Survey Response Behavior as a Proxy for Unobserved Ability: Theory and Evidence

Sonja C. de New\textsuperscript{a} and Stefanie Schurer\textsuperscript{b}

\textsuperscript{a}Centre for Health Economics, Monash University, Caulfield East, Australia, and Institute for the Study of Labor (IZA), Bonn, Germany; \textsuperscript{b}School of Economics, The University of Sydney, Sydney, Australia, and Institute for the Study of Labor (IZA), Bonn, Germany

**ABSTRACT**

An emerging literature is experimenting with using survey response behavior as a proxy for hard-to-measure abilities. We contribute to this literature by formalizing this idea and evaluating its benefits and risks. Using a standard and nationally representative survey from Australia, we demonstrate that the survey item-response rate (SIRR), a straightforward summary measure of response behavior, varies more with cognitive than with noncognitive ability. We evaluate whether SIRR is a useful proxy to reduce ability-related biases in a standard economic application. We show empirically that SIRR, although a weak and imperfect proxy, leads to omitted-variable bias reductions of up to 20%, and performs better than other proxy variables derived from paradata. Deriving the necessary and sufficient conditions for a valid proxy, we show that a strong proxy is neither a necessary nor a sufficient condition to reduce estimation biases. A critical consideration is to which degree the proxy introduces a multicollinearity problem, a finding of general interest. We illustrate the theoretical derivations with an empirical application.

**1. Introduction**

Economists view cognitive and noncognitive abilities as essential components of human capital (Heckman, Jagelka, and Kautz 2021; Lundberg 2019; Borghans et al. 2016; Almlund et al. 2011). Yet, measurement of such abilities has been a vexing problem. For instance, it is notoriously hard to get subjects to agree to a cognitive ability assessment. People are more likely to agree with personality assessments, but their self-assessed measures contain a large degree of reporting error (e.g., Schurer 2017; Kautz et al. 2014; Heckman and Kautz 2014; Cobb-Clark and Schurer 2013, 2012). In recent years, researchers started using behavioral proxies of ability to circumvent measurement problems. Kautz et al. (2014) suggest that “performance on any task or any observed behavior can be used to measure personality and other skills” (p. 16), concluding that as long as such a measure predicts behavior and can be implemented in practice, it is useful. Some studies use information derived from administrative and survey data to proxy noncognitive abilities including school attendance rates, number of suspensions (Holmlund and Silva 2014; West et al. 2016; Jackson 2018), participation in extracurricular activities (Ileras 2008), or behavior in class (Dee and West 2011; Heckman, Pinto, and Savelye 2013). Others use information derived from “paradata,” byproducts of the survey collection process (see Kreuter 2013). Some suggest survey response behavior is a useful source to construct measures of unobserved ability (Hedengren and Stratmann 2012; Hitt, Trivitt, and Cheng 2016). SIRR measures the number of items a survey respondent answers relative to all required responses. There are good reasons to believe that SIRR is driven by multiple dimensions of ability. In previous work we presented an economic model of the cognitive and noncognitive foundations of SIRR (Kassenboehmer and Schurer 2018). The model is built on decades of insights from the survey methodology literature on the factors that shape survey response (Dillman et al. 2002). The literature highlighted the importance of cognitive abilities (e.g., Sudman, Bradburn, and Schwarz 1996; Tourangeau 2003) and the stability of response styles (e.g., Weijters, Geuens, and Schillewaert 2010; Wetz et al. 2016). Survey methodologists hypothesize that “reporting errors in surveys arise from problems in the underlying cognitive processes through which respondents generate their answers to survey questions” (Tourangeau 2003, p. 5). Respondents need to understand the meaning of the question, recall relevant behavior and information, infer the appropriate answer, map the answer into the response format of the survey, and edit final answers which may then be adjusted by social desirability and privacy concerns (Schwarz and Oyserman 2001). At any of these steps, problems could arise which may lead to an incorrect or missing answer. Even apparently simple questions pose complex cognitive tasks that require memory for dates, events, and experience, and comprehension (see Joe and Mingay 1991, for an overview). Others argue that response behavior is also driven by...
personality. Some individuals simply desire to participate, while others, in contrast, worry too much about privacy or social desirability (e.g., Schwarz and Oyserman 2001; McCrae and Costa 1983).

The aim of this article is to better understand the general value, benefits and risks of using SIRR as a proxy for hard-to-measure abilities. In the first part, we test whether ability gradients in SIRR exist, considering a comprehensive battery of cognitive and noncognitive ability measures irregularly collected in national surveys. To do so, we use data from the Household, Income, and Labour Dynamics in Australia survey (HILDA), a high-quality, nationally representative survey comparable to surveys from other countries (Summerfield et al. 2017). The benefit of HILDA is that it collected at multiple points in time a rich and validated inventory of the Big Five personality traits (Losonczi 2009) and locus of control (Wilkins et al. 2010), the most widely used and accepted measures of noncognitive abilities in the context of labour market productivity (see Cobb-Clark 2015; Gensowski, Görtz, and Schurer 2021, for reviews of the literature). Importantly, HILDA also collected a high-quality, task-based inventory of cognitive ability, capturing multiple dimensions of ability such as coding speed, memory, and language (Wooden 2013).

We use both regression and nonparametric methods to document the ability gradients in SIRR. We conclude from this descriptive exercise that variations in SIRR reflect variations in cognitive ability rather than variations in personality. We find statistically significant and positive associations between SIRR and all three cognitive task measures. The strongest statistical association, with a correlation coefficient of over 0.2, is found for the symbol-digits modalities test, which captures coding speed and executive function (Kiely et al. 2014). SIRR turns out to be highly predictive of economic outcomes. Its predictive power in these outcomes is fully mediated by cognitive, but not by noncognitive, abilities.

Although not a strong proxy, we argue that SIRR could be used as a proxy variable for cognitive ability to reduce omitted variable biases (OVB). Our thinking follows a series of previous studies that have exploited other survey paradatodata to fix estimation biases (see, e.g., Heffetz and Rabin 2013; Pudney 2008; Kleinjans and van Soest 2014; Behagel et al. 2015, due to selection or heaping behavior), focusing on nonresponse of sensitive questions (Riphahn and Serfling 2005; Raessler and Riphahn 2006) or adjusting estimates of the determinants of wages, income or wealth (Zweimueller 1992; Bollinger and David 2005; Riphahn and Serfling 2005; Bollinger and Hirsch 2013).

In the second part of the article, we rigorously evaluate the benefits and risks of using SIRR as a proxy for cognitive ability. We first present a conceptual framework in which we derive the necessary and sufficient conditions under which “weak” proxy variables reduce rather than exacerbate OVB. Importantly, we allow this framework to accommodate “imperfect” proxy variables. An imperfect proxy variable correlates significantly with an important regressor of the structural model, potentially introducing a multicollinearity problem. This possibility has been pointed out by Frost (1979), but has not been discussed widely in the proxy variable literature (exceptions are Wolpin 1995, 1997). We derive these conditions in the context of a linear model with one weak and imperfect proxy variable, although extensions to multiple proxies are likely to be straightforward (see Bound, Brown, and Mathiowetz 2000, for an overview). We then use this conceptual framework to test the validity of SIRR as a proxy variable in the context of the estimated wage returns to education, a textbook example to illustrate OVB (see Gronau 2010, for a review) that goes back to the seminal work of Griliches (1977).

We show that a strong proxy variable is neither a necessary nor a sufficient condition for reducing OVB. Although this statement as such is not new (see Frost 1979, p.13), we are the first to prove it. We discover that the proxy variable breaks down if a powerful (powerless) proxy is applied in a setting where the potential for OVB is low (high), or if the potential for multicollinearity problems is high. The necessary condition for a proxy variable to reduce biases (relative to exacerbate them) requires sign equivalence of two key correlation coefficients: the partial correlation between the missing variable and the proxy (“strength”), and the partial correlation between the missing variable and the variable of interest (“potential degree of OVB”). The sufficient condition bounds the ratio of the strength of the proxy to the degree of OVB—a term we refer to as the relative strength of the proxy—by terms that only depend on the degree of multicollinearity introduced into the model by the proxy, which is always observable. A critical finding for applied research is that the proxy variable will always reduce biases if the partial correlation coefficient between the proxy and the main variable of interest in the regression model is sufficiently small.

Using this framework, we demonstrate that even though SIRR is a weak and imperfect proxy, its bias-reduction potential in the returns to education ranges between 8% for a linear measure, 14% for a nonlinear measure, and 20% for a long-term average, nonlinear measure. Item-response behaviors for the most cognitively demanding questions on computer use, time diaries, and household expenditures, are most powerful for bias reduction. The conceptional framework of imperfect proxy variables helps to explain why this is the case. In our example SIRR reduces biases, because the association between the proxy and cognitive ability has the same sign as the association between the omitted variable and main variable of interest, thus fulfilling the necessary condition. This means that the proxy variable works in the same direction as the omitted variable. Furthermore, the proxy variable is closer in nature to cognitive ability than education, the main variable of interest of the structural model. Thus, by including SIRR as an additional control variable, we do not introduce a large-enough multicollinearity problem in comparison to the relative strength of the proxy variable. Furthermore, SIRR outperforms other proxy variables derived from paradata (minutes spent on filling out the household questionnaire, interviewer rating of the respondent’s understanding of the question), which also fulfill both necessary and sufficient conditions of a valid proxy variable. However, combining SIRR and these two alternative proxy variables yields bias reductions of up to 27%, which is more than a quarter of the OVB. We illustrate the few cases where the proxy variable approach fails.

