I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17332

Noncognitive Human Capital and Misreporting Behavior in Online Surveys

Haizheng Li
Qinyi Liu
Yiting Xu

SEPTEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17332

Noncognitive Human Capital and Misreporting Behavior in Online Surveys

Haizheng Li
Georgia Institute of Technology and IZA

Qinyi Liu
University of Nottingham Ningbo China

Yiting Xu
Central University of Finance and Economics

SEPTEMBER 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Noncognitive Human Capital and Misreporting Behavior in Online Surveys*

In this study, we investigate misreporting behavior in online surveys based on the field experiments in a large-scale online training program for rural teachers. We link the digitally recorded data with survey responses and integrate randomized controlled trials (RCTs) in the survey design. Noncognitive human capital is measured using both self-reported personality traits and proxies based on observed behaviors. Our results show that the impact of observed individual characteristics varies depending on the nature of the question and survey specifics. Unobserved heterogeneity affects both survey participation and response accuracy, resulting in sample selectivity. Noncognitive human capital inferred from observed behaviors consistently shows important influence on misreporting, while that measured by self-reported personality traits suffers from the same misreporting problem. However, behavior proxy may also capture factors external to survey respondents, and it is important to separate the effect of noncognitive human capital from the external impacts. Additionally, survey design affects misreporting. Therefore, improving the efficiency of survey such as by changing the saliency and optimizing the sequence of questions, can improve survey quality. These findings carry important implications for using survey data and for improving survey data quality.

JEL Classification: D91, D83, C93, J24

Keywords: misreporting, noncognitive human capital, survey design, RCT intervention

Corresponding author:
Yiting Xu
China Center for Human Capital and Labor Market Research
Central University of Finance and Economics
39 South College Road
Haidian District
Beijing
China

E-mail: ytxu12@email.cufe.edu.cn

* Partial financial support was provided by the National Natural Science Foundation of China (Grant # 72273163 and #72203041). We have benefited greatly from the research team at the China Center for Human Capital and Labor Market Research, Central University of Finance and Economics. We are grateful for the helpful comments and suggestions from session participants at various conferences.

I. Introduction

Survey data have been widely used in research. However, respondents may not report the information accurately in a survey. Misreporting can lead to various measurement errors in the data collected.[1] Such reporting errors can result in inconsistent estimation and potentially leading to misinformed policy implications (Bertrand and Mullainathan, 2001; Bound et al., 2001; Mittag, 2019).

The rapid growth of online surveys has intensified the concerns about misreporting in the data collected. An individual's participation in an online survey is often voluntary. When to do the survey and how much time respondents spend for completing the survey are beyond the control of survey administrators. Compared to an in-person survey (face-to-face or phone interviews), online surveys are more susceptible to random influences, ranging from respondents' heterogeneity to external factors related to the survey. Therefore, understanding misreporting behavior and its implications becomes even more important in utilizing online survey data (Abay et al., 2023).

In this study, we investigate misreporting behavior in online surveys. The novel aspect is that we conduct field experiments in an ongoing large-scale online training program for rural teachers, by integrating all experiments in their routine surveys for the program evaluation. This setup creates a real-world environment for online survey similar to those widely used elsewhere. We focus on three research questions. First, is the pattern of misreporting behavior consistent across different types of survey questions and surveys? Second, does the voluntary participation of the survey create sample selectivity that the participation is correlated with misreporting? Third, how does an individual's heterogeneity affect misreporting?

This study aims to contribute to the literature in several ways. We investigate misreporting behavior of the same group of potential participants across multiple surveys, focusing on questions covering the same subject but varying in image value and saliency. It helps us get useful information about whether findings from a single survey or question in existing studies can be generalized.

Moreover, we assess the effects of individual heterogeneity on both selection into survey participation (sampling error) and misreporting in the survey (measurement error). We capture individual characteristics through a broad measure of unobserved heterogeneity as

[1] Studies find that people misreport various information, such as weight and height (Burke and Carman, 2017), education (Black et al., 2003), employment status (Hu and Schennach, 2008; Abraham et al., 2013; Feng and Hu, 2013; Hu, 2017), loan usage (Karlan, 2012), government benefits (Martinelli and Parker, 2009; Meyer et al., 2015; Mittag, 2019), investment portfolio's expected return (Tergiman and Villeval, 2023), and income (Abowd and Stinson, 2013), etc.

well as more specific noncognitive human capital, based on personality traits and observed behaviors.

Additionally, we design randomized controlled trials (RCTs) to examine how survey condition affects reporting accuracy. Our experiments focus on the efficiency of survey questions and survey settings to test cognitive and noncognitive responses. We also explore the interplay between individual heterogeneity and survey designs to further identify the true effects of noncognitive human capital using appropriate econometric techniques. Finally, our data are based on teachers, instead of the general population in other studies. The findings on misreporting behavior of this highly educated profession will provide additional implications.

Many studies have investigated reporting errors in surveys. However, in examining how individual heterogeneity affects misreporting, they often use only observed characteristics, such as education (Black et al., 2003), family background (Meyer et al., 2022), or professional identity (Cohn et al., 2014), or only focus on a single unobserved trait, such as preferences for truthfulness (Gibson et al., 2013) or survey attentiveness (Read et al., 2022). We investigate both observed and unobserved heterogeneity, using multiple measures of noncognitive human capital.

Additionally, studies have designed experiment to examine misreporting behavior, such as using signing (Shu et al., 2012), altering question formats (Eckman and Kreuter, 2018; Kuhn and Vivyan, 2022), changing survey administration condition (Chen et al., 2020), and manipulating survey fatigue (Jeong et al., 2023). Our experiments differ from existing studies by incorporating information hints and varying the framing (tone) of survey questions. They allow us to explore how an individual's cognitive and noncognitive functions respond to these interventions.

Our study is closely related to two recent studies. The first one is Dohmen and Jagelka (2024), which conducts a detailed investigation into the reliability of self-assessed measures of economic preferences and personality traits. Yet, their study lacks a true benchmark for misreporting and reporting errors, relying instead on the self-reported reliability measures. Our study creates multiple questions with true measure automatically recorded on a digital platform. The second one is Celhay et al. (2024), which provides a comprehensive study about misreporting behavior, including cognition, social desirability bias, and survey design. Our study builds on theirs by focusing on reporting errors, rather than just the probabilities of misreporting. Additionally, while their study is based on existing survey data, our study uses online surveys and expands the scope to include a broader set of noncognitive human capital.

Therefore, our study complements the existing studies in a variety of ways. Our results show that misreporting is predominantly characterized by overreporting (broadly defined as conforming to social norms), and the pattern varies across different question types and survey conditions. Survey designs, such as the saliency of the question and information hints, have a strong effect on misreporting. In addition, individual heterogeneity affects both misreporting and survey participation, resulting in sample selectivity bias. The influence of individual characteristics, observed and unobserved, varies across questions and surveys depending on their specifics and their interplay. The estimated effects of noncognitive human capital measured through observed behavior proxies show a strong and consistent influence on misreporting, while self-reported personality traits are not reliable. However, observed behavior proxy may also capture other external effects, therefore additional techniques and data are needed to identify the true effects of noncognitive human capital.

The rest of the paper is organized as follows: Section II introduces a conceptual framework, and Section III describes study design, surveys and data. Section IV examines misreporting patterns and the effects of observed individual characteristics. Section V examine the role of various measures of unobserved heterogeneity in misreporting. In Section VI, we apply various Tobit models to estimate the effect of noncognitive human capital. Section VII concludes.

II. Conceptual Framework on Misreporting

We build a simple conceptual framework for misreporting in a survey. Generally, three channels may affect misreporting: 1) an individual's ability to retrieve and report the true information, including cognitive function for recalling and noncognitive factors affecting how questions are responded; 2) an individual's willingness and preference for revealing the true information, influenced by factors such as self-image, social image/desirability concerns; 3) survey design, such as saliency of subjects, style of questions, etc. (e.g., Bound et al., 2001; Groves et al., 2011; Celhay et al., 2024). Those factors affect reporting performance in a survey by balancing the benefits from accurate reporting against its costs (Ewers and Zimmerman, 2015; Dench and Joyce, 2022).

Assume that participation in a survey is voluntary without material incentives or punishments. There is no explicit monetary reward or cost associated with responding to the survey in a specific manner. Consider a survey question, where the true value of the information is q^*, participants report it as q, and their discrepancy $q_d = q - q^*$.

Survey participants decide on whether to accurate report or misreport. First, a respondent gains utility b from accurate reporting, as honesty can bring self-satisfaction (Rosaz and Villeval, 2012; Choshen-Hillel et al., 2020; Błachnio, 2021). We define such utility gain as $b(q_d)$, and $\partial b(|q_d|)/\partial|q_d| < 0$.

Individuals also gain utility from misreporting a higher type than the real value q^*. Define the value of high type $\overline{q}$ as more socially acceptable, i.e., if an individual reports the high type $\overline{q}$, it may lead to a positively perceived image (the value of social or self- image). Individuals generally incline to conform to social norms for preserving self-esteem (Krumpal, 2013; Bašić and Quercia, 2022).[2] Thus, we express the utility gain as $I(\overline{q}(q_d)|X, q^*)$. For example, if a positive discrepancy means performing better, such as for GPA, the function $\overline{q}(q_d)$ represents overreporting ($q_d > 0$); otherwise, e.g., for reporting crime, it represents underreporting ($q_d < 0$). Therefore, we have $\partial I(\overline{q}(q_d)|X, q^*)/\partial|q_d| > 0$. The larger the discrepancy, the higher the utility. Vector *X* represents observed individual characteristics that reflect socio-economic status and affect the perceived image value.[3]

However, searching for accurate information incurs cost *C*. The larger the effort *e* needed to reduce $|q_d|$, the higher the cost and the lower the utility. The utility cost can be defined as $C(e(|q_d|)|S, R)$, where S denotes cognitive and noncognitive abilities/skills, *R* for the survey condition. The larger the reporting error, the smaller the effort, which implies $\partial C(e(|q_d|)|S, R)/\partial|q_d| < 0$. Therefore, even without any intention, misreporting still exists because of the cost of responding accurately.

The effort cost depends on the participant's cognitive and noncognitive abilities. For example, a person with good memory can recall relevant information easily. More importantly, people with a high degree of conscientiousness tend to complete a survey carefully, and similarly, those with higher levels of perseverance are more likely to spend sufficient time to answer questions accurately. Since surveys are not exams, the cognitive demands are generally secondary, as they mainly involve recalling information. Therefore, noncognitive effort is expected to play a more dominating role.

The cost of efforts also depends on the survey condition *R*, such as the length of the questionnaire (Kilic and Sohnesen, 2019) and the degrees of saliency in the subject/content of

[2] These social desirability effects are more likely to occur when an actual interviewer is present. However, in an online survey setting, where the setup guarantees respondents' anonymity, social desirability effect may still exist, if some respondents want to look good to the researcher or for self-esteem (Weisberg, 2009).

