University of Arkansas, Fayetteville
ScholarWorks@UARK

Graduate Theses and Dissertations

5-2016

Character Assessment: Three Essays

Collin E. Hitt
University of Arkansas, Fayetteville

Follow this and additional works at: https://scholarworks.uark.edu/etd

 Part of the Educational Assessment, Evaluation, and Research Commons, and the Education Policy Commons

Citation

Hitt, C. E. (2016). Character Assessment: Three Essays. *Graduate Theses and Dissertations* Retrieved from https://scholarworks.uark.edu/etd/1529

This Dissertation is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu, uarepos@uark.edu.

Character Assessment: Three Essays

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Education Policy

by

Collin Hitt
Southern Illinois University Carbondale
Bachelor of Arts in Philosophy and Political Science, 2002

May 2016
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council

_____
Jay P. Greene
Dissertation Committee Chair

_____
Patrick J. Wolf
Committee Member

_____
Gema Zamarro
Committee Member

Abstract

I propose a new approach to measuring character skills. In the following three essays, my co-authors and I measure the effort that adolescent students appear to put forward on surveys and tests. First, I examine the extent to which students simply skip questions or plead ignorance on surveys. Second, I develop new methods for detecting careless answers, those instances in which students appear to be "just filling in the bubbles." I show, using longitudinal datasets, that both measures are predictive of educational degree attainment, independent of measured cognitive ability and other demographic factors. Finally, I demonstrate that international differences in reading, math and science test scores appear in fact to partially reflect international differences in student effort on assessments. Just as some students skip questions and carelessly answer surveys, some students do the same on tests. To the extent that effort on surveys and tests reflects noncognitive skills, presumed international differences in cognitive ability (as measured by standardized tests) might in fact be driven by differences in noncognitive ability. Altogether, the measures explored in the paper present three new methods for quantifying student character skills, which can be used in future research. Throughout, my co-authors and I posit that the character skills that our measures capture are related to conscientiousness and self-control.

Acknowledgments

I must first acknowledge Dr. Jay P. Greene. He recruited me not only into graduate school, but into a scholarly life. He has encouraged this line of research from its infancy. The existence of this dissertation and its shape are due to him.

Dr. Patrick J. Wolf has, for my entire time as his student, provided encouragement and feedback on the work presented here. Indeed his research on education policy is what motivated me to examine the measurement of character skills.

Dr. Gema Zamarro joined the University of Arkansas in 2014 having conducted her own research into noncognitive skills. She has welcomed me as a co-author. This dissertation is in part an extensive of research that she began. I'm fortunate to be her student and co-author, and honored to be a co-founder with her of *Charassein:* The Character Assessment Initiative at the University of Arkansas.

My co-authors Albert Cheng and Julie Trivitt have been invaluable. The breadth and rigor of this dissertation owes much to them, as does the publication of our paper in the E*conomics of Education Review*. One couldn't ask for better or more patient colleagues.

This dissertation is comprised of three chapters, each a standalone project. Each chapter itself contains acknowledgments to individuals who have shaped and improved this work. Undoubtedly, I have omitted someone. This, and all other errors and omissions, are mine alone.

Dedication

This dissertation is dedicated to a very special soul, my wife.

Table of Contents

List of Publications

Hitt, Collin, Julie Trivitt and Albert Cheng. (2016) "When You Say Nothing At All: The Predictive Power of Student Effort on Surveys," accepted manuscript, in press, *Economics of Education Review* (doi: 10.1016/j.econedurev.2016.02.001)

Introduction

Education research is at a crossroads. For nearly twenty years, researchers and policymakers have focused primarily on test scores. It's easy to understand why. By the late 1990s, test scores showed a stark and disturbing "achievement gap" between white and minority students, in reading and math ability (Thernstrom and Thernstrom, 2004). White eighth graders tested - and continue to test - similarly to black and Hispanic twelfth graders. A reasonable consensus formed: increasing the test scores of disadvantaged students could close the opportunity gap in American society (e.g. Howell et al. 2006).

In policymaking and in research, a standard was laid out. Programs that failed to increase test scores were considered failing programs. Those that increased test scores were tagged for expansion. This certainly was the spirit of the federal No Child Left Behind Act of 2001.

But a paradox is now evident. Long-term studies have followed children after testing is over, through high school graduation, into college, and even into the workforce. Some programs have produced large, long-term gains in degree attainment and employment income but did not produce test score gains (Duncan and Magnuson, 2013; Hitt and Wolf, 2015; Elango et al. 2015). Conversely, there are programs that have produced large test score gains, only to see little impact over later outcomes (Angrist et al. 2013). This requires a re-thinking of priorities, in research and policy.

Childhood test scores and later-life outcomes are unquestionably correlated. But for many policy areas it is becoming apparent that *impacts* on test scores and *impacts* on later

outcomes are weakly correlated. Many programs failed to produce test score benefits but did produce attainment benefits - these programs clearly impacted something other than test scores. This situation begs a number of questions. What did these programs impact in children, if not test scores? What are researchers failing to measure in the early stages of policy interventions, before graduation day?

The semantic response is, "noncognitive" skills. These, by definition, are the skills and behaviors *not* captured by test scores. Some programs obviously impacted these undefined skills, and the impacts on these skills produce lasting benefits.

Rather than the term "noncognitive skills", scholars in psychology, economics and education are beginning to use "character skills" (e.g. Heckman, Humphries and Kautz, 2014; Reeves, 2015). This terminology is certainly more meaningful than "noncognitive." Still, a specific question remains. What exact character skills do these programs impact?

This question will take years to answer. One reason why the "achievement gap" was able to be so clearly documented, and a reason why the subsequent accountability movement was centered on test scores, was because standardized tests of reading and math were readily available (Heckman, Humphries and Kautz, 2014). These tests were convenient and trustworthy - it's easy to forget that testing technology took decades upon decades to develop. Generations of researchers focused on the question of how to measure reading and math ability, in order to produce the standardized tests that seem so commonplace today.

The same has not been true of noncognitive skills. While psychology has produced insights into the behaviors and skills that are important for long-term success, the means of

measuring these skills are limited (e.g. Duckworth and Yeager, 2015). I discuss these limitations in detail in the three essays that follow. But there is one concern in particular that dominates the work presented in this dissertation, and it has to do with a particular category of character skills that researchers are increasingly focusing on.

Conscientiousness is a construct studied intensively in personality psychology. This is the tendency towards self-control, orderliness, responsibility, a strong work ethic, and a respect for traditions and norms (John and Srivastava, 1999; Hill and Roberts, 2011). Conscientiousness could also be said to encompass decisiveness, punctuality and truthfulness (Jackson et al. 2010). This is not an easy concept to measure in schoolchildren.

The most reliable way to measure conscientiousness and self-control in students is through professional third party observation (e.g. Moffit et al. 2011). People who are familiar with the concepts being measured, who are trained to observe children in school, who are granted access to students and school records, are in the best position to rate children on these skills. But third party observations are logistically impossible for researchers to collect for all children. Again, we should remind ourselves that the popularity of achievement test scores in research and policymaking stems largely from the fact that these tests can be given cheaply, en masse, over a short period of time.

The closest equivalent to a standardized test in character assessment is self-reported surveys. Take, for example, the Chernyshenko Conscientiousness Scale. Respondents are asked whether they agree with statements such as "I invest little effort into my work," and "I

carry out my obligations to the best of my ability" (Hill and Roberts, 2011). Answers to these questions are aggregated into a composite scale score.

Relying on student self-reports creates many challenges, no matter what researchers are attempting to measure. But a very specific problem arises when measuring conscientiousness and self-control: students who do not possess these skills are less likely to focus on a task like a survey. In schools, some students don't turn in assignments, they don't pay attention in class, they don't follow the rules. Why would we expect these students to provide reliable reports on surveys?

This problem is intuitive and well known. Inattentive students introduce noise into survey data. But the implication for researching conscientiousness and self-control is more serious, and often ignored. Many of the students who actually lack these skills are difficult to identify in the data, because the reports they provide are inaccurate or incomplete. If researchers cannot identify students truly lacking self-control and conscientiousness, they will not have a full picture of the distribution of these character skills. This dissertation focuses on ways to identify students who are not putting forward serious effort on surveys and tests.

If it is possible for us to identify in the data students who are showing low effort, we might be able to collect information about their noncognitive skills. Indeed, that is the overall thesis of this dissertation: student character skills can be measured using their answer patterns on surveys and tests.

