. The sample excluding December excludes observations from interviews conducted in December. The final sample is limited to workers who stay at the same firm throughout survey waves.

Figure 2: Differences in reported income by type of respondent – heterogeneous effects by worker's and spouse's relative education



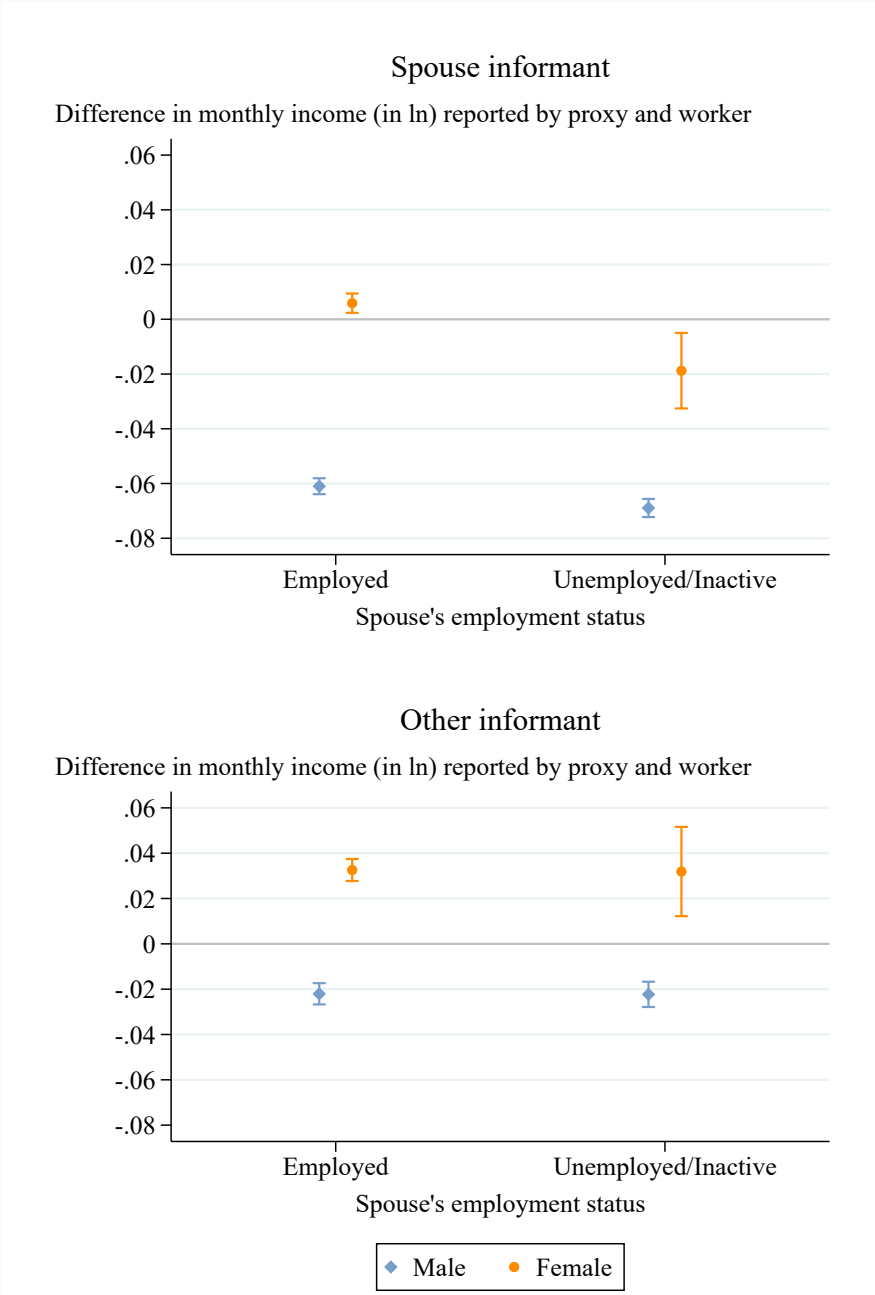
Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave or who do not cohabit with a partner. This figure presents the main coefficients and 95% confidence intervals for the estimation of equation (2) fully interacted with dummies for whether the worker has less, the same, or more education than his/her partner. The dependent variable is the monthly wage (in MEX$ of 2019 and in ln). We report the coefficients of the interaction effects for the dummy of the partner informant in the top graph, and the interactions with the dummy for whether the informant is another household member in the bottom graph. The estimates of regressions in which the worker is male (female) are presented in blue (orange).

