Validating Earnings in the German National Educational Panel Study: Determinants of Measurement Accuracy of Survey Questions on Earnings

Antoni, Manfred; Bela, Daniel; Vicari, Basha

methods, data, analyses | Vol. 13(1), 2019, pp. 59-90 DOI: 10.12758/mda.2018.08

Validating Earnings in the German National Educational Panel Study. Determinants of Measurement Accuracy of Survey Questions on Earnings

Manfred Antoni[1], *Daniel Bela*[2] & *Basha Vicari*[1]
[1] *Institute for Employment Research (IAB), Nuremberg*
[2] *LIfBi – Leibniz Institute for Educational Trajectories, Bamberg*

Abstract

Questions on earnings are counted among sensitive topics that often produce high rates of item nonresponse or measurement error. Both types of bias are well documented in the literature and are found to concentrate in the tails of the earnings distribution. In this paper, we explore whether measurement error on earnings could be explained by socially desirable reporting and whether the error is impacted by interviewer characteristics.

Using the linked dataset NEPS-SC6-ADIAB, which contains survey data from the German National Educational Panel Study, Starting Cohort "Adults", linked to administrative earnings records from the German Federal Employment Agency, we analyze the extents of over- and underreporting and the influence of respondent and interviewer characteristics on these behaviors for different quartiles of the earnings distribution.

Our results show that the average level of misreporting is relatively low (approximately 6% of median earnings). Our main logistic model reveals that female and more highly educated respondents report significantly more accurately while those with higher earnings misreport to a significantly greater extent. Regarding the impact of personality traits on reporting accuracy, we find significant positive effects for more agreeable respondents and significant negative effects for extraverted respondents. When differentiating by the direction of misreporting, we find, for instance, that women are less likely to overreport across all earnings quartiles. However, the influence of interviewer characteristics is negligible.

Keywords: Measurement error, earnings, social desirability, interviewer effects, NEPS-SC6-ADIAB

Information on earnings is among the statistics that are most pervasively collected in population surveys. It provides a basis for a wide array of research conclusions and policy decisions related to topics such as a country's overall wealth distribution and social inequality trends (Moore et al. 2000; Bound & Krueger 1991). From an individual perspective, it is often used to approximate a person's socioeconomic status in order to explain decisional or behavioral patterns. However, any survey data are prone to some kind of nonresponse or measurement error. This is especially true in regard to sensitive questions, such as questions on respondents' earnings, collected in interviewer-administered surveys (Moore et al. 2000; Groves et al. 2009).

Questions on sensitive topics often produce relatively high rates of item nonresponse and measurement error because such questions can be perceived as threatening to disclose private information or deviant behavior (Jann 2014; West & Blom 2017). Tourangeau and Yan (2007) expect high rates of item nonresponse for questions on personal income because these questions are perceived to be intrusive. They do not necessarily expect high rates of misreporting, however, because such questions are not associated with a disclosure of violation of social norms. This expectation is supported by the findings of Krumpal (2013) which show that in German population surveys the earnings question has the highest nonresponse rate among all items. Moreover, missing earnings reports are not randomly distributed; instead, the missing values are concentrated in the tails of the earnings distribution (Riphahn & Serfling 2005; Bollinger et al. 2018).

The statement of Tourangeau and Yan (2007) is, however, contrasted by a wealth of studies revealing the prevalence of misreporting in response to survey questions on earnings (see, e.g., Duncan & Hill 1985; Bound & Krueger 1991; Bound et al. 1994; Bollinger 1998; Moore et al. 2000; Pedace & Bates 2000; Gottschalk & Huynh 2005; Kapteyn & Ypma 2007; Bricker & Engelhardt 2008; Gottschalk & Huynh 2010; Kim & Tamborini 2014; Angel et al. 2017). All these studies assess the quality of earnings reports by linking survey information to auxiliary data sources, most commonly administrative records that offer more reliable measures of earnings, which are thus treated as the "true" values. Regarding the nature

Acknowledgments

We thank two anonymous reviewers and the editors whose valuable comments helped to improve the quality of our manuscript.

Direct correspondence to

Manfred Antoni, Research Data Centre (FDZ) of the Federal Employment Agency at the Institute for Employment Research (IAB), Regensburger Str. 100,
D-90478 Nuremberg
E-mail: manfred.antoni@iab.de

of earnings measurement error, these studies find a U-shaped pattern similar to that of item nonresponse: there is a negative correlation between the measurement error and the assumed true earnings value, indicating that low earners tend to over-report their earnings, while high earners tend to underreport (Bound & Krueger 1991; Bollinger 1998; Bricker & Engelhardt 2008). Nevertheless, we still know little about why respondents edit their answers depending on their positions in the earnings distribution.

We are not the first to examine the misreporting of individual earnings using a validation study, but most previous studies were conducted in the Anglo-American context, used small or restricted samples (e.g., male workers), or used a cross-sectional design. We use the linked data product called NEPS-SC6-ADIAB, which contains survey data from the German National Educational Panel Study, Starting Cohort "Adults", (NEPS SC6) – a panel survey representative of the German adult population and covering a rich set of respondent characteristics – linked to administrative earnings records from the German Federal Employment Agency. Because interviewers either can have a positive influence on participation and data quality or can cause interviewer effects (Essig & Winter 2009; Landrock 2017), we also include interviewer characteristics and paradata on the interview situation in our analysis.

Thus, we contribute to the literature in three ways: First, we provide evidence on a cultural context of money taboo, where talking about financial issues causes feelings of uneasiness (Trachtman 1999). Germany is counted among such cultural contexts (see, e.g., Kirkcaldy et al. 1992). As responding to sensitive questions is, in general, a highly context-specific behavior (Jann 2014), we assume that the cultural context of money taboo changes a merely intrusive question into one that might create embarrassment or shame. These factors make it particularly unpleasant for respondents to report very low or very high earnings (see, e.g., Bound & Krueger 1991), especially when it comes to admitting to living in poverty or in luxury in the presence of an interviewer.

Our second contribution directly derives from this fact because we combine our earnings validation study with the concept of socially desirable reporting. On the one hand, respondents might edit their reports towards some subjectively estimated norm of individual wealth. On the other hand, a competent interviewer might be able to create a trustful interviewing atmosphere and hence minimize the social desirability bias. Using a rich set of respondent and interviewer characteristics as well as variables reflecting the interview situation allows us to examine this aspect closely.

Third, we conduct several analyses that allow us to identify the direction of misreporting (over- vs. underreporting). Because the literature already documents the phenomenon of "mean-reverting measurement error" as manifested in increased misreporting in the tails of the earnings distribution (Kim & Tamborini

2014; Angel et al. 2017), we further analyze the tendencies to under- or overreport in different quartiles of the earnings distribution. This strategy also yields deeper insight into the impact of socially desirable reporting on the determinants of such tendencies.

Theoretical Background and Hypotheses

There are various reasons why collecting information on earnings is difficult. First, we should differentiate between unintentional and deliberate misreporting. Answering the question “What are your monthly gross earnings?” requires a cognitive process that passes through several stages, including interpretation of the question, retrieval of the exact amount, estimation and judgment, and response production (see, e.g., Tourangeau 1984; Moore et al. 2000; Groves et al. 2009; Kim & Tamborini 2014). Problems in interpretation/understanding, recall and response production result in unintentional misreporting. These, however, should generate an approximately randomly distributed measurement error or heaping[1] at round numbers.

In the case that respondents perceive answering an earnings question as uncomfortable or a violation of privacy, they could either refuse to answer or deliberately misreport values. This is consistent with findings that earnings questions have the highest rates of item nonresponse in general population surveys (see, e.g., Tourangeau & Yan 2007; Krumpal 2013) and that there is a substantial level of misreporting mainly in the left and right tails of the earnings distribution (Pedace & Bates 2000; Riphahn & Serfling 2005; Essig & Winter 2009; Bollinger et al. 2018). This “mean-reverting measurement error” (Bound & Krueger 1991; Bricker & Engelhardt 2008) gives rise to our assumption that such response behavior is caused by socially desirable reporting rather than by problems of understanding or recall.

According to social desirability theory, respondents reflect on an expected mainstream view in their cultural context and then edit their answers to comply with this view (see, e.g., DeMaio 1984; Krumpal 2013; Lipps & Lutz 2017). In other words, they are more likely to report desirable attributes than undesirable ones to present themselves in a positive light (Stocké & Hunkler 2007). The presence of an interviewer might either increase the tendency to edit answers, especially when the social distance between the respondent and interviewer is perceived as high (Diekmann 2008), or decrease misreporting, particularly when the interviewer is able to create a trustful atmosphere or help the respondent to interpret a question correctly (see, e.g., Landrock 2017).