Chapter 1 is titled, "When You Say Nothing At All: The Predictive Power of Student Effort on Surveys," co-authored with Dr. Julie Trivitt and Albert Cheng. We explore a simple measure of student effort: the frequency with which students skip questions or say "I don't know" to routine questions on surveys. Using six longitudinal datasets of American youth, we examine whether item nonresponse on a baseline survey is predictive of later outcomes. On these surveys, the questions asked of students are routine and knowable. Controlling for reading ability, there isn't a ready explanation of why students would fail frequently to answer basic questions. We hypothesize that item nonresponse is a proxy measure for how haphazardly students might approach the daily work of school; if this was the case, we would expect item nonresponse to be negatively predictive of later educational outcomes. Indeed that is what we find. Item response rates are predictive of later educational attainment and/or income in every dataset we examine, controlling for cognitive ability and a large set of demographic variables.

Chapter 2 is titled, "Just Filling in the Bubbles: Using Careless Answer Patterns as a Proxy Measure of Noncognitive Skills." In Chapter 1, I examine a simple measure that is easy to calculate: the rate at which students fail to respond to questions. But what about students who provide a nominal but thoughtless response? It again is common sense that some students just "fill in the bubbles." The question is, can researchers detect when students are doing this? In Chapter 2, I develop a new method for doing so, building upon simple psychometric techniques and new methods designed to flag careless answers (Meade and Craig, 2012). Insofar as previous research has attempted to measure careless answers, it has done so with the goal of removing dubious responders from the data. I take the opposite

approach, leaving students in the data in order to explore whether careless answers contain independent information about their noncognitive skills. Using two longitudinal surveys of American youth, I demonstrate that the frequency with which students provide careless answers is independently predictive of later educational attainment. Again, students inadvertently revealed something important about themselves, simply by how seriously they took a survey.

The objective of Chapters 1 and 2 is to validate proxy measures of noncognitive skills that can be used as an outcome measure in later research (e.g. Cheng and Zamarro, 2016; Cheng, 2016). I show that these measures are predictive of later life outcomes, independent of test scores. That said, I also find that these measures are at least weakly correlated with test scores. A question remains of what to make of this correlation, and what it represents. Students who score poorly on tests are more likely to skip questions and give careless answers on surveys. Is this because effort on surveys is driven by cognitive ability, or because our measures of cognitive ability are contaminated by noncognitive effort? In the final chapter, I consider the possibility that the correlation between effort on surveys and scores on standardized tests is due to the fact that noncognitive skills impact test-taking effort.

Chapter 3 is titled, "When Students Don't Care: Re-examining International Differences in Test Scores, Using Novel Measures of Student Effort on Surveys and Tests," with lead author Dr. Gema Zamarro, and Dr. Ildefonso Mendez. Student motivation and self-control impacts test scores (e.g. Duckworth et al. 2011; Wise, 2014). World-wide,

perhaps the most famous standardized tests are the Programme for International Student Assessment (PISA). Much as tests within the United States have been used to document a gap in learning between white and minority students, PISA scores have been used to show a gap between low and high performing countries. The presumption, as in the United States, is that this test score gap represents differences in math and reading ability. We test whether this gap is instead driven by differences in student effort. The PISA tests possess unique design properties, which we exploit. Each student is randomly assigned a test booklet from a larger set of booklets. Across booklets, items are randomly ordered, with difficult questions appearing at the beginning of the test in some booklets and at the end of the test in others. Following previous research, we find that on average performance declines from the beginning to the end of the test, which cannot be explained by the relative difficulty of items (Borghans and Schils, 2012; Debeer et al. 2014). Some countries see sharper rates of decline than others, signifying perhaps different levels of effort on the test. Moreover, all students taking PISA tests also take a survey afterwards, from which we calculate item nonresponse and careless answers patterns. Within countries, these measures of effort are only weakly correlated with test scores. However, across countries, between 33 and 40 percent of the variation in test scores is explained by variation in effort. Differences in effort across countries represent a number of factors, but we posit that the main driver of effort is noncognitive skills. The findings of Chapters 1 and 2 support this position. The implications are important. Generally speaking, it is presumed that countries with poor test scores have students with poor reading and math skills. A natural policy reaction then is to explore what reading and math teaching strategies can be used to improve test scores. The policy

consequences are very different if test performance is actually driven by student effort, especially if student effort is a measure of character skills.

There is a specific theme throughout this research: students tell us something about their character and values by how they approach surveys and tests. There is a more general point: student data is complicated, and researchers need to be more creative and empathetic when using student data. It is possible to retrieve information on students' noncognitive skills using creative methods of data analysis, as my co-authors and I demonstrate. Researchers will be more motivated to conduct such analyses, if they understand how it is that students view the assessment process. An adult in a position of authority gives students low-stakes tests and anonymous surveys that can take hours to complete. Without any accountability for their performance, some students put forward impressive effort on these tasks, and others don't.

Why, absent accountability and incentives, do students put forward any effort at all on these tasks? What drives them to do so? Almost by definition, students need to be conscientious in order to complete surveys and tests.

In order to identify the noncognitive skills that are crucial for children to possess to be successful throughout life, social scientists will need a much larger set of measures than what is currently available. In the following three essays, I delve into a rich, ubiquitous and previously explored source of data: answer patterns, which can tell us whether students are being conscientious as they take surveys and tests.

References: Introduction

Angrist, Joshua D., Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak, and Christopher R. Walters. "Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice." (2013), National Bureau of Economic Research, No. w19275.

Borghans, Lex, and Trudie Schils. "The Leaning Tower of Pisa." Accessed February 24, 2015. http://www.sole-jole.org/13260.pdf.

Cheng, A., & Zamarro, G. (2016). Measuring Teacher Conscientiousness and its Impact on Students: Insight from the Measures of Effective Teaching Longitudinal Database (EDRE WP No. 2016-04). University of Arkansas: Fayetteville, AR.

Cheng, A. (2016). Like Teacher, Like Student: Teachers and the Development of Student Noncognitive Skills (EDRE WP No. 2015-02). University of Arkansas: Fayetteville, AR.

Debeer, Dries, Janine Buchholz, Johannes Hartig, and Rianne Janssen. "Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment." *Journal of Educational and Behavioral Statistics* 39, no. 6 (2014): 502–23.

Duckworth, Angela Lee, Patrick D. Quinn, Donald R. Lynam, Rolf Loeber, and Magda Stouthamer-Loeber. "Role of Test Motivation in Intelligence Testing." *Proceedings of the National Academy of Sciences* 108, no. 19 (2011): 7716–20.

Duckworth, Angela L., and David Scott Yeager. "Measurement Matters Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes." *Educational Researcher* 44, no. 4 (2015): 237–51.

Duncan, Greg J., and Katherine Magnuson. "Investing in Preschool Programs." *The Journal of Economic Perspectives* 27, no. 2 (April 1, 2013): 109–32. doi:10.1257/jep.27.2.109.

Elango, Sneha, Jorge Luis García, James J. Heckman, and Andrés Hojman. "Early Childhood Education." Working Paper. National Bureau of Economic Research, November 2015. http://www.nber.org/papers/w21766.

Heckman, James J., John Eric Humphries, and Tim Kautz. *The Myth of Achievement Tests: The GED and the Role of Character in American Life*. University of Chicago Press, 2014.

Hill, Patrick L., and Brent W. Roberts. "The Role of Adherence in the Relationship between Conscientiousness and Perceived Health." *Health Psychology* 30, no. 6 (2011): 797.

Hitt, Collin and Patrick J. Wolf, (2015) "Achievement versus Attainment: Are School Choice Evaluators Looking for Impacts in the Wrong Places?", invited book chapter, under review.

William G. Howell, Paul E. Peterson, Patrick J. Wolf, and David E. Campbell. "The Education Gap: Vouchers and Urban Schools." (2006). *The Education Gap: Vouchers and Urban Schools*, (Revised Edition), Washington: Brookings.

Jackson, Joshua J., Dustin Wood, Tim Bogg, Kate E. Walton, Peter D. Harms, and Brent W. Roberts. "What Do Conscientious People Do? Development and Validation of the Behavioral Indicators of Conscientiousness (BIC)." *Journal of Research in Personality* 44, no. 4 (August 2010): 501–11. doi:10.1016/j.jrp.2010.06.005.

John, Oliver P., and Sanjay Srivastava. "The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives." *Handbook of Personality: Theory and Research* 2, no. 1999 (1999): 102–38.

Meade, Adam W., and S. Bartholomew Craig. "Identifying Careless Responses in Survey Data." *Psychological Methods* 17, no. 3 (2012): 437.

Moffitt, Terrie E., Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J. Hancox, HonaLee Harrington, Renate Houts et al. "A gradient of childhood self-control predicts health, wealth, and public safety." *Proceedings of the National Academy of Sciences* 108, no. 7 (2011): 2693-2698.

Reeves, Richard V. *Does Character Matter?: Essays on Opportunity and the American Dream*. Brookings Institution Press, 2015.

Thernstrom, Abigail, and Stephan Thernstrom. *No Excuses: Closing the Racial Gap in Learning*. Simon and Schuster, 2004.