The main contribution of our article is that we offer applied researchers a guideline on how to use SIRR in their own data application and a conceptual framework for how to think about proxy variables. Second, we contribute to a dormant and
predominantly theoretical literature that compares proxy variable biases with omitted-variable biases (Wickens 1972; McCallum 1972; Aigner 1974; Frost 1979; Kinal and Lahiri 1983; Ohtani 1981, 1985; Stahlecker and Trenkler 1993). Our findings complement a large body of literature that approaches the problem of proxy variables from an errors-in-variables perspective (Bollinger 2003; Lubotsky and Wittenberg 2006; Bollinger and Minier 2015) that goes back to Klepper and Leamer (1984). This literature has comprehensively covered measurement problems in both outcomes (e.g., wages) and treatment variables (e.g., years of schooling) (see Bound, Brown, and Mathiowetz 2000, for a review). In contrast, our article deals with measurement problems in control variables. Instead of relying on classic measurement error assumptions, we assume differential measurement error, a more realistic assumption first expressed by Frost (1979). Although the case of differential measurement error had been discussed in the literature on the errors-in-variables problem (Bound, Brown, and Mathiowetz 2000, pp. 10–15), its consequences have not been formalized or tested for in the context of a proxy variable approach. It is therefore not surprising that we observe little discussion of the validity of proxy variables despite their widespread use in microeconomic applications, a concern raised by Wolpin (1995), Wolpin (1997), and Todd and Wolpin (2003).

Finally, our conceptual framework is relevant in a broader context where causal inference is limited due to the presence of unobserved confounders (Rosenbaum and Rubin 1983; Lalonde 1986; Heckman and Hotz 1989; Dehejia and Wahba 1999; Smith and Todd 2001; Imbens 2003; Gelbach 2016). Building on Altonji, Elder, and Taber (2005), recent work by Oster (2019) provided bounds on estimation biases that depend on assumptions about the maximal degree of explained variation in an outcome of interest and the relative importance of unobservable over observable selection into treatment. Our approach requires arguably less restrictive assumptions if information on the nature of the unobserved confounder is available. Comparing the bias-reduction potential of the proxy variable approach against these methods, we conclude that in certain situations the proxy variable approach is not a bad alternative. Our work is also aligned with Pei, Pischke, and Schwandt (2019) who emphasized that, in the presence of poorly measured confounders, standard approaches used for testing the identification assumptions of a regression strategy may lead to the erroneous conclusion that OVBs are negligible.

The remainder of this article is structured as follows. In Section 2, we describe our data. In Section 3, we document the statistical associations. In Section 4, we derive necessary and sufficient conditions of a weak and imperfect proxy variable, and translate these into an empirical guideline. Section 5 presents test results of using SIRR as a proxy for cognitive ability in the context of wage returns to education. Section 6 concludes.

## 2. Data

We use data from the Household, Income, and Labour Dynamics in Australia (HILDA) survey, a nationally representative household panel study conducted annually since 2001 (Summerfield et al. 2017). All adult household members aged 15 years and above are invited to respond to an interviewer-assisted (continuing or new-person) questionnaire, in which detailed information on education, employment, income and benefits, family formation, health, views on life, satisfaction, and cognitive abilities is collected. In addition, each eligible household member is invited to complete a self-completion questionnaire (SCQ) to be filled out in private. This SCQ contains additional questions on age, sex, general health and well-being, lifestyle and living situation, personal and household finances, attitudes, values, personality, job and workplace issues, parenting, and sexual identity. The completed SCQ is collected by the interviewer at a later date or returned by mail. Few household members opt to return a completed SCQ before the face-to-face interview.

The household form takes on average 10 min, the person questionnaire 35 min for continuing members and 47 min for new members, and the SCQ 30 min to complete. HILDA pays a financial incentive of 30$ for each completed person questionnaire and a bonus of another 30$ is paid to the household if all eligible household members complete the survey. No financial incentive is paid for the completion of the SCQ, but since 2012 individuals who complete the SCQ enter a lottery for a small prize (e.g., in 2012 five iPads were offered). Household and person questionnaires are collected through computer assisted personal interviewing (Watson 2011). The most important determinant of SCQ completion is whether the household was interviewed by telephone instead of face-to-face. Telephone interviews, which make up less than 9% of all interviews, reduce the probability of SCQ completion by 17 percentage points. SCQ completion is also associated with sociodemographic characteristics, but the associations are small in magnitude (<2 percentage points) (Watson and Wooden 2015). HILDA and its SCQ component are comparable to many other international surveys (supplementary material A).

We select a sample of eligible survey participants from Wave 12 (2012) and 16 (2016), the years when cognitive ability measures were collected. Item-nonresponse is calculated from the Wave 12 and 16 SCQs. The Big-Five personality traits were collected in Waves 5, 9, 13, and 17. Locus of control data were collected in Waves 3, 4, 7, 11, and 15. In 2012 and 2016, 17,475 and 17,693 eligible adults were interviewed, respectively. Of these, 15,389 and 16,235 individuals in Waves 12 and 16 (or over 90%) returned an SCQ (Figure B.1, supplementary material). Less than 6% of this eligible sample failed to participate in the cognitive ability assessment in either 2012 or 2016. For around 7% of the eligible sample, we do not have information on personality traits from previous waves. This leaves us with a sample of around 28,000 person-year observations to study the ability gradient in SIRR. In the wage regression models, we restrict the sample to working-age adults (age 24-64). Conditional on non-missing observations in the control variables used, this leaves us with an estimation sample of 15,996 person-year observations, of which 10,754 have positive wages. Summary statistics on all variable are reported in the supplementary material (Table B.1).

### 2.1. Survey Item-Response Behavior

We calculate the survey item response rate (SIRR) from the SCQ collected in Waves 12 and 16. We considered calculating an item response proxy also from the interviewer-assisted
The reason for the extremely low SIRR in the person questionnaire is that the questionnaire is conducted through face-to-face or telephone interview (Watson and Wooden 2015). SIRR is calculated for each individual as the ratio of the number of answered questions to the total number of questions that the respondent was required to respond to. The denominator varies across individuals as some participants are asked more questions than others depending on their socio-demographic situation. One concern is that the correlation we observe between SIRR and cognitive or noncognitive ability could be an artifact due to higher-ability people having to respond to fewer questions leading to a higher SIRR. In our data, this does not seem to be the case. We find a positive correlation between cognitive ability, wages, and personality with the number of required responses (Figure B.2, supplementary material).

The number of responses in the SCQ is calculated as the number of times an individual responded to each question instead of refusing to answer. For a small sub-sample of questions, respondents could choose an option “Don’t Know.” Some argue that refusals and “Don’t Know” answers are determined by different mechanisms (e.g., Raessler and Riphahn 2006), however, in our survey this refers to only two sets of questions (neighborhood characteristics, employer entitlements) comprising 6% of all questions. We therefore conduct the analysis including these questions, but show in a robustness check that our findings are not sensitive to their inclusion.

The total range of applicable questions ranges between 149 and 264 questions in Wave 12 (Figure 1(a)) and 152 and 264 in Wave 16 (Figure 1(d)). The total number of unanswered items in Wave 12 and 16 varies between 0 and 207 (Figure 1(b)) and 0 and 208 (Figure 1(e)). The corresponding mean nonresponse count is 4.3 and 3.6, respectively. About 90% of the sample fail to respond to a maximum of 10 questions, while 1% refused to answer 55 or more questions. A 1 standard-deviation (SD) increase in item nonresponse is equivalent to 12 additional questions not responded to in the SCQ in both waves. On average, individuals respond to 97.2% in Wave 12 (Figure 1(c)) and 97.6% in Wave 16 (Figure 1(f)). The minimum response is 4% in Wave 12 and 1.5% in Wave 16. About one in three respondents have a 100% response rate in both waves.

2.2. Cognitive Ability

The HILDA survey assessed respondents’ cognitive ability in Wave 12 and Wave 16 as part of the interviewer-assisted survey. This assessment included standard tests to measure memory, executive function, and crystallized intelligence through a Backward-Digit Span Test (BDS), a Symbol-Digit Modalities Test (SDM), and a National Adult Reading Test (NART), respectively (see Wooden 2013, for an overview). The BDS measures working memory span and is a sub-component of traditional intelligence tests. The interviewer reads out a string of digits which the respondent has to repeat in reverse order. BDS measures the number of correctly remembered sequences of numbers. SDM is a test of executive function, which 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. SDM measures the number
of correctly matched symbol-number pairs. NART is assessed through a 25-item list of irregular English words, which the respondents are asked to read out loud and pronounce correctly. NART measures the number of correctly pronounced words. On average, sample members score 4 on the BDS, 53 on the SDM, and 15 on the NART tests. We use each measure individually, as we would like to identify components of cognitive ability that are most related to survey response. In a robustness check, we use a summary measure of cognitive ability, which averages scores across these three items. All measures are standardized to mean 0 and SD 1.

2.3. Noncognitive Ability

Noncognitive ability is measured with the Big Five personality traits and locus of control. HILDA collected an inventory of the Big-Five personality traits based on Saucier (1994) that can be used to construct measures for extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience. Out of these five, we would expect agreeableness and conscientiousness to be most closely related to survey response behavior because they best capture willingness to cooperate and diligence with tasks. To construct a summary measure for each trait, we use the 28 items used to measure personality on the Big-5 and conduct factor analysis (see Cobb-Clark and Schurer 2012).