Figure 3: Differences in reported income by type of respondent – heterogeneous effects by workers' income



Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. This figure presents the main coefficients and 95% confidence intervals for the estimation of equation (1) fully interacted with income quintile dummies (using workers' average income throughout all survey waves). The dependent variable is the monthly wage (in MEX$ of 2019 and in ln). The estimates of regressions in which the worker is male (female) are presented in blue (orange).

Figure 4: Differences in reported income by type of respondent – heterogeneous effects by spouse's employment status



Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave or who do not cohabit with a partner. This figure presents the main coefficients and 95% confidence intervals for the estimation of equation (2) fully interacted with dummies for whether the worker's partner is employed or unemployed/inactive. The dependent variable is the monthly wage (in MEX$ of 2019 and in ln). We report the coefficients of the interaction effects for the dummy of the partner informant in the top graph, and the interactions with the dummy for whether the informant is another household member in the bottom graph. The estimates of regressions in which the worker is male (female) are presented in blue (orange).

Table 1: Type of informant

	All	Male	Female
Panel A: Observations			
Self-reporting	0.352	0.281	0.465
Proxy respondent	0.648	0.719	0.535
Observations	3,880,117	2,388,044	1,492,073
Panel B: Individuals			
Always self-reporting	0.136	0.090	0.210
Always proxy-reporting	0.352	0.418	0.247
Both self- and proxy-reporting	0.512	0.492	0.544
Observations	946,983	582,045	364,938

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. Panel A shows the share of observations by informant type, and Panel B shows the type of informant for the individuals in our sample.

Table 2: Descriptive statistics

	Male				Female			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Demographics								
Age	40.416	9.976	25.000	64.000	40.084	9.454	25.000	64.000
Head of household	0.757	0.429	0.000	1.000	0.290	0.454	0.000	1.000
Cohabits with spouse	0.766	0.423	0.000	1.000	0.469	0.499	0.000	1.000
Educational attainment								
Incomplete primary	0.102	0.303	0.000	1.000	0.077	0.266	0.000	1.000
Complete primary	0.172	0.377	0.000	1.000	0.143	0.350	0.000	1.000
Complete lower secondary	0.335	0.472	0.000	1.000	0.333	0.471	0.000	1.000
Complete upper secondary or more	0.391	0.488	0.000	1.000	0.446	0.497	0.000	1.000
Labor market outcomes								
Salaried worker	0.869	0.338	0.000	1.000	0.879	0.326	0.000	1.000
Self-employed	0.131	0.338	0.000	1.000	0.121	0.326	0.000	1.000
Hours worked per week	48.596	14.674	1.000	168.000	40.874	13.732	1.000	168.000
Monthly wage (in MEX$ of 2019)	8377.563	5103.132	737.130	26373.455	7227.166	4828.129	737.130	26373.455
Hourly wage (in MEX$ of 2019)	47.786	41.211	1.734	5187.660	50.662	44.442	1.559	3750.000
Job characteristics								
Formal employment	0.667	0.471	0.000	1.000	0.679	0.467	0.000	1.000
Primary sector	0.089	0.285	0.000	1.000	0.008	0.088	0.000	1.000
Secondary sector	0.300	0.458	0.000	1.000	0.156	0.363	0.000	1.000
Tertiary sector	0.609	0.488	0.000	1.000	0.835	0.371	0.000	1.000
Observations	2,388,044				1,492,073			
Individuals	582,045				364,938			

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. We split the sample by gender.

