Testing the Validity of Item Non-Response as a Proxy for Cognitive and Non-Cognitive Skills

Sonja C. Kassenboehmer
Stefanie Schurer
Felix Leung

February 2015

DISCUSSION PAPER SERIES

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Testing the Validity of Item Non-Response as a Proxy for Cognitive and Non-Cognitive Skills

Sonja C. Kassenboehmer
Monash University and IZA

Stefanie Schurer
University of Sydney and IZA

Felix Leung
University of Sydney

ABSTRACT

Testing the Validity of Item Non-Response as a Proxy for Cognitive and Non-Cognitive Skills[*]

Cognitive and non-cognitive skills are important determinants of labor market outcomes, but are often unobserved. We propose a proxy for these skills derived from item non-response information and a procedure to test its validity. Exploiting a unique data-collection feature of an Australian survey, we find that fraction answered on a self-completion questionnaire fulfils all necessary requirements to be a valid proxy for cognitive skills and outperforms alternative proxies derived from paradata. Fraction answered captures a third of the effect of cognitive ability on wages and education. We provide a simple solution to reduce omitted-variable biases by up to 11%.

JEL Classification: J24, C18, C83, I20, J30

Keywords: paradata, item non-response, proxy-variable approach, cognitive ability, non-cognitive skills, wages, education

Corresponding author:

Stefanie Schurer
School of Economics
The University of Sydney
Sydney
Australia
E-mail: stefanie.schurer@sydney.edu.au

1 Introduction

One of the main goals in economic research is to identify causal relationships between policy-relevant factors and economic outcome variables. Often it is impossible to estimate causal effects because important confounding variables are not observed to the researcher – some of the most frequently mentioned in the applied microeconometric literature are cognitive ability and motivation (e.g. Wooldridge, 2010, p. 282). The reason why measures for cognitive skills are often missing from standard household surveys is that they are time intensive and invasive. Many of the most prominent nationally-representative household surveys, such as the Australian House-hold, Income and Labour Dynamics Survey (HILDA) or the German Socio-Economic Panel Survey (SOEP), have shied away from introducing cognitive ability tests up until late in their lifespan.

Measures of personality traits, sometimes referred to as non-cognitive skills (Almlund et al., 2011), are less invasive and time intensive to collect. Yet, the broadest instrument to measure non-cognitive skills in adulthood, the Big-Five personality traits, are collected rarely or too late in the lifespan of the study. For instance, two of the most comprehensive cohort studies, the National Longitudinal Study of Adolescent to Adult Health (AddHealth) and the National Child Development Study (NCDS), collected the Big-Five only in their most recent follow up, when subjects were already 30 and 50 years of age, respectively. This late collection makes it difficult for researchers to use the Big-Five measures as inputs into educational-attainment (e.g. Lundberg, 2013) or wage-regression models (e.g. Fletcher, 2013).

We propose to exploit paradata to derive a proxy for unobserved cognitive and non-cognitive skills that can help to reduce omitted-variable biases. Paradata refers to information "captured during the process of producing a survey statistic" (Kreuter, 2013, p. 3). Examples of paradata

are individual response times and response patterns, interviewer evaluations of the household, the number of contact attempts, and audio or video recordings of interviewer-respondent interactions. Until now, paradata have been used by survey methodologists to evaluate survey quality or costs (Kreuter, 2013), but more recently economists have used paradata to assess, or fix, potential estimation biases in applied economic research (Heffetz and Rabin, 2013; Kleinjans and van Soest, 2014; Behaghel et al., 2015).

In this study we posit that fraction of survey questions answered (referred to as Fraction Answered from hereon) is a valuable source of paradata to learn about a respondent's cognitive and non-cognitive skills. For many years survey methodologists have suggested that item non-response is likely to depend on the respondent's (1) knowledge about the topic, (2) judgment of the adequacy of what he or she knows relative to what the question seems to require, and (3) willingness or ability to provide complete reports (Beatty and Herrmann, 2002).[1] The first two reasons for item non-response are of a cognitive nature, while the third reason is motivational and thus dependent on the respondent's personality. A respondent's general cognitive and verbal ability plays a crucial role in understanding the question and in the decision to respond. The willingness to provide complete reports also depends on a respondent's trust in the interviewer or administering organization, perseverance to complete all questions, and care not to miss a question by accident. If cognitive factors dominate the response behavior, then Fraction Answered should strongly correlate with measures of coding speed, verbal knowledge, or memory. If motivational factors dominate response behavior, then it should strongly correlate with measures of conscientiousness, openness to experience, and agreeableness.

---

[1]Item non-response is different from unit non-response, which is affected by the social environment, the household, and the survey design, including the definition of the sampling frame, the questionnaire design, choice of mode, hiring, selection, and training of interviewers, the length of the data collection period, protocols for contacting sample units, callback rules, etc. (Groves and Couper, 1998).

The aim of this paper is to assess whether Fraction Answered is a valid proxy for either cognitive or non-cognitive skills in the context of standard wage and educational-attainment equations. We propose a three-step testing procedure based on standard statistical methods that assesses the strength, irrelevance, and independence properties of the proxy variable. Special attention will be paid to the question whether the estimation biases introduced through an imperfect proxy exceed the ones introduced through omitting relevant variables (see Wickens, 1972; McCallum, 1972; Aigner, 1974, for theoretical discussions).

For the analysis, we exploit a unique data-collection feature from the Household, Income and Labour Dynamics of Australia (HILDA) survey. The main data collection process is facilitated with professional interviewers who visit the household and interview all household members. A self-completion questionnaire (SCQ) is left with each interviewee to be filled out without help from the interviewer. We construct a measure of item non-response from this SCQ, while all of the outcome and control variables used in the analysis are derived from sections of the survey that do not feed into the item-non-response proxy. Another advantage of the data set is that the SCQ is long enough - for example, it contains up to 258 questions in wave 12 - to obtain meaningful variation in the measure of individual-specific item non-response. The data also provide other paradata that allow us to construct alternative proxy variables and evaluate the strength of Fraction Answered against these alternatives.

2 The econometric problem

2.1 Assumptions

In this section we demonstrate the conditions under which even an imperfect proxy causes less bias in the estimated parameters of interest than the omission of the unobserved variable from the model. The mechanics of the proxy-variable approach are similar to the statistical consequences of measurement error (e.g. Bollinger, 2003): the larger the error between the omitted variable and its proxy, the more noise we introduce into the system by using the proxy. Early work on the trade-off between a noisy proxy and omitting a relevant variable recommended that the former is unequivocally preferable to the latter (e.g. Wickens, 1972; McCallum, 1972), especially so in comparison to alternative, available solutions (Aigner, 1974).

We illustrate these trade-offs with an application to the Mincer wage equation adapted from Wooldridge (2003), however the same derivations hold when using educational attainment or any other outcome variable. Let log earnings (Y_i) be a function, for simplicity only, of formal training (years of education, EDU_i), non-cognitive skills (NCS_i), and cognitive skills (CS_i):

$$Y_i = \beta_0 + \beta_1 EDU_i + \beta_2 NCS_i + \beta_3 CS_i + u_i. \tag{1}$$

Let's also assume for simplicity that we do not observe cognitive skills (the same derivations hold for assuming that non-cognitive skills are unobserved), so we estimate instead:[2]

---

[2]Our application is different from e.g. Neal and Johnson (1996) in which human capital is considered to be inherently unobservable. Their study used measures from the Armed Forces Qualification Test (AFQT) and the Armed Services Vocational Aptitude Battery (ASVAB) as proxies for acquired human capital. Note, in our application acquired human capital, i.e. level of schooling, is directly and correctly observable, but cognitive ability, a subcomponent of human capital that is not necessarily shaped through the formal education process, is not observed and requires a proxy.

4

$$Y_i = \beta'_0 + \beta'_1 EDU_i + \beta'_2 NCS_i + u'_i. \tag{2}$$

Cognitive skills are likely to positively correlate with measures of non-cognitive skills (mainly openness to experience, see Fletcher and Schurer (2014)) and university education (See Wai, 2013, for an overview). Under the assumption of positive correlations between omitted and included variables and positive values of β_1 and β_2, it is straightforward to show that there is an upward bias in $\hat{\beta}'_1$ and $\hat{\beta}'_2$.

Let's assume we have the variable Fraction Answered FA_i available to be used as a proxy for CS_i to estimate log earnings as:

$$Y_i = \beta_0'' + \beta_1'' EDU_i + \beta_2'' NCS_i + \beta_3'' FA_i + u_i''. \tag{3}$$

The question is whether the estimates of β_1'' and β_2'' are closer to β_1 and β_2 than β_1' and β_2', or whether we will make things worse by including FA_i. Three assumptions must hold for the proxy-variable approach to improve upon the omitted-variable bias.

Assumption 1: Strength

The proxy must correlate strongly with the unobserved variable CS_i:

$$CS_i = \delta_0 + \delta_1 EDU_i + \delta_2 NCS_i + \delta_3 FA_i + \nu_i. \tag{4}$$

FA_i correlates with CS_i if $|\delta_3| > 0$. This condition is equivalent to the strength argument in the context of instrumental variables.

Assumption 2: Irrelevance

The proxy variable FA_i is irrelevant in the main outcome equation Eq. (1) if the unobserved variable CS_i is controlled for:

$$E(u_i|EDU, NCS, CS) = E(u_i) = 0, \tag{5}$$

and

$$E(u_i|FA) = 0. \tag{6}$$

Eq. (6) is equivalent to the validity requirement of an instrumental variable.

Assumption 3: Independence

The error term (v_i) in the proxy equation Eq. (4) is independent of education, non-cognitive skills, and Fraction Answered, or in other words, the average cognitive skill does not vary with education or non-cognitive skills once Fraction Answered is controlled for (*strong independence*).

$$E(CS_i|EDU, NCS, FA) = E(CS_i|FA) = \delta_0 + \delta_3 FA_i. \tag{7}$$

If δ_1 and δ_2 do not converge towards zero in Eq. (4), then the plug-in approach to the omitted-variable problem will incur a bias. This will almost certainly occur, because the omitted-variable problem requires fixing only if $\delta_1 \neq 0$ and $\delta_2 \neq 0$. The main question is under what condition is the incurred bias through an imperfect proxy no worse or smaller than the omitted-variable bias. It can be shown that the imperfect-proxy biases in the estimated coefficients on education and non-cognitive skills from Eq. (3) are (See Wooldridge, 2003, p. 299):

$$\text{plim}(\hat{\beta}_1'') - \beta_1 = \beta_3\delta_1, \tag{8}$$

$$\text{plim}(\hat{\beta}_2'') - \beta_2 = \beta_3\delta_2, \tag{9}$$

where δ_1 and δ_2 are obtained from Eq. (4), which we refer to from hereon as *long regression.* The coefficient β_3 measures the true effect of cognitive skills on earnings ($\beta_3 > 0$).