A measure of internal 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). Mastery refers to the extent to which an individual believes that outcomes in life are under her own control. Respondents were asked to report the extent to which they agree with each of seven statements related to the perception of control and the importance of fate. We construct a continuous measure increasing in internal locus of control using factor analysis (see Cobb-Clark and Schurer 2013; Cobb-Clark, Kassenboehmer, and Schurer 2014). To minimize measurement error in our constructs of noncognitive ability, we average personality scores across all available waves as in Cobb-Clark, Kassenboehmer, and Schurer (2014). All measures are standardized to mean 0 and SD 1.

3. Are Cognitive and Noncognitive Abilities Predictive of Survey Response Behavior?

In this section, we document the ability-gradient in survey item response behavior. For this purpose, we first estimate a separate regression model for each cognitive and noncognitive ability measure, in which SIRR is the dependent variable, and the specific cognitive or noncognitive ability measure is the independent variable. These unadjusted correlation coefficients reveal the strength of the proxy variable SIRR. We repeat this exercise by including a standard set of control variables (gender, age, education, language background, state of residence, geographic remoteness, wave, being part of the top-up sample).

Figure 2 summarizes our key findings. It plots the linear and nonlinear relationships between SIRR (vertical axis) and nine distinct ability measures (horizontal axis). The fitted solid line displays the OLS estimate of the adjusted correlation coefficient, the white dashed line plots the nonparametric kernel estimates, to allow for nonlinearities, and their 95% confidence intervals (see, e.g., Wand and Jones 1995). The nine figures demonstrate that there is a positive, significant association between SIRR and the three cognitive ability measures (Figure 2(a) to 2(c)) and a positive but weak association between SIRR and four out of the six noncognitive ability measures (Figure 2(e), 2(f), 2(h), and 2(i)). Most remarkable is the association between SIRR and the Symbol Digits Modalities test (SDM) measure. The raw and adjusted correlation coefficients between SIRR and the SDM measure are 0.19 SD and 0.22 SD, respectively. The second largest association is between SIRR and the National Adult Reading test (NART) measure (0.16 SD and 0.14 SD, respectively). In contrast, none of the adjusted correlation coefficients for noncognitive abilities are greater than 0.07 SD. The strongest associations are found for conscientiousness and locus of control of 0.068 SD and 0.064 SD, respectively.

These conclusions do not change when including in each regression model all cognitive or noncognitive ability measures simultaneously, socio-demographic variables, marital status and household composition, time availability measures such as labour supply or number of children, or interviewer ratings of the understanding of the questions (Table B.2, supplementary material). The correlation coefficient between SIRR and SDM ranges between 0.16 SD and 0.14 SD. Perhaps, the largest drop in coefficient is observed when adding the variable interviewer rating of the understanding of questions (from 0.157 SD to 0.143 SD). This is additional evidence that SIRR captures variation in cognitive processing power.

Variation in the three cognitive ability measures explain almost 5% of the total variation (as measured by R-squared) in SIRR, while noncognitive ability measures explain an additional 0.4%. While each individual cognitive ability measure makes up between 50% (BDS) and 71% (SDM) of the total explained variation in SIRR, none of the noncognitive ability measure makes up a significant amount. The only exception is conscientiousness which amounts to 25% of the total explained variation. One may argue of course that our measures of cognitive ability are reflective of other unobserved noncognitive skills such as motivation, perseverance or effort. If this was the case, a positive partial correlation between SIRR and SDM, even after controlling for observed noncognitive skills (Big Five personality traits and locus of control), might indicate that the proxy also reflects unobserved motivation, perseverance or effort. However, Wooden (2013) showed that the cognitive ability measures correlate positively, albeit weakly, with the achievement motivation scale “Hopes for success” (partial correlation coefficient of 0.056) and negatively with “Fear of failure” (−0.091), based on achievement motivation measures developed in Lang and Fries (2006). Wooden (2013) concluded that this is consistent with causation running from ability to motivation as we would have expected cognitive ability to correlate positively with both “hopes for success” and “fear of failure” if motivation was driving performance on the cognitive ability tests.

Although Figures 2(a) to 2(c) reveal some nonlinearities in the relationship between SIRR and the SDM measure, it is predominantly linear between very low and medium survey
Figure 2. Relationship between survey item-response rate (SIRR) and cognitive or noncognitive ability measures.
NOTE: Multivariate nonparametric (third degree) kernel regressions (white dashed line) and OLS regressions (gray solid line). Each model controls for age, gender, years of education, language background, geographic remoteness, state, wave and top up sample. Full estimation results—including all ability measures simultaneously—are reported in Table B.2 (supplementary material).

item-response rates. The results are not sensitive to defining nonresponse as refusals only (excluding "Don’t know"-answers) or when using a summary measure of cognitive ability that averages the scores of all three cognitive ability test scores. For instance, the unadjusted correlation coefficient between the average cognitive ability score and SIRR is 0.207, statistically significant at the 1%-level, with an adjusted R-squared contribution of 4.3 percent.

If SIRR behaves statistically in a similar way to a standard measure of cognitive ability, then individuals should not substantially change their response patterns over time beyond what can be explained by variations in survey questions and length (Figure B.3, supplementary material). Cognitive ability has been shown to moderately increase in early adulthood but to remain stable until old age when cognitive functioning begins to decline (Hertzog and Schaie 1988; Deary et al. 2000; Gow et al. 2011; Deary and Yang 2012). To test the fixed-trait hypothesis, we exploit the longitudinal nature of our data which allow us to calculate the average inter-temporal correlation coefficients (ITC) of SIRR with its own past (up until 17 waves back). If SIRR contains a strong element of an individual-specific component, then the ITC between time period $t$ and $t - 17$ should not differ substantially from the correlation calculated between time period $t$ and $t - 1$. The ITC between period $t$ and $t - 1$ is around 0.4. Although the ITCs decline the further back the lags are considered, the ITCs stabilize around 0.2 from lag 12 onwards (Figure B.4, supplementary material). We also calculated the probability of remaining an “all-item responder,” conditional
on the respondent being an "all-item responder" in the previous time period. In our data, 30% of the sample are "all-item responders" in any given time period, while this probability is twice as large—over 60%—for individuals who were an "all-item responder" in the previous period. We interpret the persistence in response behavior as evidence of a person-specific effect in SIRR.

Finally, SIRR is predictive of key economic outcomes (Table 1). An increase in SIRR by 1 SD is significantly (p < 0.01) associated with 1.5% higher wages, a 7% higher probability of graduating from high school relative to the sample mean, and better health by 0.6% relative to the sample mean, over and above the influence of a large set of control variables including noncognitive ability measures (see table notes for model specifications). Importantly, the impact of SIRR is no longer statistically significant and becomes small in magnitude once we control for the cognitive ability measures (columns 2, 4, and 6). We conclude that SIRR captures facets of cognitive ability, as the impact of SIRR on economic outcomes is mediated by cognitive ability.

Our findings are consistent with the few existing empirical articles, although our conclusions differ. Our correlation coefficient of 0.22 between SIRR and cognitive ability (as measured by SDM) is in line with the maximum correlation reported in Hitt, Trivitt, and Cheng (2016, p. 110) across six U.S.-American longitudinal datasets that followed middle and high school students into adulthood, and in Hedengren and Stratmann (2012) who found an age-adjusted standardized beta coefficient of 0.21 in the 1997 National Longitudinal Survey of Youth (1997 NLSY). Similar to our findings, Zamarro et al. (2018) estimates small associations between the item response rate and noncognitive abilities. Hedengren and Stratmann (2012), who is credited for being one of the first to estimate the association between SIRR and conscientiousness, estimate a coefficient of less than 0.04, half the size of our estimates.

### Table 1. Predictive power of SIRR.

|                  | Log hourly wages (1, 7, 6, 6) | High school (0, 1) | Health (0, 100) |
|------------------|-------------------------------|--------------------|----------------|
|                  | Including SIRR (1)            | Including CS & SIRR (2) | Including SIRR (3) | Including CS & SIRR (4) | Including SIRR (5) | Including CS & SIRR (6) |
| Survey Item Response | 0.0152*** (0.00554)         | 0.0195*** (0.00231)   | 0.0192*** (0.00210)   | 0.437** (0.134)         | 0.026 (0.136)  |
| Rate, Std.        | 0.00706 (0.00538)            | 0.00201 (0.00429)    | 0.00429 (0.00381)    | 2.024*** (0.201)        | 2.024*** (0.201) |
| Symbol Digits     | 0.0520*** (0.00544)         | 0.0389*** (0.00429)   | 0.0431*** (0.0117)    | 0.394** (0.182)         | 0.394** (0.182) |
| Modalities, Std.  | 0.031*** (0.00586)          | 0.012*** (0.00381)    | 0.057*** (0.00338)    | 0.274*** (0.149)        | 0.274*** (0.149) |
| National Adult    | 0.00732 (0.00464)           | 0.012*** (0.00338)    | 0.012*** (0.00338)    | 0.274*** (0.149)        | 0.274*** (0.149) |
| Backward Digit    | 0.00338 (0.182)             | 0.00429 (0.182)      | 0.0389*** (0.182)     | 0.394** (0.182)         | 0.394** (0.182) |
| Survey Item Response | 0.00586 (0.00464)         | 0.0437** (0.134)     | 0.0437** (0.134)     | 2.024*** (0.201)        | 2.024*** (0.201) |
| Adjusted R²       | 0.159 (0.182)               | 0.232 (0.232)        | 0.248 (0.248)        | 0.254 (0.254)           | 0.254 (0.254) |
| Akaike Info. Criteria | 27037.5 (26977.9)      | 25229.1 (23025.2)    | 0.012*** (0.00338)    | 0.274*** (0.149)        | 0.274*** (0.149) |
| Mean outcome      | 3.4470                      | 0.2695              | 187823.3 (187661.9)  |                      |                      |

NOTE: Dependent variables are Models (1) and (2): log hourly wages (range 1.6–6.6); Models (3) and (4) whether the individual completed high school (0, 1); Models (5) and (6) general health status derived from the SF-36 inventory (0, 100). Models (1) and (2) are estimated with a Tobit sample selection model, with left-censored observations set to missing. Exclusion restriction for selection equation are number of children in age brackets 0–4, 5–10, 11–14 and marital status. The Inverse Mills Ratio is 0.2381 (standard error 0.0386) for model (1) and 0.2460 (standard error 0.0389) for model (2). Models (1) and (2) are estimated on working age population aged 24–64 (N). Number of observations in selected group: 10754 and number of observations in nonselected group: 24240 and models control for male, age, age square, years of education, a dummy for casual worker, a quadratic polynomial of years of tenure with current firm, language background, state of residence dummies, regional dummies, personality traits, wave 16, and whether individual is from top up sample (joining in 2011). Models (3), (4), (5), and (6) control for male, age, age square, language background, number of children, marital status, state of residence dummies, regional dummies, personality traits, wave 16, and top up sample. Models (5) and (6) control in addition for years of education and income. Clustered standard errors (individual level) are reported in parentheses. Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01.