Table 3: Differences in reported income by type of respondent

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Proxy respondent	-0.124***	-0.061***	-0.084***	0.816***	-0.169***	0.004***	-0.024***	0.700***
	(0.001)	(0.001)	(0.001)	(0.025)	(0.001)	(0.001)	(0.001)	(0.025)
Observations	2,388,044	1,903,912	1,903,912	2,388,044	1,492,073	1,200,605	1,200,605	1,492,073
Individuals	582,045	555,636	555,636	582,045	364,938	350,829	350,829	364,938
R^2	0.505	0.785	0.767	0.632	0.496	0.817	0.789	0.709
Dep. var. mean (self)	0.885	8.910	3.740	46.785	0.913	8.632	3.678	39.479
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressor is a dummy variable taking a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: Differences in reported income by type of respondent – by marital status

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Panel A: Married workers								
Spouse informant	-0.077***	-0.064***	-0.087***	0.872***	-0.092***	0.004**	-0.014***	0.403***
	(0.001)	(0.001)	(0.001)	(0.029)	(0.001)	(0.002)	(0.002)	(0.036)
Other informant	-0.238***	-0.022***	-0.046***	0.695***	-0.220***	0.033***	-0.011***	1.050***
	(0.001)	(0.002)	(0.002)	(0.042)	(0.002)	(0.002)	(0.003)	(0.047)
Observations	1,845,093	1,502,110	1,502,110	1,845,093	722,607	605,641	605,641	722,607
Individuals	442,789	425,332	425,332	442,789	175,951	171,154	171,154	175,951
R^2	0.509	0.788	0.769	0.630	0.481	0.823	0.789	0.726
Dep. var. mean (self)	0.883	8.921	3.745	46.929	0.917	8.637	3.718	38.450
P-value (spouse=other)	0.000	0.000	0.000	0.000	0.000	0.000	0.216	0.000
Panel B: Single workers								
Proxy respondent	-0.224***	-0.079***	-0.099***	0.675***	-0.211***	-0.009***	-0.038***	0.792***
	(0.002)	(0.003)	(0.003)	(0.064)	(0.002)	(0.002)	(0.002)	(0.039)
Observations	515,581	381,302	381,302	515,581	760,429	588,504	588,504	760,429
Individuals	132,197	123,595	123,595	132,197	186,713	177,558	177,558	186,713
R^2	0.517	0.776	0.763	0.636	0.508	0.810	0.785	0.686
Dep. var. mean (self)	0.889	8.880	3.727	46.354	0.908	8.627	3.632	40.638
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. In Panel A, the sample is restricted to workers who are married or in a union and cohabit with their spouse in all periods. In Panel B, the sample is restricted to single workers or married/in union workers who do not cohabit with their spouse. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressors in Panel A are dummy variables for whether the informant is the worker's spouse or another household member (the omitted category is self-reporting by the worker). The main regressor in Panel B is a dummy variable that takes a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: Differences in reported income by type of respondent – by marital status and presence of others when self-reporting