Wise, Steven L. "The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees." *Journal of Computerized Adaptive Testing* 2, no. 3 (2014): 1–17.

Chapter One

When You Say Nothing at All:

The Predictive Power of Student Effort on Surveys

Collin Hitt
University of Arkansas

Julie Trivitt
University of Arkansas

Albert Cheng
University of Arkansas

Abstract

Character traits and noncognitive skills are important for human capital development and long-run life outcomes. Research in economics and psychology now shows this convincingly. But research into the exact determinants of noncognitive skills has been slowed by a common data limitation: most large-scale datasets do not contain adequate measures of noncognitive skills. This is particularly problematic in education policy evaluation. We demonstrate that within any survey dataset, there is important latent information that can be used as a proxy measure of noncognitive skills. Specifically, we examine the amount of conscientious effort that students exhibit on surveys, as measured by their item response rates. We use six nationally-representative, longitudinal surveys of American youth. We find that the percentage of questions skipped during the baseline year when respondents were adolescents is a significant predictor of later-life educational attainment, net of cognitive ability. Insofar as item response rates affect employment and income, they do so through their effect on education attainment. The pattern of findings

gives compelling reasons to view item response rates as a promising behavioral measure of noncognitive skills for use in future research. We posit that response rates are a measure of conscientiousness, though additional research is required to determine what exact noncognitive skills are being captured by item response rates.

Keywords: Noncognitive Skills; Educational Attainment; Labor-market Outcomes; Human Capital

Section I: Introduction

Noncognitive skills have an important influence on educational attainment, labor market outcomes, and other measures of well-being. This finding has been a key contribution of human capital research and personality psychology over the past two decades (Almlund et al. 2011; Borghans et al. 2008; Borghans, ter Weel & Weinberg, 2008; Bowles, Ginitis, & Osborne 2001; Deke & Haimson 2006; Heckman, 2000; Heckman & Rubinstein 2001; Heckman & Kautz, 2012; Heckman, Stixrud, & Urza 2006; Kaestner & Callison 2011; Lindqvist & Vestman 2011; Lundborg, Nystedt, & Rooth, 2014; Mueller & Plug, 2006). However, as researchers turn to policy questions regarding noncognitive skills, they encounter a pervasive data challenge: the large national datasets commonly used in economics, and the administrative datasets used in public policy research, do not contain adequate measures of noncognitive skills.

Some survey and administrative datasets contain no measures at all of noncognitive skills. Other survey datasets do contain just a few self-reported scales designed to capture skills such as academic effort and locus of control. But even when self-reported data are collected, scale scores based on self-reports contain poor information about students who are not conscientious enough to complete the survey. We explore a new noncognitive measure based on the effort that students seem to exhibit on the surveys.

Specifically, we examine the frequency with which students skip questions or answer "I don't know." This variable can be used in datasets that contain no other variables of noncognitive skills. And in datasets that contain at least some traditional measures such as

student self-reports, item response rates can be added to gain a fuller picture of students' noncognitive skills.

Survey methodology research (e.g. Krosnick and Presser 2010, Smith 1995) has shown that survey response rates — the rate at which respondents actually answer the questions posed to them — are driven strongly by factors other than cognitive ability. Long, low-stakes surveys require conscientious effort to complete, much like the daily busywork of school and employment. In education and human capital research, little work has been done using item response rates, or other indicators of effort on surveys, as a measure of noncognitive skills.

In our analyses of six large-scale datasets, we seek to validate item nonresponse as a control variable for noncognitive skills. We show that it is predictive of educational outcomes, after controlling for a broad range of student and household demographic characteristics. The specific datasets we examine are the National Longitudinal Survey of Youth 1979 (NLSY:79), the National Longitudinal Survey of Adolescent Health (Add Health), The National Educational Longitudinal Study of 1988 (NELS:88), High School and Beyond (HSB:80), the National Longitudinal Study of Youth 1997 (NLSY:97), and the Educational Longitudinal Study of 2002 (ELS:02). These are important datasets for social science research. All of them follow nationally representative samples of American adolescents into adulthood.

We find evidence that survey item response rates capture important behavioral traits that are largely not captured by cognitive tests. By definition, they appear to capture *non*cognitive skills. Item response rates consistently predict later educational attainment as

standalone variables in sparse models. Before controlling for cognitive ability, item response rates are significantly predictive of later educational attainment in all six datasets. In the four datasets where item nonresponse is a significant predictor of educational attainment while controlling for cognitive ability, a one standard deviation increase in item response rates is associated with completing 0.10 to 0.30 additional years of schooling. We also examine the association with employment status. Insofar as the skills captured by item response rate and self-reports influence wages and employment, they appear to do so mostly through their effect on educational attainment.

This study makes three important contributions. First, it shows that most surveys also contain a behavioral, non-self-reported measure of noncognitive skills. It is important in research to have an objective measure. What respondents say about their noncognitive skills does not always reflect how they behave; item response rates provide behavioral information about respondents who may not have otherwise provided reliable information about themselves. Second, we identify a measure that can be used in datasets that contain no other valid measures of conscientiousness or academic effort. And third, we demonstrate the importance of thinking more creatively about existing data. Surely other latent measures of noncognitive skills exist in survey data that can provide additional new information about noncognitive skills, which we urge other researchers to explore.

The article proceeds as follows. In Section II, we review the economics literature on noncognitive skills, recent work from psychology highlighting measurement challenges, and survey methodology research on the problem of item nonresponse. In Section III, we describe the national datasets used for our analysis. In Section IV we discuss our empirical

models. In Section V, we present the results of our analyses. In the final section, we discuss the results that suggest survey item response rates are a relevant source of missing information on important student noncognitive skills.

Section II: Literature Review

Survey Research in Economics and Psychology

Noncognitive skills are called *non*-cognitive for a simple reason. They are the personality factors, character traits, emotional dispositions, and social skills that tests of cognitive skills fail to capture. Both noncognitive and cognitive skills influence educational attainment and earnings. Economists have recognized that students with similar cognitive abilities vary widely in educational and labor-market outcomes later in life (Heckman and Rubinstein, 2001). However, the specific noncognitive skills that predict educational attainment and earnings are often unobserved. In such analyses, the effect of noncognitive skills on these outcomes was presumably relegated to the residual, ascribed as measurement error or as a problem of omitted variables. This measurement challenge affects program evaluation and public policy analysis: for example, preschool and school-voucher programs have been shown to improve educational attainment without improving cognitive skills. The implied effect on noncognitive skills went unmeasured in the years immediately following the intervention (Chingos & Peterson, 2015; Duncan & Magnuson, 2013).

The field of personality psychology provides key insights into the noncognitive skills that play an important role in educational attainment. A personality trait that continually reemerges in the literature is conscientiousness. It and related behavioral traits such as grit and locus of control are now understood to be independently linked to academic and labor-

market outcomes (Almlund et al. 2011). Conscientiousness is "the degree to which a person is willing to comply with conventional rules, norms, and standards" (Borghans et al. 2008; Hogan & Hogan, 2007). Facets of conscientiousness include orderliness, industriousness, responsibility and self-control (Jackson et al., 2010). With respect to educational outcomes, conscientious students are more likely to complete homework assignments, less likely to skip class, and tend to attain higher levels of education (Credé, Roch & Kieszczynka 2010; MacCann, Duckworth & Roberts 2009; Poropat, 2009; Trautwein et al. 2006; Tsukayama, Duckworth & Kim 2013). Conscientious workers are less likely to engage in counterproductive behaviors at work (Dalal 2005; Roberts et al. 2007); for example one study found that physicians rated higher in conscientiousness were less likely to miss work and falsify paperwork (Callen et al., 2015). Thus the question emerges: which policy interventions can increase conscientiousness as well as other important noncognitive skills, especially in children?

Unfortunately, the datasets used in personality psychology — often limited samples of convenience — are usually ill-suited to evaluate the relationship between noncognitive skills, social institutions, and public policy. Conversely, the massive surveys that many economists and public policy researchers depend upon rarely include noncognitive measures based on the preferred survey instruments of psychologists, which comprise lengthy questionnaires. For example, the well-regarded Revised NEO Personality Inventory is a 240-item survey designed to measure what psychologists call the Big Five Personality Traits: Conscientiousness, Agreeableness, Neuroticism, Extraversion and Openness (Costa &

McCrae, 2008). Such scales are far lengthier than the scales usually included in national longitudinal surveys projects and program evaluations.

The economics research on noncognitive skills and educational attainment, in particular, leans heavily on large longitudinal surveys of children (e.g. Coughlin & Castilla, 2014; Heckman et al., 2006). Such surveys are typically long but at most contain only short subsections about noncognitive skills. These survey design features limit the information on noncognitive skills that can be captured by the survey instruments. The short scales included in these surveys can be useful, but there are some important limitations for research. We present three examples.