1 Heaping refers to reporting numbers in increments (Zinn & Würbach 2016).

A rich literature on the influence of respondent characteristics exists, even more so for interviewer effects on item nonresponse and measurement error in the case of sensitive questions (for an overview, see West & Blom 2017). We consider both types of influences to explain the misreporting of earnings in the survey data by accounting for socially desirable reporting. Thus, we derive our three main research questions and assign several hypotheses to each of them:

1. What is the extent of earnings misreporting in the survey data?
2. How do respondent characteristics influence measurement error on earnings questions?
3. How do interviewer characteristics influence measurement error on earnings questions?

First, we are interested in the overall extent of misreporting, measured as the deviation between the two data sources. Some evidence exists that the measurement error on earnings questions is modest in panel studies (Bound & Krueger 1991; Kühne 2018). The results are inconsistent, however, with regard to whether earnings are underreported mainly by high earners (Paulus 2015; Angel et al. 2017) or low earners (Meyer & Mittag 2017), are overreported by low earners (Bollinger 1998), or are both over- and underreported depending on the characteristics of different subgroups (Pedace & Bates 2000; Kim & Tamborini 2014). Taken together, these findings lead us to expect a mean-reverting measurement error with more pronounced rates of misreporting in both tails of the earnings distribution (*hypothesis 1*).

Concerning the impact of the sociodemographic characteristics of the respondents on responses to sensitive questions, Preisendörfer and Wolter (2014, p. 126) find that female, older, and better-educated respondents are more likely to underreport delinquent behavior than male, younger, and less-educated respondents are. Regarding income questions, however, Bound and Krueger (1991) show that the average measurement error is larger for men than for women (confirmed by, e.g., Bricker & Engelhardt 2008). Bollinger (1998) finds that low-income men are most likely to overreport their earnings. Following these findings, we assume that female respondents report more accurately in general (*hypothesis 2*). The effects of age and education are less clear. Some studies find no evidence for correlations of age and education with misreporting (Bound & Krueger 1991; Gottschalk & Huynh 2005), whereas others find that the measurement error rises with reported education level (Bricker & Engelhardt 2008) or decreases with better education at higher earnings levels (Kim & Tamborini 2014). As these findings are rather ambiguous, we follow the more general study of Preisendörfer and Wolter (2014) and hypothesize that younger (*hypothesis 3*) and less-educated (*hypothesis 4*) respondents report their earnings more accurately than other groups do.

In general, respondent characteristics that are associated with the level of earnings are often considered to affect misreporting on earnings questions. Therefore, we include personality traits of the respondents in our models. Several studies confirm an effect of personality traits, as measured in terms of the "Big Five Inventory" dimensions, on earnings (e.g., Mueller & Plug 2006; Heineck & Anger 2010; Spurk & Abele 2011), although the effects differ depending on whether these traits are considered independently or in combination with sociodemographic characteristics. For our analysis, we choose two personality traits out of five that we assume to exert a significant influence on the ability to cope with the interview situation. These traits are the dimensions of "extraversion" and "agreeableness" in the Big Five Inventory.[2] In NEPS SC6, each personality trait was measured by two items on a scale from 1 through 5, as recommended by Rammstedt and John (2007), who developed this short version of the Big Five Inventory. We assume that in an interview situation, a distinctly extraverted respondent will tend to exaggerate his or her earnings and thus be more likely to overreport (*hypothesis 5*), whereas a respondent with a high score in agreeableness will tend to stay close to the true value of his or her earnings and hence report more accurately (*hypothesis 6*).

To explore how interviewers influence the measurement error on earnings, we also formulate hypotheses on the sociodemographic characteristics of the interviewer and on the interview situation. West and Blom (2017, p. 187) give an overview of the effects of the interviewer's gender, age and experience on misreporting in response to sensitive questions. The majority of studies they consider find female interviewers to elicit more accurate responses than male interviewers do. However, the interaction between the genders of the interviewer and respondent also seems to play a role (Lipps & Lutz 2017). Regarding the age of the interviewer, West and Blom (2017) find greater evidence for a positive relationship between response quality and interviewer age, although this relationship is again moderated by the interaction between the interview partners. Because the similarity between interviewer and respondent seems to be an important source of influence (cp. Diekmann 2008), we hypothesize that interviewers of the same gender (*hypothesis 7*) and of similar age (*hypothesis 8*) and educational level (*hypothesis 9*) to those of the respondent induce less misreporting.

Furthermore, West and Blom (2017) consider the effect of the interviewer's experience on response quality. These authors distinguish between overall experience as an interviewer and survey-specific experience. Although they find ambiguous evidence for both experience measures, we nevertheless hypothesize that more experienced interviewers should, in general, elicit more accurate reports (*hypothesis 10*) and that accuracy should also be positively correlated with the interviewer's

2 All Big Five Inventory dimensions seem to have some effects on earnings, however, we do not assume a significant influence of the other three dimensions (openness to experience, conscientiousness, and neuroticism) on respondent's answering behavior.

experience with the current survey. Following Preisendörfer and Wolter (2014), we thus hypothesize that reporting accuracy should increase with the number of interviews conducted within any given survey wave (*hypothesis 11*). Finally, interviewer effects inducing socially desirable reporting are stronger in face-to-face interviews, during which the interviewers' characteristics are directly observable by the respondents, than they are in telephone interviews, during which the interviewers' characteristics can only be estimated by the respondents (West et al. 2013). We therefore expect to find smaller measurement errors in computer-assisted telephone interviewing (CATI) responses than in computer-assisted personal interviewing (CAPI) responses (*hypothesis 12*).

Data and Research Strategy

Data

Survey data: NEPS SC6

For our analyses, we use survey data from the German National Educational Panel Study, Starting Cohort "Adults" (NEPS SC6, https://doi.org/10.5157/NEPS:SC6:8.0.0). An overview of the content and theoretical basis of this survey can be found in Allmendinger et al. (2011). From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. The survey data were first collected as part of the survey "Working and Learning in a Changing World" (IAB-ALWA, cp. Antoni et al. 2010). This survey consisted of a start-up survey wave, conducted in winter 2007/2008, and was followed by seven additional annual survey waves, starting in winter 2009/2010, conducted in the NEPS framework. In each of these eight waves, the employment histories of the respondents were recorded. Naturally, information for all employment episodes before each respondent's first interview had to be collected retrospectively (i.e., all past employment experiences were queried in the first interview). In each follow-up wave, all current employment episodes of each person were surveyed in real time. For episodes of the latter type, net and gross income information was queried in all survey waves except the first one; thus, our estimation sample is limited to waves 2 through 8. Additionally, more than 93% of the respondents consented for their survey information to be linked to administrative data from the German

Federal Employment Agency. The phrasing of the survey questions on earnings and linkage consent is available in the online documentation of the NEPS surveys.[3]

Linked data: NEPS-SC6-ADIAB

The longitudinal administrative data available at the Institute for Employment Research (IAB), the research unit of the German Federal Employment Agency, originate from a number of different sources within the German social security system. On the one hand, these data contain information on various aspects of unemployment insurance and assistance. On the other hand, and more importantly for our analyses, the IAB data also contain information provided by employers about all of their dependent employees.[4] These employment history data include data on every person who has been dependently employed at least once since 1975 for West Germany and since 1992 for East Germany. Information on employees is reported by employers via mandatory notifications to the social security system. For NEPS SC6 consenting respondents, the two data sources are linked. The joint data product we use for our analyses, NEPS-SC6-ADIAB (doi:10.5164/IAB.NEPS-SC6-ADIAB7515.de.en.v1), has been documented by Antoni et al. (2018).

While employers' notifications to the social security system also contain a small number of sociodemographic and job characteristics, the information most crucial to our analyses is the sum of gross earnings during each reported job episode. These earnings directly determine the contributions to social insurance. Employers who fail to issue correct notifications on earnings and to directly transfer the proper amounts to the social security system are subject to considerable sanctions, ranging from financial penalties to imprisonment of up to ten years.[5] For these reasons, the information on gross earnings contained in the administrative data at hand is considered to be highly reliable. We therefore treat the resulting measure of earnings as the true value and any deviation from that value by the respondent during an interview as measurement error. There are, however, some caveats with regard to these administrative data, some of which require us to adapt our analyses.

Sample restrictions

Our goals are to measure the accuracy of earnings reports as precisely as possible and to distinguish the different factors contributing to any deviations we find. To do so, we take several steps during data preparation to restrict our sample to rule out

3 https://www.neps-data.de/en-us/datacenter/dataanddocumentation/startingcohortadults/documentation

4 See Antoni, Ganzer, and vom Berge (2016) for more details on these administrative data and https://fdz.iab.de/en.aspx for how to access them via the Research Data Centre of the Federal Employment Agency at the IAB.