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Panel A: Married workers								
Self-reporting and spouse is present	-0.029***	-0.030***	-0.028***	0.098*	-0.019***	-0.005*	-0.006*	0.076
	(0.001)	(0.002)	(0.002)	(0.060)	(0.002)	(0.003)	(0.004)	(0.072)
Self-reporting and others are present (not spouse)	-0.052***	0.006	0.000	0.362**	-0.041***	0.001	0.004	-0.100
	(0.005)	(0.006)	(0.007)	(0.168)	(0.004)	(0.006)	(0.007)	(0.140)
Spouse informant	-0.083***	-0.070***	-0.092***	0.900***	-0.095***	0.004**	-0.015***	0.416***
	(0.001)	(0.001)	(0.001)	(0.031)	(0.001)	(0.002)	(0.002)	(0.037)
Other informant	-0.245***	-0.027***	-0.051***	0.727***	-0.224***	0.033***	-0.011***	1.052***
	(0.002)	(0.002)	(0.002)	(0.044)	(0.002)	(0.003)	(0.003)	(0.049)
Observations	1,826,709	1,486,583	1,486,583	1,826,709	699,052	585,574	585,574	699,052
Individuals	442,789	425,332	425,332	442,789	175,951	171,154	171,154	175,951
R^2	0.511	0.789	0.770	0.631	0.487	0.826	0.792	0.731
Dep. var. mean (self alone)	0.881	8.939	3.764	46.853	0.916	8.635	3.718	38.408
P-value (spouse proxy=spouse present)	0.000	0.000	0.000	0.000	0.000	0.005	0.017	0.000
P-value (other proxy=others present)	0.000	0.000	0.000	0.030	0.000	0.000	0.035	0.000
Panel B: Single workers								
Self-reporting and others are present	-0.053***	-0.033***	-0.021***	-0.165	-0.038***	-0.006	-0.005	0.157*
	(0.004)	(0.006)	(0.007)	(0.144)	(0.003)	(0.004)	(0.004)	(0.089)
Proxy respondent	-0.232***	-0.084***	-0.103***	0.650***	-0.216***	-0.010***	-0.039***	0.811***
	(0.002)	(0.003)	(0.003)	(0.067)	(0.002)	(0.002)	(0.002)	(0.041)
Observations	515,581	381,302	381,302	515,581	760,429	588,504	588,504	760,429
Individuals	132,197	123,595	123,595	132,197	186,713	177,558	177,558	186,713
R^2	0.517	0.776	0.763	0.636	0.508	0.810	0.785	0.686
Dep. var. mean (self alone)	0.890	8.886	3.731	46.384	0.908	8.628	3.634	40.601
P-value (proxy=others present)	0.000	0.000	0.000	0.000	0.000	0.300	0.000	0.000
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. In Panel A, the sample is restricted to workers who are married or in a union and cohabit with their spouse in all periods. In Panel B, the sample is restricted to single workers or married/in union workers who do not cohabit with their spouse. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressors in Panel A are dummy variables for whether the informant is the worker and their spouse is present, the informant is the worker and other household members are present (but not the worker's spouse), and the spouse or another household member is the informant (the omitted category is self-reporting by the worker with nobody else there). The main regressors in Panel B are dummy variables for whether the informant is the worker and other household members are present, and for whether another household member is the informant. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 6: Gender wage gap