First, the same scales are not used across different datasets. Because the same psychological constructs are not measured in all surveys, it is difficult to compare research on noncognitive skills across studies using different datasets. This point is illustrated in greater detail in the following data section, where we discuss six major longitudinal datasets that we use in our analysis.

Second, even within the same survey, respondents may not interpret the questions about noncognitive skills in a similar way. This is illustrated by the problem of reference group bias. Self-reports of noncognitive skills are influenced by the reference group to which respondents compare themselves. As West et al., (forthcoming) note:

> When considering whether "I am a hard worker" should be marked "very much like me," a child must conjure up a mental image of "a hard worker" to which she can then compare her own habits. A child with very high standards might consider a hard worker to be someone who does all of her homework well before bedtime and, in addition, organizes and reviews all of her notes from the day's classes. Another child might consider a hard worker to be someone who brings home her assignments and attempts to complete them, even if most of them remain unfinished the next day. (p. 6)

This is a particularly acute problem for program evaluation and public policy analysis. Educational interventions that actually increase noncognitive skills may not be measured as doing so. Two recent studies of charter schools have found large positive effects on standardized test scores, student behavior, or student educational attainment; yet the charter school students paradoxically report lower scores on self-reported measures of noncognitive skills (Dobbie & Fryer, forthcoming; West et al., forthcoming). A possible explanation of these contradictory findings is that the treatment of attending a charter school caused students to alter the standards by which they judged their own skills, reflecting different standards within the charter and comparison schools.

A third problem with survey-based methods of measuring noncognitive skills is that some respondents do not even attempt to provide accurate information. Some engage in so-called "satisficing." That is, they provide socially desirable answers, select the first attractive answer option, or simply fill in the blanks without regard to the question asked (Krosnick 1991; Krosnick, Narayan and Smith, 1996). Other respondents simply do not answer questions at all, skipping the question or pleading ignorance.

In order to avoid these problems when measuring motivation, persistence or self-control, some researchers also ask respondents to complete a task rather than answer survey questions. For example, in the 1979 & 1997 National Longitudinal Surveys of Youth, respondents were asked to complete a coding speed exercise, a sort of clerical task. Examining NLSY:79, Segal (2012) demonstrated that this was a proxy for noncognitive skills, conscientiousness in particular.

While tasks may yield interesting information, there are also practical differences between explicitly-assigned tasks and our variable of interest, item response rates. In our analysis, the survey is the task, and item response is a tacit measure of skills. The nature of assigning a task like coding speed alerts the respondent to the fact that her performance is being judged; there is no such cue for item response.

Survey Effort and Survey Response Rates

We explore a partial solution to these challenges: surveys themselves can be viewed as tasks. In taking a survey, respondents are asked to complete a tedious task on mundane topics, with no external incentives to provide accurate information. For some students, surveys must seem much like homework. In the datasets we examine, many adolescent respondents skip questions or frequently answer "I don't know," plausibly signaling a lack of effort or focus.

When students fail to answer questions, they leave holes in their survey record. Conventionally, researchers simply treat the items that respondents fail to answer as missing data or measurement errors.

We take a different approach. Instead of ignoring instances of item nonresponse, we view these so-called measurement errors as valuable pieces of information. Adolescent respondents may inadvertently show us something about how they approach the monotonous and mundane tasks of schooling and employment by how they approach a survey. Item nonresponse or its inverse, item response rates, can be revealing and used as a variable in empirical analyses. We posit that the information captured by this variable contains information specifically about noncognitive skills. Following this literature review,

we lay out a simple empirical model to estimate whether survey item response rates are predictive of educational attainment and labor-market outcomes, independent of cognitive test scores. We use this as an indirect test of whether item response rates capture noncognitive skills.

Previous literature contains only suggestive evidence on this question. For example, one can test the correlation between noncognitive scale scores and item response rates using cross-sectional data. Based upon the 2010 wave of the NLSY:97 and the 2009 wave of the German Socio-Economic Panel, Hedengren and Strattman (2012) have shown that the correlation between self-reports of conscientiousness and survey item response rates is positive. However, item response rate may be endogenous in Hedengren and Strattman's work because they examine a contemporaneous relationship. Although noncognitive ability as measured by item response rates may influence income or educational attainment, it is also possible that income or educational attainment influences response rates via the increased opportunity cost of time. This raises the possibility of simultaneity bias. Still, Hedengren and Strattman's work suggests that there are conceptual reasons to believe that survey effort as measured by item response rates is related to noncognitive skills.

Other evidence from survey methods research suggests that item nonresponse is correlated with the noncognitive skills of respondents, though research methodologists rarely venture a guess at the precise noncognitive factors that underpin item nonresponse. It has long been established within the field of survey methodology that item nonresponse on surveys is not random (Krosnick & Presser, 2010). Among adults, income and educational attainment are positively correlated with item response rates (Smith, 1982). Question salience

and survey format influence item response rates (Smith, 1995), as can incentives (Singer & Ye, 2013), suggesting strongly that item response rates are driven by individual motivation or habits — traits distinct from individual's cognitive ability to understand the questions asked.

We believe previous research provides credible evidence to consider item response as a partial measure of noncognitive skills. However, the hallmark of noncognitive skills research in education is the ability of noncognitive measures to forecast later outcomes. To our knowledge, no published research has used item response rates to forecast educational attainment and labor-market outcomes. Insofar as previous research has compared item response rates to adult outcomes such as income and educational attainment levels, it has used cross sectional data or contemporaneous correlations. Any assessments of the association with education are typically done post hoc, since most respondents are adults typically finished with school. Comparisons to income are contemporaneous. In fact, in survey methods research, educational level and income are typically used to explain the variation in item response rates, not vice versa.

It seems highly plausible to us that causation runs in the other direction. Item response rates (as a proxy for other noncognitive skills) may account for variation in educational attainment and income. Longitudinal data are needed to test this hypothesis, with item response rates measured during childhood, before respondents have attained degrees or have begun a long term career. For adolescents, a survey is a routine but mundane task, kind of like homework and financial aid applications. In adolescence one's willingness to complete these basic tasks of schooling has significant influence on educational attainment and employment earnings (Lleras, 2008; Segal, 2013). It stands to reason that item

response rates on surveys may predict later outcomes as well. Our study is the first to use panel data to determine whether item response rates predict later educational attainment and earnings.

Before we proceed to a discussion of our data, it is important to note once more that even in the face of the limitations we have discussed, researchers have made remarkable progress investigating noncognitive skills. Research to date has been possible because many (and probably most) respondents indeed provide accurate and important information about their own noncognitive skills when asked. We are essentially examining the subset of students who do not exhibit strong effort on surveys, students whose self-reported noncognitive skills are unlikely to be accurate. Therefore the aim of our study is not primarily to alter the empirical models used by noncognitive skills researchers. Rather, we investigate a measure of student effort that can be added to those models.

Section III: Data

Our study uses six major longitudinal datasets that follow American middle and high school students into adulthood. Students participating in these surveys were born between 1957 and 1987. Each survey is designed to capture a nationally representative sample of American youth. In our analyses, we always use sampling weights to account for survey design effects and sample attrition so that all results remain nationally representative. Baseline survey years ranged from 1979 to 2002. The surveys contain rich data on student demographics and household characteristics. All participants were tested at baseline for cognitive ability. In each follow up survey, respondents were asked about their educational attainment and their current income and employment status. Below we briefly discuss facets of each dataset: the

samples, survey modes, the types of item nonresponse that arise, and other explicit measures of noncognitive skills collected. We also specify the years in which outcomes were measured.

≪Table 1 Here≫

Key features of each dataset are listed in Table 1. The descriptive statistics for item response rates in each dataset are shown in Table 2. Across datasets, the average item response rate is between 95 and 99 percent. Between 14 percent and 54 percent of respondents completed every question on the survey – item response rates provide no information to distinguish between students with perfect response rates.

≪Table 2 Here≫

There is also an apparent relationship between survey mode and item response rate. The two NLSY surveys were administered one-on-one, in a face-to-face format. The response rates are far higher in the NLSY surveys than in the other surveys, which were self-administered and used pen-and-paper formats. Across all datasets, however, item response rate is negatively skewed with obvious ceiling effects, reflecting the fact that a substantial portion of respondents answered every survey item.

The National Longitudinal Study of 1979 (NLSY:79)

The NLSY:79 began with 12,686 male and female youths ranging in age from 14 to 22 as of December 31, 1978. [1] Initial surveys were conducted in-person by professional interviewers following a pen-and-paper manual. Responses were logged by the interviewer. Item nonresponse (or "missing data") in the NLSY:79 stems from three sources: the refusal

[1] Note that sample sizes in Table 1 and subsequent tables do match the original sample size in NLSY:79 and each of the other data sets as described in this section. The disparity is due to sample attrition and missing data.

to respond to a particular item, an answer of "don't know", or the incorrect skipping of an item. Interviewers were responsible for distinguishing between refusals and answers of "don't know." The distinction between these two kinds of item nonresponse is therefore blurred. Also, the incorrect skipping of an item is primarily due to interviewer error. For the NLSY:79, we therefore define item nonresponse rate as the rate of refusals and answers of "don't know."