It can also be shown that the omitted-variable biases from Eq. (2) in each case are[3]:

$$\text{plim}(\hat{\beta}_1') - \beta_1 = \beta_3\gamma_1, \tag{10}$$

$$\text{plim}(\hat{\beta}_2') - \beta_2 = \beta_3\gamma_2, \tag{11}$$

where γ_1 and γ_2 are obtained from what we refer to as the *short regression:*

$$CS_i = \gamma_0 + \gamma_1 EDU_i + \gamma_2 NCS_i + \nu_i. \tag{12}$$

The imperfect-proxy bias is smaller than the omitted-variable bias in the coefficients of education and non-cognitive skills if:

$$|\text{plim}(\hat{\beta}_1') - \beta_1| > |\text{plim}(\hat{\beta}_1'') - \beta_1| \leftrightarrow |\gamma_1| > |\delta_1|, \tag{13}$$

$$|\text{plim}(\hat{\beta}_2') - \beta_2| > |\text{plim}(\hat{\beta}_2'') - \beta_2| \leftrightarrow |\gamma_2| > |\delta_2|. \tag{14}$$

We refer to this last condition as the *weak independence* requirement, i.e. cognitive skills vary

---

[3]A simple proof can be found in Appendix 15.

less with education and non-cognitive skills once Fraction Answered is controlled for.

2.2 Test procedure

We will test the validity of the proxy Fraction Answered for two different data applications (hourly wages, educational attainment) and two different omitted variables (cognitive skills, non-cognitive skills). We can do so because we have reasonably accurate measures of both cognitive skills and non-cognitive skills available that are derived from a different data collection process than the data collection process used to construct the proxy variable (See Section 3).

The following three-step testing procedure contains a test for each of the three assumptions (strength, irrelevance, and weak independence) outlined above. We illustrate below the test procedure for hourly wages under the assumption that cognitive skills are unobserved, while education, cognitive skills, and Fraction Answered are observed. The same derivations hold if non-cognitive skills are unobserved, while cognitive skills are observed, or if the outcome variable is educational attainment.

Step 1: Strength

To test assumption (1), we estimate a variant of Eq. (4) that includes Fraction Answered as main control variable, and adds subsequently blocks of control variables (X_i') such as sex, age, educational qualification, ethnicity, and non-cognitive skills.

$$CS_i = \alpha_0 + \alpha_1 FA_i + X_i'\theta + \zeta_i, \tag{15}$$

We test the null hypothesis $H_0 : \alpha_1 = 0$ against the two-sided alternative $H_a : \alpha_1 \neq 0$.

For Fraction Answered to be a strong proxy for cognitive skills, the null hypothesis needs to be unambiguously rejected even when adding the full set of control variables. There is no standard rule-of-thumb in the proxy-variable literature to judge whether the proxy variable is weak. We therefore report the correlation coefficient on Fraction Answered, its contribution to the explained variation in the respective skill measure, and the F-statistic that refers to the above outlined null hypothesis.

Step 2: Irrelevance

To test assumption (2), we re-estimate Eq. (3) by including our observed measure of cognitive skills:

$$Y_i = \beta_0''' + \beta_1'''EDU_i + \beta_2'''NCS_i + \beta_3'''FA_i + \beta_4'''CS_i + u_i'''. \tag{16}$$

For Fraction Answered to be a valid proxy variable, the estimated coefficient on Fraction Answered in Eq. (16) should converge towards zero. We test the null hypothesis $H_0 : \beta_3''' = 0$ against the two-sided alternative $H_a : \beta_3''' \neq 0$. For the proxy variable to pass this requirement, the null hypothesis must not be rejected.[4]

Step 3: Weak independence

---

[4]It is possible, that FA_i is correlated with u_i''', hence has an independent effect on Y_i and the coefficient may be biased. We argue that this is unlikely to be a problem in our case as we find that β_3''' is zero. Anything that would bias β_3''' towards zero would have to be negatively related to Y_i and positively related to FA_i. There are not many things one can think of that may fulfill these specific requirements as in general anything positively related to Fraction Answered would also be positively related to wages. The only possibility are opportunity costs which would reduce the number of questions answered but would be positively related to wages. However, we can show that all variables associated with opportunity costs (such as ability) are consistently positively associated with Fraction Answered. Hence, we conclude that it is unlikely that there is an independent effect of Fraction Answered on wages that biases the coefficients.

To test assumption (3), we estimate both *short* (Eq. (12)) and *long regressions* (Eq. (4)), and then test whether the difference in the absolute value of coefficients between the short and long regression is greater than zero. For each pair of coefficient comparisons we have $H_0 : |\gamma_j| - |\delta_j| = 0$ ($j = 1$ for education and $j = 2$ for non-cognitive skills) against the one-sided alternative $H_a : |\gamma_j| - |\delta_j| > 0$. If we reject the null hypothesis, then the imperfect-proxy bias will be smaller than the omitted-variable bias.

3 Data

3.1 Household, Income and Labour Dynamics in Australia Survey

We use data from the Household, Income and Labour Dynamics in Australia (HILDA) Survey which is a nationally representative household panel study conducted yearly since 2001. All adult household members (aged 15 years and above) answer the continuing or new-person questionnaire which is conducted by an interviewer. In addition, each member is asked to fill out a self-completion questionnaire (SCQ) without the help of the interviewer. The completed SCQ is then either collected on the same day or at a later date by the interviewer, or returned by mail. Annually, approximately 13,000 individuals are interviewed since 2001. A top-up sample increased the respondent number to around 17,500 since 2011.

HILDA's unique data-collection feature of complementary self-completion and interviewer-assisted surveys and its longitudinal nature allow us to separate the sources from which we construct measures of cognitive and non-cognitive skills and item non-response. Therefore, the proxy variable is not affected by non-response on the cognitive or non-cognitive skills. We use data from wave 12 (2012) when measures of cognitive skills (memory, coding speed, and language

ability) were collected as part of the interviewer-assisted survey. As in 2012 no measures of non-cognitive skills were collected, we attach to wave 12 information on the Big-Five personality traits (extraversion, agreeableness, conscientiousness, emotional stability, openness to experience) and locus of control that was collected in SCQs in previous waves (waves 3, 4, 5, 7, and 9). Our proxy Fraction Answered is calculated from the 2012 SCQ.[5]

About 90% (15,380) of the respondents returned the 2012 SCQ. Conditioning on non-missing observations in the cognitive and non-cognitive skills, we obtain a sample of 10,977 individuals. This sample further drops to 7,954 individuals when conditioning on non-missing observations in all control variables, and to 4,103 individuals when conditioning on non-missing observations in hourly wages.

3.2 Fraction Answered

To construct the number of non-response in the self-completion questionnaire (SCQ), we make use of the missing value conventions in the HILDA dataset, and count "Don't know" and "Refused"/"not answered" as non-response.[6] Only questions that appear in the questionnaire are used; derived variables are therefore excluded when constructing the non-response count. Whenever a question generates more than one variable in the dataset, the number of non-responses needs to be adjusted to avoid double counting.[7]

We discard all individuals who did not return the SCQ. The non-response fraction is calcu-

---

[5]We considered calculating a non-response proxy also from the interviewer-assisted continuing/new-person questionnaire, but there was too little variation in non-response to make it a useful proxy variable.

[6]In a robustness check we excluded the response "Don't Know" from non-response. Our conclusions remain unaffected.

[7]For example, the question on height allows a response in different units, so that two variables are provided in HILDA for the question. There are a range of questions on life events, and for each the respondent was asked whether it happened in the last 12 months, and if so to indicate whether it happened 0 to 3, 4 to 6, 7 to 9, or 10 to 12 months ago (resulting in six separate variables for each event although there are essentially only two questions on each event - "Did it happen?" and if so "How many months ago?").

lated for each individual as the ratio of the number of unanswered questions to the total number of questions that the respondent was asked. The denominator varies across individuals as some people are asked more questions than others depending on whether certain questions apply to the individual's personal situation. Figure 1(a) shows that the total range in applicable questions varies from 181 to 258 questions, and there is heaping of required responses around every additional eleven required questions. The most frequent number of required questions are 207 (8%), 235 (12%), and 256 (10%). Figure 1(b) shows that the total number of unanswered items in wave 12 varies between 0 (31%) and 200. About 90% of the sample failed to respond to up to 10 items. About 1% of the sample refused to answer 50 or more questions. A 1 standard-deviation (SD) increase in item non-response is associated with 10.5 additional questions not responded to in the SCQ.

[Insert Figures 1(a) and 1(b) here]

Figure 2 describes the distribution of the non-response rate separately for male and female respondents for non-response rates of 10% or less. The distribution is highly right-skewed, and differs across the sexes (p-value of Kolmogorov-Smirnov test of equal distribution is 0.005), although we cannot find dominance of one distribution over the other.

[Insert Figure 2 here]

3.3   Cognitive skills

An individual's cognitive skills are measured by three tests conducted in Wave 12 (2012): (1) the 'Backward digits span' test (BDS), (2) a 25-item version of the 'National American Reading Test'

(NART) and (3) the 'Symbol-digit modalities' test (SDM). Participation rates were high ($> 93\%$ in each test). Details on the tests and how they were conducted are provided in Wooden (2013).

The BDS measures working memory span and is a traditional sub-component of intelligence tests. The interviewer reads out a string of digits which the respondent has to repeat in reverse order. NART is a short version of the National Adult Reading Test that measures pre-morbid intelligence. Respondents are shown 25 irregularly spelled words which they have to read out loud and pronounce correctly. SDM was originally developed to detect cerebral dysfunction but is now a recognized test for divided attention, visual scanning and motor speed. Respondents have to match symbols to numbers according to a printed key that is given to them.

We also construct a combined measure of cognitive skills by conducting a factor analysis on all three measures and predicting the first factor. All cognitive ability measures are standardized to mean 0 and standard deviation 1.

3.4   Non-cognitive skills

To construct measures of non-cognitive skills, we use data on the Big-Five personality traits and locus of control. In waves 5 and 9, HILDA collected an inventory of the Big-Five personality traits based on Saucier (1994). It consists of 30 of Saucier's original 40 items, and six additional items were added to the survey. Ultimately, 28 of these 36 items are used to construct five aggregate scores of personality. The five dimensions–extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience–represent personality at the broadest level of abstraction. Each trait is scored from 1 to 7, with a high score indicating that the personality trait describes the individual very well. The distribution of each trait is left-skewed, which means that a larger proportion of the sample agrees with the statements about their personality underlying

each trait (See Cobb-Clark and Schurer, 2012, for a detailed description).