[3] For example, studies find underreporting in receiving government benefits for the poor, because reporting a low income may incur shame of low financial standing (Meyer and Mittag, 2019; Stephens and Unayama, 2019). Wealthy people may underreport wealth as they want to downplay their financial success and mitigate negative social reactions.

the question (Celhay et al., 2024). Saliency refers to the perceived importance or relevance of a question's content to the respondent (Stern et al., 2012). Survey design can influence the effort required to answer questions (Bound et al., 2001).

Therefore, a respondent can maximize the utility U by choosing the value of q_d:

$$U(q_d) = b(|q_d|) + I(\overline{q}(q_d)|X, q^*) - C(e(|q_d|)|S, R). \quad (1)$$

The optimal misreporting level is determined by $q_d^* = f(X, S, R, q^*)$.[4] Given the focus of this paper, we conduct empirical analysis to explore the factors affecting misreporting behavior.

In the model, X and S are individual characteristics (both observed and unobserved) that affect misreporting. As discussed above, X denotes observed characteristics that affect the social or self-image value, and it also partially represents abilities, for example, education and job are indicators of an individual's abilities.

More importantly, S includes both cognitive and noncognitive abilities. Noncognitive abilities refer to a broad set of personality attributes, behaviors, and attitudes that are not directly related to cognitive intelligence (e.g., not measured by IQ tests) but are essential for success in life (Heckman et al., 2021). Noncognitive abilities include grit (perseverance and motivation), self-discipline, emotional resilience, etc. (Humphries and Kosse, 2017; Hoeschler et al., 2018). In a general sense, abilities are both inherited and learned, connotating different properties such as "traits" and "skills" (Heckman et al., 2018). These abilities belong to the broader concept of noncognitive human capital. Therefore, we refer to them as noncognitive human capital in this study to capture both unobserved abilities and acquired skills.

Noncogntive human capital governs how an individual generally treats tasks at hand and affects the attitude and behavior in a survey setting. To represent the noncognitive component of S, studies commonly use personality measures, such as Big Five personality traits. However, some noncognitive attributes—like time and risk preference, motivation and leadership—do not fit neatly within the Big Five framework (Becker et al., 2012; Hoeschler et al., 2018). Recent studies also utilize observed behavior as a proxy for noncognitive human capital. Observed behavior has advantages over self-reported personality measures, while its limitation is that it can also be influenced by external factors.

Misreporting involves the propensity to misreport and the magnitude of reporting error. The above model includes accurate reporting ($q_d = 0$), overreporting ($q_d > 0$) and

[4] It is possible to solve the above optimization problem with various assumptions of specific functional forms.

underreporting ($q_d < 0$). We will consider those different reporting behaviors in the empirical estimation.

III. Study Design, Surveys and Data

Based on the framework above, we include some key elements in designing this study. In particular, we create questions with different degrees of saliency, and then conduct multiple surveys with the same questions for the same pool of potential participants. Moreover, we incorporate random interventions.

We design this study and collect the data from a large-scale online training program for young teachers in rural China, called the Young Teacher Empowerment Program (hereafter referred to as "YTEP").[5] This annual training program aims for new teachers in rural schools to improve their teaching skills. The program started in 2017, and all courses are free of charge. It is not an official training program from the government. However, the program works with the local county bureau of education to promote participation. Participating teachers register for the program officially through the bureau, and most of them are recommended by their schools. Details about the program are listed in Appendix A.

We work with the research team invited to do third-party evaluation for the program. The research team can conduct independent surveys, design experiments, and interview training participants. Our data are collected from the YTEP program 2021. The program started in September 2021 and completed in June 2022. As usual, this year's program offers two compulsory courses: *Career Development* (in Fall semester) and *Teacher Ethics* (in Spring semester). Each compulsory course consists of 10 live lectures and 5 assignments, with one lecture per week.[6] Both courses have similar contents related to teaching, though with different titles. The program also offers elective courses in different subjects, such as teaching Math, Chinese, etc.

Participation in any of the lectures is voluntary, and there is no penalty for non-participation. Trainees will be eligible to receive a certificate when they get 75 credits or more out of 150. The credits can be earned by submitting the assignments of the compulsory courses, as well as through other program activities.

[5] The survey data from the YTEP have been used in various studies, e.g., Li et al. (2022).
[6] The liver lectures are offered on Wednesday evenings, starting at 7:00 pm and lasting for 1.5 hours.

3.1. Key questions for testing reporting behavior

We focus on the two compulsory courses, as they are designed for all trainees. We create two questions related to the courses: the number of assignments submitted, and the number of live lectures attended. These two questions have different levels of saliency and image value. Assignments are required for receiving the training certificate (accounting for up to 2/3 of the total credits), while lecture attendance is not. Additionally, recalling the number of lectures (total 10) is relatively more difficult than that of assignments (total 5). Therefore, assignments have a much higher level of saliency and image value than lectures for the trainees. When the event is more salient, the cognitive and noncognitive efforts required to retrieve accurate information are lower (Celhay et al., 2024), and the information is more likely to be accurately recalled (Groves, 2005; Weisberg, 2009; Adua and Sharp, 2010). Since the courses spread across two semesters, we can explore how the responses vary across surveys at different stages of the program.

More importantly, the true information about those two questions is available, as the training broadcast platform automatically records both assignment submission and lecture attendance digitally. The survey participants also know that their true values about those questions are available in the system.

3.2. Survey design

We designed three waves of survey at the beginning, mid-term, and the end of the program as part of a routine program evaluation, as shown in Appendix Table A1. Wave 1 is to collect trainees' individual information. We separate the survey for collecting personal information from the ones collecting responses on the two key questions above to mitigate the simultaneity problem. Wave 2 collects responses to the key questions in the first semester, and Wave 3 collects the same information in the second semester. Because Wave 3 is at the end of the program, the trainees might become less serious and motivated about the program. Therefore, we can examine how the changes in attitude may influence the responses.

In three survey waves, many other questions are asked for the purpose of program evaluation. Participation in any survey is voluntary with neither monetary awards nor penalties. The survey data are only accessible to the third-party evaluation team, and not shared with the training program organizer and the participants (including their employers).

Those surveys were implemented through an online survey program called SurveyStar, which is not related to the training program.[7] The questionnaires were sent to all registered in the training program through a link. Questions cannot be skipped when responding, i.e., it can move next question only after responding the question before it. The surveys usually remain open for 2-4 weeks.

3.3. Intervention experiments

We design several interventions to change the survey condition by altering the effort required to search for and recall information. We randomly assign the survey participants into different groups for various treatments. The random assignments were carried out by the SurveyStar system with its built-in function. In all waves, the respondents are unaware of any experiments in the survey.

3.3.1. RCT: Information hints

To assess the influence of cognitive processes on misreporting, we add information hints into the survey questions about the total number of assignments and lectures. Information hints may help improve recall and lessen cognitive strain, potentially reducing the incidence of misreporting (Kashner et al., 1999).

We employ both positively and negatively framed questions. The detailed questions are described in Appendix Table A2. Studies show that framing significantly affects decision-making (Abraham et al., 2020). Negatively framed questions are more effective in some cases because they can invoke negative emotional responses, which may motivate individuals to act to alleviate these feelings (Agrawal and Duhachek, 2010; Duhachek et al., 2012). As a result, questions framed negatively are often re-read more frequently and for a longer time than those framed positively.

We conduct the balance tests in Appendix Table A3. The results confirm that the control and treated groups are balanced for the demographic variables included in the model. Additionally, the correlation of the treatment groups assigned between Wave 2 and 3 is small, less than 0.08.

3.3.2. The sequence of questions in a survey

In addition to the RCT experiments, we also alter the sequence in which the questions were presented. In Wave 2, those key questions regarding the assignment completion and lecture attendance are placed as the 3rd and 4th questions among 122 questions. In Wave 3,

[7] SurveyStar is a web-based application for conducting online surveys, website: https://www.wjx.cn/.

they appear as the 37th and 38th out of 125 questions. Questions at the beginning of a survey are generally treated more carefully compared to those in the middle of a survey (Galesic and Bosnjak, 2009).

The extended duration in Wave 3 compared to Wave 2 before reaching the questions may have induced survey fatigue among respondents. The fatigue can impair both cognitive and noncognitive functioning, making respondents less efficient in recalling details and less careful in reading the survey questions. This experiment enables us to explore the interplay between individual heterogeneity and survey conditions.

3.4 Measuring noncognitive human capital—The Big Five personality traits

We employ the widely used Big Five Personality Model to measure noncognitive human capital. Based on Costa and McCrae (1992), the Big Five model comprises of five dimensions: Conscientiousness, Openness, Extraversion, Agreeableness, and Emotional Instability (or Neuroticism). For example, Conscientiousness influences how well an individual understands survey tasks and the level of care he/she takes in providing accurate responses. Sutin and Terracciano (2016) finds that individuals with a high level of perceived Conscientiousness did not under- or overreport their weight or height. Dohmen and Jagelka (2024) shows that four of the Big Five personality traits (except for Extraversion) are positively associated with the reliability of reported answers.

We assess the Big Five personality traits in both Wave 1 and Wave 3 surveys. The reason for doing such an assessment twice within six months is that we can investigate the consistency of self-reported personality measures. The details about the tests are reported in Appendix A. Table A4 lists the description of the five personality traits and the questions used in the Big Five Scale (BFI-S).

IV. Misreporting across Questions and Surveys

4.1 Samples and misreporting patterns

We plot the distributions of reporting errors in Figure 1 and Figure 2. We find a few distinct features of misreporting. First, misreporting exists in all cases and is much more prevalent for lectures than for assignments, probably due to their different saliency levels. Second, the error distributions are asymmetric with overwhelmingly overreporting. The reporting errors vary dramatically, ranging from -10 to 30 for assignments, and from -19 to exceeding 1 billion for lectures, indicating intentional misreporting. Third, misreporting is more prominent in Wave 3 compared to Wave 2.

To further investigate misreporting behavior in regression, we restrict the sample to those aged between 20-45, as the YTEP focuses on young teachers (less than 0.20% observations fall outside this range). We also limit the sample to those teaching in elementary and middle schools, as they are the main target of the training program (accounting for more than 94% of the participating trainees). Additionally, we drop extreme reporting errors greater than 50 for assignments and lectures, and those who completed the survey exceeding one hour (the average time is 13 minutes in Wave 2 and 15 minutes in Wave 3).[8]

The final sample and descriptive statistics are shown in Table 1. In the samples, more than 80% are female, the average age is around 25, over 80% of them hold at least a bachelor's degree. Approximately 65% of the participants teach in elementary school, and more than 76% are non-tenured teachers. Additionally, the same individuals participated in different waves of survey, and thus form panel data samples.

Table 2 shows that the probability of misreporting is very high, ranging from 11% to 87% in the sample. The average reporting errors span from 6% to 159% of the true values. Misreporting is more prevalent in the low saliency subject, such as lectures, compared to assignments, both in terms of probability and magnitude. This finding is consistent with the literature that higher saliency reduces misreporting (Celhay et al., 2024) and enhances response reliability (Stern et al., 2012). Moreover, the probability of overreporting is at least twice, and in some cases up to 14 times, that of underreporting.