Regarding measures of noncognitive skills, respondents in the initial round of the NLSY:79 were asked a series of 23 questions adapted from the Rotter (1966) Locus of Control scale for adults. Higher scores indicate a high feeling of individual control over the events of one's life, while lower scores indicate a high level of external control.

High School and Beyond, 1980 (HSB:80)

High School and Beyond (HSB:80) followed two cohorts of students: the sophomore and senior classes from a nationally representative sample of US high schools in 1980. Data was collected by the US Department of Education. The analysis of HSB:80 begins with nearly 12,000 members of the senior-class cohort. We limit our analysis to this senior-class cohort; adult outcomes of the sophomore-class cohort are unavailable as they had barely completed undergraduate work at the final wave of data collection. The final year of the survey is five to six years after the end of high school, meaning that a substantial portion of the population has yet to enter the workforce after college. Thus, we include HSB:80 in only our educational attainment models. The survey mode was a self-administered pen-and-paper survey, with a proctor present. Questions were primarily multiple-choice or fill-in-the-blank format. "Don't know" or "refuse" were answer options for very few questions. The most

common instances of item nonresponse are when students skipped questions altogether. Some questions were asked only to a subset of students, conditional on answers to previous questions. For HSB:80, we define item nonresponse rate as the proportion of missing answers to all the questions that students should have answered conditional on answers to previous questions. HSB:80 also included two student-reported measures of noncognitive skills: the Rosenberg (1965) Self-Esteem Scale and the Rotter (1966) Locus of Control scale.

Several other longitudinal studies bear strong resemblance to the HSB:80. Among the datasets in our analysis, the National Education Longitudinal Study of 1998 and the Educational Longitudinal Study of 2002 are part of the same longitudinal study project administered by the U.S. Department of Education. We calculate item response rates similarly across those datasets.

The National Educational Longitudinal Study of 1988 (NELS:88)

NELS:88 interviewed about 12,000 eighth-graders during the spring semester of 1988, immediately before most students matriculated to high school. NELS:88 followed students until 2000, twelve years after their eighth grade year. NELS:88 used a self-administered, pen-and-paper survey instrument, similar to that used in HSB:80. Here again we calculate item nonresponse rates as the percentage of questions skipped by respondents. Similar to HSB:80, NELS:88 contains locus of control scale scores, as well as scores on a self-concept scale.

National Longitudinal Study of Adolescent Health (Add Health)

Add Health is a longitudinal survey of US middle and high school students (Harris & Udry, 2009). We use a publicly available version of the Add Health dataset. The public-use

version contains roughly 6,000 student records that were randomly-selected from the full sample. These students completed a 45-minute, in-school pen-and-paper survey. The baseline survey year was 1994-1995. About 4,700 of the students were additionally selected for in-home follow up surveys. For our analysis, we use data from those who participated in the in-home surveys because key information such as educational attainment and labor-market outcomes, which are collected in 2007-2008, are available only for this subsample. Survey response rates, however, are based upon the in-school, pen-and-paper survey since in-home interviews were primarily conducted using a computer adaptive system that largely removed the possibility of skipping survey questions. As with other pen-and-paper surveys in our analyses, the primary source of item nonresponse comes from skipping items that should have been answered. For Add Health, we calculate item nonresponse rates as the percentage of questions that respondents were supposed to answer but skipped altogether. Add Health also contains items from the Rosenberg (1965) self-esteem scale, which we incorporate into our analysis.

The National Longitudinal Survey of Youth 1997 (NLSY:97)

NLSY:97 is a survey of 8,984 American youths aged 12 to 17 in 1997. Surveys were computer-adaptive, administered in home with the assistance of a professional interviewer. Questions were primarily multiple-choice and "unsure" was a frequent answer option. Refusal to answer was also a response option, though prompts from computer software and the interviewer made outright refusal a less likely response than in the NLSY:79. We calculate item nonresponse as the rate at which interviewees answer "unsure" or refuse to answer items.

The NLSY:97 is rare among longitudinal datasets in that it includes a behavioral task that has been shown to measure noncognitive skills. As part of the Armed Services Vocational Aptitude Battery, participants are asked to match words to a numeric code, according to a key. This is a clerical task. Respondents are scored based on the speed and accuracy of their responses. Hitt and Trivitt (2013) found that coding speed is correlated with both item response rates and noncognitive ability in NLSY:97. As discussed in the literature review above, Segal (2013) found that coding speed is a plausible measure of conscientiousness.

The Educational Longitudinal Study of 2002 (ELS:02)

ELS:02 followed a nationally representative sample of over 15,000 tenth graders from 2002 through 2012. Like HSB:80 and NELS:88, the survey mode for the baseline year was a self-administered pen-and-paper survey. Similar to those surveys, "don't know" or "unsure" were rarely offered as response options in the multiple choice questions that constitute most of the survey. We calculate a respondent's item nonresponse rate in ELS:02 as the percentage of questions left unanswered among questions that the respondent should have answered based on responses to previous questions. ELS:02 also contains various self-reported measures of self-regulation. In particular, we use the general effort and persistence scale and the control expectations scale, which were used in the 2000 Program for International Student Assessment. These items were also field tested before use in ELS:02 as well as used in other research (Burns et al. 2003; Pintrich et al., 1993).

Summary

The surveys used in each of the six datasets above have common design features. They are supposed to be easily understandable. The pen-and-paper surveys are designed to be readable, even for students with reading skills well below grade level. The surveys are long, averaging more than 300 items, which to some students is undoubtedly boring and tedious.

We hypothesize that item response rates are driven by student motivation and effort, and not just cognitive abilities. Response rates are, at most, only moderately correlated with cognitive ability, ranging from null to 0.21. These figures indicate that item response rates are not simply explained by cognitive ability. This alone does not mean that item response rates capture other abilities. Item response rates may largely not capture any abilities at all; they could simply be noise. Thus, in the following section, we turn to our empirical strategy, which aims to establish whether item response rates – as a measure of effort on the survey – capture information about noncognitive skills. A hallmark of noncognitive skills research has been the fact that noncognitive skills are predictive of later-life outcomes, independent of cognitive ability. We examine whether that is the case for item response rates.

Section IV: Empirical Strategy

Empirical Models

Our study is concerned with a previously unexploited control variable for noncognitive skills. Failing to control for noncognitive abilities can be problematic when estimating human capital models. Consider the following model that specifies employment income, (Y) as a function of cognitive ability (A), educational attainment and work experience ($\boldsymbol{E}$), and demographic and household characteristics ($\boldsymbol{H}$):

$$Y_i = f(A, \mathbf{E}, \mathbf{H}; \boldsymbol{\beta}) + \nu, \quad (1)$$

where $\boldsymbol{\beta}$ is a vector of parameters to be estimated and ν is the error term. In these models, noncognitive ability is not specified and is therefore relegated to ν. Insofar as noncognitive skills are correlated with other independent variables, insufficiently controlling for noncognitive skills leads to biased estimates of $\boldsymbol{\beta}$ (Heckman & Kautz, 2012). Additionally, the importance of noncognitive skills on employment income cannot be identified based on this theoretical formulation.

In our analysis, we explicitly include noncognitive skills as an independent variable in our human capital models. That is, we specify, for example, employment income as

$$Y_i = g(A^c, A^n, \mathbf{E}, \mathbf{H}; \boldsymbol{\gamma}) + \mu, \quad (2)$$

where A^c captures cognitive ability, A^n captures noncognitive ability, $\boldsymbol{\gamma}$ is a vector of parameters to be estimated, and μ is the error term. Analogous models where employment status or educational attainment is the dependent variable can be specified as well. As discussed above, the difficulty in estimating (2) is that noncognitive skills are difficult to observe and most datasets do not have adequate measures of such skills.