A measure of locus of control is derived from seven available items from the Psychological Coping Resources component of the Mastery Module developed by Pearlin and Schooler (1978), which were collected in waves 3, 4, and 7. Mastery refers to beliefs about the extent to which life's outcomes are under one's own control. Specifically, respondents were asked the extent to which they agree with each of the seven statements described in Table 7 in the Online Appendix. Possible responses range from 1 (strongly disagree) to 7 (strongly agree). We construct a continuous measure of internal locus of control using factor analysis (See Cobb-Clark and Schurer, 2013; Cobb-Clark et al., 2014, for a description).

To minimize measurement error in the measures of non-cognitive skills, we follow Cobb-Clark et al. (2014) by averaging the scores for each of the Big-Five personality traits from 2005 and 2009, and the scores for locus of control from 2003, 2004, and 2007. All personality variables are standardized to mean 0 and standard deviation 1.

3.5   Other variables

When assessing the association between cognitive or non-cognitive skills and Fraction Answered, we also control for the respondent's gender, age, education and ethnicity as we would like to see the associations net of these effects. We follow Leigh and Ryan (2008) and Li and Powdthavee (2014) to construct a continuous years of education measure from categorical variables that indicate the highest completed level of education.[8]

---

[8]Someone who completed a Masters or Doctoral degree is assigned 17 years of education; Graduate diploma or certificate is assigned 16 years; Bachelor or honours is assigned 15 years; Advanced diploma or diploma or just finishing year 12 is assigned 12 years; Finishing a Certificate III/IV is also assigned 12 years; Finishing Year 11 is assigned 11 years; Minimum schooling is assigned 9 years; and dropping out of school before Year 9 is assigned 8 years.

We also investigate the association between the interviewer's assessment of the interviewee's understanding of the questions, another variable derived from the survey's paradata, and Fraction Answered to understand pathways through which item non-response and cognitive/non-cognitive skills are related. The interviewer judges the level of understanding on a five-point scale where 1 represents "very poor" and 5 "excellent". To make the strength of the associations comparable, we standardize this variable also to mean 0 and standard deviation 1.

There is more paradata information available in our data source which could be exploited as alternative proxies for cognitive or non-cognitive skills. These are, for instance, whether the individual failed to return a SCQ in 2012 (around 10% of the sample, or 5% of our sample that includes individuals who previously returned a SCQ), the minutes an individual spent on completing the person-specific or the household-specific questionnaires, or the number of days that elapsed between the receipt and the completion of the SCQ. In an extension to the main analysis we evaluate the suitability of these measures as proxy variables relative to Fraction Answered.

In the wage equations, we additionally account for the type of employment with a binary indicator of whether the respondent is a casual worker. As a measure for wages we use information on the respondent's log hourly wages that are constructed from gross annual salaries and wages and annual hours worked. All variables used in the subsequent analysis are described (before standardization) in Table 7 in the Online Appendix.

4   Estimation results

In the subsequent sections we test whether Fraction Answered in a self-completion questionnaire is a valid proxy for either cognitive or non-cognitive skills. We test the validity of this proxy variable in the context of wage and educational-attainment regression models. To test the validity,

we follow the three-step test procedure outlined in Section 2.2. First, we test whether the correlation between Fraction Answered and individual measures of cognitive/non-cognitive skills is statistically and economically meaningful, even when controlling for a large set of confounding factors. This first step will help us to identify the particular skills with which Fraction Answered correlates the strongest (assumption (1)).

Then we move on to estimate standard wage and educational-attainment regression models that include either Fraction Answered, or both cognitive/non-cognitive skills. From these regression models we can directly test assumptions (2). The proxy variable must be redundant in the wage or educational-attainment regression when both the proxy, Fraction Answered, and the unobserved variable, either cognitive or non-cognitive skills, are included (assumption (2)). By regressing the 'unobserved' variable (e.g. cognitive skills) on both the proxy variable and the control variables used in the wage or educational-attainment regressions, we are able to test assumption (3). If the control variables of the wage or education-outcome equations are partially correlated in this long regression with the omitted variable, then the proxy-variable approach will yield biased and inconsistent estimates. However, if the partial correlation between omitted variable and the control variables becomes smaller in magnitude in the long relative to the short regression, the bias introduced by the imperfect proxy is smaller than the bias of omitting the proxied variable (Wooldridge, 2010, p.67-72).

4.1 Correlation between Fraction Answered and skills

In this section we test whether Fraction Answered is a strong proxy (assumption (1)) for cognitive or non-cognitive skills. Model (1) in Table 1 reports the bivariate correlation coefficients between Fraction Answered and each measure of cognitive and non-cognitive skills, and the contribution

of Fraction Answered to the explained variation in each specific skill. In models (2)-(4) we report the multivariate correlation coefficients between the same variables but we subsequently add control variables such as sex, age, education, and ethnicity. In model (5) we add cognitive skills as control variables into the non-cognitive skill equations and non-cognitive skills into the cognitive skill equations. The results in each row in Table 1 stem from separate regression models that differ only by their outcome variables. All outcome variables and Fraction Answered are standardized to mean 0 and standard deviation 1.

[Insert Table 1 here]

The correlation coefficients are the strongest between measures of cognitive skills and Fraction Answered, with a maximum association obtained for the Symbol-Digits-Modalities test (SDM). Fraction Answered explains 4.5% of the variation in the SDM score. A 1 SD increase in Fraction Answered, which corresponds to an additional 10.5 questions answered on the SCQ, is associated with a 0.212 SD increase in the SDM score. Fraction Answered is also associated with higher scores on the National Adult Reading (NART) (0.129 SD) and the Backward-Digit-Span (BDS) test scores (0.114 SD), and it explains 1.7% and 1.3% of the variation in NART and BDS, respectively. All correlations are highly statistically significant, with F-statistics in the order of magnitude of 113 (BDS), 394 (SDM), and 145 (NART).

When controlling for sex, age, education, ethnicity and non-cognitive skills, the associations are reduced by 30-45%, but they are still sizable and statistically significant (F-statistics are for BDS: 57, SDM: 203, NART: 100). Fraction Answered passes unambiguously the minimum requirement to be a valid proxy for cognitive ability.[9]

---

[9]The tests are passed equally strongly when using the first predicted value of all three cognitive ability measures. The correlation coefficient with Fraction Answered is 0.200 SD, Fraction Answered explains 4.1% of its variation, and the F-statistics in model (1) and (5) are 364 and 191, respectively.

Fraction Answered also correlates, although less strongly, with some of the non-cognitive skills. In order of magnitudes, the strongest correlation coefficients are obtained for openness to experience (0.09 SD), locus of control (0.07 SD), agreeableness (0.05 SD), and conscientiousness (0.03 SD). In some cases, the correlation coefficient becomes stronger when controlling for the full set of control variables (model (5)) as in the case of conscientiousness (0.04 SD) and agreeableness (0.06 SD). For openness to experience and locus of control, the correlation coefficient is reduced by two-thirds when controlling for cognitive skills. These correlation coefficients are all significant at the 1 percent level, but their F-statistics vary substantially. It is only for agreeableness, openness to experience, and locus of control that the F-statistic is large, between 20 and 70 in both models (1) and (5).[10] Fraction Answered is not related to extraversion and emotional stability. In contrast to cognitive skills, Fraction Answered explains less than 1% of the variation in each of the considered non-cognitive skills.

To ascertain that the association between Fraction Answered and cognitive skills is not due to other unobserved factors (e.g. opportunity costs), but due to a better understanding of the questions, we also report the correlations between Fraction Answered and other paradata available in our dataset (see lower panel). We find a strong correlation between the interviewer's assessment of the respondent's understanding of the questions. A 1 SD increase in the Fraction Answered is associated with a 0.134 SD increase in the interviewer's rating of the subject's understanding of the questions, which drops to 0.068 SD when controlling for both cognitive and non-cognitive skills. The F-statistics in model (1) and (5) are large (155, 101), and Fraction Answered explains almost 2% of the variation in the interviewer's rating. Fraction Answered is also

---

[10]The F-statistics in model (1) and (5) are for agreeableness: 22 and 41; for openness to experience: 669 and 20; for locus of control: 55 and 21; and for conscientiousness: 6 and 21, respectively.

positively correlated with the total time taken to answer the household questionnaire. A one-standard-deviation increase in Fraction Answered is associated with a 0.034 standard deviation increase in the minutes spent on completing the household questionnaire, however, the association is entirely explained by observable factors.[11]

That Fraction Answered is likely to capture cognitive ability can be observed in Figure 3(a), in which we plot the relationship between cognitive ability and age and compare it to the relationship between Fraction Answered and age. Both figures are based on non-parametric kernel regression methods to allow for non-linearities (See e.g. Wand and Jones, 1995). Cognitive ability, which is here measured as the first predicted factor, strongly declines from age 40 onward, and declines even more strongly from age 65 onward. Fraction Answered also starts declining from about age 40, but at a flatter profile, but the age gradient steepens substantially from age 65 onward. Figure 3(b) clarifies that the similarities in the age gradients of both Fraction Answered and cognitive ability are mainly driven by the Backward Digit Span.

[Insert Figures 3(a) and 3(a) here]

In a further robustness check, we re-estimate the bivariate relationships from model (1) also using a non-parametric kernel regression method to allow for non-linearities between Fraction Answered and skills. The estimated results are plotted in Figure 4, where on the horizontal and vertical axes we report the skill measures and the distribution of Fraction Answered, respectively.

[Insert Figures 4 here]

We conclude that the relationship between Fraction Answered and cognitive skills is approximately linear over a wide range of values for both the BDS and NART. Some non-linearities

---

[11]We could not find any association between the time spent on completing the personal questionnaire or the days elapsed between receiving and completing the SCQ. Results provided upon request.

emerge for the SDM test, openness to experience and locus of control mainly in the left tail of the skill distribution ($< -2\text{SD}$ away from the mean). For most observations in the data that lie within the range of -2 to 2 SD around the mean, the relationship is approximately linear. We no longer find a discernable relationship between Fraction Answered and agreeableness or conscientiousness. The strong non-linearities emerging in extraversion and emotional stability are deceptive; they oscillate around the zero mean of Fraction Answered, given their wide confidence intervals.

We infer from these results that ignoring the possible non-linearities between cognitive/non-cognitive skills and Fraction Answered is unlikely to weaken the strength (assumption (1)) of Fraction Answered as a proxy for cognitive skills, openness to experience, and locus of control.