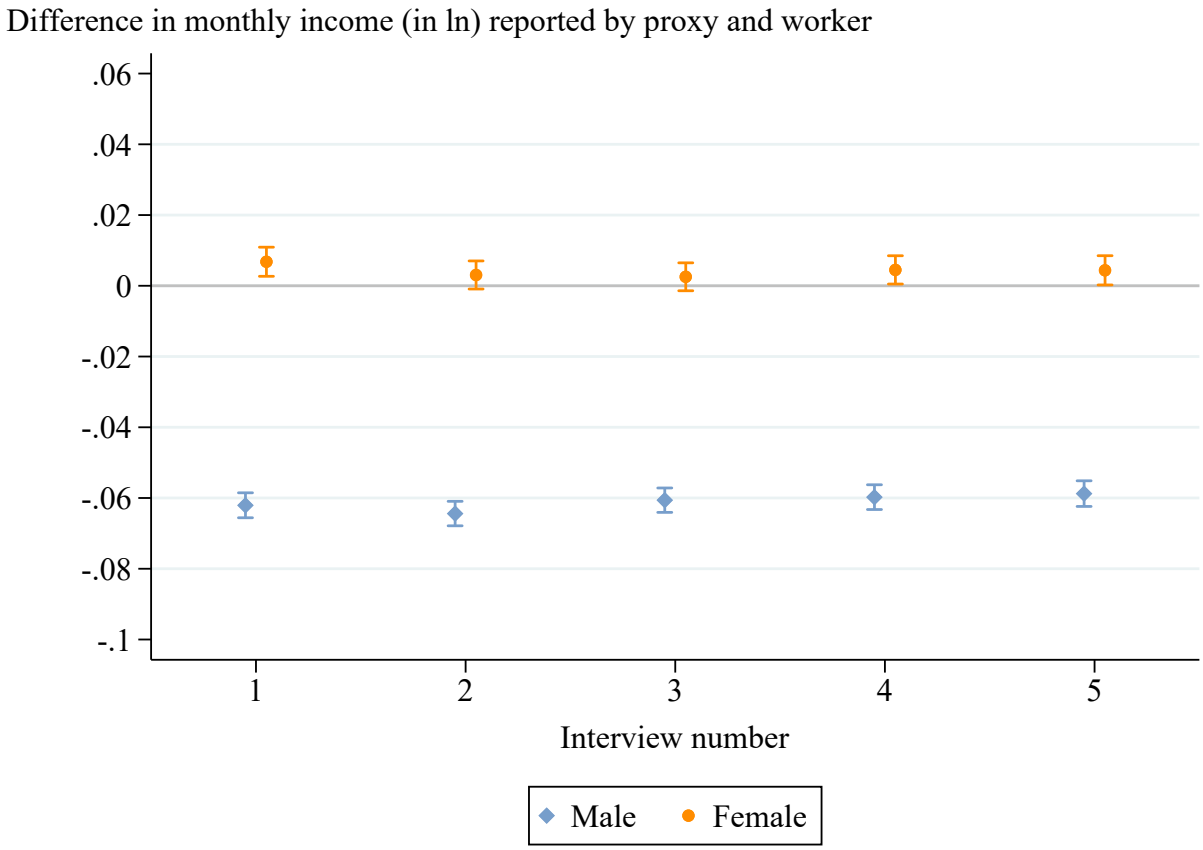
	All	Self-reporting	Proxy respondent
Female	-0.068***	-0.109***	-0.038***
	(0.002)	(0.002)	(0.002)
Observations	1,330,956	585,753	745,203
R^2	0.286	0.299	0.277
Controls	✓	✓	✓
Time FE	✓	✓	✓

Note: The sample is composed of all full-time (30 or more hours worked per week) salaried and self-employed workers aged 25–64 who have at least one observation that is proxy-reported and at least one observation that is self-reported. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods and observations with zero working hours. In columns 2 and 3, we split the sample by whether the observation is self-reported or proxy-reported, respectively. The dependent variable in all regressions is the worker's hourly wage (in MEX$ of 2019 and in ln). Our main regressor is a dummy for whether the worker is female. We also control for age, age squared, educational attainment dummies, a dummy for whether the worker lives in an urban area, state fixed effects, and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

ONLINE APPENDIX

Appendix A Appendix Figures and Tables

Figure A.1: Differences in reported income by type of respondent – heterogeneous effects by interview number



Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. This figure presents the main coefficients and 95% confidence intervals for the estimation of equation (1) fully interacted with dummies for the interview number from which the observation comes. The dependent variable is the monthly wage (in MEX$ of 2019 and in ln). The estimates of regressions in which the worker is male (female) are presented in blue (orange).

Table A.1: Descriptive statistics by type of informant

	Male			Female		
	Always self-reporting	Always proxy	Self and proxy	Always self-reporting	Always proxy	Self and proxy
Demographics						
Age	42.447	39.560	40.778	41.480	37.312	40.741
Head of household	0.941	0.646	0.815	0.578	0.107	0.268
Cohabits with spouse	0.410	0.748	0.834	0.414	0.319	0.549
Educational attainment						
Incomplete primary	0.103	0.102	0.102	0.110	0.051	0.076
Complete primary	0.147	0.186	0.165	0.165	0.121	0.145
Complete lower secondary	0.273	0.344	0.337	0.331	0.304	0.346
Complete upper secondary or more	0.476	0.368	0.396	0.394	0.522	0.433
Labor market outcomes						
Salaried worker	0.768	0.914	0.848	0.737	0.967	0.891
Self-employed	0.232	0.086	0.152	0.263	0.033	0.109
Hours worked per week	46.340	50.002	47.841	39.159	43.293	40.467
Monthly wage (in MEX$ of 2019)	9093.871	8041.119	8505.699	6824.732	7327.540	7342.596
Hourly wage (in MEX$ of 2019)	56.552	43.586	49.414	51.659	46.880	51.556
Job characteristics						
Formal employment	0.635	0.660	0.678	0.559	0.740	0.694
Primary sector	0.072	0.082	0.098	0.010	0.007	0.008
Secondary sector	0.234	0.329	0.288	0.136	0.179	0.153
Tertiary sector	0.692	0.587	0.614	0.853	0.813	0.838
Observations	184,146	962,054	1,241,844	286,678	347,707	857,688
Individuals	52,539	243,388	286,118	76,470	90,008	198,460