4.2 Baseline results: observed factors and misreporting

We specify an empirical model for misreporting $q_d = f(X, S, R, q^*)$ as,

$$q_d = W\boldsymbol{\theta} + u = X\beta + S\delta + R\gamma + \varphi q^* + u. \quad (2)$$

For the regressor $W = (X, S, R, q^*)$, vector X denotes individual characteristics, and S includes cognitive and noncognitive human capital, R refers to the survey design, and q^* is the true value. The model is similar to the one used in Bound et al. (2001).

The model can be specified as the propensity of misreporting (binary variable), or as the magnitude of misreporting. The sign of q_d represents overreporting or underreporting. It is desirable to separate overreporting and underreporting in the estimation. However, because the data are dominated by overreporting, in our main regression, we use $|q_d|$ to represent misreporting. It helps increase sample size, since observations with underreporting will be

[8] The respondent might just keep the survey questionnaire open without actually working on it.

dropped in the overreporting model. We also run overreporting models separately to check the robustness.

Moreover, the zero-valued error should generally be treated differently from other errors. Changing from zero to a non-zero error means different behaviors (from accurate reporting to misreporting). We will consider those issues in the regression in Section VI.

Misreporting is affected by both individual attributes and survey design. Individual characteristics such as gender, age, education and job type, affect a person's self- and social image values due to the socio-economic status (Bursztyn and Jensen, 2017). For example, studies show that self-employed workers underreport their earnings to tax authorities in household surveys (Hurst et al., 2014), and lower levels of education are associated with increased misreporting (Ljungvall et al., 2015; Evans et al., 2023). We include tenure status in our analysis because non-tenured teachers, who face less job security, may either be more meticulous in their survey responses or may tend to overreport to enhance their perceived standing.[9]

Another reason for including the above individual characteristics is that their true values are available in the official registration file. We first estimate the models using OLS to examine the consistency of the results. Additionally, OLS does not need distribution assumption for the regression error. Table 3 reports the results, with Panel A for the full sample and Panel B for the panel sample.

Panel A shows that the effects of the individual characteristics vary across reporting subjects and across waves. There is no clear pattern in sign and statistical significance. Panel B displays similar results, even for the same individuals in both surveys. The results concur with our hypothesis that misreporting behavior varies depending on subjects and survey specifics. Therefore, it is likely that empirical findings based on one reporting subject, or one survey may not be generalized.

On the contrary, the information hints show a robust effect in reducing misreporting, with statistically significant effects (except for misreporting lectures in Wave 3). Moreover, true value displays a robust effect: the higher the true value, the lower the reporting errors.

V. Unobserved Heterogeneity and Misreporting Behavior

The results from the baseline model show that observed individual and job characteristics do not have a strong effect in misreporting. A natural question is whether

[9] Although survey participants are assured that the survey will not be accessible to their employers, some respondents may still believe that there is a chance that their responses might influence decisions on their promotion.

unobserved individual heterogeneity not controlled in the model helps explain the misreporting behavior. It is known that unobserved heterogeneity may affect an individual's education attainment, job, and the participation in the training program. Thus, they can cause omitted variable bias in the regression.

In this section, we first examine whether participating in the survey affects the reporting behavior, i.e., the sample selectivity due to unobserved heterogeneity. We then investigate what components of the unobservable exert the influences using various measures of noncognitive human capital.

5.1 Unobserved heterogeneity and sample selectivity

Given that all waves of the survey are voluntary, it's possible that participation in a survey is affected by individual heterogeneity. If the unobservable not only influences whether an individual participates in a survey but also how he/she responds questions in the survey, then it causes a sample selectivity problem. Studies using survey data consider either sampling error or non-sampling error such as measurement error (Meyer and Mittag, 2021). In this study, in addition to misreporting itself, we also investigate sampling errors by examining sample selectivity in a survey. More specifically, non-sampling measurement error and sampling error could both be affected by individual heterogeneity, i.e., the unobservable that drives survey participation also affects reporting accuracy.

Therefore, those who participated in the survey may differ from those not in terms of unobservable, and the estimation based on the sample of survey respondents may result in sample selectivity bias. We apply the Heckman two-step procedure (Heckman, 1979) for sample selection. Following equation (2), the first step about survey participation can be modelled as:

$$p^* = Z\eta + W\pi + \epsilon, \tag{3}$$

where p^* is the latent variable. Survey participation is given by $p = 1$ if $p^* > 0$, i.e., $p = 1[Z\eta + W\pi + \epsilon > 0]$. Z is a vector of selection variables that affect participation but not reporting error, ϵ is the disturbance term. Because the reporting discrepancy q_d is observed only for survey participants, i.e., when $p = 1$, the regression model becomes $E(q_d|Z, W, \epsilon > -Z\eta - W\pi)$. When the error terms (ϵ, u) are correlated, i.e., the unobservable affects both survey participation and misreporting, it can be shown that the correlation results in an omitted variable, i.e., the inverse Mills ratio (*IMR*). Because *IMR* is also related with included regressors, the estimation without control for the sample selectivity

will be inconsistent. The key to the Heckman two-step procedure is to add the inverse Mills ratio in the model to correct for the omitted variable bias.

We separate participating trainees into two samples based on whether they participated in a survey. To satisfy the exclusion restriction in the Heckman approach, we need instruments that influence survey participation but not reporting behavior. Two variables about the participants are selected, "whether majoring in Education in college" and "whether teaching in elementary school." Individuals with a college degree in Education are less likely to participate in program activities, and thus less likely to join the surveys. Similarly, teaching in an elementary school may have different demand for training compared to in middle school, and thus may have different level of participation. These characteristics are presumed to have no direct impact on reporting accuracy.

Additionally, the Heckman's procedure requires that the variables in the outcome model (reporting errors) be a strict subset of those in the selection model (participating in a survey). However, the interventions included in the misreporting model are not available for those who did not participate in the survey in the first-step estimation. Yet, it can be shown that, if the intervention variables are independent of other regressors (as is likely the case for our RCTs), the Heckman's two-step procedure remains valid even without including those variables in the first step.

The results of the first step probit model are reported in Panel B of Table 4. As expected, those with a major in Education in college are less likely to take part in a survey, and the effect is mostly statistically significant. Those teaching in elementary schools appear to be more likely to join Wave 2 but less likely to join Wave 3, while the effect is only significant in Wave 2 for assignments. Those results support the choice of the instruments.

Table 4 Panel A reports the second step results from the Heckman model. First, the sample selection term *IMR* is positive in almost all cases, indicating a positive correlation between participating in a survey and misreporting. This finding is similar to that in Celhay et al. (2024) that higher response rate has a negative effect on survey accuracy. However, it is only statistically significant in Wave 3 for assignments. It seems that the sample selection only occurs for the high-saliency subject (assignments) but not for the low-saliency subject (lectures). One possible explanation is that, near the program's end in Wave 3, trainees care more on earning the training certificate, making them more likely to engage in program activities (such as survey) and at the same time overstate the number of assignments submitted (required for credits).

Accordingly, the estimated effects of individual characteristics in the model also change substantially, as all of them become statistically insignificant (compared with Table 3). It indicates that the omitted variable bias caused by the sample selectivity has a major effect in this model. For other models, we can still see changes in the estimates of the individual characteristics, but the changes are moderate.

Another major change is the effect of the true value on reporting assignment after controlling for sample selectivity. It becomes statistically insignificant in both waves, while no major changes for reporting lectures. Finally, as expected, the effects of interventions remain largely unchanged with the control of sample selectivity, possibly due to their random nature.

Overall, our results show that the sample selection plays a role in misreporting and results in sampling errors. The selection is generally positive, implying a dilemma for a survey, i.e., increasing survey response rate may worsen survey quality.

5.2 Personality traits and misreporting

It is challenging to control for sample selectivity for a survey because data on non-participants are generally not available. As the individual heterogeneity influences both survey participation and reporting accuracy, one approach is to include the measures for unobservable in the model. It can help mitigate the sample selectivity, as the correlation caused by unobservable heterogeneity is accounted for in the model (instead of leaving it in the error term). More importantly, it provides direct estimates to assess the impact of the unobservable on misreporting.

We use the Big Five Model to measure personality traits, the specific components of unobservable heterogeneity. The measures are obtained in Wave 1 and Wave 3 survey as discussed in the data section. However, one concern is that, as documented in many studies, self-reported measures of personality traits have many problems (Feng et al., 2022). For example, Gnambs (2015) finds that measurement errors in self-reports of personality encompass various transient, item- and scale-specific error components. Chen et al. (2020) shows that self-reported noncognitive abilities are sensitive to survey conditions. If the self-reported measures of personality traits are influenced by the individual heterogeneity that also affects the responses to other questions, the self-reported traits are endogenous, and their estimated effects will be inconsistent.

We first examine the self-reported Big Five scores by comparing them between Wave 1 and 3 for the panel sample. In Appendix Table A5, even for the same individuals, the average

self-reported scores and their standard error are uniformly higher in Wave 1 compared to Wave 3, with all differences statistically significant. It is known that most personality traits change very slowly (Graham et al., 2020), especially in midlife (Damian et al., 2019; Seifert et al., 2022). Given Wave 1 and 3 is only six months apart, significant changes should be unlikely.

Furthermore, we plot the distributions of Emotional Stability and Conscientiousness in Appendix Figure A1 (as their difference between waves is the largest). As can be seen, their distributions differ significantly between Wave 1 and 3. The distribution of Conscientiousness in Wave 1 shows a very strange shape. Therefore, our data suggest that the self-reported personality traits may not be unreliable. This result is expected, given the widespread misreporting found in the two key variables discussed above.

Nevertheless, we include the self-reported measures of personality traits (standardized) in the model. Note that we are unable to control for sample selectivity because the Big Five scores are not available for non-participants. To avoid simultaneity in reporting Big Five traits and in responding the key survey questions, we use the Big Five scores reported in Wave 1 in the regressions of the subsequent waves. In this case, their reported values are ex-ante to Wave 2 and 3. For comparison, we also run models for Wave 3 using the self-reported traits from Wave 3 to assess the potential simultaneity effect. We report the results based on the full set of Big Five measures. The results using a subset are similar (not reported). For comparison, we run the models using the panel sample in the two waves.

Panel A in Table 5 presents the results using the Big Five measures from Wave 1. The results are similar to the baseline models, i.e., the effects of Big Five traits vary across subjects and waves. Some results are even counterintuitive; for example, Conscientiousness displays a positive and statistically significant effect on misreporting, and similarly for Openness (Dohmen and Jagelka, 2024). Extraversion and Emotional Stability display a statistically significant effect in a scattered manner. Finally, only Agreeableness does show a negative effect, consistent with the literature, e.g., those who are more agreeable may be more inclined to help the researchers and thus provide more reliable responses (de New and Schurer, 2023).

Moreover, we compare the results for Wave 3 using reported personality measures from Wave 1 and Wave 3 separately (Panel A vs Panel B), and find some noticeable differences. For example, the statistical significance changes for two of the five traits for reporting lectures, while the results are quite consistent for reporting assignments. Therefore, the

simultaneity in reporting personality traits and in responding other survey questions affects the regression results. The effect seems to be stronger in reporting low-saliency subjects.