4.2 Alternative proxies for cognitive and non-cognitive skills

Given the specific protocol that governs the data collection process of the HILDA Survey, we are able to derive alternative proxy variables such as 'Interviewer rating', 'No SCQ returned in 2012', 'Minutes spent on completing the unique personal or household questionnaires', and 'Days elapsed to complete SCQ since receipt'. Each of these variables could be good proxies for e.g. conscientiousness and cognitive ability. Individuals with higher cognitive abilities may be able to complete the surveys faster by being able to read, comprehend, and respond to the questions. However, less time spent on the survey could also be the result of careless behavior. Very conscientious individuals may take more time to respond to the questions by carefully reading the questions and discussing with the interviewer which information is required for a valid response. Individuals who do not return the SCQ may either be very careless (low on conscientiousness), or uncooperative (low on agreeableness). On the contrary, they may not return the SCQ because

of very high opportunity costs.

In this section we identify and discuss the strength of the correlation between these alternative proxies, relative to Fraction Answered, and both cognitive and non-cognitive skills. Table 8 in the Online Appendix displays a correlation matrix for all proxy variables, including Fraction Answered, and all measures of cognitive and non-cognitive skills. All measures are standardized so that a one-unit change reflects a move from the bottom to the top of the distribution to make the unit of analysis comparable to the binary indicator 'No SCQ'. We find that 'Minutes to complete the unique person questionnaire' (or to complete the household questionnaire) is most strongly positively correlated with openness to experience (0.09) and cognitive ability (0.06). 'No SCQ in 2012' is most strongly and negatively correlated with cognitive ability (0.06) and conscientiousness (0.08). Interviewer rating is strongly correlated with cognitive ability (0.282), openness to experience (0.09), and locus of control (0.16). No significant correlations emerge between 'Days elapsed' and any of the skill measures.

Since many of these alternative proxy variables derived from paradata also fulfill requirement (1), we test their validity as proxy in the subsequent analysis. This will allow us to evaluate the relative strength of Fraction Answered as proxy variable for unobserved skills. The strongest contender will be interviewer rating of the respondent's understanding of the questions as it correlates more strongly with cognitive ability than Fraction Answered (0.28 versus 0.21).

4.3 Application 1: Hourly wages

Having shown that Fraction Answered correlates strongly with all cognitive and some non-cognitive skill measures, we now test whether Fraction Answered passes assumptions (2) and (3) in a standard wage regression. In Table 2 we report the multivariate relationships between

log hourly wages[12] and Fraction Answered. All models are estimated with OLS and control for gender, age, age squared, age cubed, a measure for casual worker and educational attainment on a sample of 3,999 working individuals.

In model (1) we report the estimated coefficients obtained from a model that includes the basic set of control variables plus either Fraction Answered or one of the individual cognitive or non-cognitive skill measures. To test assumption (2), we include into the model with the basic set of control variables Fraction Answered, cognitive skills, and non-cognitive skills (model (2)). To pass assumption (2) the coefficient on Fraction Answered has to become zero in both statistical and economic terms, as we move from model (1) to (2).

[Insert Table 2 here]

Fraction Answered is strongly correlated with hourly wages. A 1 SD increase in the Fraction Answered is associated with a 4.2% increase in hourly wages, which is (almost) identical to the effect of an equivalent increase in the BDS test score (5.7%), conscientiousness (6.3%), and emotional stability (3.6%) on wages. The relationships with wages for the SDM and NART test scores tend to be stronger (12%).

Overall, the coefficient on Fraction Answered is no longer statistically significant at the 5% level or better when controlling for both cognitive and non-cognitive skills as is displayed in model (2) of Table 2, and therefore passes assumption (2). The estimated coefficient on Fraction Answered drops significantly by 45% from 0.042 to 0.023 (p-value < 0.001).

However, the estimation results from model (2) do not allow us to evaluate whether Fraction Answered is a valid proxy for cognitive or non-cognitive skills. We therefore estimate two further

---

[12]The same results hold when using gross annual earnings. Robustness checks are provided upon request.

models, in which we include only non-cognitive (model (3)) or cognitive skills (model (4)) in addition to Fraction Answered.[13]

When controlling for both Fraction Answered and non-cognitive skills (model (3)), the effect of Fraction Answered on wages does not drop (from 0.042 to 0.039) significantly, and is still statistically significant at the 1% level. We may have over-estimated this coefficient, because we have omitted cognitive skills, which positively correlate with Fraction Answered. Therefore, we cannot say whether Fraction Answered passes assumption (2) if proxying non-cognitive skills with certainty.

In stark contrast, when controlling for Fraction Answered and cognitive skills in model (4), the estimated coefficient on Fraction Answered decreases significantly by 40% from 0.042 to 0.025 (p-value < 0.001) and is no longer statistically significant at the 5% level or better. Since we are likely to overestimate the coefficient on Fraction Answered due to omitting non-cognitive skills, with which Fraction Answered is positively correlated, we have certainty that Fraction Answered passes assumption (2) when used as a proxy for cognitive skills.

Thus, from here onward we restrict our testing procedure for assumption (3) to Fraction Answered as a valid proxy for cognitive skills. The test results are presented in Table 3. In model (1) we present the estimated coefficients from the short regression of cognitive skills (first predicted factor standardized to mean 0, standard deviation 1) on the control variables used in the wage regression, while in model (2) we present the estimated coefficients from the long regression that

---

[13]Technically speaking, both model (3) and (4) yield biased estimates on the coefficient of Fraction Answered, because we omit either cognitive or non-cognitive skills, of which we know correlate strongly with Fraction Answered. However, we know that Fraction Answered positively correlates with all cognitive and non-cognitive skills, and so, at worst, we are likely to over-estimate the coefficient on Fraction Answered. To pass assumption (2) the coefficient on Fraction Answered requires to be zero. To commit a Type II error, that is erroneously accepting a null hypothesis, would happen in our case only if Fraction Answered negatively correlates with cognitive or non-cognitive skills. In this case, the coefficient on Fraction Answered would be downward biased.

includes also the proxy variable Fraction Answered. In column 3 we report the estimated differences between the absolute values of coefficients of models (1) and (2). Column 4 reports the relevant p-value of a test statistic which tests under the null that the coefficients in model (1) are the same as the coefficients in model (2), against the alternative that the coefficients in model (1) are larger (in absolute values). The proxy-variable approach is preferable to omitting the unobserved variable if at least some of the coefficients on the potentially endogenous variables decrease in a statistically meaningful way, while no others increase in magnitude.

[Insert Table 3 here]

We consider the change in magnitude of the estimated coefficients for both years of education and the non-cognitive skill variables that were used as predictors in the hourly-wage equation. The p-values from column 4 in Table 3 show that we reject the null hypothesis against the alternative in the case of years of education, openness to experience, and locus of control. Fraction Answered therefore passes the third requirement to be a valid proxy for cognitive skills.

Although not a perfect proxy, its bias will be smaller than the bias caused by omitting cognitive skills from the hourly wage regression. In our specific sample, the bias reduction lies between 3% for the estimated coefficient on years of education and almost 11% for the estimated coefficient on internal locus of control (See column 5 of Table 3). We conclude that Fraction Answered predominantly captures the coding-speed element of cognitive skills. Approximately 36% (0.042/0.118 in Table 2, model (1)) of the effect of the Symbol-Digits-Modalities test is explained by Fraction Answered.

4.4 Application 2: Years of education

In this section we test whether Fraction Answered passes assumption (2) and (3) in the context of educational-attainment models, following the same procedural steps as for hourly wages. We start out with a baseline OLS model in which we regress years of education on a dummy variable for being male and a quadratic polynomial of age (N=7,954). Estimation results for testing assumption (2) are reported in Table 4.

In model (1) we report the estimated coefficients obtained from a model that includes the basic set of control variables plus either Fraction Answered or one of the individual cognitive or non-cognitive skill measures. To test assumption (2), we include into the model with the basic set of control variables Fraction Answered, cognitive skills, and non-cognitive skills (model (2)). To pass assumption (2) the coefficient on Fraction Answered has to become zero in both statistical and economic terms, as we move from model (1) to (2).

[Insert Table 4 here]

Fraction Answered strongly correlates with years of education. A one-standard deviation increase in Fraction Answered is associated with a 0.1 standard-deviation increase in years of education (model (1)). The estimated coefficient on Fraction Answered is no longer statistically significant at the 1% level when controlling for both cognitive and non-cognitive skills when moving from model (1) to model (2) in Table 4. The estimated coefficient drops significantly by 76% from 0.094 SD to 0.023 SD (p-value < 0.001).

Again, model (2) does not allow us to evaluate whether Fraction Answered is a valid proxy for cognitive or non-cognitive skills. We therefore estimate two further models, in which we include, apart from the baseline control variables and Fraction Answered, either non-cognitive

skills (model (3)) or cognitive skills (model (4)). Note again, in both models (3) and (4) we are likely to overestimate the coefficient on Fraction Answered due to omitting either cognitive or non-cognitive skills. When controlling for both Fraction Answered and non-cognitive skills in model (3), the effect of Fraction Answered on years of education, although decreasing slightly, is still highly statistically significant. In stark contrast, when moving from model (1) to model (4), which controls for Fraction Answered and cognitive skills, the estimated coefficient on Fraction Answered decreases significantly by 71% from 0.094 SD to 0.027 SD (p-value < 0.001). Although still statistically significant at the 1% level, Fraction Answered is likely to pass assumption (2) when used as a proxy for cognitive skills, because of its dramatic drop in magnitude. We therefore continue our test procedure only for Fraction Answered to be used as proxy for cognitive skills.

Column 3 in Table 5 reports the estimated difference in the absolute value of coefficients of the endogenous non-cognitive skill variables between the short (model (1)) and the long regression (model (2)). The results indicate that the estimated coefficients drop significantly in magnitude for extraversion, openness to experience, and locus of control, while the difference is insignificant for agreeableness, conscientiousness and emotional stability.

[Insert Table 5 here]

Again, although not a perfect proxy, its bias will be smaller than the bias caused by omitting cognitive skills from the years-of-education regression. In our specific sample, the bias reduction in the estimated coefficients on extraversion, internal locus of control, and openness to experience are in the order of magnitude of 7.2%, 4.7%, and 2.6%, respectively (See column 5 of Table 5). These findings emphasize that our measure of Fraction Answered predominantly captures education-related variation in cognitive skills. Approximately 31% (0.094/0.304 in Table 4, model 1) of the

26

effect of the Symbol-Digits-Modalities test is explained by Fraction Answered.