5 As of § 266a of the German Criminal Code, "StGB".

any factors in the two data-generating processes that might contribute to deviations that respondents cannot understand or influence.

Because the administrative employment histories include only dependent employees whose earnings are subject to mandatory social security contributions, these data do not contain information on civil servants (the German “Beamte”) or the self-employed. Any such employment episodes are therefore not considered in our analyses, even when they are reported in the survey data. Moreover, employer notifications do not include the working hours corresponding to employment episodes. This prevents us from calculating hourly earnings, which is particularly problematic for part-time workers. For this reason, we restrict all analyses to full-time employment episodes.

Another reason for this restriction to full-time employment episodes is that the record linkage procedure merely identifies the administrative data corresponding to a given person from the survey data. The linkage process does not extend to the assignment of every single employment episode from one dataset to its exact counterpart in the other dataset. We therefore restrict our estimation sample to employment episodes that were either ongoing on the date of the interview or had ended shortly before the interview. In this way, we can ensure that we are actually comparing earnings measures related to the same employment episode. If we did not remove part-time employment episodes from the analyses, our sample could include respondents with two parallel part-time jobs at the time of the interview. This would strongly increase the risk of assigning two unrelated job episodes to each other and, thus, of comparing the wrong earnings measures.

For observations with administrative earnings beyond the social security contribution ceiling (“Beitragsbemessungsgrenze”), the measure is truncated at this threshold value. Because it would be impossible to determine a valid administrative earnings measure in these cases, we eliminate them from the estimation sample in accordance with the procedures recommended by Drews, Groll, and Jacobebbinghaus (2007, p. 32).

Finally, some special payments made to employees (e.g., end-of-year bonuses) may be reported in separate but parallel notifications, usually with a much shorter duration than that of the main employment episode. We do not add the wage sums reported in such notifications to our administrative earnings measures because it would be impossible to determine whether a given respondent considered such a payment when reporting on his gross earnings. Because such special payments may introduce natural deviations between the earnings measures that are unrelated to the response behavior during the interview, we include a number of variables in our estimations that at least allow us to control for the existence of such factors.

In addition to these deliberate exclusions of cases, we also drop observations for which any of the dependent or independent variables are missing. The greatest loss in observations for our complete case analyses results from the fact that the

Big Five personality traits were surveyed only in waves 5 and 8 of the NEPS SC6 survey. However, the stability of the Big Five instruments over time is documented in the literature, especially for the adult population (cp. Cobb-Clark & Schurer 2012; Rantanen et al. 2007). This allows us to transfer the reported data for a given person to all corresponding interviews from other waves without measurements of these traits. Ultimately, because we exclude all respondents without any personality trait measurements, our sample is reduced to all respondents who answered the relevant questions in at least one of waves 5 and 8.

Due to omitting such a large part of the original NEPS SC6 sample, and especially due to restricting the estimation sample to full-time dependently employed persons, we have to acknowledge that our remaining estimation sample is no longer representative. In Table A1[6] in the Appendix, we compare the subsamples (estimation sample vs. non-estimation sample) for the main respondent characteristics used in our analyses. Unsurprisingly, nearly all characteristics show significant biases between the two groups. As a result, we cannot claim representativeness for our estimations.

Dependent Variables

Our main focus is on (deliberate) deviations in reported earnings relative to the administrative measure. As our main dependent variable, we chose an indicator that reflects deviation of reported earnings from the administrative measure by more than 20%.[7]

Additionally, we extended the dependent variable to a multinomial indicator reflecting the direction of deviation (“underreporting”, “no deliberate deviation”, or “overreporting”). Again, we chose a threshold of deviation by more than 20% in each direction. Because the three categories are mutually exclusive by nature, this generated variable can be used as a dependent variable in multinomial logit models without violating the assumption of the independence of irrelevant alternatives. This indicator separates our estimation sample into 1464 instances of underreporting and 760 instances of overreporting, corresponding to 10% and 5%, respectively, of the total number of observations.

6 All tables in this paper were generated using the user-written Stata routine estout (Jann 2005).

7 Three alternative variations of all models, using a threshold of 10%, a full standard deviation of the earnings distribution, and one-half standard deviation of the earnings distribution, have been calculated. The results are available by request.

Respondent and Interviewer Characteristics

The estimation sample comprises 14065 observations from 4087 respondents, i.e., the average number of observations per respondent is approximately 3.4 (see Table 1 for a tabular overview). A descriptive analysis of the respondents' characteristics reveals that 70% of the observations correspond to male respondents. This overrepresentation is attributed to the fact that we consider only full-time employment episodes, which are still more common among men than among women in Germany. Most respondents were aged between 30 and 49 years (52%); only 9% were younger.[8] A small minority of 4% of the respondents reported no vocational degree after schooling; the education level of the majority (38%) corresponded to intermediate schooling with vocational training. Approximately one-fifth of the observations were collected from respondents who had completed lower secondary education and vocational training (20%), another one-fifth to respondents who had completed

Table 1 Respondent characteristics

	Mean	SD	Min	Max
Resp. gender				
Male	0.70	0.46	0	1
Female	0.30	0.46	0	1
Resp. age				
Aged 29 and lower	0.09	0.28	0	1
Aged 30-49	0.52	0.50	0	1
Aged 50 or older	0.39	0.49	0	1
Resp. education				
Schooling, no training	0.04	0.20	0	1
Lower secondary, voc. train.	0.20	0.40	0	1
Intermediate, voc. training	0.38	0.49	0	1
Upper secondary, voc. train.	0.18	0.38	0	1
Higher education degree	0.21	0.41	0	1
Personality traits				
Big 5: Extraversion	3.32	0.92	1	5
Big 5: Agreeableness	3.54	0.59	1.3	5
Survey mode				
CAPI	0.49	0.50	0	1
CATI	0.51	0.50	0	1

Source: NEPS-SC6-ADIAB, own calculations.
Notes: Number of observations: 14065 of 4087 respondents.

8 We classified respondent age to brackets similar to those available for interviewers.

upper secondary education and vocational training (18%), and a similar number are associated with respondents holding a higher education degree (21%). The two personality traits considered in our analyses, "extraversion" and "agreeableness", show means of 3.32 and 3.54, respectively. Approximately one half of the interviews included in our estimation sample were conducted via CAPI; the other half were conducted via CATI.

Table 2 presents a comparison of the interviewer characteristics for each survey mode. Most of the available interviewer attributes show significant differences between the CATI and CAPI modes, as shown by t-tests of the differences between the means for the two groups. The most striking findings are that the interviewers in the CAPI group were significantly older and more experienced than those in the CATI group. On the other hand, the CAPI interviewers performed significantly fewer interviews on average than the CATI interviewers did. This finding is not surprising, considering that CAPI interviewers must travel to their respondents' locations before conducting interviews, while a CATI interviewer may be assigned

Table 2 Interviewer characteristics, t-test by interview mode

	CAPI	CATI	Difference	t
Interviewer's gender				
I: male	0.569	0.528	0.041***	4.907
I: female	0.431	0.472	-0.041***	-4.907
Interviewer's age				
I: aged 29 and lower	0.008	0.311	-0.303***	-53.148
I: aged 30-49	0.151	0.372	-0.221***	-30.604
I: aged 50-65	0.607	0.277	0.330***	41.765
I: aged older than 65	0.234	0.041	0.194***	35.054
Interviewer's education				
I: lower secondary	0.165	0.082	0.082***	15.001
I: intermediate	0.246	0.185	0.061***	8.790
I: upper secondary	0.589	0.732	-0.143***	-18.142
Interviewer's experience				
I: exp. less than 2 years	0.144	0.285	-0.141***	-20.653
I: exp. 2-3 years	0.294	0.291	0.003	0.454
I: exp. 4-5 years	0.200	0.235	-0.034***	-4.948
I: exp. 6 or more years	0.362	0.189	0.172***	23.361
I: no. of interviews conducted so far	28.574	47.179	-18.605***	-28.102

Source: NEPS-SC6-ADIAB, own calculations.
Notes: Number of observations: 14065. Number of interviewers: 800. *** indicates significance at the 0.1% level.

another interview immediately after the previous one without leaving the telephone studio. We expect the interviewer's knowledge of this specific NEPS SC6 questionnaire to be a possible factor in reducing the risk of eliciting socially desirable answers to the earnings question. Thus, we include a variable reflecting the interviewer's individual familiarity with the specific survey instrument to test our *hypothesis 11*. This variable counts the number of interviews an interviewer has conducted in each wave up to and including the current one.