### 4. Using SIRR as a Proxy Variable for Cognitive Ability: A Conceptual Framework

We have shown that SIRR is significantly associated with cognitive ability, with unadjusted and adjusted correlation coefficients ranging between 0.19 and 0.22. The question arises as to whether SIRR has the potential to reduce omitted variable biases (OVB) when used as a proxy variable for unobserved cognitive ability. Although robust, the estimated association between SIRR and cognitive ability is not strong. Therefore, we are asking under what conditions this noisy proxy would reduce OVB.

The early theoretical literature on proxy variables suggested that it is always preferable to use a proxy variable as long as its measurement error is random and uncorrelated with the missing variable and other covariates in the structural model (Wickens 1972; McCallum 1972; Aigner 1974). The results presented in Wickens (1972) and McCallum (1972) are based on asymptotic derivations. Aigner (1974), and later Kinal and Lahiri (1983), argued that a proxy variable could introduce higher variances in the model estimates. Aigner (1974) derived the relative biases in small samples, demonstrating that the tradeoff between the two methods, expressed in terms of mean squared error, depends on the sample size, the proportion of measurement error and the correlation between the missing variable and the main variable of interest. The article concludes that the proxy variable approach is preferable in samples of 50 or more observations, even if the potential for OVB and measurement error is high.

Both Maddala (1977) and Frost (1979) criticized the assumption of random measurement error. Frost (1979) was explicit, saying that “in general, the difference between the unmeasurable variable and the proxy variable is not a random variable independent of the true regressors” (p. 323). He warned that substantial biases may occur when the proxy variable is also “imperfect,” which means that its measurement error is correlated with the
true, unobserved variable, and therefore it is correlated with the key variable of interest in the structural model. Thus, proxy variables should not be used "indiscriminately" (Frost 1979, p. 325). Similar theoretical criticisms have been expressed by Ohtani (1981), Ohtani (1985) and Stahlecker and Trenkler (1993) in the context of conditional prediction. Wolpin (1995) was the first to criticize indiscriminate applications of proxy variables in the context of microeconomic applications, for example, job search models. His work furthermore argued that an imperfect proxy variable may confound the interpretation of other important variables in the model (Wolpin 1995, 1997; Todd and Wolpin 2003).

With the exception of Wooldridge (2010), little empirical guidance is available on how to assess the risk of proxy variables in an empirical context. We contribute to the previous literature by deriving both the necessary and sufficient conditions under which an "imperfect" proxy variable model would yield less biased estimates than the omitted variable biased model, building on Frost (1979). In the second step, we illustrate this conceptual framework applied to SIRR as a proxy variable for unobserved cognitive ability in a standard wage regression model.

We start out with a correctly specified linear model of hourly wages \( Y_i \), which are a function of years of education \( (X_i) \) and variables \( (Z_i) \) that are commonly associated with productivity (e.g., experience, type of work contract, etc.):

\[
Y_i = \alpha_0 + \alpha_1X_i + \alpha_2M_i + Z_i'\theta + u_i. \tag{1}
\]

In addition, we assume that hourly wages depend on two components of ability as in Oster (2019). We define these two components as cognitive ability \( M_i \) and noncognitive ability \( N_i \), two complementary components of a person's human capital that have market returns (Almlund et al. 2011; Lundberg 2019; Heckman, Jagelka, and Kautz 2021). The most widely adopted measures for noncognitive ability in this context are the Big Five personality traits (see Gensowski, Gortz, and Schurer 2021, for a review) and locus of control (see Cobb–Clark 2015, for a review). For ease of illustration, we include noncognitive ability in the vector \( Z_i \).

We furthermore assume that \( u_i \) satisfies strict exogeneity conditions \( E(u_i|X_i, M, Z) = 0 \) and that the included variables have no or classical measurement error. The parameter \( \alpha_1 \) is of main interest to our inquiry. It measures the true wage returns for an extra year of education under the assumption of the model.

Let's consider the case that cognitive ability \( M_i \) is unobserved, while noncognitive ability \( N_i \) is observed, which is a common scenario in empirical research. As researchers we would have to work with one of the following two misspecified models. First, we could simply estimate equation (2) with \( M_i \) omitted:

\[
Y_i = \beta_0 + \beta_1X_i + Z_i'\theta' + \mu_i. \tag{2}
\]

It is straightforward to show the OVB in \( \beta_1 \), the estimated wage returns of education in the misspecified model (supplementary material C).

Second, we could estimate equation (3) adding the variable \( P_i \) (in our case: SIRR) as a proxy for \( M_i \):

\[
Y_i = \gamma_0 + \gamma_1X_i + Z_i'\theta'' + \gamma_2P_i + \nu_i, \tag{3}
\]

where \( \phi = M_i + \phi_i \) is measured with error. The critical question is what assumption should be made about measurement error \( \phi_i \). In this article, we deviate from the classical measurement error assumption \( \text{cov}(M_i, \phi_i) = 0 \) and \( \text{cov}(u_i, \phi_i) = 0 \). Instead, we assume \( \text{cov}(M_i, \phi_i) \neq 0 \) and \( \text{cov}(u_i, \phi_i) = 0 \). This means that while \( \phi \) still does not depend on unobservable determinants of the outcome in the structural model (here: hourly wages), it is now allowed to depend on variables of the structural model (here: \( X_i, Z_i \)). Although not explicitly written out in this way in Frost (1979), it reflects the assumption that the measurement error depends on observable determinants of the outcome of interest (see Kinal and Lahiri 1983). This assumption is sometimes referred to in the errors-in-variables literature as "differential measurement error" (Bound, Brown, and Mathiowetz 2000), a middle ground between the unrealistic but convenient classical measurement error assumption and the more realistic but inconvenient assumption of differential measurement error, where \( \phi_i \) is correlated with \( u_i \). Bound, Brown, and Mathiowetz (2000) argues that "in many cases, assuming that measurement error is classical is a simple (and potentially dangerous) expedient when we have little a priori reason to believe that any other particular assumption would be more plausible. In other situations, we have good reason to believe that the errors are differential, and the basis for this belief can help us write down relatively detailed but still manageable models" (p. 10).

Thus, under the assumption of differential measurement error, \( \phi \) depends on variables of the structural model. In our data setting, the most reasonable contender is education \( (X_i) \). Higher levels of education improve people's access to information and an ability to complete tasks, and therefore they are expected to have a higher SIRR, the proxy \( P_i \). In our setting, it is also reasonable to assume that people with higher levels of education are more likely to perform well on ability assessments captured in \( M_i \), as higher education trains people's assessment skills. Of course other assumptions could be made about what is captured in \( \phi_i \), depending on the empirical application. For instance, measurement error \( \phi \) could depend on other variables captured in \( Z_i \) and our model is flexible in allowing for this possibility.

Our model would break down if \( \phi_i \) depends on the dependent variable \( Y_i \), and therefore on \( u_i \) (non-differential measurement error). This would be the case if SIRR is driven by opportunity costs, whereby these costs are a function of hourly wages. People with higher hourly wages would have less incentive to spend time on survey questions, which would lead to a lower SIRR. People with higher opportunity costs would also be scoring higher on ability \( M_i \), Although theoretically possible, we have ruled out this case empirically (see Table 1). SIRR has no independent association with hourly wages once cognitive ability is controlled for, providing evidence against the assumption that opportunity costs drive measurement error \( \phi_i \). Supplementary material C presents a derivation of the structural model in this more complicated case.

Under the assumption of differential measurement error, an imperfect proxy variable is bound to lead to estimation biases (supplementary material, C). The imperfect proxy variable bias (IPB) may be smaller or larger than the biases which we would obtain by omitting a relevant confounder (OVB). To understand what determines the tradeoff between the IPB and the OVB,
we follow Frost (1979) to express the relative squared biases (IPB²/OVB²) in terms of three partial correlation coefficients, when measurement error ϕ depends on a variable from the structural equation X (in our case: education).

\[
\lambda = \frac{(E(\hat{\gamma}_1 - \alpha_1))^2}{(E(\hat{\beta}_1 - \alpha_1))^2} = \frac{(r_{XMIZ} - r_{MPIZ}r_{XPZ})^2}{r_{XMIZ}(1 - r_{XPZ}^2)^2}. \tag{4}
\]

These three partial correlation coefficients conditional on Z are defined as follows:

- \( r_{MPIZ} \): the partial correlation coefficient of the omitted variable \( M \) and the proxy \( P \);
- \( r_{XMIZ} \): the partial correlation coefficient of education \( X \) and the omitted variable \( M \); and
- \( r_{XPZ} \): the partial correlation coefficient of education \( X \) and the proxy variable \( P \).