We use a simple empirical strategy to estimate the effect of noncognitive abilities. We begin with educational attainment as our outcome of interest. We model years of schooling as an individual utility maximization decision where the costs and benefits can vary with cognitive and noncognitive ability. The costs of schooling include tuition and foregone wages, and the opportunity costs of effort. This model also allows marginal productivity of time spent to vary with cognitive and noncognitive abilities. We assume linearity in the parameters and estimate the following empirical model:

$$Y_i = \boldsymbol{\alpha}\mathbf{X_i} + \boldsymbol{\beta}\mathbf{H_i} + \gamma_c A_i^c + \boldsymbol{\gamma_n}\mathbf{A_i^n} + \epsilon_i \quad (3)$$

where Y_i is the years of education for individual *i*. $\mathbf{X_i}$ is a vector of control variables to detect regional differences in the costs of acquiring additional education (explicit and opportunity costs). The control variables we include in $\mathbf{X_i}$ are gender, and indicator variables for birth year and census region. $\mathbf{H_i}$ is a vector of individual characteristics that influence previously accumulated human capital, expected increase in the benefits gained in the marriage market, and the benefits of household production. The specific variables included in $\mathbf{H_i}$ are the highest grade completed by household head, race, and an indicator for living in a two-parent household[2]. A^c_i is standardized observed cognitive ability, as measured by math and verbal standardized tests included in each dataset. $\mathbf{A^n_i}$ is the observed noncognitive ability of individual *i* as measured by standardized response rate as well as the scores on a variety of scales designed to measure noncognitive skills (e.g. Rotter [1966] Locus of Control Scale). Finally, ϵ_i is a normally distributed error term. All equations are initially estimated using ordinary least squares with sampling weights to correct for sampling methods utilized.[3]

[2]To the degree that discrimination exists in labor markets or households make different investments in male and female offspring, many of our control variables could arguably be included either in X or H or both. We recognize the coefficients we estimate are reduced form, but are primarily interested in $\mathbf{A}_n$.

[3] To address the possibility that diploma effects exist, we also use multinomial logit to run models that treat educational attainment as a categorical dependent variable instead of a continuous variable as in our main models. For each of our six datasets, we estimate equation (3) via multinomial logit using the same explanatory variables but with six diploma levels of education: no degree, GED, high school diploma, some postsecondary education, bachelor's degree, and more than bachelor's degree. The only exception to this is the HSB:80 dataset which does not have a separate category for GED and more than bachelor's degree as educational attainment outcomes. Respondents to HSB are 12th graders so many of them are already on track to receive a high school diploma, while many high school dropouts and eventual GED earners are out of sample. Furthermore, the last wave of data collection for HSB occurred 6 years after the initial wave of data collection, making it uncommon to

Next, we estimate the impact of noncognitive ability as proxied by item response rates on labor-market outcomes, specifically employment income and employment status. We first use ordinary least squares to estimate equation (3) where log of employment income is the dependent variable. The covariates remain the same, except we additionally include measures of educational attainment and work experience.[4] For our employment status models, we use probit regression to estimate a model similar to equation (3), except the dependent variable is a binary indicator equal to one if the respondent is employed and equal to zero if the respondent is unemployed. As in the employment income model, we control for educational attainment and work experience when estimating the impact of noncognitive skills on employment status.

Summary statistics for the number of years of education completed by respondents in each dataset are listed in the second column of Table 3. The remaining columns of Table 3 show summary statistics for employment income and the percentage of respondents who are employed but broken down by gender. We turn to the results of these analyses in the next section.

≪Table 3 Here≫

Section V: Results

---

observe respondents who have already obtained a graduate degree. Importantly, results are similar whether we estimate equation (3) using ordinary least squares or estimate the multinomial logit model. We only present results based on the ordinary least square analysis for simplicity.

[4] It is possible income is influenced by the decision to enter the workforce. As robustness checks, we use Full Information Maximum Likelihood procedures to estimate models that account for selection that occurs as some people opt out of the labor market. These results are equivalent to those based on ordinary least squares estimations of equation (3) and are available upon author request.

To reiterate, our objective is to document the relationship between survey response rate and three life outcomes: educational attainment, employment, and income. All models control for our full spectrum of the respondent's baseline household and demographic characteristics: age, race, gender, parental household income, parent's education, single-parent household and census region. Additional controls for alternative measures of noncognitive skills and demographic characteristics such as mother's age at birth, are included when available.[5] Given this set of control variables, our results likely represent conservative estimates for the importance of noncognitive skills. Many of the variables we control for likely influence noncognitive skill formation, educational attainment and adult earnings.

Educational Attainment

Table 4 shows the estimates of our empirical models where the number of years of education is the dependent variable. All samples are restricted to observations present in our full model (column 5) as missing data is prevalent for many of our covariates.[6] As depicted in column 1, response rates are positively correlated with educational attainment across all

[5] In the baseline year, we use log of household income and dummy variables indicating the highest grade level of education attainment completed by the head of the household when available. Some data sets, such as ELS:2002, provided categorical instead of continuous measures of household income. In these cases, dummy variables were used to control for household income. Mother's age at birth was included for the HSB:80, NELS:88, Add Health and ELS:02. In order to give more uniform sample sizes in the NLSY:79 and NLSY:97, mother's age at birth was not used as a control variable.

[6] As it turns out, restricting the sample is a more conservative test of whether item response rates are noise or capturing something systematic. We are excluding the group of respondents whose covariates are missing because they failed to answer those questions. That is, we are setting out to determine whether item response rates are predictive of later life outcomes, amongst a sample of people who at least answered basic demographic questions.

six datasets, before including cognitive ability and survey responses on noncognitive skills. A one-standard-deviation increase in response rates is associated with attaining 0.11 to 0.33 years of additional education in this basic model, all statistically significant at the 0.05 level.

Cognitive ability is a much stronger predictor of educational attainment in all six datasets, per column 2, which includes cognitive ability as the only variable of interest. Comparing column 2 to column 1, the effect size for cognitive ability is larger than that of item response rate by several orders of magnitude, except in HSB:80. This remains the case once noncognitive controls are included in the same models as cognitive ability, per the remaining columns. An important matter to keep in mind when comparing cognitive and noncognitive effect sizes is that cognitive test scores are composite measures created from subtests of math, reading, verbal and other cognitive skills. The noncognitive measures, including item response rate, are parsimonious measures, capturing only a part of the larger body of noncognitive abilities.

When including both response rate and cognitive ability as explanatory variables to predict educational attainment, response rate remains significant in four of the six datasets. As depicted in column 3, when significant, effect sizes range from 0.10 to 0.30 additional years of education for every one-standard-deviation increase in response rate. By comparison, a one standard deviation increase in cognitive test scores is associated with a 0.10 to 1.44 year increase in additional years of education attained. The co-variation between item response rate and cognitive test scores influences the relationship between item response rate and educational attainment. We discuss concerns about this cause of attenuation in Section VI.

As mentioned, the specification in Column 3 contains no other noncognitive variables. We have argued that item response rates can serve as a measure of noncognitive skills, particularly in datasets that contain no other measure. The specification in Column 3 includes a set of regressors that resembles the data typically available in education program evaluations: test scores, household information, and item response rates (which are available but ignored). Researchers using administrative data typically have no explicit measure of students' noncognitive skills. In this respect, the NLSY97 resembles administrative data often used in program evaluation: it contains no baseline-year, self-reported measure of noncognitive skills (the only baseline measure of noncognitive skills is coding speed, which is behavioral and not self-reported). Item response rate is consistently a significant predictor of educational attainment in that dataset, providing new and relevant information about participants' noncognitive skills.

Column 4 contains the model without nonresponse but with self-reported measures of noncognitive skills (or in NLSY:97, the coding speed task). In every instance, self-reported scales are predictive of educational attainment, independent of cognitive ability. Comparing Column 3 to Column 4, the addition of self-reported noncognitive skills adds relatively little to the overall R-squared, no more than 0.017 in the case of ELS:2002. Nevertheless, the coefficients for self-reported noncognitive skills are largely significant. Similarly, when comparing Column 2 to Column 3 the addition of item response rates does not substantially increase the R-squared. For self-reported and behavioral measures of noncognitive skills, this suggests that part of the effect was previously hidden within demographic control variables.

Column 5 of Table 4 displays estimates of a full model in which we include self-reported measures of noncognitive skills along with item response rates. Item response rates in these models remain statistically significant in HSB:80, Add Health, and NLSY:97. Response rate remains positive but falls short of significance in the remaining datasets. The coefficients on self-reported noncognitive skills rarely change when including item nonresponse, i.e. comparing columns 4 and 5. This suggests that item nonresponse can provide additional information about noncognitive skills, rather than serving as a substitute for traditional measures. This is also consistent with our assertion that item response rates capture information not captured by self-reports. For adolescents with low item response rates, the answers on self-reported measures may be so unreliable that they constitute random noise.

≪Table 4 Here≫

Employment and Income

We now turn to the results for employment and income. We first examine whether respondents reported being employed during the most recent survey year. Table 5 shows probit results. These estimates test whether the association of employment with item nonresponse is independent not only of measures collected during childhood but also of educational attainment, workforce experience and marital status. We have already demonstrated that item response rates are associated with later educational attainment.

≪Table 5 Here≫

Item response rate has no additional association with employment. This is also largely true of cognitive ability. Insofar as cognitive ability and noncognitive ability impact later employment, our results suggest they do so via educational attainment.