Control Variables for Multivariate Analyses

As mentioned earlier, our aim is to reduce all deviations between the two earnings measures to only those that can be considered deliberate misreporting, to the greatest possible extent. Thus, all regression analyses are performed on a set of control variables that can potentially support this distinction. Most importantly, we introduce four dummy variables based on the survey data that may influence the accuracy of the earnings measures. First is an indicator of paid overtime, complemented by a second indicator of other special payments. Third is a dummy that indicates whether a child benefit ("Kindergeld") is integrated into the earnings report. These variables act as approximations of factors that make deviations more likely because they represent monetary benefits that may be counted as earnings in one data source but not the other. A fourth dummy variable indicates whether the person is working for a public or private employer. This may influence how accurately respondents recall their gross earnings because public employees are assigned to highly standardized wage schemes, whereas employees of private companies have considerably more individual bargaining power over their earnings. We also control for the region of birth (West Germany, East Germany, or outside of Germany) as a proxy to reduce potential cultural differences in reporting. Finally, we include indicators of the panel wave in which an interview was performed.

Results

Extent and Determinants of Item Nonresponse of the Earnings Question

Before beginning a detailed analysis of the measurement error on reported earnings, we substantiate one of our central assumptions, namely, that information on earnings is considered sensitive, at least for respondents from a cultural context in which money is a taboo topic, such as in Germany. This is corroborated by the following findings: First, we encounter a substantially higher share of item nonresponse on the gross earnings question compared to questions on more generic information, e.g., a respondent's job. The number of answers in the categories "don't know" and "refuse to answer" together represent more than 11% of the responses to the earnings question, a much higher share than those for questions on, for instance, part-time vs. full-time employment (0.7%) or attendance of training courses during a given job (less than 0.5%). Only one-half of the item nonresponse on the gross earnings question correspond to recall problems (i.e., answering "don't know"), either claimed or true.

We derive our second confirmation of question sensitivity by estimating the effects of the respondents' and interviewers' characteristics on the propensity to validly answer the open-ended question[9] about gross earnings. The results of this standard logit model are presented in Table A2 in the Appendix. The lack of a face-to-face presence of the interviewer during a telephone interview (in contrast to the CAPI mode) reduces the risk of item nonresponse, which is consistent with our assumption that the mere presence of an interviewer might lead respondents to avoid answering the earnings question. On the other hand, we see a significant influence of interviewer experience, with field personnel with at least two years of experience successfully reducing the risk of item nonresponse on the gross earnings question. Using the administrative data source, we are able to classify the survey participants by their "true" earnings, even if they did not respond to the earnings question. To do so, we include the variable that reflects the appropriate quartile of the earnings distribution. The results show that respondents in the lowest earnings quartile are the most likely to provide a valid answer to the open gross earnings question, although only the marginal effects of the second and fourth quartile are significant. These results are consistent with the findings in the existing literature, which indicate that respondents with lower earnings levels are more likely to answer the earnings question, whereas persons with higher earnings levels are more likely to refuse to answer. Overall, these findings are well consistent with our assumption regarding the sensitive nature of the earnings question.

Descriptive Comparison of Earnings Measures

A comparison of the survey gross earnings measure and the more reliable administrative gross earnings measure illustrated in Table 3 shows considerable deviation between these two central attributes. When inspecting the two earnings measures separately, we find both of their means to be slightly above 3000 Euros (rows 1 and 2). From row 3 onward, Table 3 shows the deviation of the survey earnings measure from the administrative earnings measure based on the difference computed for

9 For each relevant job, the question on gross earnings was first asked using an open-ended question to elicit the exact value. Only if the respondent was unable or unwilling to answer that question would he or she be asked to at least classify his or her earnings relative to a list of earnings brackets.

Table 3 Descriptive statistics of survey and administrative gross earnings measures and individual deviations (in Euro)

	Mean	Median	Min	Max	N
Gross earnings measures:					
Administrative data	3353.1	3247.3	1216.8	6050.5	14065
Survey data	3162.5	3000.0	980.0	18000.0	14065
Deviation of survey measure from administrative measure:					
Overall	-190.6	-194.8	-3707.8	12532.3	14065
>20% underreporting	-1041.6	-980.6	-3707.8	-297.6	1464
>20% overreporting	1268.9	955.9	257.3	12532.3	760

Source: NEPS-SC6-ADIAB, own calculations.

Notes: All values in the three bottom rows are calculated based on individual observations, not by calculating the difference between the first two table rows. In the two bottom rows, a difference is counted as under- or overreporting if the survey measure is more than 20% below or above the administrative measure, respectively.

each individual observation. Both mean and median of the overall deviation are roughly 200 Euros (190.6 Euros and 194.8 Euros, respectively). The absolute value of the mean deviation represents underreporting by almost 6% relative to the mean of administrative earnings. The two bottom rows of Table 3 present additional details on observations with under- or overreporting by more than 20% compared to the given administrative earnings measure. Among these observations, the mean absolute deviations are more than 1000 Euros in each direction.

By comparing the whole distribution of each of the two variables, we find typical heaping structures in the survey data. In addition, the distribution of the survey measure is slightly but visibly shifted towards the lower side of the distribution (see Figure 1).

A closer look at the differences between the two earnings measures shows an interesting pattern: While the deviation is balanced in the lower earnings groups, it becomes broader with higher earnings. For illustration, Figure 2 visualizes this pattern across the four quartiles of the earnings distribution, as drawn from the administrative data. While the standard deviation of the difference might be expected to increase for higher earnings groups because respondents with higher earnings have a broader range of possible answers to which they could deviate, the positions of the quartiles are emphasized here. The higher the earnings quartile is, the more likely is underreporting compared to the more reliable administrative earnings. This result, although only initial descriptive evidence, supports parts of *hypothesis 1.*

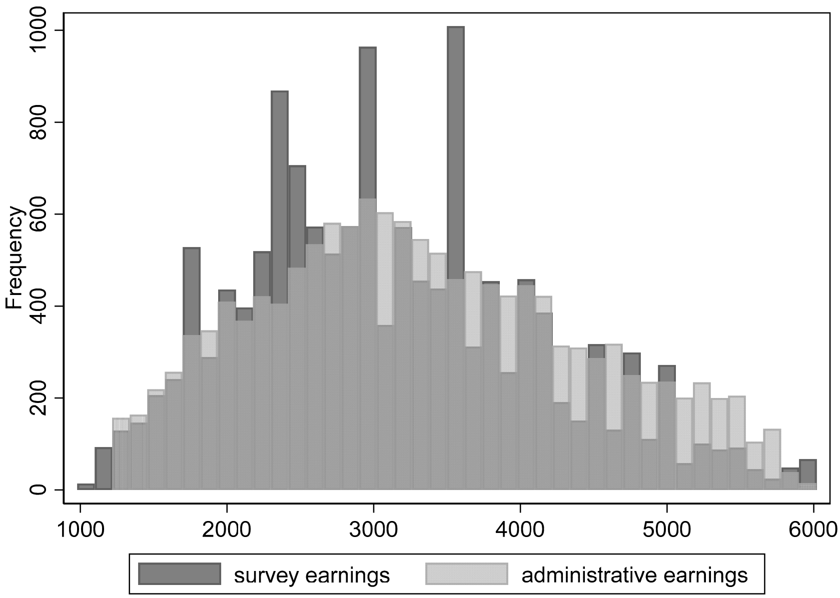

Figure 1 Histograms of reported and administrative monthly gross earnings (excluding outliers; in Euro)

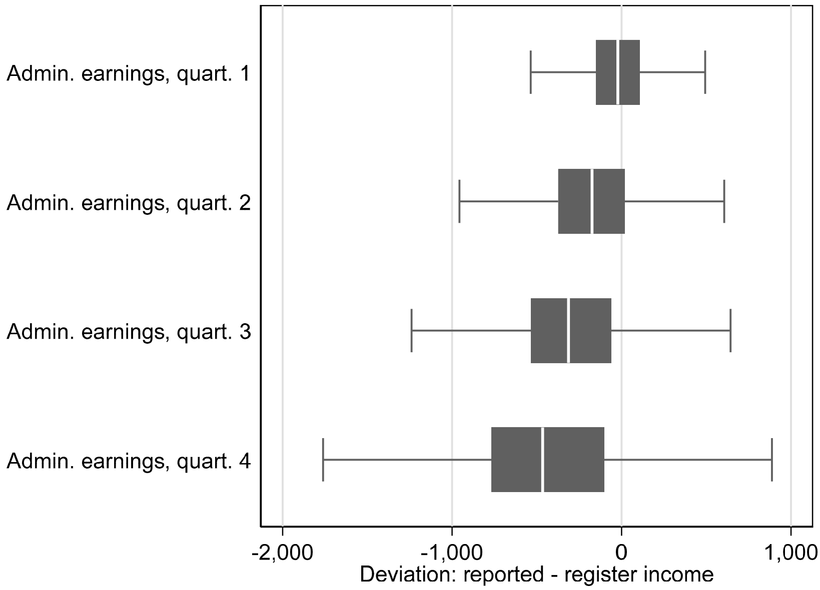

Figure 2 Box plots of deviations between the reported and administrative monthly gross earnings for each quartile of the administrative earnings (excluding outliers; in Euro)

Bias in Reported Earnings

We next examine the extent of bias in case that the open-ended earnings question is answered. Therefore, we calculate two basic model specifications with the dependent binary variable "reported monthly gross earnings differ from administrative earnings by more than 20%". The first specification ("restricted model", left column of Table 4) displays the average marginal effects for respondent and interviewer characteristics without controlling for quartiles of the earnings distribution. As existing literature reflects the relevance of the earnings quartiles (e.g., Angel et al. 2017; Meyer & Mittag 2017), we add respective dummies to the specification "basic model" (right column of Table 4). The results of both standard logit models are presented and discussed in the following.