4.5 Relative performance of Fraction Answered

We have shown so far that Fraction Answered passes all three requirements to be a valid proxy for cognitive skills in both wage and educational-attainment regressions, and that its application reduces the omitted-variable bias in some coefficients by up to 11%. We have also shown that alternative proxy variables also derived from paradata could be powerful solutions to the omitted-variable problem. In this section we evaluate the relative performance of Fraction Answered against these alternatives.

Table 6 summarizes the validity test results of all four proxy variables (Fraction Answered, Interviewer Rating, No matched SCQ (in 2012), and Time take for completing the household questionnaire (HHQ)).[14] For brevity, we consider the validity of these alternatives as proxies for unobserved cognitive ability in the context of a wage-regression model.

[Insert Table 6 here]

All four proxy variables unambiguously pass assumption (1). All of them correlate significantly with cognitive skills, although the strongest correlation coefficients are obtained for Fraction Answered and Interviewer Rating. However, Interviewer Rating does not pass assumption (2), nor does Time taken to complete the HHQ. The proxy variable 'No matched SCQ' technically passes assumption (2), but it does not correlate with hourly wages even when omitting cognitive and non-cognitive skills from the wage regression. It also does not pass assumption (3).

Finally, both Interviewer Ratings and Time spent to complete the HHQ pass assumption (3) and entail a reduction in the omitted variables for some variables. However, the bias-reduction

---

[14]Full estimation results are provided in the Online Appendix.

in estimated coefficients on e.g. years of education, does not exceed the bias reduction obtained through Fraction Answered.

Therefore, we are able to show that Fraction Answered is the only proxy variable derived from paradata that unambiguously passes all three requirements to be a valid proxy variable for cognitive ability.

5   Conclusions

In this study, we develop a simple three-step procedure to test the validity of a proxy variable derived from paradata that can be applied to many settings and research questions. Following the procedural steps, we are able to show that item-non-response information obtained from an unassisted self-completion questionnaire is strongly correlated with highly-validated measures of cognitive skills, and to a lesser extent with measures of non-cognitive skills. Put to the test, Fraction Answered turns out to be a valid proxy for cognitive skills, but not for non-cognitive skills, in both wage and educational-attainment regressions. It also turns out to be the only proxy variable among several alternatives equally derived from paradata to fulfill all requirements of a valid proxy variable.

This conclusion has valuable implications for the modeling strategy of applied researchers. Although not perfect, we provide a simple solution to reduce omitted-variable biases in estimated coefficients on years of education or cognitive skills in the magnitude of 3 to 11% when cognitive ability is the omitted variable.

We find the strongest associations between Fraction Answered and the Symbol-Digit-Modalities test score, a measure of coding speed that exhibits a strong age-gradient, and the National Adult Reading Test score. Favoring Fraction Answered as a proxy for coding speed and language ability

over openness to experience or locus of control is surprising, but this finding is corroborated by the interviewer's assessment about the respondent's understanding of the questions. Item non-response is largely driven by the subject's poor (cognitive) understanding of the questions, and not due to a personality-related unwillingness to complete all survey questions.

Our results may prove useful to researchers using data collected with unassisted self-completion questionnaires. The strong association between item non-response and cognitive ability points to the strong degree of self-selection of cognitively more able individuals to provide complete information. This selection may potentially bias the researcher's analysis when using most-affected questions from such self-completion questionnaires. Depending on the individual research question and variables used, the researcher may want to account for this self-selection. These results may also prove valuable to survey collectors who would like to boost their response rates. Survey collectors may need to rethink their survey design to allow people with lesser cognitive skills to understand and answer the questions.

References

Aigner DJ. 1974. Mse dominance of least squares with errors of observation. *Journal of Econometrics* **2**: 365–372.

Almlund M, Lee Duckworth A, Heckman JJ, Kautz T. 2011. Personality psychology and economics. In Eric A Hanushek SM, Woessmann L (eds), *Handbook of the Economics of Education*, volume 4. North-Holland, 1–181.

Beatty P, Herrmann DJ. 2002. To answer or not to answer: Decision processes related to survey

item nonresponse. In Groves RM, Dillman DA, Eltinge JL, Little RJA (eds), *Survey Nonresponse.* New York: John Wiley and Sons, 71–86.

Behaghel L, Crepon B, Gurgand M, Le Barbanchon T. 2015. Please call again: Correcting non-response bias in treatment effect models. *Review of Economics and Statistics* **Forthcoming**.

Bollinger C. 2003. Measurement error in human capital and the black-white wage differential. *Review of Economics and Statistics* **85**: 578–585.

Cobb-Clark DA, Kassenboehmer S, Schurer S. 2014. Healthy habits: What explains the connection between diet, exercise, and locus of control? *Journal of Economic Behavior & Organization* **98**: 1–28.

Cobb-Clark DA, Schurer S. 2012. The stability of big-five personality traits. *Economics Letters* **115**: 11–15.

Cobb-Clark DA, Schurer S. 2013. Two economists' musings on the stability of locus of control. *The Economic Journal* **123**: F358–F400.

Fletcher JJ, Schurer S. 2014. The effect of childhood health and mistreatment on adulthood personality: Evidence from siblings data. Unsubmitted manuscript, University of Sydney.

Fletcher JM. 2013. The effects of personality traits on adult labor market outcomes: Evidence from siblings. *Journal of Economic Behavior & Organization* **89**: 122–135.

Groves R, Couper M. 1998. *Nonresponse in Household Interview Surveys.* New York: Wiley.

Heffetz O, Rabin M. 2013. Conclusions regarding cross-group differences in happiness depend on difficulty of reaching respondents. *American Economic Review* **103**.

Kleinjans KJ, van Soest A. 2014. Rounding, focal point answers and nonresponse to subjective probability questions. *Journal of Applied Econometrics* **29**: 567–585.

Kreuter F. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information.* John Wiley & Sons.

Leigh A, Ryan C. 2008. Estimating returns to education using different natural experiment techniques. *Economics and Education Review* **27**: 2.

Li J, Powdthavee N. 2014. Does more education lead to better health habits? evidence from the school reforms in australia. *Social Science & Medicine* **Forthcoming**.

Lundberg S. 2013. The college type: Personality and educational inequality. *Journal of Labor Economics* **31**: 421–441.

McCallum BT. 1972. Relative asymptotic bias from errors of omission and measurement. *Econometrica* **40**: 757–758.

Neal DA, Johnson WR. 1996. The role of premarket factors in black-white wage differences. *Journal of Political Economy* **104**: 869–895.

Pearlin LI, Schooler C. 1978. The structure of coping. *Journal of Health and Social Behavior* **19**: 2–21.

Saucier G. 1994. Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment* **63**: 506–516.

Wai J. 2013. Investigating America's elite: Cognitive ability, education, and sex differences. *Intelligence* **41**: 203–211.

Wand MP, Jones M. 1995. *Kernel smoothing. Monographs on Statistics and Applied Probability.* London: Chapman and Hall.

Wickens MR. 1972. A note on the use of proxy variables. *Econometrica* **40**: 759–760.

Wooden M. 2013. The measurement of cognitive ability in wave 12 of the hilda survey. HILDA Project Discussion Paper Series 1/13, Melbourne Institute of Applied Economic and Social Research.

Wooldridge JM. 2003. *Introductory Econometrics - A modern Approach.* Mason, OH: Souther-Western.

Wooldridge JM. 2010. *Econometric Analysis of Cross Sectiona And panel Data.* The MIT Press.

(a) Required responses   (b) Non-response
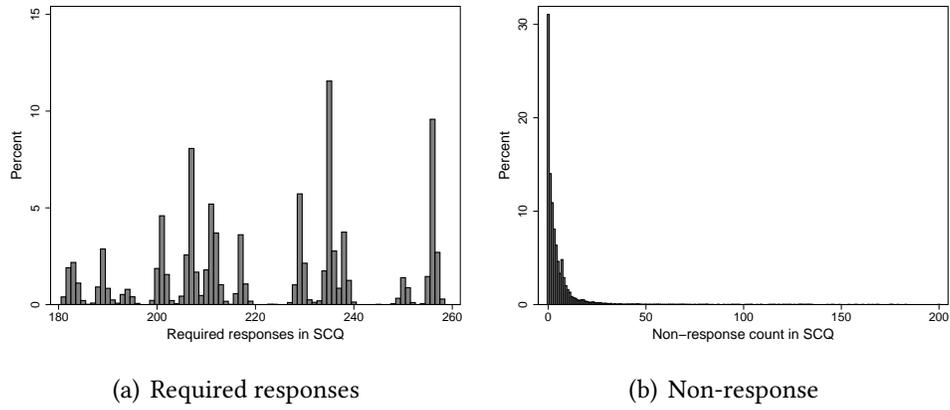
Figure 1: Distribution of required number of questions and of count of non-responded questions in SCQ (wave 12)
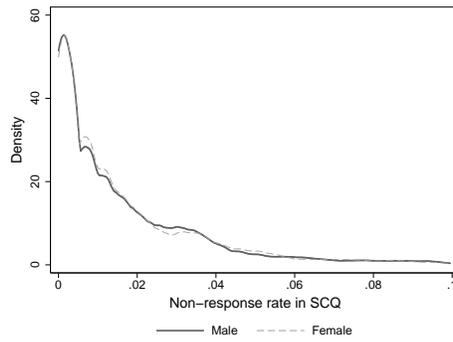
Figure 2: Distribution of non-response rate in SCQ (wave 12), by sex (p-value of Kolmogorov-Smirnov test of equal distribution: p=0.005)

Table 1: Effect of 'Fraction Answered' on cognitive and non-cognitive skills

Outcome	(1) No controls		(2) Sex/Age		(3) Education		(4) Ethnicity		(5) CS or NCS	
	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2
BDS	.114*** (.011)	.013	.101*** (.011)	.024	.081*** (.011)	.064	.079*** (.011)	.067	.074*** (.011)	.077
SDM	.212*** (.011)	.045	.143*** (.009)	.394	.122*** (.009)	.431	.121*** (.009)	.432	.115*** (.009)	.445
NART25	.129*** (.011)	.017	.144*** (.011)	.030	.099*** (.010)	.211	.093*** (.010)	.254	.084*** (.010)	.300
Pred. Factor[a]	.200*** (.011)	.041	.171*** (.011)	.106	.133*** (.010)	.238	.130*** (.010)	.255	.121*** (.010)	.288
Extraversion	.006 (.011)	-.000	-.001 (.011)	.023	-.003 (.011)	.024	-.003 (.011)	.024	-.006 (.011)	.024
Agreeableness	.050*** (.011)	.002	.067*** (.011)	.086	.062*** (.011)	.088	.063*** (.011)	.088	.061*** (.011)	.089
Conscientiousn.	.031** (.011)	.001	.060*** (.011)	.056	.051*** (.011)	.063	.051*** (.011)	.064	.039*** (.011)	.071
Emotional stab.	-.020 (.011)	.000	.014 (.011)	.077	.008 (.011)	.081	.006 (.011)	.086	-.009 (.011)	.094
Openness	.089*** (.011)	.008	.074*** (.011)	.022	.045*** (.011)	.092	.045*** (.011)	.095	.027* (.011)	.131
Locus of control	.071*** (.011)	.005	.059*** (.011)	.015	.045*** (.011)	.031	.042*** (.011)	.040	.018 (.011)	.058
Interv. rating	.134*** (.011)	.018	.122*** (.011)	.027	.111*** (.011)	.037	.108*** (.011)	.048	.068*** (.011)	.100
Min. complete HHQ	.039*** (.011)	.001	.026* (.011)	.015	.021 (.011)	.019	.020 (.011)	.019	.015 (.012)	.022

Models are estimated with OLS (N=7,930 individuals); all measures of cognitive and non-cognitive skills are standardized to mean 0 and standard deviation 1. [a]: First predicted factor of the three measures of cognitive ability. In model (5) cognitive skills are included as control variables in all non-cognitive skill equations, non-cognitive skills are included as control variables in all cognitive skill equations, and both are included in the interviewer rating equation. ** p<0.05, *** p<0.01.