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. We split the sample by gender and by the types of informant the worker has across survey waves.

Table A.2: Differences in reported income by type of respondent – nominal income

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Proxy respondent	-0.124***	-0.061***	-0.084***	0.816***	-0.169***	0.004***	-0.024***	0.700***
	(0.001)	(0.001)	(0.001)	(0.025)	(0.001)	(0.001)	(0.001)	(0.025)
Observations	2,388,044	1,903,912	1,903,912	2,388,044	1,492,073	1,200,605	1,200,605	1,492,073
Individuals	582,045	555,636	555,636	582,045	364,938	350,829	350,829	364,938
R^2	0.505	0.795	0.775	0.632	0.496	0.823	0.795	0.709
Dep. var. mean (self)	0.885	8.627	3.456	46.785	0.913	8.349	3.394	39.479
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the nominal monthly (hourly) wage in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressor is a dummy variable that takes a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Differences in reported income by type of respondent – full-time workers

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Proxy respondent	-0.124***	-0.065***	-0.078***	0.560***	-0.172***	-0.007***	-0.011***	0.118***
	(0.001)	(0.001)	(0.001)	(0.026)	(0.001)	(0.001)	(0.002)	(0.026)
Observations	1,789,273	1,423,762	1,423,762	1,789,273	909,342	714,357	714,357	909,342
Individuals	431,900	412,110	412,110	431,900	218,751	209,144	209,144	218,751
R^2	0.501	0.782	0.782	0.626	0.493	0.790	0.810	0.649
Dep. var. mean (self)	0.887	8.956	3.671	50.652	0.903	8.747	3.538	46.949
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all full-time (30 or more hours worked per week) salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressor is a dummy variable that takes a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Differences in reported income by type of respondent – salaried workers

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Proxy respondent	-0.124***	-0.064***	-0.087***	0.879***	-0.173***	-0.006***	-0.034***	0.753***
	(0.001)	(0.001)	(0.001)	(0.026)	(0.001)	(0.001)	(0.001)	(0.024)
Observations	2,074,199	1,683,134	1,683,134	2,074,199	1,311,631	1,049,308	1,049,308	1,311,631
Individuals	505,850	486,109	486,109	505,850	321,544	308,626	308,626	321,544
R^2	0.495	0.771	0.769	0.638	0.500	0.826	0.812	0.693
Dep. var. mean (self)	0.907	9.004	3.817	47.025	0.919	8.766	3.784	39.276
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried workers aged 25–64. We exclude workers who do not have a salaried position in all periods, observations with zero working hours, and workers who only appear in one survey wave. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressor is a dummy variable that takes a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: Differences in reported income by type of respondent – salaried workers who are only paid a salary

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Proxy respondent	-0.129***	-0.056***	-0.079***	0.831***	-0.176***	-0.002	-0.031***	0.782***
	(0.001)	(0.002)	(0.002)	(0.039)	(0.001)	(0.002)	(0.002)	(0.034)
Observations	930,617	741,045	741,045	930,617	711,501	561,170	561,170	711,501
Individuals	234,897	223,455	223,455	234,897	179,051	170,544	170,544	179,051
R^2	0.516	0.803	0.804	0.663	0.515	0.841	0.828	0.699
Dep. var. mean (self)	0.898	8.995	3.841	45.789	0.915	8.748	3.809	37.890
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried workers aged 25–64 who are only paid a salary (instead of commissions, performance bonuses, etc.). We exclude workers who do not have a salaried position in all periods, observations with zero working hours, and workers who only appear in one survey wave. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressor is a dummy variable that takes a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.6: Differences in reported income by type of respondent – interviews conducted in January–November