In the restricted model we find an effect of gender that confirms our *hypothesis 2*: female respondents have a lower likelihood of deviating from their "true" earnings when answering the survey question. We do not find any effect regarding age (*hypothesis 3*). However, there is a clear educational effect. Highly educated respondents are less likely to report earnings that differ from their administrative earnings. This result contradicts our *hypothesis 4*. Considering the influence of personality traits, we find that extroverted persons show a higher tendency to inaccurately report their earnings, while the opposite is true for persons with a high score in agreeableness. These findings are in accordance with *hypotheses 5* and *6*. We also find significantly higher accuracy for interviews conducted via telephone, which supports our *hypothesis 12*. However, we do not find any significant effects for interviewer characteristics (*hypotheses 7* to *11*).

In the basic model most of these effects persist, except the effect of respondents' gender. Additionally, we find a clear tendency for respondents with the highest level of earnings to misreport, which is partly consistent with *hypothesis 1*. Both calculated measures of model fit (Pseudo R^2 and AIC) indicate that the basic model that controls for earnings quartiles is more appropriate. Thus, all subsequent analyses are based on this model specification.

Direction of Bias in Reported Earnings

To gain deeper insight into the topic, we analyze the direction of misreporting as a categorical dependent variable (underreporting vs. no deviation vs. overreporting) in a multinomial logit model. Table 5 presents the results of this model, with non-deviating respondents being the baseline category for the multinomial calculation.

Table 4 Logit regressions, basic model and restricted model without monthly gross earnings quartiles as control variables, results displayed as average marginal effects

	Restricted model		Basic model	
Resp. gender (ref.: male)				
Female	-0.024**	(-3.17)	-0.015	(-1.94)
Resp. age (ref.: aged 29 and lower)				
Aged 30-49	-0.006	(-0.51)	-0.021	(-1.67)
Aged 50 or older	0.005	(0.38)	-0.014	(-1.07)
Resp. education (ref.: schooling, no training)				
Lower secondary, voc. training	0.005	(0.25)	0.007	(0.38)
Intermediate, voc. training	-0.010	(-0.57)	-0.016	(-0.85)
Upper secondary, voc. training	-0.012	(-0.70)	-0.026	(-1.43)
Higher education degree	-0.036*	(-2.13)	-0.064***	(-3.46)
Admin. earnings (ref.: quart. 1)				
Admin. earnings, quart. 2			0.010	(1.03)
Admin. earnings, quart. 3			0.022	(1.91)
Admin. earnings, quart. 4			0.085***	(6.75)
Personality traits				
Big 5: Extraversion	0.010**	(2.72)	0.010**	(2.68)
Big 5: Agreeableness	-0.015*	(-2.42)	-0.013*	(-2.18)
Survey mode (ref.: CAPI)				
CATI	-0.030*	(-2.01)	-0.031*	(-2.13)
I: gender (ref.: male)				
I: female	0.004	(0.52)	0.004	(0.46)
I: age (ref.: aged 29 and lower)				
I: aged 30-49	-0.018	(-1.62)	-0.018	(-1.69)
I: aged 50-65	0.003	(0.24)	0.004	(0.32)
I: aged older than 65	-0.012	(-0.69)	-0.012	(-0.69)
I: education (ref.: lower secondary)				
I: intermediate	0.019	(1.17)	0.019	(1.20)
I: upper secondary	-0.001	(-0.09)	-0.001	(-0.06)
I: experience (ref.: exp. below 2 years)				
I: exp. 2-3 years	0.002	(0.19)	0.001	(0.14)
I: exp. 4-5 years	-0.008	(-0.70)	-0.008	(-0.67)
I: exp. 6 or more years	0.001	(0.07)	0.001	(0.06)
I: no. of interviews cond. so far	0.000	(1.84)	0.000	(1.77)

	Restricted model	Basic model
Pseudo R^2	0.013	0.020
AIC	12179	12105
Observations	14065	14065

Source: NEPS-SC6-ADIAB, own calculations.
*Note*s: Indicator for absolute deviation by >20% of administrative monthly gross earnings as dependent variable, z-statistics in parentheses. The constant and the following control variables are omitted from the table: region of birth, panel wave, public employer, paid overtime, special payments and child benefits. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 800 interviewers.

As expected, we find strongly diverging effects for underreporters and overreporters.

For instance, female respondents show a significantly higher tendency to underreport their earnings than their male peers, whilst they simultaneously show a clear tendency to overreport less often. By differentiating the direction of bias in reported earnings, we can see that the former support of *hypothesis 2* was driven mainly by females' non-overreporting behavior.

In contradiction to the results from Table 4, we find significant effects for the respondent's age: Compared to persons below the age of 30 years, older respondents are less likely to underreport their earnings. Simultaneously, the older age groups show a higher tendency to overreport their earnings. This finding now partly supports *hypothesis 3*, while the basic model did not reveal such an effect. The same is true for the impact of education on misreporting. We see that persons with upper secondary or higher educational degrees are less likely to underreport but more likely to overreport their earnings. This result partly supports *hypothesis 4*.

The enhanced model also shows that the effects of the respondents' personality traits, as found earlier, are driven only by the overreporting respondents. Extraverted persons show a significantly higher likelihood to overstate their earnings, and agreeable respondents show a reduced likelihood to do so. This supports our *hypotheses 5* and *6*. Nevertheless, the former effect of a lower likelihood of misreporting in CATI mode vanishes (*hypothesis 12*). Again, we do not see any effects of interviewer characteristics (*hypotheses 7* to *11*).

Table 5 Multinomial logit regressions to differentiate between over- and underreporting, results displayed as average marginal effects

	Underreporting		Overreporting	
Resp. gender (ref.: male)				
Female	0.024**	(3.18)	-0.038***	(-10.44)
Resp. age (ref.: aged 29 and lower)				
Aged 30-49	-0.049***	(-3.97)	0.016**	(3.02)
Aged 50 or older	-0.054***	(-4.31)	0.027***	(4.44)
Resp. education (ref.: schooling, no training)				
Lower secondary, voc. training	-0.000	(-0.02)	0.006	(0.70)
Intermediate, voc. training	-0.024	(-1.26)	0.005	(0.56)
Upper secondary, voc. training	-0.052**	(-2.92)	0.024*	(2.42)
Higher education degree	-0.087***	(-4.75)	0.029**	(2.89)
Admin. earnings (ref.: quart. 1)				
Admin. earnings, quart. 2	0.050***	(8.14)	-0.056***	(-6.75)
Admin. earnings, quart. 3	0.090***	(9.81)	-0.082***	(-10.27)
Admin. earnings, quart. 4	0.166***	(15.88)	-0.081***	(-9.71)
Personality traits				
Big 5: Extraversion	-0.005	(-1.56)	0.015***	(5.88)
Big 5: Agreeableness	-0.005	(-0.87)	-0.008**	(-2.61)
Survey mode (ref.: CAPI)				
CATI	-0.022	(-1.66)	-0.009	(-1.39)
I: gender (ref.: male)				
I: female	-0.001	(-0.15)	0.005	(1.41)
I: age (ref.: aged 29 and lower)				
I: aged 30-49	-0.015	(-1.75)	-0.003	(-0.47)
I: aged 50-65	0.003	(0.29)	0.001	(0.08)
I: aged older than 65	-0.007	(-0.50)	-0.004	(-0.47)
I: education (ref.: lower secondary)				
I: intermediate	0.021	(1.41)	-0.004	(-0.53)
I: upper secondary	-0.001	(-0.12)	-0.000	(-0.03)
I: experience (ref.: exp. below 2 years)				
I: exp. 2-3 years	-0.002	(-0.24)	0.003	(0.59)
I: exp. 4-5 years	0.001	(0.08)	-0.008	(-1.42)
I: exp. 6 or more years	0.008	(0.84)	-0.007	(-1.27)
I: no. of interviews conducted so far	0.000	(1.43)	0.000	(1.60)

	Underreporting	Overreporting
Pseudo R^2	0.065	
AIC	14291	
Observations	14065	

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Indicator for over-/underreporting by >20% of administrative monthly gross earnings as dependent variable, z-statistics in parentheses. The constant and the following control variables are omitted from the table: region of birth, panel wave, public employer, paid overtime, special payments and child benefits. ***, **, ***** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 800 interviewers.