We then turn to the question of income from employment, per Table 6. Simply regressing the log of income on the same set of covariates as above, we find again that item response rates have no additional association with employment income, except in NELS:88, where a one standard deviation increase in item response rates is associated with a 3.5 percentage point increase in employment income.

≪Table 6 Here≫

Section VI: Discussion and Conclusion

The importance of our findings rests first upon whether we have made a convincing argument that survey response rates capture noncognitive ability. This study began by considering the perspectives of adolescents participating in a survey, who are asked to answer hundreds of boring questions about everyday life. There is strong presumption in the field of survey methodology that item nonresponse signals disinterest or disengagement in the survey process. We have argued that, seemingly, survey completion mirrors the routine work of school, which in psychological research has consistently been linked to noncognitive skills.

We then test whether item response rates independently predict outcomes that have a well-established relationship with both cognitive and noncognitive skills. We find that item response rates are a significant predictor of educational attainment in every dataset, before controlling for cognitive ability. Once including cognitive test scores, the effect of item

response rates attenuates, but remains significant in four of six datasets. Our results show that item response rate is not predictive of employment income and employment status, but our models include educational attainment as a control variable. Previous work suggests that labor-market benefits attributable to noncognitive skills operate through the effect of noncognitive skills on educational attainment (Cawley, Heckman, & Vytlacil, 2001). According to the simple definition of noncognitive skills as "not cognitive skills," survey response rates possess the characteristics of noncognitive skills that are related to later life outcomes.

It is worth noting that our estimates show that the effect of noncognitive skills attenuates when cognitive test scores are included.[7] Just like surveys, low-stakes cognitive tests require effort. Students showing low effort on surveys might be showing low effort on the accompanying cognitive test as well, leading to an artificially low estimate of their cognitive abilities. In controlling for test scores, part of what we attribute to cognitive ability is simply effort on the test. From previous literature, we know with confidence that test scores are affected by student motivation and noncognitive skills (e.g. Duckworth et al. 2011, Levitt et al. 2012).

Thus, any correlation between test scores and item response rates causes attenuation in some of our results. The implication could be that cognitive ability affects response rates. But the correlation between cognitive tests and item response rates could just as easily

[7] Another source of attenuation in our estimates is the inclusion of demographic and human capital variables in our regression models. While this attenuation makes it difficult to measure the impact of noncognitive skills on later outcomes, it also illustrates that some of the effect attributed to demographic factors is associated with specific behaviors or noncognitive skills.

indicate the opposite: item response rates partly capture the motivation (or lack of motivation) of students to complete low-stakes tests and other mundane tasks[8]. Item response rates are admittedly a noisy measure of noncognitive skills such as student effort. But even a cleaner measure of student effort, if used as a predictor of later outcomes, would still suffer from attenuation when cognitive test scores were included – because test scores themselves are affected by student effort.

In both of our educational attainment analyses, the statistical significance of item response rate is rarely influenced by the inclusion of self-reported noncognitive skills of locus of control or self-control. Conversely, the estimates of these self-reported noncognitive skills rarely attenuate substantially upon the inclusion of item response rates. The exception to this pattern is ELS:02, which contains self-reported measures of persistence and effort. Perhaps item response rate measures a particular set of conscientious behaviors. Or perhaps item response rate measures noncognitive skills similar to what the scales were designed to capture, and that item response rate contains information from respondents whose self-reports were essentially just noise, due to a lack of attention to the survey. Ultimately, in future research, survey effort should be compared to performance on other tasks or to third-party skills assessments.[9]

[8] Nascent work on this topic was begun over a decade ago in an unpublished manuscript by Boe, May and Baruch (2002) which examined the relationship between student scores on the Trends International Mathematics and Science Study and item response rates on a corresponding survey. Our findings strongly suggest that such work should be revisited.
[9] This is a topic for future research, where the data and methods of psychologists and experimental economists are of considerable value. Under laboratory conditions, it has been shown that financial incentives and fatiguing exercises have temporarily altered a person's observed self-control or conscientiousness (Hagger et al. 2010; McGee and McGee 2011; Segal 2012). Similar experiments could be conducted on survey effort. Evidence from field

We must also acknowledge that item nonresponse is a limited measure in some ways. Item response rate is undoubtedly a noisy measure with ceiling effects. Estimates based solely on item response rate will be prone to false negatives. Relying on it as the sole noncognitive measure is not advisable, but sometimes the data give no other choice. The NLSY:97, for example, contains no self-reported noncognitive skills in the baseline year. Used by itself, item response rate is of course limited in value, but this is true of any single measure of noncognitive skills, including short, self-reported scales. For this reason, it is common for researchers to build composite indices of noncognitive skills (e.g. Heckman et al., 2006). Our results suggest that item response rates could be included in such composite measures.

In future research, we will explore how item response rates can be combined with other measures to form stronger, more comprehensive measures of noncognitive skills, whether these measures of noncognitive skills serve as key dependent variables of interest or as control variables. It is possible that the inclusion of item response rate as a control variable, when no other noncognitive skill measures are available, could alter estimates of other variables of interest. It is our hope that other researchers join this effort. The object of this paper is to demonstrate that item response rates, and other measures of survey effort, are worthy of further attention.

---

experiments would also be instructive. Experimental programs have been shown to improve student study habits and focus in school; it would be instructive to learn whether treatment effects also exist on measures of survey effort. Such research could provide considerable insight into what psychological constructs in particular underlie survey effort.

In conclusion, we summarize three important contributions of this article. Primarily, our work establishes that response rates capture noncognitive skills that are important to future educational attainment, which ultimately affects other longer-run outcomes, such as labor-market outcomes (Cawley et al., 2001). While self-reported measures of noncognitive skills may show what attitudes and character traits are associated with those outcomes, our measure is behavioral. Self-reported noncognitive measures tell us that people who say that they have higher noncognitive skills on balance do better in life. Our findings provide further clues into how people with higher educational attainment behave: they complete mundane tasks given to them by relative strangers in positions of authority, even if the immediate incentive to complete that task is unclear.

Second, the noncognitive variable that we validate can be used in hundreds of existing datasets that do not contain better measures of noncognitive skills. The information captured by item response rates can be used to evaluate the impact of certain policies on those skills. Moreover, even in datasets with explicit measures of noncognitive skills, item response rates do not suffer from the problems of reference group bias and satisficing that plague those measures. That said, as with other measures of noncognitive skills, it should also be noted that this measure has limited viability as a way to evaluate noncognitive skills in data collected for high-stakes evaluations, especially in cases where participants would be aware that item response rate is a performance measure. It is also worth noting that recent digital survey designs that force respondents to answer all questions before they can proceed to the next section are eliminating this latent noncognitive skill measure in many datasets — which may incidentally introduce measurement error by generating forced, careless answers.

Third, and perhaps most importantly, our findings show the benefit of thinking more creatively about the data used in economics and education research. In our case, we examine long surveys completed by adolescents. Item response rates are a latent source of data that has been available for decades, but missing answers have been treated simply as measurement errors even though it has long been understood that item nonresponse is not random. If simple item nonresponse can be shown to be a measurement of other noncognitive skills, then social scientists and psychometricians should begin to explore other latent measures of noncognitive skills that are perhaps more difficult to measure.

The field of economics has made crucial contributions to the understanding of noncognitive skills' importance to education, employment and well-being. The single greatest challenge faced by this research program is the omission of noncognitive measures from key datasets. Discovering and exploiting new and latent measures of noncognitive skills will only enhance future noncognitive skills research. This is what we have set out to do.

Two decades ago, noncognitive skills were "dark matter," relegated to the residual in economic models (Heckman & Rubenstein, 2001, p. 149). Bit by bit, researchers have brought the role of noncognitive skills into clearer view. In an incremental step, our research helps rescue noncognitive skills from the error term.

Acknowledgements

We wish to thank Martin West, Darren Lubotsky, Anna Egalite, and participants at the 2014 Association for Public Policy Analysis and Management Conference; Angela Duckworth and members of the Character Lab at the University of Pennsylvania; Laura Crispin and participants at the 2014 Southern Economic Association Conference; participants in the University of Arkansas Department of Economics Seminar Series; participants in the University of Arkansas Department of Education Reform Brown Bag Seminar Series; the editors and reviewers at *Economics of Education Review* for their comments and suggestions on earlier versions of this article. Credit for any remaining or omissions errors belongs to the authors.

References

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. D. (2011). Personality Psychology and Economics. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (pp. 1–181). Amsterdam: Elsevier.

Boe, E.E., May, H., & Boruch, R.F. (2002). Student task persistence in the Third International Mathematics and Science Study: A major source of achievement differences at the national, classroom, and student levels. Center for Research and Evaluation in Social Policy, University of Pennsylvania: Philadelphia, PA.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas Ter Weel. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources* 43 (4): 972–1059.

Borghans, L., ter Weel, B., & Weinberg, B.A. (2008). Interpersonal styles and labor market outcomes. *Journal of Human Resources*, *43*(4), 815–858.