Overall, we find that the self-reported Big Five personality traits, although commonly used in the studies on reporting behavior in surveys (e.g., Lee et al., 2020; Hilbig, 2022; Dohmen and Jagelka, 2024), do not provide a consistent story. Investigating reporting accuracy based on other self-reported variables in the same survey or in a related survey is subject to the same influences of unobserved heterogeneity.

5.3 Noncognitive human capital inferred from observed behavior

Given the problems with the self-reported measures on personality traits, we further explore other measures of noncognitive human capital. As discussed in Heckman et al. (2021), abilities/skills can be inferred not only by questionnaires and experiments but also from observed behaviors. Noncogntive human capital can thus be measured by task performances, because performance depends on skills and incentives. Studies have utilized the information about behaviors in a survey to measure personal characteristics. For example, Hitt et al. (2016) uses the percentage of items a respondent skipped to measure effort and conscientiousness, and de New and Schurer (2023) uses survey item-response rate as a proxy for unobserved ability.

One advantage of our data is that the information of participation is recorded automatically in the digital platforms. Therefore, we use the behaviors in the survey and in the training program as proxies for unobserved noncognitive human capital. Those measures represent their true values and are not subject to misreporting.

The first behavior proxy we consider is the time spent completing the survey. Intuitively, spending more time suggests greater effort and, consequently, less misreporting (Galesic and Bosnjak, 2009). The time spent also reflects a respondent's conscientiousness, and attitude toward the task at hand (Meade and Craig, 2012). As shown in Table 1, the average time for completing the survey is 13 minutes for Wave 1 and 2, and 15 minutes for Wave 3.

Another proxy is the survey submission time (whether during working hours). Because the surveys remain open for up to two weeks, the timing of doing the survey may reflect one's time management skill, e.g., getting work done during working hours. It is also possible that completing the survey during work hours exposes respondents to distractions. In Wave 1, 29% of respondents completed the survey during work, and it increased to 52% in Wave 2 and 64% in Wave 3.

Furthermore, individuals who regularly join live lectures during the training program before the starting time may have higher levels of noncognitive human capital, such as strong motivation, effective time management and a high degree of self-discipline. These noncognitive abilities may enhance an individual's capacity to regulate their behaviors, thoughts, and emotions to achieve goals (DeLisi, 2014). The training course delivery platform digitally recorded a trainee's participation up to every minute, including the exact time joining a lecture. We measure a trainee's early attendance to a lecture as arriving at least five minutes before the starting time. However, attending just one lecture early may not sufficiently capture a participant's personality, so we also consider the frequency of early attendance. More specifically, we define early lecture attendance as a dummy variable equal to 1 if the participant joined at least five minutes before the starting time for more than the average number of live lectures attended in a semester. As shown in Table 2, the average number of live lecture attendance in the first semester is 5.2, while in the second semester it is 3.0. The definition results in 22% of early attendance in the first semester, and 18% in the second semester (Table 1).

The regression results based on the panel sample are reported in Table 6. Compared to results using self-reported personality traits, a notable difference is that the effects of the behavior variables are consistent across subjects in terms of both sign and statistical significance. The time spent on the survey reduces misreporting, while submitting a survey during working hours increases it only in Wave 3. However, early lecture attendance does not show a statistically significant effect, although generally negative. Because the variables related to survey behavior are not available for non-participants, we cannot run the Heckman's approach to test the potential sample selectivity. However, because the unobservable that affects participation and misreporting is partially controlled for in the model, the sample selectivity should be much smaller.

Overall, noncognitive human capital based on observed behaviors exhibit strong and consistent influences on misreporting. One limitation is that those behaviors may be influenced by external factors specific to the survey. In next section, we will further examine this issue.

VI. Further Investigation: What Can We Learn?

In this section, we further investigate potential factors that affect misreporting behavior based on the findings above. In particular, we will use the individual characteristics reported

in the administrative file and the behavior proxies for noncognitive human capital recorded in digital format, as they do not suffer from measurement errors.

6.1 Tobit estimation for misreporting models

The regression above is based on a linear model because the estimator needs weak assumptions. However, it may not be efficient, and moreover, the linear specification may not accurately capture the mechanism of misreporting. For example, a respondent may first decide whether to report accurately then how much to misreport. In this case, the nonlinear Tobit model is more suitable, as it distinguishes accurate reporting from other reporting errors in its estimation. Moreover, reporting errors can be viewed as censored data and fit the Tobit estimation, because accurate reporting is censored to zero, while misreporting is expressed in different values.

The Tobit model separates the marginal effects into two parts: on the propensity of misreporting and on the magnitude of reporting errors conditional on misreporting. For simplicity of notation, in the case of overreporting, following equation (2), the total marginal effect of variable w_j in W can be expressed as:

$$\partial E(q_d|W)/\partial w_j = \partial P(q_d > 0|W)/\partial w_j \cdot E(q_d|q_d > 0, W) + P(q_d > 0|W) \cdot \partial E(q_d|q_d > 0, W)/\partial w_j . \quad (4)$$

The term $\partial P(q_d > 0|W)/\partial w_j$ represents the marginal effect of w_j on the propensity to misreport, and $\partial E(q_d|q_d > 0, W)/\partial w_j$ represents the marginal effect on the magnitude of reporting error, conditional on misreporting.

Table 7 presents the coefficient estimates of the Tobit model using the full sample. Panel A shows the results on misreporting (based on $|q_d|$). The observed individual characteristics do not have any statistically significant effect in Wave 2. However, in Wave 3, females are less likely to misreport the question of high saliency (assignments) but not for that of low saliency (lectures). Individuals with lower education are more likely to misreport subject of low saliency (lectures). Other characteristics, such as age and job security, do not have a significant effect. These results suggest again that the impact of observed personal characteristics on misreporting is limited and varies across subjects and surveys.

The behavior variable, time spent on the survey, shows a negative and statistically significant effect across waves and subjects. We calculate the two marginal effects based on the sample average. In Wave 2, an additional 10 minutes on the survey reduces the propensity of misreporting assignments by 4.1 percentage points and lectures by 1.9 percentage points. In Wave 3, the marginal effects increase to 6.6 and 2.9 percentage points, respectively. The

effect seems to be larger for a high saliency question. Similar patterns can be found on the magnitude of misreporting.

Regarding the other two behavior variables, the time of survey submission has no statistically significant impact in Wave 2 but a positive significant effect in Wave 3. Attending lectures early shows mixed negative effects. We will further discuss these results in Section 6.2 based on the panel data estimation.

Information hints significantly reduce misreporting across subjects and surveys. For example, in Wave 2, with the positive hint, the misreporting propensity decreases by 11.6 percentage points for assignments, much higher than 2.1 percentage points for lectures. In Wave 3, the reductions are even higher, at 33.3 and 2.7 percentage points, respectively. A similar pattern is found for magnitude of reporting errors.

This pattern suggests that information hints are more effective for high-saliency questions and have a larger impact in Wave 3. It is possible that, because survey participants become motivated as the training program approaches to the end, the information hints become more influential. Additionally, as the survey progresses, the fatigue intensifies, causing information hints more helpful in reducing the cognitive effort of retrieving information.

Our results show that the information hint with a negatively framed question has generally a similar impact as the positive hint, although Timmons et al. (2021) finds a stronger effect of negative framing. As an exception, in Wave 3, the negatively framed question shows no effect on reporting lectures. One possible reason is that this question requires recalling the number of lectures attended and then calculating the number missed, a more cognitively demanding task. The complexity of this particular question may have diminished the effectiveness of the hint, especially in Wave 3.

Finally, because the pattern of misreporting is dominated by overreporting, the estimation results above should generally represent the effect of overreporting behavior. There is one exception, i.e., for misreporting assignments in Wave 2, the overreporting probability is only twice that of underreporting (other cases are more than three times). For comparison, we run a separate model for overreporting assignments in Wave 2.

The results are reported in Panel B of Table 7. Compared with Column 1, the results are similar in terms of significance and sign. However, the variables generally have a larger marginal effect. The largest difference happens for the negative hint: it reduces the propensity to overreport by 15.1 percentage points and the magnitude of misreporting by 0.917, larger than that for overall misreporting (11.4 percentage points and 0.548, respectively). It suggests

that the negatively framed question is much more effective in reducing the magnitude of overreporting, compared to the overall misreporting.

6.2 Panel data estimation: how much does unobserved human capital matter?

In the above estimations, the observed behaviors show a strong influence on misreporting. It raises a natural question: how much do the behaviors capture noncognitive human capital? Based on the Proxy Variable approach, if the true measures are included in the model, the proxy variables would be redundant. Otherwise, those proxy variables may capture some other unrelated information.

To investigate this issue, we apply the Tobit Fixed Effects (FE) estimation to examine the impacts of the proxy variables. In the panel data sample, an individual's unobserved human capital can be viewed as constant between Wave 2 and 3. Therefore, individual fixed effects can effectively control for both observed and unobserved time-invariant heterogeneity. The disadvantage of the FE estimation is that time-invariant variables including some observable characteristics cannot be identified.

In the panel data model, given the dramatic differences between the two waves in misreporting, we interact the Wave 3 dummy with all variables to allow their marginal effects to differ in two waves, except for the variable of early attendance. The definition of early lecture attendance has already accounted for wave-specific differences (using wave specific average).

For comparison, we also estimate the Tobit model using pooled data (without including individual fixed effects). The panel data results are reported in Table 8. First, we find statistically significant estimates for Wave 3 dummy (not reported) and its interaction terms, supporting our model specification. Secondly, in the FE estimation, the behavior variables remain statistically significant after removing individual heterogeneity. It indicates that they are not perfect proxy but also capture some other influences.

To further explore the effect of noncognitive human capital, we compare the differences between the Pooled and FE Tobit estimates. The Pooled Tobit estimates capture the effects of both individual heterogeneity and external factors, whereas the FE Tobit estimates remove the effect of individual heterogeneity. Therefore, the difference between the two reflects the impact of individual heterogeneity. This difference can be interpreted as largely capturing the effect of unobserved heterogeneity, since the observed characteristics generally do not show a significant effect in the estimations above. Moreover, those behaviors generally reflect some

specific attributes such as motivation, self-discipline, perseverance, and time management. Therefore, they can be viewed as representing the effect of noncognitive human capital.

In particular, in Wave 2 for reporting assignments, the estimated total effect (including both external factors and individual unobservable) for the variable time spent on the survey is -1.466, while the FE estimate of the external effects (after removing individual heterogeneity) is -0.547. Thus, the difference of -0.919 could be viewed as the impact of noncognitive human capital on reducing misreporting, and it accounts for 63% of the total effect. The external effect in this case can be viewed as a pure input effect, i.e., the more time spent, the more accurate the answers are. The results are consistent for reporting lectures and across both waves. It shows an individual with higher noncognitive human capital spend more time on the survey and do it more carefully.

Following the same exercise, we find the consistent results for the impact of early lecture attendance. The noncognitive human capital imbedded in the behavior reduce misreporting for both assignments and lectures. The attributes that drive an individual to attend class early also reduce misreporting. The effect is strong, almost double the external effect for assignments.