34
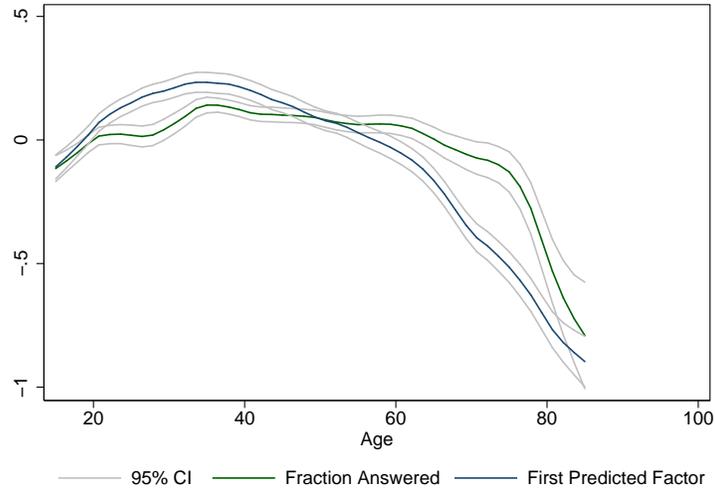
(a) First predicted factor
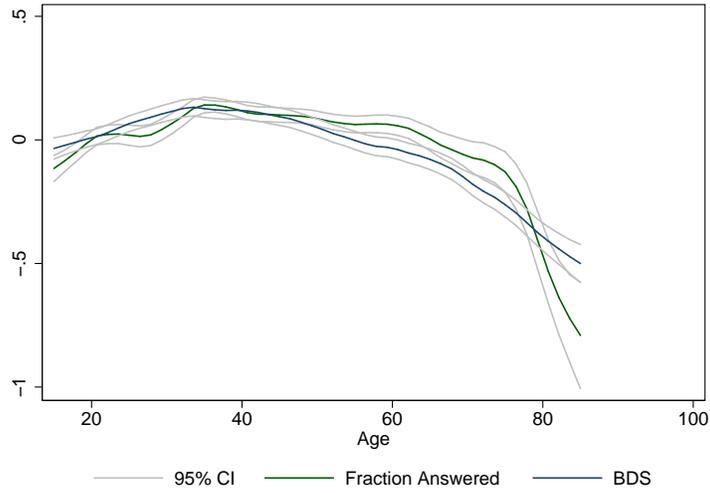
(b) Backward-Digit Span

Figure 3: Non-linearities in the relationship of Fraction Answered and Cognitive Ability with Age (15 to 85 only) (bivariate non-parametric kernel regression)

(a) BDS　　　　　(b) SDM　　　　　(c) NART 25

(d) Extraversion　　　(e) Agreeableness　　　(f) Conscientiousness

(g) Emotional stability　(h) Openness to experience　(i) Locus of control
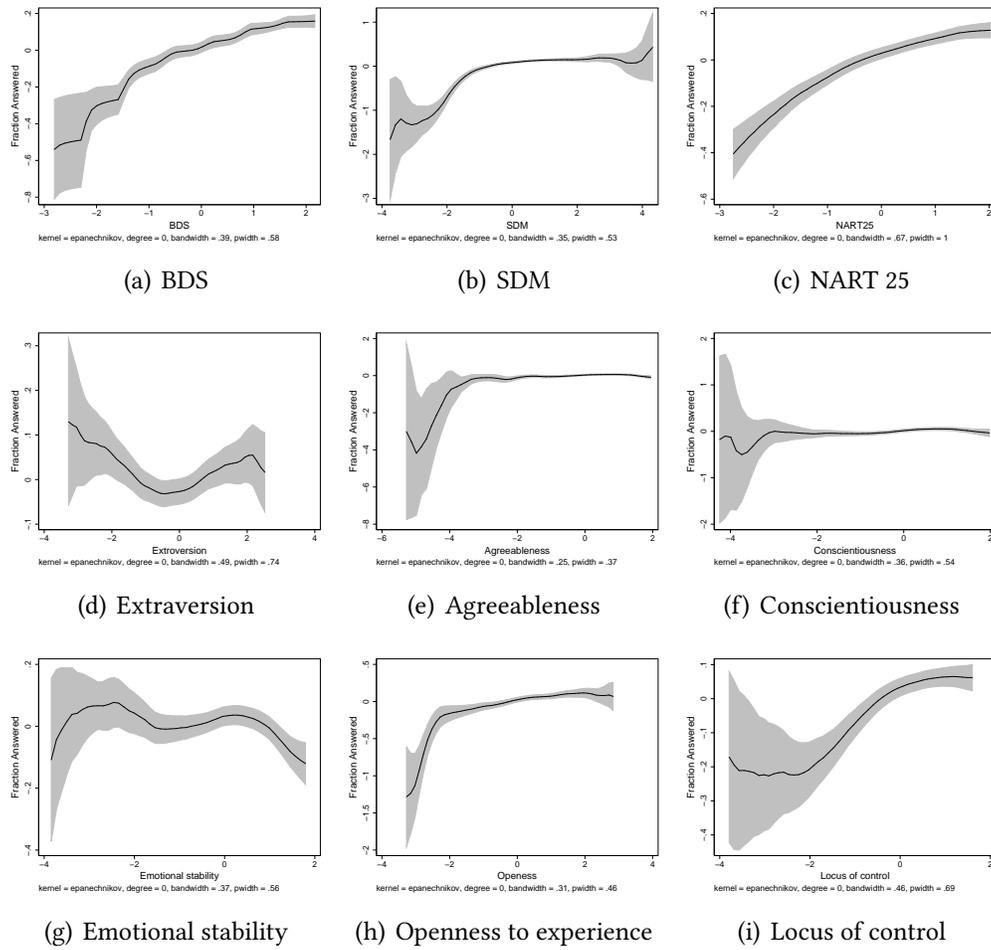
Figure 4: Non-linearities in the relationship between cognitive or non-cognitive skills and Fraction Answered (bivariate non-parametric kernel regression)

36

Table 2: Effect of Fraction Answered on log hourly wages

Variables	Incl. NCS/CS indiv.		Incl. FA jointly with NCS and/or CS					
			NCS & CS		NCS		CS	
	Model (1)		Model (2)		Model (3)		Model (4)	
	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2
Fraction Answered	.042*** (.014)	.259	.023 (.014)	.284	.039*** (.014)	.268	.025 (.014)	.277
BDS	.057*** (.014)	.261	.006 (.015)				.008 (.015)	
SDM	.124*** (.015)	.270	.093*** (.016)				.098*** (.016)	
NART25	.118*** (.015)	.269	.090*** (.016)				.089*** (.016)	
Extraversion	.031** (.014)	.258	.022 (.014)		.014 (.014)			
Agreeableness	-.019 (.014)	.258	-.038** (.015)		-.049*** (.015)			
Conscientiousness	.063*** (.014)	.261	.054*** (.015)		.052*** (.015)			
Emotional stability	.036** (.014)	.259	-.010 (.016)		.007 (.016)			
Openness to exp.	.012 (.014)	.257	-.004 (.015)		.025 (.015)			
Locus of control	.080*** (.014)	.264	.059*** (.015)		.067*** (.015)			

All models are estimated with OLS (N=3,999 working individuals observed in Wave 12) and control for gender, age, age squared, age cubed, a measure for casual worker, and educational attainment (Base level: Dropping out of high school of finishing year 11). The basic model without cognitive or non-cognitive skills explains 25.7% of the variation in hourly wages. A test of equality of coefficients on Fraction Answered between column (1) and (2) is not rejected (p-value: 0.1206), while equality of coefficients between columns (1) and (3) is rejected (p-value < 0.001). All measures of cognitive ability and non-cognitive skills are standardized to mean 0 and standard deviation 1. Standard errors are reported in parentheses. ** $p<0.05$, *** $p<0.01$.

Table 3: Test of validity of proxy variable Fraction Answered in a wage-regression application

	Model (1)	Model (2)	Diff[a]	p-value[b]	Bias[c] ↓ (%)
Years of education	.277***	.269***	.00794	.000512	2.870
	(.0155)	(.0154)			
Extraversion	-.0410***	-.0403***	.000719	.280	1.755
	(.0153)	(.0152)			
Agreeableness	-.0594***	-.0618***	-.00240	.872	-4.044
	(.0166)	(.0165)			
Conscientiousness	-.00578	-.00553	.000244	.429	4.229
	(.0160)	(.0159)			
Emotional stability	.101***	.103***	-.00141	.818	-1.394
	(.0170)	(.0169)			
Openness to exp.	.175***	.172***	.00379	.00612	2.158
	(.0163)	(.0162)			
Locus of control	.0523***	.0467***	.00561	.00159	10.73
	(.0162)	(.0161)			
Fraction Answered		.103***			
		(.0146)			
R^2	0.160	0.170			
Observations	3999	3999			

All models are estimated with OLS and control for a cubic polynomial of age, gender, and casual worker. The dependent variable is the first predicted factor of three measures of cognitive ability, and both dependent variable, years of education, and non-cognitive skills are standardized to mean 0, and standard deviation of 1. [a] Estimated differences and standard errors between absolute values of coefficients obtained from model (1) and model (2). [b] the p-value for rejecting the null hypothesis of equality of the coefficients in absolute terms against the alternative hypothesis that the coefficient in Model(1) is greater than that in Model(2), in absolute term. A statistically-significant positive difference indicates that the omitted-variable bias would be larger than the imperfect-proxy bias. [c] Bias reduction is calculated as the percentage change in the omitted variable bias, calculated as: $\frac{(OVB - IPB)}{OVB} * 100$. ** p<0.05, *** p<0.01.