Finally, we find substantive effects of the earnings quartiles. Compared to persons from the lowest quartile (the reference category), belonging to a higher quartile of the earnings distribution is correlated with a higher tendency of underreporting and a lower tendency of overreporting. This finding again only partly supports our *hypothesis 1.* Moreover, it suggests that there may be different mechanisms of report bias, and different directions of bias, in different parts of the earnings distribution. As the bias does not appear to be randomly distributed, we interpret it as evidence for deliberate misreporting.

Bias in Reported Earnings Across the Earnings Distribution

To account for the strong impact of the earnings quartiles in the former models and to follow an approach similar to Kim and Tamborini (2014), we recalculate our multinomial model separately for each of the four quartiles of the earnings distribution. These models are presented in Table 6. We find a consistent gender effect, indicating that females are less likely to overreport across all earnings quartiles, supporting *hypothesis 2*. Yet, in the two earnings groups below the median, the model reveals an increased likelihood for females to underreport their earnings.

The effects of the respondent's age are somehow contradictory. While older persons are less likely than respondents below the age of 30 years to underreport in earnings quartiles 2 and 4, this result does not hold for the other quartiles. Older respondents also show a higher tendency to overreport but only in the tail quartiles. Thus, *hypothesis 3* again appears to be only partly supported. However, these effects foster the idea of a U-shaped pattern of misreporting across the earnings distribution, dependent on the age groups, which indirectly supports *hypothesis 1.*

The effect of higher education making respondents less likely to underreport but more likely to overreport is clearly driven only by the second quartile of the

Table 6 Multinomial logit regressions to differentiate between over- and underreporting, estimated separately by quartiles of administrative earnings, results displayed as average marginal effects

	Earnings quart. 1		Earnings quart. 2		Earnings quart. 3		Earnings quart. 4	
	Underrep.	Overrep.	Underrep.	Overrep.	Underrep.	Overrep.	Underrep.	Overrep.
Resp. gender (ref.: male)								
Female	0.023 **	-0.058***	0.037**	-0.037***	0.004	-0.027***	0.021	-0.032***
	(2.902)	(-5.801)	(2.774)	(-4.677)	(0.263)	(-5.358)	(1.208)	(-5.423)
Resp. age (ref.: aged 29 and lower)								
Aged 30-49	-0.003	0.026*	-0.070***	0.018	-0.050	-0.003	-0.082*	0.029**
	(-0.252)	(2.043)	(-3.525)	(1.753)	(-1.808)	(-0.305)	(-2.000)	(2.997)
Aged 50 or older	-0.019	0.045**	-0.049*	0.029*	-0.044	0.017	-0.113**	0.030**
	(-1.483)	(2.984)	(-2.249)	(2.466)	(-1.588)	(1.485)	(-2.691)	(2.936)
Resp. education (ref.: schooling, no training)								
Lower secondary, voc. training	-0.008	-0.003	-0.047	0.036**	0.009	-0.001	0.053	-0.007
	(-0.519)	(-0.162)	(-1.266)	(2.841)	(0.192)	(-0.042)	(1.150)	(-0.333)
Intermediate, voc. training	-0.005	0.006	-0.084*	0.032**	-0.010	-0.007	0.007	-0.001
	(-0.279)	(0.277)	(-2.396)	(2.716)	(-0.239)	(-0.413)	(0.159)	(-0.044)
Upper secondary, voc. training	-0.010	0.027	-0.092**	0.049***	-0.037	0.009	-0.064	0.019
	(-0.570)	(1.062)	(-2.688)	(3.334)	(-0.839)	(0.473)	(-1.534)	(0.945)
Higher education degree	-0.008	0.052	-0.121***	0.041*	-0.091*	-0.005	-0.103*	0.026
	(-0.413)	(1.784)	(-3.671)	(2.559)	(-2.106)	(-0.298)	(-2.495)	(1.259)

	Earnings quart. 1		Earnings quart. 2		Earnings quart. 3		Earnings quart. 4	
	Underrep.	Overrep.	Underrep.	Overrep.	Underrep.	Overrep.	Underrep.	Overrep.
Personality traits								
Big 5: Extraversion	-0.002	0.024***	-0.000	0.012**	-0.017**	0.011***	-0.002	0.011*
	(-0.689)	(3.597)	(-0.002)	(2.582)	(-2.960)	(3.330)	(-0.254)	(2.574)
Big 5: Agreeableness	-0.002	-0.024**	0.001	0.009	-0.025*	-0.016**	0.001	-0.003
	(-0.278)	(-3.082)	(0.059)	(1.161)	(-2.021)	(-3.023)	(0.057)	(-0.457)
Pseudo R^2	0.053		0.066		0.071		0.057	
AIC	3402		3514		3396		4029	
Observations	3517		3516		3516		3516	

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Indicator for over-/underreporting by >20% of administrative monthly gross earnings as dependent variable, z-statistics in parentheses. The constant, all interviewer variables from the basic model in table 4 and the following control variables are omitted from the table: region of birth, public employer, paid overtime, special payments, child benefits, survey mode and panel wave. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 800 interviewers.

earnings distribution. The model shows no significant effects for the corresponding variables in the lowest earnings group and only a slight similar effect for highly educated respondents in the two upper quartiles. This result still supports our *hypothesis 4* but only for one earnings group. Also, the results contradict the findings of Kim and Tamborini (2014).

Regarding the personality trait of "extraversion", the analyses show results consistent with those of the joint estimation. The tendency of extraverted persons to be more likely to overreport their earnings can be seen across all earnings quartiles, aided by a small decrease in the likelihood of underreporting in the third quartile, which supports *hypothesis 5*. More "agreeable" persons, however, do not show a consistently smaller likelihood to overreport across all earnings groups. By contrast, only persons in the first and third earnings quartiles show an effect of this kind, conveyed by the lower tendency to underreport in the third quartile. This result only weakly supports *hypothesis 6*.

Interviewer Effects

In the literature on interviewer effects, several studies suggest estimating multilevel models to calculate the extent of the interviewer's impact (see, e.g., O'Muircheartaigh & Campanelli 1999; Lipps & Pollien 2011; Korbmacher & Schroeder 2013). In the next step, we follow this approach to validate our previous findings of any interviewer effects. Given that we have thus far found little evidence for interviewer effects on reporting accuracy, it is not surprising that the intraclass correlation coefficient (ICC) is very low (0.0234). This result indicates that very little of the variance in our misreport indicator is explained by interviewer characteristics.

Moreover, the literature suggests possible influences not only of the interviewer characteristics themselves but also of the similarity in socio-demography between the interviewer and the respondent, as we have stated in *hypotheses 7* to *9* (e.g., Diekmann 2008; Lipps & Lutz 2017; West & Blom 2017). Thus, we recalculate both the basic and multinomial models after introducing dummy variables representing similarity in gender, age and education between both interview counterparts. This newly introduced indicators presented in Table 7 suggest an effect of the educational similarity between the interviewer and respondent, showing that interviewers who are more or less educated than their respective respondents are more likely to elicit underreported answers to the earnings question. This notable effect corroborates *hypothesis 9*. For the similarity in gender and age, we however do not find any support (*hypotheses 7* and *8*). Additionally, we do not see evidence to confirm our assumptions that the overall interviewing experience (*hypothesis 10*) or the interviewers' familiarity with the NEPS SC6 survey instrument (*hypothesis 11*) reduce the tendency to misreport.