However, the submission time of a survey has a mixed result. More specifically, the unobserved heterogeneity associated with the behavior displays somewhat opposite effects across subjects (i.e., positive for assignments and negative for lectures). It may indicate conflicting noncognitive human capital that drive an individual to do the survey during work. For example, because completing a survey is not part of the routine job, doing it at work may reflect the attitude of rushing though tasks at hand; or in contrast, it indicates a higher level of self-discipline that an individual wants to complete job-related tasks (like the surveys) at work. Those traits have opposite effects on misreporting, and their relative strength depends on the context of the question such as saliency and the specifics of a survey.

For the true value, it also incorporates the effect of noncognitive human capital. The difference between the two estimates shows an effect of reducing misreporting for assignments but no effect for lectures (the estimates are very close to each other). One possible explanation is that the traits associated with assignments may differ from that associated with attending more lectures. For example, because assignments are tied to course credits and certificate, it may reflect more on conscientiousness and desire for rewards, while attending more lectures may reflect curiosity or desire for social activities. Additionally, after removing individual fixed effects, the true value shows a robust effect in reducing

misreporting. Therefore, it confirms that, mechanically, more assignments submitted or more lectures attended leave less room for misreporting.

Additionally, the FE estimation isolates the impact of external factors by controlling for individual fixed effects. The estimates for time spent on a survey are mostly negative, showing the input effect. External factors associated with completing the survey at work exhibit a negative effect in Wave 2 but shift to a positive effect in Wave 3, probably due to the interplay between workplace distractions and the respondent's work-related focus during survey completion. For example, Wave 3 takes place just before the summer break, a particularly busy period for teachers, which likely increasing distractions. Interestingly, external factors related to early lecture attendance are associated with an increase in misreporting. One possibility is that those attending lecture early may actively engage in other activities, such as socializing with trainees or interacting with teaching assistants, which could divert their attention and lead to less accurate survey responses.

The panel data results further demonstrate that an individual's reporting behavior varies across question subjects and survey specifics. The behavior-based proxies generally capture both noncognitive human capital attributes and the external influences specific to the question subject and survey condition. After removing the external factors, the noncognitive human capital associated with the behavior displays a quite consistent effect on misreporting.

VII. Conclusions

In this study, we design a field experiment in a large-scale online training program for rural teachers to investigate misreporting behaviors in online surveys. We create questions with different saliency levels and self-/social image connotations and put them in multiple surveys for the same group of potential respondents. We also design random interventions to alter survey conditions. We leverage the true information from both digitally recorded data and administrative records.

In the estimation, we apply the Heckman's two-step procedure for sample selectivity. Moreover, we employ the Tobit model to fit better the misreporting data and utilize both cross-sectional and panel samples in the Tobit Pooled and Fixed Effects models to deal with various data issues.

We find very useful results about misreporting behavior in online surveys. First, misreporting always exists in an online survey and is dominated by overreporting (defined as conforming to social norms). However, the pattern varies across question types and survey

specifics. Information hints, question saliency, and the sequence in which the questions are presented all affect misreporting.

Moreover, we investigate the impact of individual heterogeneity on misreporting at different layers. First, we examine generally unobserved heterogeneity on both participating in the survey and misreporting in responding questions in the survey. Secondly, we explore specifically how personality traits, a component of individual unobserved heterogeneity, affect misreporting using self-reported measures of Big Five. Finally, we infer an individual's specific noncognitive human capital attributes based on the observed behaviors.

We find that online surveys not only suffer from reporting error but also from sampling error, because participation in a survey is likely to be correlated with misreporting behavior. Additionally, individual heterogeneity affects misreporting. However, self-reported measures of personality traits suffer from the same problem of misreporting.

Furthermore, the noncognitive human capital inferred from observed behavior has a strong and consistent effect on misreporting. Yet, the behaviors are generally not perfect proxies for unobserved human capital attributes but also capture some external factors in the survey. Therefore, their estimates vary depending on the relative effects of noncognitive human capital versus external factors related to the question and the survey.

Our findings provide useful implications for empirical work using survey data and for survey design. More specifically, given the complexity of misreporting, conclusions about misreporting behavior drawn from a single survey, or a particular subject may not be generalized. Moreover, given the self-selection in participating in a survey, there may be a trade-off between increasing participation rates and the quality of survey responses.

While our field experiment is situated within a large-scale online training program, the patterns of misreporting behavior we observe are likely relevant across various types of surveys, especially those administered online. Factors such as question saliency, information hints, and the sequence of questions in the survey are not unique to our context and may similarly influence response accuracy in other types of surveys. Our findings indicate that the reporting accuracy will improve by making a survey more efficient, such as with low cost in searching/recalling information. Additionally, the issue of self-selection and its relationship to misreporting behavior extends beyond this particular training program, suggesting that researchers using survey data should be cautious of potential sampling bias.

Finally, this study focuses on teachers, a group with higher human capital than the general population. Therefore, our findings are likely to be a lower bound of the results in terms of the impact of human capital on misreporting. Moreover, a teacher's behavior in the

online surveys could be related to his/her teaching activities. As role models for students who are developing noncognitive skills, teachers' actions, shaped by their noncognitive human capital, can significantly impact students. Therefore, it is crucial for teachers to be mindful of these influences in their teaching. In this sense, our findings offer valuable information for improving training programs for teachers in general.

References

Abay, K. A., Barrett, C. B., Kilic, T., Moylan, H., Ilukor, J., & Vundru, W. D. (2023). Nonclassical measurement error and farmers' response to information treatment. *Journal of Development Economics*, *164*(June), 103136.

Abowd, J. M., & Stinson, M. H. (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics*, *95*(5), 1451–1467.

Abraham, K. G., Filiz-Ozbay, E., Ozbay, E. Y., & Turner, L. J. (2020). Framing effects, earnings expectations, and the design of student loan repayment schemes. *Journal of Public Economics*, *183*, 104067.

Abraham, K. G., Haltiwanger, J., Sandusky, K., & Spletzer, J. R. (2013). Exploring differences in employment between household and establishment data. *Journal of Labor Economics*, *31*(S1), S129-S172.

Adua, L., & Sharp, J. (2010). Examining survey participation and response quality: The significance of topic salience and incentives. *The Oxford Handbook of Political Methodology*, *1*(36), 95–109.

Agrawal, N., & Duhachek, A. (2010). Emotional compatibility and the effectiveness of antidrinking messages: A defensive processing perspective on shame and guilt. *Journal of Marketing Research*, *47*(2), 263–273.

Bašić, Z., & Quercia, S. (2022). The influence of self and social image concerns on lying. *Games and Economic Behavior*, *133*, 162–169.

Becker, A., Deckers, T., Dohmen, T., Falk, A., & Kosse, F. (2012). The relationship between economic preferences and psychological personality measures. *Annual Review of Economics*, 4, 453 ‑ 478.

Bertrand, M., & Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. *The American Economic Review*, *91*(2), 67–72.

Błachnio, A. (2021). Be happy, be honest: The role of self-control, self-beliefs, and satisfaction with life in honest behavior. *Journal of Religion and Health*, *60*(2), 1015-1028.

Black, D., Sanders, S., & Taylor, L. (2003). Measurement of higher education in the census and current population survey. *Journal of the American Statistical Association*, *98*(463), 545–554.

Bound, J. ., Brown, C. C., & Mathiowetz, N. A. (2001). Chapter 59: Measurement error in survey data. In *Handbook of Econometrics* (pp. 3705–3843). Elsevier.

Burke, M. A., & Carman, K. G. (2017). You can be too thin (but not too tall): Social desirability bias in self-reports of weight and height. *Economics and Human Biology*, *27*, 198–222.

Bursztyn, L., & Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, *9*(1), 131-153.

Celhay, P., Meyer, B. D., & Mittag, N. (2024). What leads to measurement errors? Evidence from reports of program participation in three surveys. *Journal of Econometrics*, *238*(2), 105581.

Chen, Y., Feng, S., Heckman, J. J., & Kautz, T. (2020). Sensitivity of self-reported noncognitive skills to survey administration conditions. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(2), 931–935.

Choshen-Hillel, S., Shaw, A., & Caruso, E. M. (2020). Lying to appear honest. *Journal of Experimental Psychology: General*, *149*(9), 1719.

Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516(7529), 86-89.

Costa, P. T., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorder*, *6*(4), 343–359.

Damian, R. I., Spengler, M., Sutu, A., & Roberts, B. W. (2019). Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *Journal of Personality and Social Psychology*, 117(3), 674.

de New, S. C., & Schurer, S. (2023). Survey response behavior as a proxy for unobserved ability: Theory and evidence. *Journal of Business and Economic Statistics*, *41*(1), 197–212.

DeLisi, M. (2014). Low self-control is a brain-based disorder. *The Nurture Versus Biosocial Debate in Criminology: On the Origins of Criminal Behavior and Criminality*, 172-184.

Dench, D., & Joyce, T. (2022). Information and credible sanctions in curbing online cheating among undergraduates: A field experiment. *Journal of Economic Behavior and Organization*, *195*, 408–427.

Dohmen, T., & Jagelka, T. (2024). Accounting for individual-specific reliability of self-assessed measures of economic preferences and personality traits. *Journal of Political Economy Microeconomics*, 2(3), 399-462.

Duhachek, A., Agrawal, N., & Han, D. (2012). Guilt versus shame: Coping, fluency, and framing in the effectiveness of responsible drinking messages. *Journal of Marketing Research*, *49*(6), 928–941.

Eckman, S., & Kreuter, F. (2018). Misreporting to looping questions in surveys: Recall, motivation, and burden. *Survey Research Methods*, *12*(1), 59-74.

Evans, A., Gray, E., & Reimondos, A. (2023). How tall am I again? A longitudinal analysis of the reliability of self-reported height. *SSM - Population Health*, *22*, 101412.