For the proxy variable to improve upon OVB, we require \( \lambda < 1 \). The relative bias depends on the strength of the proxy (\( r_{MPIZ} \), short: strength) and the strength of the relationship between \( M_i \) and \( X_i \) (\( r_{XMIZ} \)), which indicates the potential for OVB (short: POVB). It also depends on the correlation between \( X_i \) and \( P_i \) (\( r_{XPZ} \)). Large values for \( r_{XPZ} \) imply that the proxy variable is closer in nature to education rather than the underlying omitted variable, thus there is the potential for a multicollinearity problem (short: PMCP). While \( r_{MPIZ} \) and \( r_{XMIZ} \) are usually unobserved by the researcher, \( r_{XPZ} \) is always observed.

We propose that the relative performance of the proxy variable approach depends on the sign equivalence of \( r_{MPIZ} \) and \( r_{XMIZ} \) and the ratio between \( r_{MPIZ} \) and \( r_{XMIZ} \) relative to \( r_{XPZ} \). The necessary and sufficient conditions for an imperfect proxy variable to reduce OVB are as follows (for proofs, see supplementary material C):

**Theorem 1.** A necessary condition for the imperfect proxy variable to reduce OVB is that

\[
\text{sign}(r_{MPIZ}) = \text{sign}(r_{XMIZ}) \quad \text{if} \quad r_{XPZ} > 0, \quad \text{and} \quad \text{sign}(r_{MPIZ}) \neq \text{sign}(r_{XMIZ}) \quad \text{if} \quad r_{XPZ} < 0.
\]

**Theorem 1** implies that if \( r_{XPZ} > 0 \), then for the imperfect proxy variable approach to improve upon omitting a relevant variable, it must be the case that the relative strength of the proxy variable must be positive. This means that the sign of the partial correlation between the proxy and the omitted variable must be the same sign as the partial correlation between the missing and the main variable of interest in the model (in our case: education). If \( r_{XPZ} < 0 \), then the two partial correlation coefficients must be of opposite signs.

**Theorem 2.** A sufficient condition for the imperfect proxy variable to reduce OVB is that:

\[
\begin{align*}
r_{XPZ} < r_{MPIZ} &< \frac{2 - r_{XPZ}^2}{r_{XPZ}} \quad \text{if} \quad r_{XPZ} > 0, \quad \text{and} \\
r_{XPZ} > r_{MPIZ} &> \frac{2 - r_{XPZ}^2}{r_{XPZ}} \quad \text{if} \quad r_{XPZ} < 0.
\end{align*}
\]

**Theorem 2** states that to improve upon the omitted variable approach, the relative strength of the proxy variable must lie within an interval bounded between \( r_{XPZ} \) and \( \frac{2 - r_{XPZ}^2}{r_{XPZ}} \).

We illustrate these tradeoffs in a simulation exercise. Let’s assume \( r_{XPZ} > 0 \) and \( r_{XPZ} \in \{0.05, 0.20, 0.40, 0.60, 0.80\} \). Figure 3(a) depicts the relative strength of the proxy variable (\( \frac{r_{MPIZ}}{r_{XMIZ}} \)) on the horizontal axis. Although the values for this ratio can become indefinitely large, we restrict its possible values between \(-1 \) to \(+3 \) for ease of illustration. Negative values on the x-axis indicate that the sign of the partial correlation coefficients are opposite in sign. \( \lambda \) is expressed on the vertical axis. Values of \( \lambda < 1 \) imply a reduction in the OVB, while values of \( \lambda > 1 \) imply an increase in the OVB.

The analytical and simulation results emphasize that a strong proxy variable is neither a necessary nor a sufficient condition for reducing OVB. The proxy variable needs to be strong only relative to the POVB and relative to the PMCP. For instance, if the latter is small (\( r_{XPZ} = 0.05 \)), then the IPB is smaller than the OVB as long as the relative strength of the proxy variable is greater than 0.05 and smaller than 0.39.95. In contrast, if the POVB is large (\( r_{XPZ} = 0.80 \)), then the relative strength of the proxy variable must lie within a small window of 0.80 and 1.7.

To illustrate this point further, we depict in Figure 3(b) the regions of bias when varying values of the three relevant partial correlation coefficients. The figure shows in two-dimensional
space the strength of the proxy variable and POVB on the horizontal and vertical axis, respectively. Superimposed onto this two-dimensional space are the minimum values of PMCP for which OVB will be exacerbated for every combination of strength and POVB. The proxy variable approach is most likely to increase biases if the proxy variable is weak (e.g., partial correlation coefficient (PCC) < 0.1) but the potential for OVB is large (e.g. PCC > 0.8), or where a strong proxy (e.g., PCC > 0.8) is paired with a negligible POVB (e.g., PCC < 0.1). In this situation, a researcher can expect an increase in OVB when using the proxy variable even if the PCC of the PMCP is tending toward zero (dark blue area). Although theoretically possible, these two cases are unlikely empirically. They imply for instance that the information in the omitted variable (in our case cognitive ability) contains almost no information from the proxy variable but almost identical information from the main variable in the structural model (in our case education). Figure 3(b) shows that in more realistic scenarios, where the main variable in the structural model (in our case education). Figure 3(b) shows that in more realistic scenarios, where the main variable in the structural model (in our case education).

We now evaluate the validity of SIRR as a proxy for cognitive ability in the context of estimated wage returns to education. The estimation sample consists of 10,754 person-year observations with positive hourly wages, and the model is based on column (1) of Table 1. Table 2 reports the three relevant partial correlation coefficients, the ratio of the squared biases (λ) and the relative strength of the proxy variable (PMCP). SIRR is a valid proxy if λ is smaller than 1. We therefore test for whether λ is equal to one against the one-sided hypothesis that it is smaller than one.

In Panel A, we report the test results for linear measures of SIRR and its variations, respectively. In all considered cases, the proxy variable passes the test. For the benchmark measure of SIRR (row 1), the partial correlation coefficient (PCC) indicates that the strength of the proxy is weak (r_{MP|Z} = 0.10). However, applied researchers can use knowledge of r_{MP|Z}, which is always observed, to make informed choices about the risk of their proxy variable approach.

5. Empirical Illustration: Wage Returns to Education

5.1. The Validity of SIRR as a Proxy for Cognitive Ability

Table 2. Calculations of λ and percent bias reductions in the estimated wage returns to education.

| Strength | POVB | PMCP | Relative strength | λ^a | Bias reduction in % of OVB^b | N |
|----------|------|------|------------------|-----|--------------------------|---|
| r_{MP|Z} | r_{XM|Z} | r_{XP|Z} | r_{MP|Z} | r_{XM|Z} | r_{XP|Z} | r_{MP|Z} | r_{XM|Z} | r_{XP|Z} | λ | Bias reduction in % of OVB | N |
| Refused and don't know | 0.1036 | 0.1335 | 0.1136 | 0.7758 | 0.8534 | 7.62 | 10754 |
| Refused | 0.1057 | 0.1335 | 0.0665 | 0.7916 | 0.9055 | 4.84 | 10754 |
| Average across waves | 0.2137 | 0.1385 | 0.0980 | 1.5430 | 0.7344 | 14.30 | 11190 |
| Questions that all answered | 0.0799 | 0.1335 | 0.0662 | 0.5989 | 0.9304 | 3.54 | 10754 |
| Questions with max bias reduction | 0.1113 | 0.1335 | 0.1164 | 0.8338 | 0.8379 | 8.46 | 10754 |
| Nonlinear measure (binary or dummy variables for each decile) | | | | | | | |
| Binary Variable (=1 if SIRR=1, =0 if SIRR <1) | | | | | | | |
| Dummy for deciles | 0.8069 | 0.1385 | 0.0665 | 0.7916 | 0.9055 | 4.84 | 10754 |
| Average across waves, by deciles | 0.6626 | 0.1385 | 0.0665 | 0.7916 | 0.9055 | 4.84 | 10754 |
| Questions that all answered, by deciles | 0.9257 | 3.96 | 10754 |
| Questions max bias reduction, by deciles | 0.7804 | 11.66 | 10754 |
| Panel B: Variation on cognitive ability | | | | | | | |
| BDS | 0.0693 | 0.1286 | 0.1114 | 0.5391 | 0.9058 | 4.82 | 11279 |
| NART 25 | 0.1064 | 0.2625 | 0.1136 | 0.5004 | 0.9129 | 4.45 | 10655 |
| Cognitive ability all factors | 0.1262 | 0.2276 | 0.1113 | 0.5545 | 0.9026 | 4.99 | 10443 |
| Panel C: Variations in sample restrictions | | | | | | | |
| Age 20-69 | 0.1005 | 0.1294 | 0.1044 | 0.7765 | 0.8632 | 7.09 | 12069 |
| SIRR >0.5 | 0.1050 | 0.1328 | 0.1200 | 0.7904 | 0.8434 | 8.16 | 10727 |
| English language background | 0.0972 | 0.1358 | 0.1113 | 0.7157 | 0.8684 | 6.81 | 9510 |
| Interviewers say excellent understanding | 0.0784 | 0.1118 | 0.1126 | 0.7012 | 0.8702 | 6.71 | 8903 |
| Panel D: Variations in alternative proxies or combinations (each Std) | | | | | | | |
| Interviewer rating of understanding | 0.1440 | 0.1385 | 0.0645 | 1.0403 | 0.8775 | 6.32 | 11189 |
| Minutes spent on person questionnaire | −0.0330 | 0.1392 | 0.0180 | −0.2374 | 1.0092 | −0.46 | 11171 |
| Minutes spent on HH questionnaire | 0.0538 | 0.1391 | 0.2554 | 0.3868 | 0.9295 | 3.59 | 11006 |
| Days taken to return SCQ | 0.0020 | 0.1335 | −0.0100 | 0.0150 | 1.0005 | −0.03 | 11063 |
| Average number of times no SCQ | −0.0839 | 0.1384 | −0.0479 | −0.5993 | 0.9477 | 2.65 | 11190 |
| Top 3 proxies, linear | 0.2137 | 0.1384 | 0.0980 | 1.5438 | 0.6335 | 20.41 | 11189 |
| Top 3 proxies, nonlinear | | | | | | | |

NOTE: r_{MP|Z}, r_{XM|Z}, and r_{XP|Z} are the partial correlation coefficients of cognitive skills M and the proxy variable P (strength of proxy), of the education variable X and the proxy variable P (multicollinearity potential), and of the education variable X and cognitive skills M (omitted variable potential), netting out the effect of all other control variables used in the wage regression model.