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Proxy respondent	-0.121***	-0.061***	-0.083***	0.812***	-0.166***	0.005***	-0.023***	0.692***
	(0.001)	(0.001)	(0.001)	(0.027)	(0.001)	(0.001)	(0.001)	(0.026)
Observations	2,241,786	1,793,071	1,793,071	2,241,786	1,390,233	1,122,640	1,122,640	1,390,233
Individuals	582,017	551,885	551,885	582,017	364,910	348,404	348,404	364,910
R^2	0.518	0.792	0.774	0.639	0.509	0.822	0.794	0.716
Dep. var. mean (self)	0.886	8.901	3.732	46.765	0.914	8.626	3.672	39.485
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. We also exclude interviews conducted in December. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressor is a dummy variable that takes a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.7: Differences in reported income by type of respondent – observations with usual working hours

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Proxy respondent	-0.123***	-0.062***	-0.079***	0.683***	-0.168***	0.004***	-0.021***	0.653***
	(0.001)	(0.001)	(0.001)	(0.025)	(0.001)	(0.001)	(0.001)	(0.025)
Observations	2,186,513	1,746,863	1,746,863	2,186,513	1,395,875	1,123,736	1,123,736	1,395,875
Individuals	580,190	549,168	549,168	580,190	364,102	348,016	348,016	364,102
R^2	0.517	0.796	0.792	0.669	0.506	0.822	0.808	0.731
Dep. var. mean (self)	0.887	8.929	3.724	47.786	0.914	8.651	3.662	40.323
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. We also exclude observations in which the worker is not working his/her usual hours. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressor is a dummy variable that takes a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.8: Differences in reported income by type of respondent – workers employed in the same company across survey waves

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Proxy respondent	-0.119***	-0.062***	-0.084***	0.811***	-0.163***	0.004**	-0.023***	0.659***
	(0.001)	(0.001)	(0.002)	(0.036)	(0.001)	(0.002)	(0.002)	(0.035)
Observations	1,234,375	995,078	995,078	1,234,375	762,836	621,985	621,985	762,836
Individuals	318,558	305,372	305,372	318,558	197,693	191,117	191,117	197,693
R^2	0.514	0.795	0.777	0.639	0.503	0.822	0.797	0.718
Dep. var. mean (self)	0.891	8.900	3.727	46.841	0.921	8.625	3.665	39.640
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. We also restrict our sample to individuals who work for the same company in the period between their first interview and their last interview with the extended questionnaire. We drop interviews conducted after the last interview with the extended questionnaire. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressor is a dummy variable that takes a value of 1 if the informant is another household member, and 0 if the worker self-reports. We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.9: Relationship between proxy informant and worker

	All	Male	Female
Panel A: Married workers			
Spouse	0.759	0.799	0.607
Child	0.171	0.140	0.291
Parent	0.015	0.012	0.026
Sibling	0.005	0.005	0.007
Other relative	0.044	0.039	0.061
Non-relative	0.001	0.001	0.001
Unknown	0.004	0.004	0.006
Observations	1,706,302	1,352,979	353,323
Panel B: Single workers			
Child	0.423	0.370	0.464
Parent	0.231	0.249	0.218
Sibling	0.175	0.195	0.159
Other relative	0.101	0.120	0.087
Non-relative	0.035	0.031	0.039
Unknown	0.034	0.035	0.033
Observations	775,811	338,592	437,219

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. In Panel A, the sample is restricted to workers who are married or in a union and cohabit with their spouse in all periods. In Panel B, the sample is restricted to single workers or married/in union workers who do not cohabit with their spouse. This table shows the relationship between the worker and his/her proxy for all observations in which the informant is a proxy.