Ewers, M., & Zimmermann, F. (2015). Image and misreporting. *Journal of the European Economic Association*, *13*(2), 363–380.
Feng, S., Han, Y., Heckman, J. J., & Kautz, T. (2022). Comparing the reliability and predictive power of child, teacher, and guardian reports of noncognitive skills. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(6), 1–8.
Feng, S., & Hu, Y. (2013). Misclassification errors and the underestimation of the US unemployment rate. *American Economic Review*, *103*(2), 1054-1070.
Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360.
Gerlitz, J.-Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. Dokumentation der Instrumentenentwicklung BFI-S auf Basis des SOEP-Pretests 2005. *DIW Research Notes*, *4*, 1–36.
Gibson, R., Tanner, C., & Wagner, A. F. (2013). Preferences for truthfulness: Heterogeneity among and within individuals. *American Economic Review*, *103*(1), 532–548.
Gnambs, T. (2015). Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences*, *84*, 84–89.
Graham, E. K., Weston, S. J., Gerstorf, D., Yoneda, T. B., Booth, T. O. M., Beam, C. R., ... & Mroczek, D. K. (2020). Trajectories of big five personality traits: A coordinated analysis of 16 longitudinal samples. *European Journal of Personality*, 34(3), 301-321.
Groves, R. M. (2005). *Survey Errors and Survey Costs*. John Wiley & Sons.
Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey Methodology*. John Wiley & Sons.
Hahn, E., Gottschling, J., & Spinath, F. M. (2012). Short measurements of personality - Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality*, *46*(3), 355–359.
Heckman, James. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153-161.
Heckman, J. J., Jagelka, T., & Kautz, T. (2021). Some contributions of economics to the study of personality. In O. P. John & R. W. Robins (Eds.), *Handbook of Personality: Theory and Research* (4th ed., pp. 853–892). The Guilford Press.
Hilbig, B. E. (2022). Personality and behavioral dishonesty. *Current Opinion in Psychology*, *47*, 101378.
Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, *52*, 105–119.
Hoeschler, P., Balestra, S., & Backes-Gellner, U. (2018). The development of non-cognitive skills in adolescence. *Economics Letters*, 163, 40-45
Hu, Y. (2017). The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics. *Journal of Econometrics*, *200*(2), 154-168.
Hu, Y., & Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, *76*(1), 195–216.
Humphries, J. E., & Kosse, F. (2017). On the interpretation of non-cognitive skills–What is being measured and why it matters. *Journal of Economic Behavior & Organization*, 136, 174-185.
Hurst, E., Li, G., & Pugsley, B. (2014). Are household surveys like tax forms? Evidence from income underreporting of the self-employed. *Review of Economics and Statistics*, *96*(1), 19-33.
Jeong, D., Aggarwal, S., Robinson, J., Kumar, N., Spearot, A., & Park, D. S. (2023). Exhaustive or exhausting? Evidence on respondent fatigue in long surveys. *Journal of Development Economics*, *161*, 102992.
Karlan, D. S., & Zinman, J. (2012). List randomization for sensitive behavior: An application for measuring use of loan proceeds. *Journal of Development Economics*, *98*(1), 71-75.
Karlan, D., & Zinman, J. (2008). Lying about borrowing. *Journal of the European Economic Association*, *6*(2–3), 510–521.
Kashner, T. M., Suppes, T., Rush, A. J., & Altshuler, K. Z. (1999). Measuring use of outpatient care among mentally ill individuals: A comparison of self reports and provider records. *Evaluation and Program Planning*, *22*(1), 31–40.
Kilic, T., & Sohnesen, T. P. (2019). Same question but different answer: experimental evidence on questionnaire design's impact on poverty measured by proxies. *Review of Income and Wealth*, *65*(1), 144-165.
Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*(4), 2025-2047.
Kuhn, P. M., & Vivyan, N. (2022). The misreporting trade-off between list experiments and direct questions in practice: Partition validation evidence from two countries. *Political Analysis*, *30*(3), 381-402.
Lee, S. D., Kuncel, N. R., & Gau, J. (2020). Personality, attitude, and demographic correlates of academic dishonesty: A meta-analysis. *Psychological Bulletin*, *146*(11), 1042.
Li, H., Ma, M., & Liu, Q. (2022). How the COVID-19 pandemic affects job sentiments of rural teachers. *China*

Economic Review, *72*, 101759.
Ljungvall, Å., Gerdtham, U. G., & Lindblad, U. (2015). Misreporting and misclassification: implications for socioeconomic disparities in body-mass index and obesity. *European Journal of Health Economics*, *16*(1), 5–20.
Martinelli, C., & Wendy Parker, S. (2009). Deception and misreporting in a social program. *Journal of the European Economic Association*, *7*(4), 886–908.
Meade, Adam W, and S Bartholomew Craig. (2012). Identifying careless responses in survey data. *Psychological Methods* 17 (3): 437.
Meyer, B. D., Mittag, N., & Goerge, R. M. (2022). Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. *Journal of Human Resources*, 57(5), 1605-1644.
Meyer, B. D., & Mittag, N. (2021). An empirical total survey error decomposition using data combination. *Journal of Econometrics*, *224*(2), 286-305.
Meyer, B. D., & Mittag, N. (2019). Misreporting of government transfers: How important are survey design and geography? *Southern Economic Journal*, *86*(1), 230–253.
Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199–226.
Mittag, N. (2019). Correcting for misreporting of government benefits. *American Economic Journal: Microeconomics*, *11*(2), 142–164.
Read, B., Wolters, L., & Berinsky, A. J. (2022). Racing the clock: Using response time as a proxy for attentiveness on self-administered surveys. *Political Analysis*, *30*(4), 550-569.
Rosaz, J., & Villeval, M. C. (2012). Lies and biased evaluation: A real-effort experiment. *Journal of Economic Behavior & Organization*, *84*(2), 537-549.
Seifert, I. S., Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2022). The development of the rank-order stability of the Big Five across the life span. *Journal of Personality and Social Psychology*, 122(5), 920.
Shibata, I. (2022). Reassessing classification errors in the analysis of labor market dynamics. *Labour Economics*, *78*, 102252.
Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences of the United States of America*, 109(38), 15197.
Stephens, M., & Unayama, T. (2019). Estimating the impacts of program benefits: Using instrumental variables with underreported and imputed data. *Review of Economics and Statistics*, *101*(3), 468–475.
Stern, M. J., Smyth, J. D., & Mendez, J. (2012). The effects of item saliency and question design on measurement error in a self-administered survey. *Field Methods*, *24*(1), 3–27.
Sutin, A. R., & Terracciano, A. (2016). Five-factor model personality traits and the objective and subjective experience of body weight. *Journal of Personality*, *84*(1), 102–112.
Tergiman, C., & Villeval, M. C. (2023). The way people lie in markets: Detectable vs. deniable lies. *Management Science*, *69*(6), 3340-3357.
Timmons, S., McGinnity, F., Belton, C., Barjaková, M., & Lunn, P. (2021). It depends on how you ask: Measuring bias in population surveys of compliance with COVID-19 public health guidance. *Journal of Epidemiology and Community Health*, *75*(4), 387–389.
VandenBos, G. R. (2015). APA dictionary of psychology. In *American Psychological Association*.
Weisberg, H. F. (2009). *The Total Survey Error Approach: A Guide to the New Science of Survey Research.* University of Chicago Press.

Tables and Figures

Table 1: Descriptive statistics for observed characteristics

Variable	Wave 1	Wave 2	Wave 3
Observed characteristics			
Female	0.863	0.839	0.848
Age	25.213	25.406	25.450
Low education (1=Junior college degree, 0=Bachelor's degree or above)	0.149	0.155	0.146
Low job security (1=non-tenured teacher)	0.764	0.783	0.796
Major in Education	0.555	0.548	0.534
Teaching in elementary school	0.648	0.654	0.642
Proxy for noncognitive human capital			
Time spent on survey (unit: 10 mins)	1.298	1.303	1.471
	(0.619)	(0.704)	(0.753)
Submit survey during work (8am-6pm)	0.290	0.516	0.644
	(0.454)	(0.500)	(0.479)
Early lecture attendance	-	0.217	0.182
	-	(0.412)	(0.386)
Obs.	4,342	3,990	2,970
Panel sample in Wave 2 and Wave 3	-	1,757	1,757
Panel sample in all three waves	1,201	1,201	1,201

Notes:

1. Observed characteristics are from the official registration records. Behavior variables are from digital data automatically recorded in the training delivery platform and the survey implementation platform. The numbers in parentheses are standard errors.
2. The trainees' lowest education level is junior college (3-year college). Most people in the sample have a bachelor's degree, and very few with master's degree, e.g., their proportions are are 84.2% and 0.9%, respectively, in Wave 1.
3. "Early lecture attendance" is defined as a dummy variable equal to 1 if the participant arrived at least 5 minutes before the lecture start time for more than the average number of lectures in the semester (5 in Wave 2 and 3 in Wave 3).

Table 2: Misreporting pattern (Full sample)

Variable	Wave 2		Wave 3	
	Assignment (Total 5)	Lecture (Total 10)	Assignment (Total 5)	Lecture (Total 10)
Misreporting probability				
Misreporting	0.108	0.841	0.434	0.871
	(0.310)	(0.366)	(0.496)	(0.335)
Overreporting	0.071	0.786	0.338	0.786
	(0.257)	(0.41)	(0.473)	(0.41)
Underreporting	0.037	0.055	0.096	0.085
	(0.188)	(0.229)	(0.294)	(0.278)
Reporting error				
True value	4.742	5.204	4.269	3.003
	(0.643)	(3.407)	(0.997)	(3.43)
Reported value	4.844	8.975	5.288	7.432
	(1.156)	(2.848)	(2.309)	(3.622)
Absolute error	0.272	4.064	1.291	4.776
	(1.001)	(3.557)	(2.152)	(3.837)
Ratio of absolute error to true value (%)	5.736	78.094	30.241	159.041
Obs.	3,990	3,985	2,970	2,967

Note: For lectures, we dropped 5 observations with absolute errors over 50 as extreme values in Wave 2, and 3 observations in Wave 3.

Table 3: Observed characteristics and reporting error—Baseline results

	Panel A: Full sample				***Panel B: Panel data sample***			
Dependent Var:	Wave 2		Wave 3		Wave 2		Wave 3	
Absolute Error	Assignment	Lecture	Assignment	Lecture	Assignment	Lecture	Assignment	Lecture
Observed characteristics								
Female	-0.066	-0.154	-0.180**	-0.260	-0.049	-0.021	-0.157	-0.257
	(0.048)	(0.117)	(0.089)	(0.164)	(0.068)	(0.138)	(0.112)	(0.198)
Age	-0.008	0.011	-0.031**	-0.045*	0.005	0.012	-0.037**	-0.051*
	(0.005)	(0.015)	(0.014)	(0.024)	(0.008)	(0.020)	(0.018)	(0.029)
Low education	0.010	0.008	0.148	0.627***	-0.057	0.028	0.202	0.527**
	(0.044)	(0.100)	(0.106)	(0.176)	(0.058)	(0.127)	(0.150)	(0.229)
Low job security	0.059*	0.041	-0.095	0.069	0.126***	-0.023	-0.034	0.141
	(0.035)	(0.101)	(0.085)	(0.155)	(0.047)	(0.137)	(0.116)	(0.194)
Interventions								
Info hints positive	-0.377***	-0.318***	-1.857***	-0.466***	-0.284***	-0.280**	-1.716***	-0.404**
	(0.040)	(0.106)	(0.100)	(0.161)	(0.060)	(0.141)	(0.137)	(0.202)
Info hints negative	-0.332***	-0.300***	-2.170***	0.134	-0.280***	-0.246*	-2.096***	0.194
	(0.044)	(0.107)	(0.092)	(0.152)	(0.061)	(0.138)	(0.123)	(0.190)
True value	-0.143***	-0.743***	-0.419***	-0.587***	-0.116*	-0.776***	-0.467***	-0.563***
	(0.037)	(0.013)	(0.036)	(0.015)	(0.060)	(0.018)	(0.051)	(0.019)
Obs.	3,990	3,985	2,970	2,967	1,757	1,757	1,757	1,757

Notes: The absolute error is defined as the absolute value of the self-reported value minus the true value.