^a λ measures the squared relative bias.

^b Bias reduction is calculated as the percentage change in the omitted variable bias, calculated as \left(\frac{OVB_{PB}}{OVB}\right) \times 100.

^c Nonlinear measure of SIRR is captured by four dummy variables that indicate nonresponse rates within the 5th percentile, between the 5th and 10th percentile, between the 10th and 25th percentile, and above the 25th percentile. The comparison group is zero nonresponse. All proxy variables are standardized to mean 0 and standard deviation 1.
there are only minor concerns over the PMCP and POVB, as the two partial correlation coefficients are small too ($r_{XPZ} = 0.11, r_{XMIZ} = 0.13$).

SIRR fulfills the necessary condition to be a valid proxy variable for estimating the returns to education, because sign($r_{MPZ}$) = sign($r_{XMIZ}$), given $r_{XPZ} > 0$. It also fulfills the sufficient condition since the relative strength of the proxy variable is greater than the potential multicollinearity problem, and it is always smaller than the maximum upper bound $\frac{1 - r^2_{XPZ}}{r_{XMIZ}}$. Thus, including SIRR as a proxy variable in the wage regression model would yield bias reductions for the estimated returns to education, which is reflected in $\lambda = 0.85$.

Theorems 1 and 2 are useful for discussing the bias-reduction potential of SIRR, even in the absence of observable information on the omitted variable. Knowledge about the PMCP is enough to make an informed risk assessment regarding use of the proxy. We observe in the data that the partial correlation coefficient between education $X$ and proxy $P$ is positive (0.11). To fulfill the necessary condition of sign equivalence (Theorem 1), we need to argue that the sign of the partial correlation coefficient between the unobserved variable $M$ (cognitive ability) and $X$ (education) is the same sign as the partial correlation coefficient between $M$ and the proxy variable $P$. In our case this means that cognitive ability improves education and cognitive ability improves survey item response rates. The former has been widely shown to be true in the literature. The latter has been shown theoretically in Kassenboehmer and Schurer (2018), building on a mature survey response methodology literature.

Furthermore, to fulfill the sufficient condition (Theorem 2), we would need to assess whether it is reasonable to assume that the relative strength of the proxy lies within an interval that can be calculated from knowledge of $r_{XPZ}$. In our data setting, this interval is: $0.11 < \frac{r_{MPZ}}{r_{XMIZ}} < 16.6$. In other words, for SIRR to be an invalid proxy, the strength of the proxy would have to be either 17 times larger or just one eighth or less than the potential for omitted variable problem ($r_{MPZ} \geq 16.6r_{XMIZ}$ or $r_{MPZ} \leq 0.11r_{XMIZ}$). Given that this is a wide interval, one could reasonably assume that the proxy variable approach is associated with a very low risk of exacerbating the bias. A more straightforward rule of thumb is that we expect OVB to be reduced if $r_{XPZ} < 0.2$. This is the case in our setting.

### 5.2. Magnitude of Bias Reductions

SIRR is a valid proxy for cognitive ability. How strong is its OVB-reduction potential? Panels A and B of Table 2 report the bias-reduction potential of linear and nonlinear variations in SIRR. The benchmark linear measure reduces OVB by 8% (calculated as $\frac{OVB}{\text{SDM}}$). The largest bias reductions of 14% (linear) and 19% (nonlinear) are obtained when using a long-term SIRR measure which averages out wave-specific variations in SIRR.

In comparison to other cognitive ability measures (Backward-digit span, National English reading test), the Symbol digits modalities test has the highest bias reduction potential (Panel B, 8% versus 5%). Our findings are robust to using a larger sample spanning ages 20-69 (7%), dropping individuals who responded less than 50% of the SCQ items (8%) or dropping individuals with a non-English speaking background (7%) or individuals who did not understand the questions in the person questionnaire according to the interviewer (7%) (Panel C). Bias reductions are slightly larger when using only Wave 12 data than when using only Wave 16 data (Table B.3, supplementary material).

### 5.3. Relative Performance of SIRR to Other Proxy Variables

SIRR performs well relative to alternative proxy variables also derived from para data (interviewer’s ratings of the participants’ understanding of the questions, minutes taken for completing the personal questionnaire or the household questionnaire, number of times failed to return an SCQ across all waves, and days elapsed until the self-completion questionnaire is returned). We find that three other proxies reduce OVB, albeit in smaller magnitudes. These are in order of relevance: (1) the interviewer’s rating of the participants’ understanding of the questions (6%), minutes spent on household questionnaire (4%) and number of times failed to return SCQ across all waves (3%) (Panel D).

In the case of “minutes spent on personal questionnaire,” we observe a small increase in OVB (-0.5%). The reason is that the necessary condition of a valid proxy variable is not fulfilled. Since the partial correlation coefficient for education and SIRR is positive (column (3)), it must be true that the partial correlation coefficients for the strength (column (1)) and the POVB (column (2)) have the same sign. This is not the case. Individuals who spend more time on the person questionnaire have lower levels of cognitive ability, while cognitive ability positively affects education. Thus, the proxy variable operates in the opposite direction as with education. Even though the proxy introduces almost no multicollinearity problem, it exacerbates the OVB.

### 5.4. Heterogeneity by Block of Questions

Our SIRR measure captures the average relationship between item response and cognitive ability (SDM) across over 260 survey questions. Figure 4 scatterplots the bias reduction in the returns to education for each individual survey question when used as a binary proxy variable (vertical axis) and the difference in cognitive ability between respondents and nonrespondents (horizontal axis) for both Waves 12 and 16. (see Tables B.4 and B.5, supplementary material, for the full list). We find a positive relationship between bias reduction and differences in cognitive ability, with a correlation coefficient of 0.3 for Wave 12 and 0.02 for Wave 16.

As can be seen, there are some questions which exacerbate biases, while others reduce OVB by up to 4.5%. Per wave, there are about 30 survey items for which each individual question reduces OVB in the estimated returns to education by between 1% and 4.5%. These are, for instance, seven questions on the usefulness of computer use to learn more skills, 14 questions on weekly time use, three questions on household expenditures, and five questions on achievement motivation and satisfaction regarding domestic life. These are also the questions that are most strongly associated with differences in SDM between respondents and nonrespondents. For instance,
Table 3. Bias-adjusted wage returns to education under various assumptions.

|                      | True Complete list controls (1) | OVB Omit cognitive ability (2) | IPB Add SIRR proxy | Oster (2019) method Assume $\delta = 1$ | $R^2_{\text{max}} = 1.3R^2$ |
|----------------------|----------------------------------|---------------------------------|---------------------|------------------------------------------|---------------------------|
| Education            | 0.0809*** (0.0028)               | 0.0841*** (0.0028)              | 0.0833*** (0.0028)  | 0.0803*** (0.0028)                      | 0.1113                    |
| Adjusted R-squared   | 0.2908                           | 0.2839                          | 0.2849              | 0.2906                                   |                           |
| MSE of estimator $\hat{\beta}$ | 0.0000181                      | 0.0000166                       | 0.0000166          |                                          | 1.0000                    |
| Max R-squared        | 1.0000                           | 1.0000                          | 1.0000              |                                          |                           |
| Difference to true estimate (1) | 0.0032                         | 0.0024                          | -0.0006             |                                          |                           |
| Bias in percent of true | 3.95                          | 2.94                           | -0.78               |                                          |                           |
| Bias reduction in percent of OVB | 25.68                        | 118.75                         |                     |                                          |                           |
| Wave-person observations | 10768                       | 10768                          | 10768               | 10768                                    | 10768                     |