Table 4: Effect of Fraction Answered on years of education

Variables	Incl. NCS/CS indiv.		Incl. FA jointly with NCS and/or CS					
			NCS & CS		NCS		CS	
	Model (1)		Model (2)		Model (3)		Model (4)	
	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2
Fraction Answered	.094*** (.011)	.090	.023** (.010)	.295	.071*** (.010)	.179	.027*** (.010)	.268
BDS	.197*** (.011)	.119	.013 (.011)				.014 (.011)	
SDM	.304*** (.014)	.136	.106*** (.014)				.120*** (.014)	
NART25	.425*** (.010)	.259	.329*** (.011)				.383*** (.011)	
Extraversion	.027** (.011)	.082	-.007 (.010)		-.025** (.011)			
Agreeableness	.042*** (.011)	.083	-.046*** (.011)		-.081*** (.012)			
Conscientiousness	.079*** (.011)	.087	.034*** (.011)		.023** (.011)			
Emotional stability	.068*** (.011)	.085	.033*** (.011)		.091*** (.012)			
Openness to exp.	.262*** (.010)	.148	.175*** (.011)		.294*** (.011)			
Locus of control	.134*** (.011)	.099	.045*** (.011)		.100*** (.011)			

The dependent variable is years of education, standardized to mean 0, and SD 1. All models are estimated with OLS (N=7,954 individuals observed in Wave 12) and control for gender, and a cubic polynomial of age. [α]: First predicted factor of the three measures of cognitive ability. The basic model without cognitive or non-cognitive skills explains 8.13% of the variation in years of education. Standard errors are reported in parentheses. A test of equality of coefficients on Fraction Answered between column (1) and (2), (1) and 3, or (2) and (3) is rejected (all p-values are less than 0.001). ** $p<0.05$, *** $p<0.01$.

Table 5: Test of validity of proxy variable Fraction Answered in an educational-attainment regression application

	Model (1)	Model (2)	Diff[a]	p-value[b]	Bias[c] ↓ (%)
Extraversion	-.0393***	-.0365***	.00285	.0225	7.244
	(.0107)	(.0106)			
Agreeableness	-.0745***	-.0792***	-.00477	.993	-6.410
	(.0116)	(.0115)			
Conscientiousness	.000670	-.00348	-.00281	.550	–
	(.0113)	(.0112)			
Emotional stability	.138***	.137***	.000722	.328	.522
	(.0122)	(.0121)			
Openness to exp.	.255***	.249***	.00656	.00000162	2.571
	(.0111)	(.0110)			
Locus of control	.132***	.125***	.00622	.00206	4.723
	(.0114)	(.0113)			
Fraction Answered		.132***			
		(.0103)			
R^2	0.184	0.201			
Observations	7954	7954			

All models are estimated with ordinary least squares and control for a cubic polynomial of age and gender. The dependent variable is the first predicted factor of three measures of cognitive ability, standardized to mean 0, and standard deviation of 1. [a] Estimated differences and standard errors between absolute values of coefficients obtained from model (1) and model (2). [b] the p-value for rejecting the null hypothesis of equality of the coefficients in absolute terms against the alternative hypothesis that the coefficient in Model(1) is greater than that in Model(2), in absolute term. A statistically-significant positive difference indicates that the omitted-variable bias would be larger than the imperfect-proxy bias. [c] Bias reduction is calculated as the percentage change in the omitted variable bias, calculated as: $\frac{(OVB-IPB)}{OVB} * 100$. ** p<0.05, *** p<0.01.

Table 6: Relative performance of Fraction Answered as proxy variable for cognitive ability

Proxy	Ass. (1)	Ass. (2)	Ass. (3)	Bias[c] ↓ (%)
Fraction Answered	0.205***	✓	✓	Yrs education 3%
				Openness 2.2%
				Locus of control 10.7%
Interviewer Rating	0.282***	✗	✓	Yrs education 3%
				Emo. Stab. 5.8%
				Openness 3%
				Locus of control 13%
No matched SCQ	.0574***	✓	✗	✗
Time taken for HHQ	0.0691***	✗	✓	Yrs education <1%
				Locus of control 4.3%

[c] Bias reduction is calculated as the percentage change in the omitted variable bias, calculated as: $\frac{(OVB - IPB)}{OVB} * 100$. ** p<0.05, *** p<0.01.

A Online Appendix

Table 7: Summary statistics

Variable definition	N	Mean	SD	Min	Max
Count non-response	7954	4.34	10.48	0	200
Count required responses	7954	223.71	20.95	181	258
Non-response rate	7954	0.02	0.05	0	0.958
Fraction Answered	7954	0.98	0.05	0.042	1
Age	7954	49.54	17.19	20	100
Male	7954	0.46	0.50	0	1
Backward-Digit-Span (BDS)	7954	3.95	1.41	0	7
Symbol-Digit-Modalities (SDM)	7954	48.31	12.93	0	104
National-Adult-Reading-Test (NART)	7954	14.39	5.24	0	25
Extraversion (Avg Wave 5 and 9)	7954	4.43	1.02	1.083	7
Agreeableness (Avg Wave 5 and 9)	7954	5.38	0.83	1	7
Conscientiousness (Avg Wave 5 and 9)	7954	5.09	0.96	1	7
Emotional stability (Avg Wave 5 and 9)	7954	5.22	0.99	1.42	7
Openness to experience (Avg Wave 5 and 9)	7954	4.22	0.98	1	7
Locus of control (Avg Wave 3, 4 and 7)					
Little control over life	7954	2.752	1.244	1	7
No way to solve problems	7954	2.560	1.282	1	7
Cannot change the important things	7954	2.571	1.253	1	7
Feel helpless	7954	2.508	1.267	1	7
Pushed around	7954	2.613	1.325	1	7
Future depends on me	7954	5.569	1.174	1	7
Can do just about everything	7954	5.348	1.187	1	7
Hourly wages	3999	32.650	17.667	1.611	312.209
Wages & salary	3999	63567	44147	991	700483
Casual worker	3999	0.184	0.388	0	1
Years of education	7954	12.35	2.28	8	17
Finished Year 11 or less	7954	0.26	0.44	0	1
Finished Year 12	7954	0.15	0.35	0	1
Certificate III/IV	7954	0.22	0.42	0	1
Diploma (TAFE)	7954	0.10	0.30	0	1
Bachelors degree	7954	0.15	0.36	0	1
Postgraduate qualification	7954	0.12	0.32	0	1
Interviewer rating of understanding					
Excellent	7954	.788	.408	0	1
Good	7954	.194	.396	0	1
Fair	7954	.015	.124	0	1
Poor	7954	.001	.025	0	1
Very poor	7954	.0002	.015	0	1

Table 8: Correlation matrix

	Cogn. abil.	Extra version	Agree able	Consc tious	Emot. Stab.	Open- ness	LoC	Frac. Answ.	Days elapsed	HQ length	PQ length	No SCQ
Cogn. ability[a]	1											
Extraversion	.0459*	1										
Agreeableness	.0377*	.160*	1									
Conscientiousness	.0273*	.104*	.306*	1								
Emotional stability	.0332*	.170*	.160*	.314*	1							
Openness to exp.	.241*	.0678*	.284*	.0655*	-.220*	1						
Locus of control	.198*	.224*	.128*	.232*	.336*	.0475*	1					
Fraction Answered	.205*	.00930	.0524*	.0314*	-.0127	.0937*	.0919*	1				
Days btw PQ and SCQ	.0107	.0233	-.00364	-.0253	-.0201	.0220	.0131	-.0123	1			
Min. complete HHQ	.0607*	.0326*	.0112	-.00999	-.0359*	.0152	.0461*	.0313*	-.0341*	1		
Min. complete PQ	.0601*	.0474*	.0457*	-.00517	-.0316*	.0987*	.0130	.0266*	-.0667*	.202*	1	
Interv. rating	.282*	.0526*	.0427*	.0491*	.0375*	.0920*	.163*	.170*	.0198	-.0146	.00665	1
SCQ matched	.0574*	-.0355*	.0382*	.0808*	.0542*	.0137	.0677*	0	0	.0360*	.0717*	.132*	1

Cognitive ability is measured as the first predicted factor of the three measures of cognitive ability (BDS, SDM, and NART 25). * p<0.05.

Table 9: Effect of Returning SCQ on log hourly wages

Variables	Incl. NCS/CS indiv.		Incl. FA jointly with NCS and/or CS					
			NCS & CS		NCS		CS	
	Model (1)		Model (2)		Model (3)		Model (4)	
	Coeff (SE)	Adj. R²	Coeff (SE)	Adj. R²	Coeff (SE)	Adj. R²	Coeff (SE)	Adj. R²
SCQ matched	.088 (.059)	.262	.048 (.058)	.288	.066 (.059)	.271	.066 (.058)	.281
BDS	.060*** (.013)	.265	.008 (.014)				.010 (.014)	
SDM	.126*** (.014)	.275	.096*** (.015)				.101*** (.015)	
NART25	.119*** (.015)	.273	.094*** (.016)				.091*** (.016)	
Extraversion	.033** (.013)	.263	.024 (.014)		.017 (.014)			
Agreeableness	-.020 (.014)	.262	-.036** (.015)		-.047*** (.015)			
Conscientiousness	.062*** (.014)	.265	.052*** (.014)		.049*** (.015)			
Emotional stability	.036*** (.013)	.263	-.014 (.015)		.004 (.015)			
Openness to exp.	.007 (.014)	.262	-.010 (.015)		.020 (.015)			
Locus of control	.082*** (.013)	.268	.062*** (.015)		.071*** (.015)			

All models are estimated with OLS (N=4,228 working individuals observed in Wave 12) and control for gender, age, age squared, age cubed, a measure for casual worker, and educational attainment (Base level: Dropping out of high school of finishing year 11). All measures of cognitive ability and non-cognitive skills are standardized to mean 0 and standard deviation 1. Standard errors are reported in parentheses. ** $p<0.05$, *** $p<0.01$.