Table 4: Survey participation and misreporting with sample selectivity

Dependent Var: Absolute Error	***Panel A: Second step results***			
	Wave 2		Wave 3	
	Assignment	Lecture	Assignment	Lecture
Observed characteristics				
Female	-0.066	-0.195	-0.095	-0.201
	(0.048)	(0.150)	(0.103)	(0.191)
Age	-0.007	-0.004	-0.006	-0.031
	(0.011)	(0.032)	(0.020)	(0.034)
Low education	0.011	0.028	0.056	0.562***
	(0.045)	(0.109)	(0.120)	(0.209)
Low job security	0.066	-0.044	0.217	0.224
	(0.071)	(0.188)	(0.211)	(0.323)
IMR (inverse Mills ratio)	0.069	-0.962	2.280*	1.168
	(0.613)	(1.943)	(1.382)	(2.132)
Interventions				
Info hints positive	-0.377***	-0.319***	-1.858***	-0.467***
	(0.040)	(0.106)	(0.100)	(0.161)
Info hints negative	-0.332***	-0.300***	-2.169***	0.134
	(0.044)	(0.107)	(0.092)	(0.152)
True value	-0.126	-0.812***	-0.002	-0.515***
	(0.151)	(0.139)	(0.257)	(0.132)
Obs.	3,990	3,985	2,970	2,967
Dependent Var: Survey Participation Probability	***Panel B: First step results for selection variables***			
Major in education	-0.014	-0.050*	-0.069**	-0.077***
	(0.026)	(0.026)	(0.028)	(0.028)
Teaching in elementary school	0.081***	0.027	-0.001	-0.012
	(0.027)	(0.027)	(0.029)	(0.029)
Other variables	Yes	Yes	Yes	Yes
Obs.	11,369	11,369	10,467	10,467

Notes:

1. In the first step, we do separate probit estimation for assignments and lectures, rather than estimating a single model for participating in Wave 2 (or Wave 3), because their true values are different.
2. The variables of interventions are included in the second step because they are only available for survey participants.

Table 5: Self-reported Big Five personality and reporting error (Panel data sample)

Dependent Var: Absolute Error	*Panel A: Using Big Five from Wave 1*				*Panel B: Using Big Five from Wave 3*	
	Wave 2		Wave 3		Wave 3	
	Assignment	Lecture	Assignment	Lecture	Assignment	Lecture
Big Five personality traits						
Conscientiousness	0.038	0.130	0.060	0.336**	0.033	0.378***
	(0.032)	(0.090)	(0.075)	(0.135)	(0.072)	(0.114)
Openness	-0.003	-0.080	0.211***	0.145	0.193***	0.470***
	(0.033)	(0.075)	(0.078)	(0.115)	(0.068)	(0.105)
Extraversion	0.055*	0.045	0.081	-0.023	-0.032	-0.101
	(0.032)	(0.087)	(0.073)	(0.110)	(0.068)	(0.102)
Agreeableness	-0.067*	-0.045	-0.025	-0.225*	-0.040	-0.399***
	(0.039)	(0.076)	(0.061)	(0.120)	(0.057)	(0.105)
Emotional Stability	0.026	0.049	-0.055	0.335***	0.055	0.156
	(0.025)	(0.074)	(0.065)	(0.106)	(0.062)	(0.100)
Other variables	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	1,201	1,201	1,201	1,201	1,201	1,201

Notes:
1. In the regression, we standardized the variables of self-reported personality traits within the sample.
2. We control the observed characteristics, interventions and true value as in Table 3.

Table 6: Behavior proxy of noncognitive human capital and reporting error (Panel data sample)

Dependent Var:	Wave 2		Wave 3	
Absolute Error	Assignment	Lecture	Assignment	Lecture
Proxy for noncognitive human capital				
Time spent on survey	-0.130***	-0.141*	-0.264***	-0.435***
	(0.029)	(0.080)	(0.052)	(0.094)
Submit survey during work	-0.044	-0.070	0.149*	0.324**
	(0.045)	(0.104)	(0.088)	(0.151)
Early lecture attendance	-0.024	-0.062	-0.127	0.002
	(0.052)	(0.105)	(0.100)	(0.183)
Other variables	Yes	Yes	Yes	Yes
Obs.	1,757	1,757	1,757	1,757

Note: We control the observed characteristics, interventions and true value as in Table 3.

Table 7: Misreporting results using Tobit estimation

Dependent Var:	***Panel A: Absolute Error***				***Panel B: Overreporting Error***
	Wave 2		Wave 3		Wave 2
	Assignment	Lecture	Assignment	Lecture	Assignment
Proxy for noncognitive human capital					
Time spent on survey	-1.183***	-0.270***	-0.696***	-0.565***	-1.669***
	(0.230)	(0.066)	(0.103)	(0.091)	(0.306)
Submit survey during work	-0.194	0.055	0.256*	0.256*	-0.109
	(0.259)	(0.091)	(0.154)	(0.142)	(0.323)
Early lecture attendance	0.023	-0.485***	-0.891***	-0.151	0.184
	(0.329)	(0.138)	(0.207)	(0.251)	(0.410)
Observed characteristics					
Female	-0.397	-0.108	-0.511***	-0.265	-0.587
	(0.340)	(0.124)	(0.196)	(0.185)	(0.418)
Age	0.021	0.019	-0.037	-0.025	-0.003
	(0.053)	(0.019)	(0.031)	(0.029)	(0.066)
Low education	0.221	0.040	0.247	0.720***	0.584
	(0.352)	(0.126)	(0.203)	(0.189)	(0.424)
Low job security	0.181	-0.026	-0.216	0.070	0.283
	(0.318)	(0.111)	(0.177)	(0.167)	(0.397)
Interventions					
Info hints positive	-3.319***	-0.311***	-3.494***	-0.521***	-3.599***
	(0.337)	(0.111)	(0.175)	(0.164)	(0.399)
Info hints negative	-3.269***	-0.309***	-4.137***	0.153	-6.505***
	(0.340)	(0.112)	(0.180)	(0.163)	(0.614)
True value	-1.654***	-0.835***	-1.155***	-0.679***	-2.175***
	(0.170)	(0.017)	(0.071)	(0.028)	(0.207)
Obs.	3,990	3,985	2,970	2,967	3,844

Notes:

1. For the model of overreporting errors, we do not include the observations who underreported.
2. The two marginal effects of Tobit model can be expressed as: $\partial P(q_d > 0|W)/\partial w_j = \theta_j/\sigma \cdot \phi(W\boldsymbol{\theta}/\sigma)$, and $\partial E(q_d|q_d > 0, W)/\partial w_j = \theta_j \cdot \{1 - \lambda(W\boldsymbol{\theta}/\sigma)[W\boldsymbol{\theta}/\sigma + \lambda(W\boldsymbol{\theta}/\sigma)]\}$. The sign of the coefficients θ_j reflects the direction of the marginal effect.

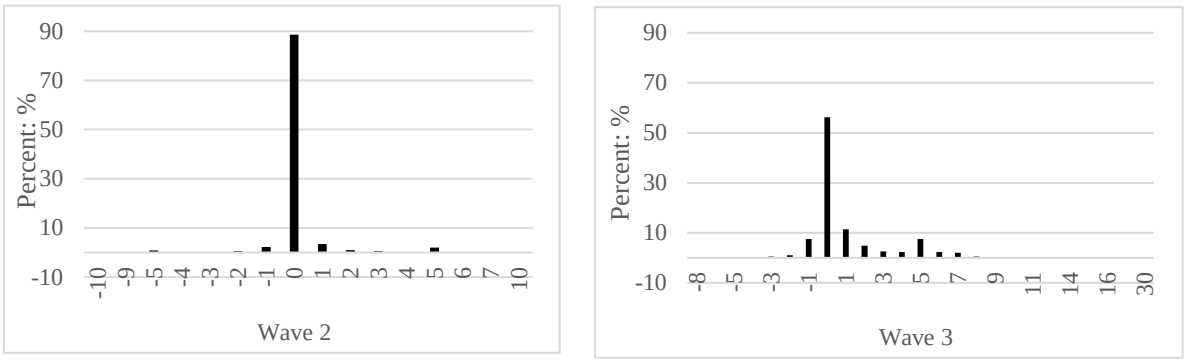
Table 8: Misreporting results with Tobit pooled and fixed effect models (Panel data sample)

Dependent Var: Absolute Error	Assignment		Lecture	
Method	Tobit	Tobit FE	Tobit	Tobit FE
Proxy for noncognitive human capital				
Time spent on survey	-1.466***	-0.547**	-0.147	0.068
	(0.335)	(0.246)	(0.111)	(0.109)
Time spent on survey × Wave 3	0.606	0.113	-0.354**	-0.512***
	(0.372)	(0.251)	(0.147)	(0.118)
Submit survey during work	-0.502	-0.698***	0.021	0.062
	(0.339)	(0.262)	(0.151)	(0.140)
Submit survey during work × Wave 3	0.718*	0.692*	0.404*	0.520***
	(0.418)	(0.365)	(0.217)	(0.199)
Early lecture attendance	-0.593**	0.568*	-0.373**	0.022
	(0.236)	(0.321)	(0.162)	(0.177)
True value	-1.530***	-0.706***	-0.921***	-0.971***
	(0.251)	(0.187)	(0.026)	(0.029)
True value × Wave 3	-0.008	-0.001	0.300***	0.345***
	(0.272)	(0.225)	(0.032)	(0.024)
Interventions	Yes	Yes	Yes	Yes
Interventions × Wave 3	Yes	Yes	Yes	Yes
Wave 3	Yes	Yes	Yes	Yes
Observed characteristics	Yes	-	Yes	-
Individual fixed effects	-	Yes	-	Yes
Obs.	3,514	3,514	3,514	3,514

Notes:

1. The data used are the panel sample of Wave 2 and Wave 3, with the number of observations of 1,757 in each wave.
2. We do not include the interaction term between early lecture attendance and Wave 3, because the early attendance is defined with different frequency for each wave and thus it has taken into account the different effects between the two waves.

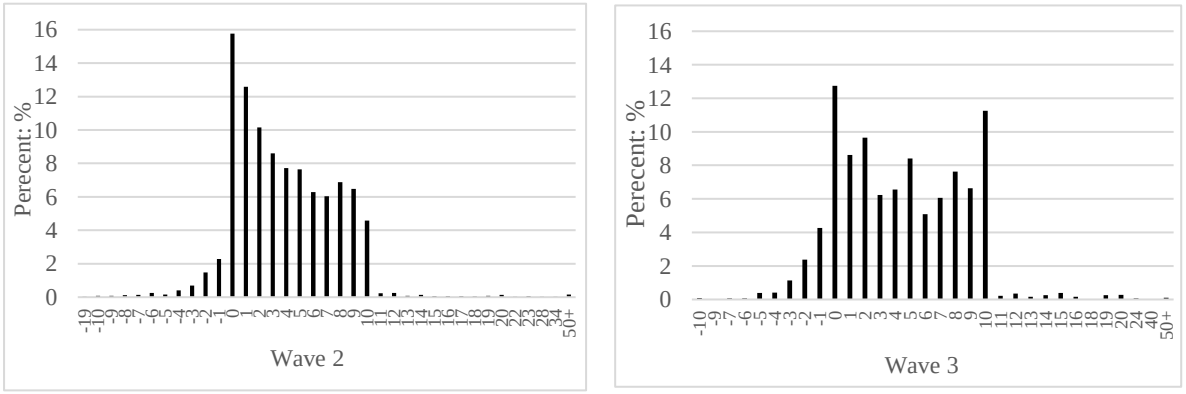
Figure 1. Reporting error distribution for assignments submitted



Notes:

1. The sample includes all participants surveyed in each wave, with a size of 4,290 and 3,202, respectively.
2. The total number of assignments per semester is 5.
3. Reporting error = self-reported value-true value.