NOTE: Reported are estimates of the wage returns to education. The dependent variable is log hourly wage. The model is estimated with OLS but adjusted for selection into positive wages (inclusion of estimated Inverse Mills Ratio obtained from a Tobit sample selection model, with left-censored observations set to missing. Exclusion restrictions for the selection equation are number of children in age brackets 0–4, 5–10, 11–14 and marital status. The models are estimated on working age population (ages 24-64). Control variables are male, age, age square, years of education, a dummy for casual worker, a quadratic polynomial of years of tenure with current firm, language background, state of residence dummies, regional dummies, personality traits, wave 16 dummy, and whether the individual is from the top up sample (joining in 2011). Model (1) is considered as the “true” model which additionally controls for cognitive ability. Model (2) omits cognitive ability and thus presents the returns to education estimate with omitted variable bias (OVB). Models (3) and (4) control for a linear and nonlinear measure of SIRR. Under the assumption of a weak and imperfect proxy, we expect (3) and (4) to yield returns to education estimates with imperfect proxy bias (IPB). Mean squared error of estimator with OVB and IPB estimators are, respectively: $MSE(\hat{\beta}_1) = \text{var}(\hat{\beta}_1) + (\text{Bias}(\hat{\beta}_1, \alpha_1))^2$ and $MSE(\hat{\gamma}_1) = \text{var}(\hat{\gamma}_1) + (\text{Bias}(\hat{\gamma}_1, \alpha_1))^2$ (see Eq. (6) Aigner 1974, p. 367). Smaller numbers indicate a better combined outcome of variance and bias. Models (5) and (6) show the bias-adjusted returns to education when making specific assumptions about the degree of unobserved heterogeneity and the maximum R-squared that can be achieved (Oster 2019). Starting out from the perspective of the OVB model (2), Model (5) presents returns to education estimates under the assumption that the degree of selection on unobservables is equal to the degree of selection on observables ($\delta = 1$) and the maximum R-squared that can be achieved is 1.3 times the R-Squared of Model (2). Model (6) also assumes $\delta = 1$ and a maximum achievable R-squared of 1. The presented numbers in (5) and (6) show the minimum upper bound of the returns to education under the assumption of the model. Numbers are generated using the STATA command psaalc written by Emily Oster and available through ssc. Clustered standard errors (individual level) are reported in parentheses. Significance levels: $p < 0.10$, $** p < 0.05$, $*** p < 0.01$. 

5.5. Relative Performance of the Proxy Variable Approach to Other Methods

How does the proxy variable approach fare against other methods to bound OVB? The most widely used method uses information on the likely degree of self-selection into treatment by observable and unobservable confounders (Oster 2019; Altonji, Elder, and Taber 2005). One could argue that there are many unobserved confounders that cause OVB in the returns to education, as the explained variation in hourly wage models averages around 40% (Oster 2019). Altonji, Elder, and Taber (2005) suggest that the explained variation could reach 100%. Oster (2019) argues that the explained variation can never reach 100% because of measurement error. This suggestion is not unfounded. Empirical evidence by Keane and Wolpin (1997) and Keane and Wolpin (2001) shows that up to 42% and 82%, respectively, of the variation in (ln) hourly wages, observed in the NLSY, could be due to measurement error. Collating evidence from 27 studies published in the top general-
interest journals in economics on wage returns, Oster (2019) demonstrated that the explained variation in hourly wages can be at best 1.3 times the explained variation in the omitted variable model. The advantage of this method is that it makes no assumption about the number and type of unobservable covariates, but it comes at the cost of the strong assumption that selection on unobservable is as strong as selection on observable characteristics.

We re-estimate our main wage regression model from Table 1 in the following way: Model (1) is the true model which includes cognitive ability; Model (2) is the omitted variable model which excludes cognitive ability; and Model (3) includes the imperfect proxy variable but excludes cognitive ability. Table 3 shows that the true returns to education (8.1%, column 1) are over-estimated when excluding cognitive ability (8.4%, column 2) by almost 4%. A model, which includes the linear version of the imperfect proxy variable (SIRR), yields a return to education estimate of 8.3%, an over-estimate of 3% (Model (3)). A nonlinear version of the proxy variable, yields a returns to education estimate of 8%, which implies an under-estimate of 1% (Model (4)). Thus, in the real world, the proxy variable reduces OVB by 26% for the linear measure and 119% for the nonlinear measure. Another way of assessing the quality of the IPB estimator relative to the OVB estimator is to compare their mean squared errors, which consider both the efficiency of the estimator and its squared bias, a metric used for comparisons in Aigner (1974), Ohtani (1981), and Ohtani (1985). A smaller number implies a higher quality of the estimator. Table 3 shows that the models using the nonlinear proxy (MSE: 0.0000165) or the linear proxy (MSE: 0.0000166) yield smaller MSEs than the model omitting the relevant variable (MSE: 0.0000181).

In comparison, using the bounding method that assumes that the maximum explained variation is no more than 1.3 the explained variation reported for Model (2) and that the degree of selection on unobservables is as strong as selection on observables, then we would obtain a returns to education estimate of 8%, which implies an under-estimate of 4%. A model, which includes the linear version of the imperfect proxy variable (SIRR), yields a return to education estimate of 8.3%, an over-estimate of 3% (Model (3)). A nonlinear version of the proxy variable, yields a returns to education estimate of 8%, which implies an under-estimate of 1% (Model (4)). Thus, in the real world, the proxy variable reduces OVB by 26% for the linear measure and 119% for the nonlinear measure. Another way of assessing the quality of the IPB estimator relative to the OVB estimator is to compare their mean squared errors, which consider both the efficiency of the estimator and its squared bias, a metric used for comparisons in Aigner (1974), Ohtani (1981), and Ohtani (1985). A smaller number implies a higher quality of the estimator. Table 3 shows that the models using the nonlinear proxy (MSE: 0.0000165) or the linear proxy (MSE: 0.0000166) yield smaller MSEs than the model omitting the relevant variable (MSE: 0.0000181).

In comparison, using the bounding method that assumes that the maximum explained variation is no more than 1.3 the explained variation reported for Model (2) and that the degree of selection on unobservables is as strong as selection on observables, then we would obtain a returns to education estimate of at least 11% (Model (5)). This estimate is a larger over-estimate than what our imperfect-proxy variable approach yields. When making the assumption that the explained variation can be 100% (see Altonji, Elder, and Taber 2005), then the returns to education estimate would be at least 298% (Model (6)), which is implausible. Thus, we conclude that in certain situations the proxy variable approach is a good alternative to the bias-adjustment methods proposed in Oster (2019) and Altonji, Elder, and Taber (2005).

### 6. Conclusions

What do our results mean for applied researchers? First, the survey item response rate (SIRR) is a good candidate to proxy cognitive ability, but less well suited to proxy noncognitive ability. In particular, SIRR has moderate statistical associations with the symbol-digits modalities (SDM) test scores. The SDM screening instrument has wide applicability in clinical research settings to identify neurological dysfunction and cognitive aging. Performance on the SDM test is determined by attention, perceptual speed, motor speed, and visual scanning. Although the SDM test is unable to differentiate between specific disorders, impaired performance has been associated with traumatic brain injury, concussion in athletes, multiple sclerosis, Huntington’s disease, Parkinson’s disease, and stroke (see Kiely et al. 2014, for a review). It is therefore not surprising to discover that SIRR has predictive power in wages, education and health, and that its influence is fully mediated by cognitive ability. Using SIRR as a proxy for attention, perceptual speed, motor speed, and visual scanning has no risks of exacerbating omitted variable biases (OVB) in the context of applications on the wage returns to education. Even though SIRR does not eliminate OVB, it reduces it by up to 20%.

Our proxy performs well relative to other proxies derived from paradata, such as interviewer ratings, minutes taken for completing the personal and household questionnaire and days taken to return the self-completion questionnaire. We identify a subset of survey questions for which item response is most strongly associated with differences in cognitive ability. These are questions that require recall ability and tedious coding such as questions on household expenditures or time-use that are widely used as inputs or outcomes of economic decision models over the lifecycle (see Aguiar, Hurst, and Karabarbounis 2012, for an overview). Survey data on these types of questions are more complete for participants with better cognitive function. Researchers working with time-use or household-expenditure data may need to account for this self-selection as was proposed and demonstrated in Heffetz and Rabin (2013) in the context of life satisfaction. We also show that in certain situations—for example, if information on the nature of omitted confounders and measures of proxies are available—the proxy variable approach is a good alternative to the bias-adjustment methods that rely on assumptions about the likely degree of self-selection into treatment (Altonji, Elder, and Taber 2005; Oster 2019).

Although some may argue that our results are specific to the Australian data context or the specific nature of the self-completion questionnaire, we propose that our findings are generalizable. The HILDA survey is comparable to other household panel surveys. It has many similarities with both the British Household Panel Survey (BHPS) and the German Socio-Economic Panel (SOEP) (Watson and Wooden 2014, pp. 503-504). The HILDA is not unique in having a self-completion questionnaire either. For example the SOEP typically collects all information using one mode, but one option is self-administration (Watson and Wooden 2014). The Understanding Society survey, which was built on the British Household Panel Survey (BHPS), currently adopts a push-to-web mixed-mode design allowing in-person interviews as well as self-administered online completion of the survey (d’Ardenne et al. 2017). In principle, SIRR can be calculated for any survey that allows for nonresponse. Hitt, Trivitt, and Cheng (2016) have demonstrated the applicability of such a proxy measure, for example, for the 1997 National Longitudinal Survey of Youth, and many of the self-administered surveys such as the 1980 High School and Beyond survey, the 1988 National Educational Longitudinal Study, the Add Health study, and the 2002 Educational Longitudinal Study. Previous research has shown the predictive power of item response rates for many U.S. datasets (Hedengren and Stratmann 2012; Hitt, Trivitt, and Cheng 2016; Zamarro et al. 2018), a demonstration of the
wide applicability of the proxy variable. The recent trend of incorporating web-based self-completion surveys into large, representative panel studies such as the U.K. Innovation Panel of Understanding Society opens up unlimited opportunities for deriving proxies for cognition from survey response patterns.

Our conceptual framework extends Frost (1979) and Wolpin (1995), the first to caution that the use of proxy variables may exacerbate OVB if their measurement error depends on important covariates of the model including the outcome of interest. This insight was not acknowledged in the early work of Wickens (1972), McCallum (1972), and Aigner (1974) who assumed classical measurement error. Valid under the more flexible assumption of differential measurement error, we accommodate the possibility that the measurement error in the proxy variable is a function of key control variables in the structural model. The assumption is reasonable, as it allows for the possibility that the unobserved variable (here cognitive ability) depends on education. The assumptions keep the model manageable while at the same time being more flexible than the models considered in the previous literature (Bound, Brown, and Mathiowetz 2000).