Table 10: Test of validity of proxy variable Returning SCQ in a wage-regression application

	Model (1)	Model (2)	Diff[a]	p-value[b]	Bias[c] (%)
Years of education	.280***	.280***	.000246	.244	.0877
	(.0150)	(.0150)			
Extraversion	-.0396***	-.0391***	.000493	.193	1.244
	(.0149)	(.0149)			
Agreeableness	-.0561***	-.0560***	.0000915	.359	.163
	(.0161)	(.0161)			
Conscientiousness	-.00951	-.0101	-.000641	.811	-6.742
	(.0156)	(.0156)			
Emotional stability	.100***	.100***	.000269	.237	.268
	(.0166)	(.0166)			
Openness to exp.	.173***	.174***	-.0000621	.606	-.0358
	(.0159)	(.0159)			
Locus of control	.0535***	.0531***	.000460	.205	.859
	(.0157)	(.0157)			
SCQ matched		.0593			
		(.0630)			
R^2	0.160	0.160			
Observations	4228	4228			

All models are estimated with OLS and control for a cubic polynomial of age, gender, and casual worker. The dependent variable is the first predicted factor of three measures of cognitive ability, and both dependent variable, years of education, and non-cognitive skills are standardized to mean 0, and standard deviation of 1. [a] Estimated differences and standard errors between absolute values of coefficients obtained from model (1) and model (2). [b] the p-value for rejecting the null hypothesis of equality of the coefficients in absolute terms against the alternative hypothesis that the coefficient in Model(1) is greater than that in Model(2), in absolute term. A statistically-significant positive difference indicates that the omitted-variable bias would be larger than the imperfect-proxy bias. [c] Bias reduction is calculated as the percentage change in the omitted variable bias, calculated as: $\frac{(OVB-IPB)}{OVB}*100$. ** p<0.05, *** p<0.01.

Variables	Incl. NCS/CS indiv.		Incl. FA jointly with NCS and/or CS					
	Model (1)		Model (2)		Model (3)		Model (4)	
			NCS & CS		NCS		CS	
	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2	Coeff (SE)	Adj. R^2
Interv. rating	.158*** (.032)	.266	.091*** (.032)	.290	.143*** (.032)	.274	.102*** (.032)	.283
BDS	.060*** (.013)	.265	.008 (.014)				.010 (.014)	
SDM	.126*** (.014)	.275	.093*** (.015)				.097*** (.015)	
NART25	.119*** (.015)	.273	.088*** (.016)				.084*** (.016)	
Extraversion	.033** (.013)	.263	.023 (.014)		.015 (.014)			
Agreeableness	-.020 (.014)	.262	-.036** (.015)		-.046*** (.015)			
Conscientiousness	.062*** (.014)	.265	.053*** (.014)		.050*** (.014)			
Emotional stability	.036*** (.013)	.263	-.015 (.015)		.002 (.015)			
Openness to exp.	.007 (.014)	.262	-.010 (.015)		.018 (.015)			
Locus of control	.082*** (.013)	.268	.061*** (.015)		.068*** (.015)			

All models are estimated with OLS (N=4,228 working individuals observed in Wave 12) and control for gender, age, age squared, age cubed, a measure for casual worker, and educational attainment (Base level: Dropping out of high school of finishing year 11). All measures of cognitive ability and non-cognitive skills are standardized to mean 0 and standard deviation 1. Standard errors are reported in parentheses. ** p<0.05, *** p<0.01.

Table 12: Test of validity of proxy variable Interviewer Rating in a wage-regression application

	Model (1)	Model (2)	Diff[a]	p-value[b]	Bias[c] (%)
Years of education	.280***	.272***	.00826	.000418	2.946
	(.0150)	(.0149)			
Extraversion	-.0396***	-.0412***	-.00157	.785	-3.969
	(.0149)	(.0148)			
Agreeableness	-.0561***	-.0535***	.00262	.125	4.669
	(.0161)	(.0160)			
Conscientiousness	-.00951	-.00964	-.000133	.525	-1.401
	(.0156)	(.0155)			
Emotional stability	.100***	.0945***	.00585	.00743	5.828
	(.0166)	(.0165)			
Openness to exp.	.173***	.168***	.00522	.0109	3.011
	(.0159)	(.0157)			
Locus of control	.0535***	.0466***	.00698	.00291	13.03
	(.0157)	(.0156)			
Interv. rating		.316***			
		(.0337)			
R^2	0.160	0.177			
Observations	4228	4228			

All models are estimated with OLS and control for a cubic polynomial of age, gender, and casual worker. The dependent variable is the first predicted factor of three measures of cognitive ability, and both dependent variable, years of education, and non-cognitive skills are standardized to mean 0, and standard deviation of 1. [a] Estimated differences and standard errors between absolute values of coefficients obtained from model (1) and model (2). [b] the p-value for rejecting the null hypothesis of equality of the coefficients in absolute terms against the alternative hypothesis that the coefficient in Model(1) is greater than that in Model(2), in absolute term. A statistically-significant positive difference indicates that the omitted-variable bias would be larger than the imperfect-proxy bias. [c] Bias reduction is calculated as the percentage change in the omitted variable bias, calculated as: $\frac{(OVB-IPB)}{OVB} * 100$. ** p<0.05, *** p<0.01.

Table 13: Effect of HHQ length on log hourly wages

Variables	Incl. NCS/CS		Incl. FA jointly with NCS and/or CS					
	indiv.		NCS & CS		NCS		CS	
	Model (1)		Model (2)		Model (3)		Model (4)	
	Coeff	Adj.	Coeff	Adj.	Coeff	Adj.	Coeff	Adj.
	(SE)	R^2	(SE)	R^2	(SE)	R^2	(SE)	R^2
Min. complete HHQ	.048***	.264	.039***	.290	.045***	.273	.042***	.283
	(.014)		(.014)		(.014)		(.014)	
BDS	.059***	.265	.006				.008	
	(.013)		(.014)				(.014)	
SDM	.128***	.275	.097***				.102***	
	(.015)		(.015)				(.015)	
NART25	.121***	.273	.094***				.091***	
	(.015)		(.016)				(.016)	
Extraversion	.031**	.262	.020		.012			
	(.013)		(.014)		(.014)			
Agreeableness	-.019	.262	-.035**		-.046***			
	(.014)		(.015)		(.015)			
Conscientiousness	.062***	.265	.055***		.052***			
	(.014)		(.014)		(.015)			
Emotional stability	.036***	.263	-.013		.006			
	(.014)		(.015)		(.016)			
Openness to exp.	.008	.262	-.008		.023			
	(.014)		(.015)		(.015)			
Locus of control	.082***	.268	.061***		.069***			
	(.013)		(.015)		(.015)			

All models are estimated with OLS (N=4,228 working individuals observed in Wave 12) and control for gender, age, age squared, age cubed, a measure for casual worker, and educational attainment (Base level: Dropping out of high school of finishing year 11). All measures of cognitive ability and non-cognitive skills are standardized to mean 0 and standard deviation 1. Standard errors are reported in parentheses. ** $p<0.05$, *** $p<0.01$.

Table 14: Test of validity of proxy variable HHQ length in a wage-regression application

	Model (1)	Model (2)	Diff[a]	p-value[b]	Bias[c] (%)
Years of education	.280***	.278***	.00221	.0121	.788
	(.0151)	(.0151)			
Extraversion	-.0406***	-.0420***	-.00147	.958	-3.629
	(.0150)	(.0150)			
Agreeableness	-.0573***	-.0567***	.000603	.203	1.052
	(.0162)	(.0162)			
Conscientiousness	-.0102	-.00916	.00105	.0691	10.28
	(.0157)	(.0157)			
Emotional stability	.101***	.103***	-.00165	.965	-1.636
	(.0167)	(.0167)			
Openness to exp.	.174***	.175***	-.00129	.945	-.743
	(.0160)	(.0160)			
Locus of control	.0557***	.0533***	.00243	.00751	4.353
	(.0158)	(.0158)			
Min. complete HHQ		.0411***			
		(.0147)			
R^2	0.159	0.161			
Observations	4189	4189			

All models are estimated with OLS and control for a cubic polynomial of age, gender, and casual worker. The dependent variable is the first predicted factor of three measures of cognitive ability, and both dependent variable, years of education, and non-cognitive skills are standardized to mean 0, and standard deviation of 1. [a] Estimated differences and standard errors between absolute values of coefficients obtained from model (1) and model (2). [b] the p-value for rejecting the null hypothesis of equality of the coefficients in absolute terms against the alternative hypothesis that the coefficient in Model(1) is greater than that in Model(2), in absolute term. A statistically-significant positive difference indicates that the omitted-variable bias would be larger than the imperfect-proxy bias. [c] Bias reduction is calculated as the percentage change in the omitted variable bias, calculated as: $\frac{(OVB-IPB)}{OVB} * 100$. ** p<0.05, *** p<0.01.

Proof 15: Omitted Variable Bias in the Multivariate Case

$$Y_i = \beta_0 + \beta_1 EDU_i + \beta_2 NCS_i + \beta_3 CS_i + u_i. \tag{17}$$

If we would omit CS_i, we would estimate the following equation instead:

$$Y_i = \beta_0' + \beta_1' EDU_i + \beta_2' NCS_i + u_i' \tag{18}$$

However, CS_i is related to the other regressors:

$$CS_i = \gamma_0 + \gamma_1 EDU_i + \gamma_2 NCS_i + v_i. \tag{19}$$

By plugging in CS_i in Eq. (17), we get the following:

$$
\begin{aligned}
Y_i = \quad & \beta_0 + \beta_1 EDU_i + \beta_2 NCS_i + \beta_3(\gamma_0 + \gamma_1 EDU_i + \gamma_2 NCS_i + v_i) + u_i. \\
= \quad & \underbrace{(\beta_0 + \beta_3\gamma_0)}_{\hat{\beta_0'}} + \underbrace{(\beta_1 + \beta_3\gamma_1)}_{\hat{\beta_1'}} EDU_i + \underbrace{(\beta_2 + \beta_3\gamma_2)}_{\hat{\beta_2'}} NCS_i + \underbrace{\beta_3 v_i + u_i}_{\hat{u_i'}}.
\end{aligned}
$$

Hence, $\hat{\beta_0'}$ is an estimate of $(\beta_0 + \beta_3\gamma_0)$ instead of β_0', $\hat{\beta_1'}$ is an estimate of $(\beta_1 + \beta_3\gamma_1)$ instead of β_1, $\hat{\beta_2'}$ is an estimate of $(\beta_2 + \beta_3\gamma_2)$ instead of β_2 and $\hat{u_i'}$ is an estimate of $(\beta_3 v_i + u_i)$ instead of u_i'. The biases for our coefficient of interest are therefore:

$$\hat{\beta_1} - \beta_1 = \beta_1 + \beta_3\gamma_1 - \beta_1 = \beta_3\gamma_1 \tag{20}$$

$$\hat{\beta_2} - \beta_2 = \beta_2 + \beta_3\gamma_3 - \beta_2 = \beta_3\gamma_2 \tag{21}$$