Table 7 Logit and multinomial logit regressions, basic model estimated with additional indicators for difference between respondent and interviewer, results displayed as average marginal effects

	Logit		Underreporting		Overreporting	
I: experience (ref.: exp. below 2 years)						
I: exp. 2-3 years	0.002	(0.17)	-0.002	(-0.18)	0.003	(0.59)
I: exp. 4-5 years	-0.005	(-0.46)	0.003	(0.27)	-0.008	(-1.34)
I: exp. 6 or more years	0.001	(0.12)	0.009	(0.87)	-0.007	(-1.21)
I: no. of interviews conducted so far	0.000*	(1.98)	0.000	(1.70)	0.000	(1.69)
I: gender disparity (ref.: same gender as resp.)						
I: different gender than respondent	0.001	(0.16)	-0.003	(-0.58)	0.005	(1.41)
I: age disparity (ref.: same age class as resp.)						
I: higher age class than respondent	-0.005	(-0.56)	-0.005	(-0.59)	-0.001	(-0.14)
I: lower age class than respondent	-0.004	(-0.41)	-0.007	(-0.99)	0.003	(0.48)
I: educational disparity (ref.: same schooling as resp.)						
I: higher schooling than respondent	0.023*	(2.31)	0.023**	(2.59)	-0.000	(-0.01)
I: lower schooling than respondent	0.022*	(2.27)	0.027**	(2.91)	-0.005	(-0.94)
Survey mode (ref.: CAPI)						
CATI	-0.038*	(-2.44)	-0.027	(-1.92)	-0.011	(-1.65)
Pseudo R^2	0.019		0.065			
AIC	12110		14290			
Observations	14065		14065			

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Indicator for absolute deviation by >20% or for over-/underreporting by >20%, respectively, of administrative monthly gross earnings as dependent variable, z-statistics in parentheses. Interviewer characteristics sex, age and schooling are removed from the model. The constant and the remaining independent variables from the basic model in Table 4 are omitted from the table. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 800 interviewers.

Summary and Conclusions

We used linked survey data from the German National Education Panel Study's Starting Cohort "Adults" (NEPS SC6) and administrative data from the German Federal Employment Agency to estimate and analyze the drivers giving rise to measurement error in monthly gross earnings based on a sequence of logistic and multinomial logistic models. Constraints in comparability between the earnings measures in both data sources lessen the generalizability of our results. Following the latest validation studies (see, e.g., Kim & Tamborini 2014; Angel et al. 2017), we are able to classify inaccurate responses as either over- or underreporting. Gaining insight into the different mechanisms driving these two kinds of misreporting, we show that the higher response accuracy of female respondents is driven by a reduced tendency to overreport, while the inaccuracy effects for older and better-educated respondents are primarily driven by a reduced likelihood to underreport earnings. Moreover, the reporting inaccuracy of extraverted persons results from a higher tendency to overreport, whereas agreeable respondents are less likely to follow this pattern.

In regressions separated by earnings groups, we find mainly consistent effects of gender, age and personality traits. The education level effect persists only in the second quartile of the earnings distribution.

None of our calculations indicate important direct effects of interviewer characteristics on either reducing or amplifying the tendency to misreport. This may be an indication of highly competent field personnel and, if so, is good news in general for users of NEPS survey data. However, we find evidence that interviewers with education levels similar to those of their respective respondents may elicit more accurate results and, especially, reduce the risk of underreporting. This can be interpreted as the result of respondents' tendency to provide socially desirable answers.

In addition to all aspects covered by this article, cognitive factors may also affect the reporting of income. To validly answer a question about earnings, the respondent must at least pass through the cognitive stages of interpretation or understanding, retrieval, and response production (cp. Tourangeau 1984; Groves et al. 2009). Cognitive effects, if present, might be misinterpreted as an influence of socially desirable behavior. Thus, further analyses should aim to make use of competency assessment data to approximate these cognitive aspects and to narrow down the subset of misreporting that is truly due to social desirability.

References

Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., Leuze, K., Matthes, B., Pollak, R. & Ruland, M. (2011). Adult education and lifelong learning. *Zeitschrift für Erziehungswissenschaft*, 2 (Special Issue 14), 283–299. https://doi.org/10.1007/s11618-011-0197-0

Angel, S., Heuberger, R., & Lamei, N. (2017). Differences between household income from surveys and registers and how these affect the poverty headcount: evidence from the Austrian SILC. *Social Indicators Research*. https://doi.org/10.1007/s11205-017-1672-7

Antoni, M., Bachbauer, N., Eberle, J., & Vicari, B. (2018). *NEPS-SC6 survey data linked to administrative data of the IAB (NEPS-SC6-ADIAB 7515)* (FDZ-Datenreport No. 02/2018 EN). Institut für Arbeitsmarkt- und Berufsforschung (IAB). Nürnberg. https://doi.org/10.5164/IAB.FDZD.1802.en.v1

Antoni, M., Drasch, K., Kleinert, C., Matthes, B., Ruland, M., & Trahms, A. (2010). *Working and learning in a changing world. Part I: overview of the study* (FDZ Methodenreport No. 05/2010). Institut für Arbeitsmarkt- und Berufsforschung (IAB). Nürnberg.

Antoni, M., Ganzer, A., & vom Berge, P. (2016). *Sample of integrated labour market biographies (SIAB) 1975-2014* (FDZ-Datenreport No. 04/2016(en)). Institut für Arbeitsmarkt- und Berufsforschung (IAB). Nürnberg.

Education as a Lifelong Process: The German National Educational Panel Study (NEPS). (2011). In H.-P. Blossfeld, H. G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft (Vol. 2, Special Issue 14)*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bollinger, C. R. (1998). Measurement error in the current population survey: a nonparametric look. *Journal of Labor Economics*, 16(3), 576–594. https://doi.org/10.1086/209899

Bollinger, C. R., Hirsch, B., Hokayem, C. M., & Ziliak, J. P. (2018). *Trouble in the tails? What we know about earnings nonresponse thirty years after Lillard, Smith, and Welch* (IZA Discussion Paper No. 11710). IZA – Institute of Labor Economics. Bonn.

Bound, J., Brown, C., Duncan, G. J., & Rodgers, W. L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 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? *Journal of Labor Economics*, 9 (1), 1–24.

Bricker, J. & Engelhardt, G. V. (2008). Measurement error in earnings data in the health and retirement study. *Journal of Economic and Social Measurement*, 33(1), 39–61.

Cobb-Clark, D. A. & Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, 115(1), 11–15. https://doi.org/10.1016/j.econlet.2011.11.015

DeMaio, T. J. (1984). Social desirability and survey measurement: a review. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 257–282). Russell Sage Foundation.

Deutscher Bundestag. (1998). Criminal Code in the version promulgated on 13 November 1998, Federal Law Gazette [Bundesgesetzblatt] I p. 3322, last amended by Article 1 of the law of 24 September 2013, Federal Law Gazette I p. 3671 and with the text of Article 6(18) of the law of 10 October 2013, Federal Law Gazette I p 3799.

Diekmann, A. (2008). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen* (19. Aufl.). Reinbek bei Hamburg: Rowohlt.

Drews, N., Groll, D., & Jacobebbinghaus, P. (2007). *Programmierbeispiele zur Aufbereitung von FDZ Personendaten in Stata* (FDZ Methodenreport Nr. 06/2007). Institut für Arbeitsmarkt- und Berufsforschung (IAB). Nürnberg.

Duncan, G. J. & Hill, D. H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, 3(4), 508–532.

Engel, U., Jann, B., Lynn, P., & Scherpenzeel, A. (Eds.). (2014). *Improving survey methods: Lessons from recent research*. European Association of Methodology Series. London: Routledge.

Essig, L. & Winter, J. (2009). Item non-response to financial questions in household surveys: an experimental study of interviewer and mode effects. *Fiscal Studies*, 30(4), 367–390.

Gottschalk, P. & Huynh, M. (2005). Validation study of earnings in the SIPP – do older workers have larger measurement error? *Center for Retirement Research Working Papers*, (2005-07).

Gottschalk, P. & Huynh, M. (2010). Are earnings inequality and mobility overstated? The impact of nonclassical measurement error. *Review of Economics and Statistics*, 92(2), 302–315. https://doi.org/10.1162/rest.2010.11232

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, New Jersey: Wiley.

Heineck, G. & Anger, S. (2010). The returns to cognitive abilities and personality traits in Germany. *Labour Economics*, 17 (3), 535–546. https://doi.org/10.1016/j.labeco.2009.06.001

Jann, B. (2005). Making regression tables from stored estimates. *Stata Journal*, 5(3), 288–308.

Jann, B. (2014). Asking sensitive questions: overview and introduction. In U. Engel, B. Jann, P. Lynn, & A. Scherpenzeel (Eds.), Improving survey methods: Lessons from recent research (pp. 101–105). European Association of Methodology Series. London: Routledge.

Kapteyn, A. & Ypma, J. Y. (2007). Measurement error and misclassification: a comparison of survey and administrative data. *Journal of Labor Economics*, 25(3), 513–551.