Figure 2 Reporting error distribution for lectures attended



Notes:

1. The sample includes all participants surveyed in each wave, with a size of 4,290 and 3,202, respectively.
2. The total number of lectures per semester is 10.
3. Reporting error = self-reported value-true value.
4. The proportion of errors larger than 50 is 0.16% in Wave 2 and 0.09% in Wave 3.

Appendix A: Survey Design and Data

1. YTEP Program and Participants

YTEP (Young Teacher Empowerment Program) was initiated in 2017 by the YouChange China Social Entrepreneur Foundation and is sponsored by various non-profit organizations, universities, and corporations in China. The training program runs annually, begins in September and ends in the following June. The program provides training courses online from a broadcast platform called WeLink. Participation in the program is mostly voluntary, while some schools may require their new teachers to get the training. More details can be found at their official website: http://www.youcheng.org/project_detail.php?id=837.

Most of the teachers participating in the YTEP are the so-called special-post teachers, hired through a national program. This national hiring program aims at recruiting college graduates to teach in rural areas for three years to improve rural education. After three years of service, those who pass the assessment can become tenured teachers (more details can be found at http://www.moe.gov.cn/srcsite/A10/s7151/202103/t20210331_523712.html.)

The YTEP training program offers two compulsory courses and many other field courses covering subjects commonly taught in primary and junior middle schools. Participants are required to attend the two compulsory courses and select at least one field course. All courses are offered online live or recorded.

In the YTEP 2021 program, a total of 12,826 teachers registered in the program. The attendance of live lectures is not mandatory and recorded lecture videos are available to watch. Some of registered trainees never participated in any program activities after registration. Therefore, we include only those who attended at least one live lecture or submitted at least one assignment in a semester, otherwise, they are treated as non-participants. The average attendance of live lectures is generally lower than 50% (39.7% in the first semester and 20.7% in the second semester for the compulsory course).

2. Surveys

Three surveys were designed and conducted at different stages of the YTEP 2021, as shown in Table A1. Wave 1 was around the middle of the first semester, Wave 2 at the end of the first semester (middle of the training program), and Wave 3 at the end of the program. Survey questionnaires were distributed to all registered trainees.

Because the surveys are voluntary, they have different participation rates, between 28-40%, consistent with the average rates of participating in live lectures (Table A1). To increase survey participation, the YTEP administrative team sent several reminders.

3. RCT Experiments

In implementing the experiments, we alter the two specific questions to test reporting accuracy and randomly assign them to control group and treatment groups in each wave. The detailed questions for various groups are shown in Table A2.

4. Big Five Personality Traits

Big Five personality traits are commonly assessed through self-reports, utilizing instruments like the Big Five Inventory (BFI). Gerlitz and Schupp (2005) develop a 15-item Big Five Scale (BFI-S). The BFI-S has been adopted widely, because each dimension is easily measured by three items. Hahn et al. (2012) find that this short version provides reasonable consistencies with its longer counterpart. We also use the BFI-S in our surveys (detailed descriptions and related questions are listed in Table A4).

In particular, participants rate the statements on a 11-point Likert scales (from 0 "Completely disagree" to 10 "Completely agree"). To ensure consistency, negatively worded items (e.g., "I tend to be lazy") are transformed with 10 minus the original value to reverse the score (Hahn et al., 2012; Chen et al., 2020). Accordingly, we also code the trait Neuroticism as "Emotional Stability" so that a higher value indicates more stability and lower levels of Neuroticism. We average the scales of all items within each Big Five dimension to create a composite index to measure the trait. In the regression, we standardize each dimension within the sample.

Table A1: Participation in compulsory course and surveys

	Compulsory course		
	Fall 2021		Spring 2022
Time	Sep. 15-Dec. 6		Mar. 16-May. 25
Total participating trainees	12,117		11,151
Average attendance rate per lecture (%)	39.66		20.67
Average # of attendants per lecture	4,805		2,305
	Surveys		
	Wave 1 Fall 2021	Wave 2 Fall 2021	Wave 3 Spring 2022
Time	Nov. 10-19	Dec. 6-Jan. 6	May. 25-Jun. 22
Survey participation rate (%)	38.58	35.29	28.36
Participants	4,675	4,276	3,162
Join both Wave 1 and 2	2,619	2,619	-
Join both Wave 1 and 3	1,802	-	1,802
Join both Wave 2 and 3	-	1,882	1,882
Join all waves	1,291	1,291	1,291
Total # of questions in the survey	82	122	125

Notes:

1. The total enrolment in the 2021 YTEP training program is 12,826. The total participating trainees exclude those who either attended a lecture of a compulsory course nor submitted an assignment in a semester.
2. The participation rate is calculated as the ratio of participants to the total participating trainees in the semester.

Table A2: Two key questions for testing reporting accuracy (Assignments completion and lecture attendance)

RCT Group	Wave 2	Wave 3
Control	Q3. How many live-streamed lectures for the *Career Development* course did you attend this semester?	Q37. How many live-streamed lectures for the *Teacher Ethics* course did you attend this semester?
	Q4. How many assignments for this course did you submit this semester?	Q38. How many assignments for this course did you submit this semester?
Treatment 1	Q3. Out of the 10 live-streamed lectures for the *Career Development* course this semester, how many did you attend?	Q37. Out of the 10 live-streamed lectures for the *Teacher Ethics* course this semester, how many did you attend?
	Q4. Out of the 5 assignments for this course this semester, how many did you submit?	Q38. Out of the 5 assignments for this course this semester, how many did you submit?
Treatment 2	Q3. Out of the 10 live-streamed lectures for the *Career Development* course this semester, how many did you miss?	Q37. Out of the 10 live-streamed lectures for the *Teacher Ethics* course this semester, how many did you miss?
	Q4. Out of the 5 assignments for this course this semester, how many did you not submit?	Q38. Out of the 5 assignments for this course this semester, how many did you not submit?

Table A3: Balance test of characteristics between treatment and control groups

	Wave 2				Wave 3			
	Treat1 – Control		Treat2 – Control		Treat1 – Control		Treat2 – Control	
Variable	Diff	p-value	Diff	p-value	Diff	p-value	Diff	p-value
Female	-0.008	0.561	-0.040	0.005	-0.037	0.020	-0.031	0.049
Age	0.144	0.134	0.189	0.051	0.078	0.463	0.041	0.704
Low education	0.015	0.303	-0.007	0.633	-0.020	0.215	0.001	0.965
Low job security	-0.000	0.992	0.002	0.884	0.025	0.165	0.011	0.539
Major in education	0.001	0.971	0.000	0.992	-0.008	0.724	-0.012	0.588
Teach in elementary school	0.036	0.048	0.015	0.422	-0.001	0.970	-0.011	0.620
Obs.	1,343	1,321	1,326	1,321	991	982	997	982

Table A4: Questions to measure Big Five personality traits

Big Five	Descriptions	Statements
Conscientiousness	The tendency to be organized, responsible, and hardworking	1) I do a thorough job. 2) I tend to be lazy. 3) I do things effectively and efficiently.
Openness	The tendency to be open to new aesthetic, cultural, or intellectual experiences	1) I am original, come up with new ideas. 2) I value artistic experience. 3) I have an active imagination.
Extraversion	One's interests and energies toward the outer world of people and things	1) I am communicative, talkative. 2) I am outgoing, sociable. 3) I am reserved.
Agreeableness	The tendency to act in a cooperative, unselfish manner	1) I am sometimes somewhat rude to others. 2) I have a forgiving nature. 3) I am considerate and kind to others.
Emotional Stability	A tendency to exhibit unpredictable and rapid changes in emotions	1) I worry a lot. 2) I get nervous easily. 3) I am relaxed, and handle stress well.

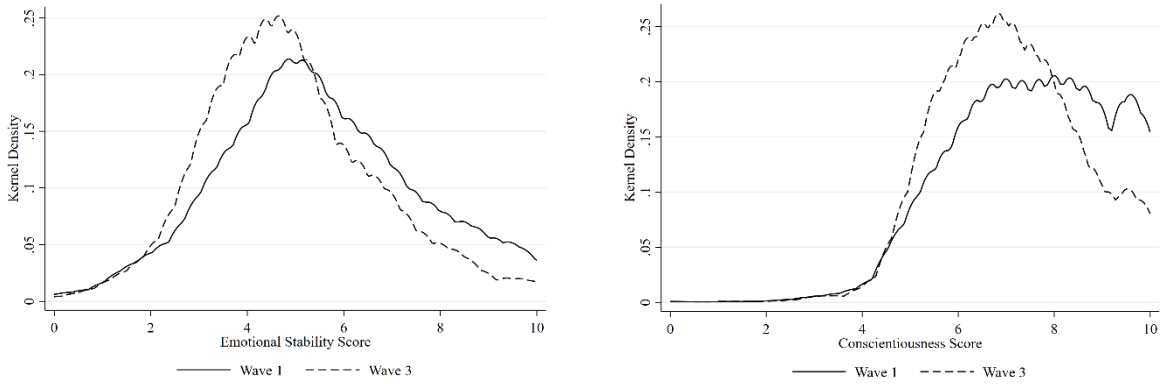
Note: The description of the Big Five personality traits is from the APA Dictionary of Psychology (VandenBos, 2015). The test statements are from Hahn et al. (2012).

Table A5: Descriptive statistics for Big Five personality traits (Panel data sample)

Variable	Wave 1	Wave 3	Diff
Conscientiousness	7.611	7.124	0.487***
	(1.635)	(1.474)	
Openness	7.077	6.797	0.280***
	(1.823)	(1.801)	
Extraversion	5.823	5.634	0.189***
	(1.828)	(1.690)	
Agreeableness	7.931	7.518	0.413***
	(1.522)	(1.424)	
Emotional Stability	5.502	4.909	0.593***
	(2.069)	(1.806)	
Obs.	1,201	1,201	

Note: The scores are the average of all items within each category. Each item is answered on a eleven-point Likert-type scale, ranging from 0="Completely disagree" to 10="Completely agree." To ensure consistency, negatively worded items are reverse scored (i.e., transformed by subtracting the original value from 10).

Figure A1. Distribution of Big Five Measures (Panel data sample)



Note: The sample size is 1,201 for the panel sample as in Table A5. The scores are the average of all three items within each personality category.