We demonstrate that the necessary condition for a proxy variable to reduce OVB is sign equivalence between two key correlation coefficients that measure the strength of the proxy variable and the potential for omitted variable problems. Although this condition can never be tested formally it can be rationalized through knowledge of the context, theoretical arguments, and previous empirical findings. We also show that a sufficient condition for the proxy variable to reduce OVB is that the ratio of the strength of the proxy to the potential for omitted variable problem—which we refer to as the relative strength of the proxy—needs to be bound by terms that only depend on the partial correlation between the variable of interest and the proxy, which we refer to as the potential for multicollinearity. These bounds can easily be calculated. Although these conditions do not allow for a water-tight testing procedure, they equip researchers with an empirical guideline based on prior information for informed risk evaluation of using a weak and/or imperfect proxy variable approach. Such guidelines can be used in a wide array of empirical settings.

Acknowledgments

The authors thank the Associate Editor and two anonymous referees for valuable feedback. We also thank the following individuals for their helpful feedback: Frauke Kreuter, Joachim Winter, Jeffrey Wooldridge, Kenneth Wolpin, Colin Cameron, James J. Heckman, Mark Wooden, Denzil Fiebig, Zhihua Pei, Michael P. Keane, Jana Mareckova, Peter Siminski, David W. Johnston, Agne Suziedelyte; and Felix Leung for research assistance. This article uses unit record data from the Household, Income and Labour Dynamics in Australia Survey (HILDA) conducted by the Australian Government Department of Social Services (DSS). The findings and views reported in this article, however, are those of the authors and should not be attributed to the Australian Government, DSS, or any of DSS’ contractors or partners. DOI: doi:10.26193/YP7MNU

Previous versions of this article were published under the title “Survey item-response behavior as an imperfect proxy for unobserved ability: Theory and application” published as an IZA Discussion Paper 11449 in 2018 and under the title “Testing the Validity of Item Non-Response as a Proxy for Cognitive and Non-Cognitive Skills” (IZA Discussion Paper 8874, 2015).

Funding

The authors acknowledge financial support from an Australian Research Council Early Career Discovery Program Grant (DE140100463), an Australian Research Council Discovery Project Grant (DP190102765), and the Australian Research Council Centre of Excellence for Children and Families over the Life Course (project number CE14010027).

References

Aguilar, M., Hurst, E., and Karabarbounis, L. (2012), “Recent Developments in the Economics of Time Use,” Annual Review of Economics, 4, 373–398. [209]

Aigner, D. J. (1974), “MSE Dominance of Least Squares With Errors of Observation,” Journal of Econometrics, 2, 365–372. [198,203,208,209,210]

Almlund, M., Lee Duckworth, A., Heckman, J. J. and Kautz, T. (2011), “Personality Psychology and Economics,” in Handbook of the Economics of Education (Vol. 4), eds. S. M. Eric, A. Hanushek, and L. Woessmann, pp. 1–181, Amsterdam, The Netherlands: North-Holland. [197,204]

Altonji, J. G., Elder, T. E. and Taber, C. R. (2005), “Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools,” Journal of Political Economy, 113, 151–184. [199,208,209]

Behaghel, L., Crepon, B., Gurgand, M., and Le Barbanchon, T. (2015), “Please Call Again: Correcting Non-Response Bias in Treatment Effect Models,” Review of Economics and Statistics, 97, 1070–1080. [198]

Bollinger, C., and David, M. H. (2005), “I Didn’t Tell and I Won’t Tell: Dynamic Response Error in the SIPP,” Journal of Applied Econometrics, 20, 563–569. [198]

Bollinger, C., and Hirsch, B. (2013), “Is Earnings Response Ignorable,” Review of Economics and Statistics, 95, 407–416. [198]

Bollinger, C. R. (2003), “Measurement Error in Human Capital and the Black-White Wage Gap,” Review of Economics and Statistics, 85, pp. 578–585. [199]

Bollinger, C. R., and Minier, J. (2015), “On the Robustness of Coefficient Estimates to the Inclusion of Proxy Variables,” Journal of Econometric Methods, 4, 101–122. [199]

Borghans, L., Golsteyn, B. H. H., Heckman, J. J., and Humphries, J. E. (2016), “What Grades and Achievement Tests Measure,” PNAS, 113, 13354–13359. [197]

Bound, J., Brown, C., and Mathiowetz, N. (2000), “Measurement Error in Survey Data,” Research Report, 00-450 Population Studies Center at the Institute for Social Research. [198,199,204,210]

Cobb-Clark, D. A., and Kassenboehmer, S., and Schurer, S. (2014), “Healthy Habits: What Explains the Connection Between Diet, Exercise, and Locus of Control?” Journal of Economic Behavior & Organization, 98, 1–28. [201]

Cobb-Clark, D. A., and Schurer, S. (2012), “The Stability of Big-Five Personality Traits,” Economics Letters, 115, 11–15. [197,201]

d’Ardenne, J., Collins, D., Gray, M., Jessop, C., and Pilley, S. (2017), Assessing the Risk of Mode Effects: Review of Proposed Survey Questions for Waves 7–10 of Understanding Society, Understanding Society Working Paper, 2017-04, Essex: University of Essex. [209]

Deary, I. J., J. Yang et al. (2012), “Genetic Contributions to Stability and Change in Intelligence From Childhood to Old Age: Follow-up of the 1932 Scottish Mental Survey,” Intelligence, 48, 212–215. [202]

Deary, I.J., Whalley, L., Lemmon, H., Crawford, J., and Starr, J. (2000), “The Stability of Individual Differences in Mental Ability From Childhood to Old Age: Follow-up of the 1932 Scottish Mental Survey,” Intelligence, 28, 49–55. [202]

Dee, T. S., and West, M. (2011), “The Non-Cognitive Returns to Class Size,” Educational Evaluation and Policy Analysis, 33, 23–46. [197]

Dehejia, R., and Wahba, S. (1999), “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs,” Journal of the American Statistical Association, 94, 1053–1062. [199]
Schurer, S. (2017), “Does Education Strengthen Life Skills of Adolescents?,” *IZA World of Labor*, 366. [197]

Schwarz, N., and Oyserman, D. (2001), “Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction,” *American Journal of Evaluation*, 22, 127–160. [197,198]

Smith, J., and Todd, P. (2001), “Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods,” *American Economic Review (Papers & Proceedings)*, 91, 112–118. [199]

Stahlecker, P., and Trenkler, G. (1993), “Some Further Results on the Use of Proxy Variables in Prediction,” *The Review of Economics and Statistics*, 75, 707–711. [199,204]

Sudman, S., Bradburn, N., and Schwarz, N. (1996), *Thinking About Answers. The Application of Cognitive Processes to Survey Methodology*, San Francisco, CA: Jossey Bass Publishers. [197]

Summerfield, M., Bevitt, A., Fok, Y., Hahn, M., La, N., Macalalad, N., O’Shea, M., Watson, N., Wilkins, R. and Wooden, M. (2017), “HILDA User Manual Release 17,” Melbourne: Melbourne Institute: Applied Economic and Social Research, The University of Melbourne. [198,199]

Todd, P., and Wolpin, K.I. (2003), “On the Specification and Estimation of the Production Function for Cognitive Achievement,” *The Economic Journal*, 113, F3–F331. [199,204]

Tourangeau, R. (2003), “Cognitive Aspects of Survey Measurement and Mismeasurement,” *International Journal of Public Opinion Research*, 15, pp. 3–7. [197]

Wand, M. P., and Jones, M. (1995), *Kernel Smoothing. Monographs on Statistics and Applied Probability*, London: Chapman and Hall. [201]

Watson, N. (2011), *Methodology for the HILDA top-up Sample*, HILDA Project Technical Paper Series, 1/11, Melbourne: Melbourne Institute of Applied Economic and Social Research. [199]

Watson, N., and Wooden, M. (2014), “Re-Engaging With Survey Non-Respondents: Evidence From Three Household Panels,” *Journal of the Royal Statistical Society, Series A*, 177, 499–522. [209]

Wolpin, K.I. (1995), “Empirical Methods for the Study of Labor Force Dynamics,” in *Fundamentals of Pure Applied Economic Series* (Vol. 60), F. Welch, ed. Labor Economics Section, Luxembourg: Harwood Academic Publishers. [198,199,204,210]

Wolpin, K.I. (1997), “Determinants and Consequences of the Mortality and Health of Infants and Children,” in *Handbook of Population and Family Economics* (Vol. 60), eds. M. Rosenzweig and O. Stark, Amsterdam: Elsevier. [198,199,204]

Wooden, M. (2013), “The Measurement of Cognitive Ability in Wave 12 of the HILDA Survey,” HILDA Project Discussion Paper Series, 1/13, Melbourne: Melbourne Institute of Applied Economic and Social Research. [198,200,201]

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section A and panel Data*, The MIT Press. [204]

Zamarro, G., Cheng, A., Shakeel, M. D., and Hitt, C. (2018), “Comparing and Validating Measures of Non-Cognitive Traits: Performance Task Measures and Self-Reports From a Nationally Representative Sample,” *Journal of Behavioral and Experimental Economics*, 72, 51–60. [197,203,209]

Zweimueller, J. (1992), “Survey Non-Response and Biases in Wage Regressions,” *Economics Letters*, 39, 105–109. [198]