Kim, C. & Tamborini, C. R. (2014). Response error in earnings: an analysis of the survey of income and program participation matched with administrative data. *Sociological Methods & Research*, 43(1), 39–72. https://doi.org/10.1177/0049124112460371

Kirkcaldy, B. D., Furnham, A. F., & Lynn, R. A. (1992). National differences in work attitudes between the UK and Germany. *European Work and Organizational Psychologist*, 2(2), 81–102. https://doi.org/10.1080/09602009208408537

Korbmacher, J. M. & Schroeder, M. (2013). Consent when linking survey data with administrative records: the role of the interviewer. *Survey Research Methods*, 7(2), 115–131. https://doi.org/10.18148/srm/2013.v7i2.5067

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47 (4), 2025–2047.

Kühne, S. (2018). From strangers to acquaintances? Interviewer continuity and socially desirable responses in panel surveys. *Survey Research Methods*, 12(2), 121–146. https://doi.org/10.18148/srm/2018.v12i2.7299

Landrock, U. (2017). How interviewer effects differ in real and falsified survey data: using multilevel analysis to identify interviewer falsifications. *methods, data, analyses*, 11(2), 163–188. https://doi.org/10.12758/mda.2017.03

Lipps, O. & Lutz, G. (2017). Gender of interviewer effects in a multi-topic centralized CATI panel survey. *methods, data, analyses*, 11(1), 67–86. https://doi.org/10.12758/mda.2016.009

Lipps, O. & Pollien, A. (2011). Effects of interviewer experience on components of nonresponse in the European Social Survey. *Field Methods*, 23(2), 156–172. https://doi.org/10.1177/1525822x10387770

Meyer, B. D. & Mittag, N. (2017, August). Using linked survey and administrative data to better measure income: implications for poverty, program effectiveness and holes in the safety net (IZA Discussion Paper No. 10943). IZA Institute of Labor Economics. Bonn.

Moore, J. C., Stinson, L. L., & Welniak, Jr., E. J. (2000). Income measurement error in surveys: a review. *Journal of Official Statistics*, 16(4), 331–361.

Mueller, G. & Plug, E. (2006). Estimating the effect of personality on male and female earnings. *ILR Review*, 60(1), 3–22. https://doi.org/10.1177/001979390606000101

National Research Council. (1984). *Cognitive aspects of survey methodology* (T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau, Eds.). Washington, D. C.: National Academies Press. https://doi.org/10.17226/930

O'Muircheartaigh, C. & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(3), 437–446. https://doi.org/10.1111/1467-985X.00147

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

Pedace, R. & Bates, N. (2000). Using administrative records to assess earnings reporting error in the survey of income and program participation. *Journal of Economic and Social Measurement*, 26(3,4), 173–192.

Preisendörfer, P. & Wolter, F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, 78(1), 126–146. https://doi.org/10.1093/poq/nft079

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: a 10-item short version of the Big Five inventory in english and german. *Journal of Research in Personality*, 41(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rantanen, J., Metsäpelto, R.-L., Feldt, T., Pulkkinen, L., & Kokko, K. (2007). Long-term stability in the big five personality traits in adulthood. *Scandinavian Journal of Psychology*, 48(6), 511–518. https://doi.org/10.1111/j.1467-9450.2007.00609.x

Riphahn, R. & Serfling, O. (2005). Item non-response on income and wealth questions. *Empirical Economics*, 30(2), 521–538.

Spurk, D. & Abele, A. E. (2011). Who earns more and why? A multiple mediation model from personality to salary. *Journal of Business and Psychology*, 26(1), 87–103. https://doi.org/10.1007/s10869-010-9184-3

Stocké, V. & Hunkler, C. (2007). Measures of desirability beliefs and their validity as indicators for socially desirable responding. *Field Methods*, 19 (3), 313–336. https://doi.org/10.1177/1525822X07302102

Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology*. Washington, D. C.: National Academies Press. https://doi.org/10.17226/930

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Trachtman, R. (1999). The money taboo: its effects in everyday life and in the practice of psychotherapy. *Clinical Social Work Journal*, 27(3), 275–288. https://doi.org/10.1023/A:1022842303387

Turner, C. F. & Martin, E. (Eds.). (1984). *Surveying subjective phenomena*. Russell Sage Foundation.

West, B. T. & Blom, A. G. (2017). Explaining interviewer effects: a research synthesis. *Journal of Survey Statistics and Methodology*, 5(2), 175–211. https://doi.org/10.1093/jssam/smw024

West, B. T., Kreuter, F., & Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: a function of sampling, measurement error, or nonresponse? *Journal of Official Statistics*, 29 (2), 277–297. https://doi.org/10.2478/jos-2013-0023

Zinn, S. & Würbach, A. (2016). A statistical approach to address the problem of heaping in self-reported income data. *Journal of Applied Statistics*, 43(4), 682–703. https://doi.org/10.1080/02664763.2015.1077372

Appendix

Table A1 T-test of characteristics of respondents within and outside the estimation sample

	Not in est. sample	In est. sample	Difference	t
Resp. gender				
Male	0.444	0.675	-0.231***	-26.215
Female	0.556	0.325	0.231***	26.215
Resp. age				
Aged 29 and lower	0.061	0.046	0.015***	3.592
Aged 30-49	0.388	0.460	-0.072***	-8.177
Aged 50 or older	0.551	0.494	0.057***	6.370
Region of birth				
West Germany	0.708	0.660	0.048***	5.772
East Germany	0.180	0.250	-0.070***	-9.759
Abroad	0.112	0.090	0.022***	3.954
Resp. education				
Schooling, no training	0.095	0.046	0.049***	9.868
Lower secondary, voc. training	0.204	0.190	0.014	1.910
Intermediate, voc. training	0.294	0.360	-0.066***	-7.966
Upper secondary, voc. training	0.142	0.174	-0.032***	-4.912
Higher education degree	0.265	0.230	0.035***	4.475
Personality traits				
Big 5: Extraversion	3.411	3.343	0.068***	3.899
Big 5: Agreeableness	3.597	3.534	0.063***	5.532
Survey mode				
CAPI	0.314	0.250	0.064***	7.791
CATI	0.686	0.750	-0.064***	-7.791

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Number of respondents: 16873. Unlike all other tables, this analysis only includes the most recent observation for each respondent instead of including all valid observations of respondents within the observation period. *** indicates significance at the 0.1% level.

Table A2 Logit regression on the availability of an open gross earnings report, results displayed as average marginal effects

	Dep. var.: open earnings	
Resp. gender (ref.: male)		
Female	-0.018**	(-3.05)
Resp. age (ref.: aged 29 and lower)		
Aged 30-49	-0.021**	(-2.80)
Aged 50 or older	-0.033***	(-4.49)
Resp. education (ref.: schooling, no training)		
Lower secondary, voc. training	-0.018	(-1.51)
Intermediate, voc. training	-0.005	(-0.43)
Upper secondary, voc. training	0.000	(0.04)
Higher education degree	0.009	(0.75)
Admin. earnings (ref.: quart. 1)		
Admin. earnings, quart. 2	-0.017**	(-2.63)
Admin. earnings, quart. 3	-0.010	(-1.67)
Admin. earnings, quart. 4	-0.035***	(-4.94)
Personality traits		
Big 5: Extraversion	-0.000	(-0.12)
Big 5: Agreeableness	0.006	(1.50)
Survey mode (ref.: CAPI)		
CATI	0.048***	(4.08)
I: gender (ref.: male)		
I: female	0.008	(1.16)
I: age (ref.: aged 29 and lower)		
I: aged 30-49	-0.012	(-1.33)
I: aged 50-65	-0.010	(-0.99)
I: aged older than 65	-0.003	(-0.21)
I: education (ref.: lower secondary)		
I: intermediate	-0.019	(-1.55)
I: upper secondary	-0.013	(-1.42)
I: experience (ref.: exp. below 2 years)		
I: exp. 2-3 years	0.020*	(2.11)
I: exp. 4-5 years	0.020	(1.91)
I: exp. 6 or more years	0.023*	(2.13)
I: no. of interviews conducted so far	-0.000	(-1.87)

	Dep. var.: open earnings
Pseudo R^2	0.035
AIC	7659
Observations	15162

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Indicator for the availability of a valid response to the open question on gross earnings as dependent variable, z-statistics in parentheses. The constant and the following control variables are omitted from the table: region of birth, panel wave, public employer, paid overtime, special payments and child benefits. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 808 interviewers. Contrary to all other regressions, this regression also includes observations without valid responses to the open-ended question on gross earnings. The number of observations and interviewers is therefore higher than in all other tables. The administrative monthly gross earnings quartiles as control variables are recalculated to accommodate the different sample size.