Table A.10: Differences in reported income by type of respondent – by marital status and relationship with informant

	Male				Female			
	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked	Reports wage	Monthly wage (in ln)	Hourly wage (in ln)	Hours worked
Panel A: Married workers								
Spouse informant	-0.078***	-0.065***	-0.087***	0.872***	-0.092***	0.004**	-0.014***	0.402***
	(0.001)	(0.001)	(0.001)	(0.029)	(0.001)	(0.002)	(0.002)	(0.036)
Child informant	-0.223***	-0.007***	-0.034***	0.762***	-0.204***	0.040***	-0.003	1.010***
	(0.002)	(0.002)	(0.002)	(0.046)	(0.002)	(0.003)	(0.003)	(0.051)
Other informant	-0.288***	-0.072***	-0.089***	0.467***	-0.285***	-0.000	-0.047***	1.187***
	(0.003)	(0.003)	(0.004)	(0.074)	(0.004)	(0.005)	(0.005)	(0.089)
Observations	1,839,948	1,498,948	1,498,948	1,839,948	720,534	604,346	604,346	720,534
Individuals	442,642	425,045	425,045	442,642	175,864	170,995	170,995	175,864
R^2	0.510	0.789	0.769	0.630	0.482	0.823	0.789	0.726
Dep. var. mean (self)	0.883	8.921	3.745	46.929	0.917	8.637	3.718	38.450
Panel B: Single workers								
Child informant	-0.225***	-0.094***	-0.111***	0.546***	-0.209***	-0.007***	-0.037***	0.793***
	(0.003)	(0.004)	(0.004)	(0.078)	(0.002)	(0.002)	(0.002)	(0.045)
Parent informant	-0.223***	-0.124***	-0.141***	0.635***	-0.204***	-0.048***	-0.077***	0.900***
	(0.003)	(0.004)	(0.005)	(0.089)	(0.003)	(0.003)	(0.004)	(0.062)
Sibling informant	-0.223***	-0.044***	-0.072***	0.887***	-0.209***	0.009***	-0.020***	0.770***
	(0.003)	(0.004)	(0.005)	(0.089)	(0.003)	(0.003)	(0.004)	(0.064)
Other informant	-0.228***	-0.043***	-0.066***	0.720***	-0.230***	0.001	-0.027***	0.586***
	(0.004)	(0.006)	(0.006)	(0.124)	(0.004)	(0.004)	(0.005)	(0.093)
Observations	503,904	373,602	373,602	503,904	746,087	579,221	579,221	746,087
Individuals	131,453	122,450	122,450	131,453	186,165	176,590	176,590	186,165
R^2	0.522	0.779	0.765	0.639	0.512	0.812	0.786	0.689
Dep. var. mean (self)	0.889	8.880	3.727	46.354	0.908	8.627	3.632	40.638
Individual FE	✓	✓	✓	✓	✓	✓	✓	✓
Time FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The sample is composed of all salaried and self-employed workers aged 25–64. We exclude workers who do not have the same occupation status (salaried or self-employed) in all periods, observations with zero working hours, and workers who only appear in one survey wave. In Panel A, the sample is restricted to workers who are married or in a union and cohabit with their spouse in all periods. In Panel B, the sample is restricted to single workers or married/in union workers who do not cohabit with their spouse. The regressions in columns 1–4 are restricted to male workers, and those in columns 5–8 to female workers. The dependent variable in columns 1 and 5 is a dummy for whether a wage is reported. The dependent variable in columns 2 and 6 (3 and 7) is the monthly (hourly) wage, both in MEX$ of 2019 and in ln. The dependent variable in columns 4 and 8 is the weekly hours worked. The main regressors in Panel A are dummy variables for whether the informant is the worker's spouse, child, or another household member (the omitted category is self-reporting by the worker). The main regressors in Panel B are dummy variables for whether the informant is the worker's child, parent, sibling, or another household member (the omitted category is self-reporting by workers). We control for individual and survey-wave fixed effects. Standard errors clustered at the individual level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.