Bowles, S., Gintis, H., & Osborne, M. (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature*, 1137–1176.

Burns, L. J., R. Heuer, S. J. Ingels, J. Pollack, D. J. Pratt, …, & Stutts, E. (2003). Education Longitudinal Study of 2002 base year field test report (No. NCES 2003-03). National Center for Education Statistics: Washington, DC.

Callen, M., Gulzar, S., Hasanain, A. Khan, Y., & Rezaee, A. (2015). Personalities and Public Sector Performance: Evidence from a Health Experiment in Pakistan (NBER Working Paper No. 21180). National Bureau of Economic Research: Cambridge, MA.

Cawley, J., Heckman, J.J., Vytlacil, E. (2001). Three observations on wages and measured cognitive ability. *Labour Economics,* 8(4), 419-442.

Chingos, M. M., & Peterson, P. E. (2015). Experimentally Estimated Impacts of School Vouchers on College Enrollment and Degree Attainment. *Journal of Public Economics*, *122*, 1–12.

Costa, P.T., & McCrae, R.R. (2008). The revised Neo Personality Inventory (neo-Pi-R). In G.J Boyle, G. Matthews, & D.H. Saklofske (Eds.), *The SAGE Handbook of Personality Theory and Assessment* (2nd ed.) (pp. 179–98). SAGE Publications Ltd: London, UK.

Coughlin, C., & Castilla, C. (2014). The effect of private high school education on the college trajectory. *Economics Letters*, 125(2), 200-203.

Credé, M., Roch, S.G., & Kieszczynka, U.M. (2010). Class attendance in college a meta-

analytic review of the relationship of class attendance with grades and student characteristics. *Review of Educational Research, 80*(2), 272–295.

Cunha, F, & Heckman, J.J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources, 43*(4), 738–782.

Dalal, R.S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90*(6): 1241-1255.

Deke, J., & Haimson, J. (2006). *Valuing student competencies: Which ones predict postsecondary educational attainment and earnings, and for whom? Final report.* Mathematica Policy Research, Inc: Princeton, NJ.

Dobbie, W., & Fryer, R.G. (Forthcoming). The medium-term impacts of high-achieving charter schools. *Journal Political Economy.*

Duckworth, A.L., Quinn, P.D., Lynam, D.R., Loeber, R., Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences 108*(19), 7716–7720.

Duncan, G.J., & Magnuson, K. (2013). Investing in preschool programs. *The Journal of Economic Perspectives, 27*(2), 109–132.

Hagger, M.S., Wood, C., Stiff, C., & Chatzisarantis N.L. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological bulletin, 136*(4) 495-525.

Harris, K.M., & Udry, J.R. (2009). National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]. ICPSR21600-v16. Chapel Hill, NC: Carolina Population Center, University of North Carolina-Chapel Hill/Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors].

Heckman, J.J. (2000). Policies to foster human capital. *Research in Economics*, *54*(1), 3–56.

Heckman, J.J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics, 19*(4), 451–464.

Heckman, J.J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review*, 91(2), 145–149.

Heckman, J.J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior." *Journal of Labor Economics*, *24*(3): 411–482.

Hedengren, D., & Strattman, T. (2012). The dog that didn't bark: What item nonresponse shows about cognitive and noncognitive ability. Unpublished Manuscript. Retrieved from: http://ssrn.com/abstract=2194373

Hitt, C., & Trivitt, J. Don't know? Or don't care? Predicting educational attainment using survey response rates and coding speed tests as measures of conscientiousness (EDRE Working Paper 2013-05). Department of Education Reform, University of Arkansas: Fayetteville, AR.

Hogan, R., & Joyce, H. (2007). *Hogan Personality Inventory Manual,* (3rd ed.). Hogan Assessment Systems: Tulsa, OK.

Jackson, J.J., Wood, D., Bogg, T., Walton, K.E., Harms, P.D., Roberts, B.W. (2010). What do conscientious people do? Development and validation of the behavioral indicators of conscientiousness (BIC). *Journal of Research in Personality*, *44*(4), 501–511.

Kaestner, Robert, and Kevin Callison. (2011). Adolescent cognitive and noncognitive correlates of adult health. *Journal of Human Capital, 5*(1), 29–69.

Krosnick, Jon A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236.

Krosnick, J.A., Narayan, S., & Smith, W.R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation, 1996*(70), 29–44.

Krosnick, J.A., & Presser, S. (2010). Question and questionnaire design. In P.V. Marsden & J.D. Wright (Eds.), *Handbook of survey research (2nd ed.)* (pp. 263–314). Emerald Group Publishing Limited: Bingley, UK.

Levitt, S., List, J., Neckerman, S., Sadoff, S. (2012). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance (NBER Working Paper 18165). National Bureau of Economic Research: Cambridge, MA.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics*, 101–128.

Lleras, C. (2008). Do skills and behaviors in high school matter? The contribution of noncognitive factors in explaining differences in educational attainment and earnings. *Social Science Research,* 37(3), 888–902.

Lundborg, P., Nystedt, P., & Rooth, D. (2014). Height and earnings: The role of cognitive and noncognitive skills. *Journal of Human Resources*, *49*(1), 141–166.

MacCann, C., Duckworth, A.L., & Roberts, R.D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences,* 19(4), 451–458.

McGee, A., & McGee, P. (2011). Search, effort, and locus of control (IZA Discussion paper No. 5948). Retrieved from: http://www.econstor.eu/handle/10419/55119.

Morris, R. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Mueller, G., & Plug, E. (2006). Estimating the effect of personality on male and female earnings. *Industrial and Labor Relations Review*, *60*(1), 3–22.

Pintrich, P.R., Smith, D.A., García, T., McKeachie, W.J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ) *Educational and Psychological Measurement, 53*(3), 801–813.

Poropat, A.E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322-338.

Roberts, B.W., Harms, P.D., Caspi, A., & Moffitt, T.E. (2007). Predicting the counterproductive employee in a child-to-adult prospective study. *Journal of Applied Psychology*, *92*(5), 1427–1436.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80*(1), 1-28.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58*(8), 1438–1457.

Segal, C. (2013). Misbehavior, education, and labor market outcomes. *Journal of the European Economic Association*, *11*(4), 743–779.

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 112–141.

Smith, T.W. (1982). Educated don't knows: An analysis of the relationship between education and item nonresponse. *Political Methodology*, 47–57.

Smith, T.W. (1995). Little things matter: A Sampler of how differences in questionnaire format can affect survey responses. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 1046–1051. American Statistical Association Alexandria, VA.

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98*(2), 438-456.

Tsukayama, E., Duckworth, A.L. & Kim, B. (2013). Domain-specific impulsivity in school-age children. *Developmental Science, 16*(6), 879–893.

West, M.R., Kraft, M.A, Finn, A.S., Martin, R., Duckworth, A.L., Gabrieli, C. &, Gabrieli, J. (Forthcoming). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis.*

Table 1: Datasets

Dataset	Years of Data Collection	Respondent Age Range at Initial Year of Data Collection	National-Representativeness	Measure of Cognitive Ability
NLSY:79	1979 to 1992	14 to 22	Adolescents who were 14 to 22 as of December 31, 1978	Armed Forces Qualification Test (AFQT) Percentile
HSB:80	1980 to 1986	15 to 21	Twelfth-grade students in public and private schools during the 1979-1980 school year	Scores on standardized tests of math, reading, and vocabulary
NELS:88	1988 to 2000	12 to 15	Eighth-grade students in public and private schools during the 1987-1988 school year	Scores on standardized tests of math and reading
Add Health	1994 to 2008	10 to 19	Seventh- through twelfth-grade students in public and private schools during the 1994-1995 school year	Scores on an abridged version of Peabody Picture Vocabulary Test
NLSY:97	1997 to 2010	12 to 16	Adolescents who were 12 to 16 years old as of December 31, 1996	Armed Services Vocational Aptitude Battery (ASVAB) Math and Verbal Percentile
ELS:02	2002 to 2012	14 to 19	Tenth-grade students in public and private schools during the 2001-2002 school year.	Scores on standardized tests in math and reading

Table 2: Summary Statistics for Item Response Rate

	Observations	Mode of Survey	Item Response Rate			Questions Faced			
			Avg. %	SD	"Perfect"	Avg.	SD	Min.	Max.
NLSY:79	8,230	live interview	99.72	0.43	36.68	750.41	64.50	603	1,094
HSB:80	6,073	pen and paper	96.44	5.88	14.89	370.03	9.58	343	375
NELS:88	9,989	pen and paper	97.10	7.21	38.69	320.00	0.00	320	320
Add Health	2,458	pen and paper	94.86	14.51	54.47	97.16	2.88	87	105
NLSY:97	5,158	live interview	99.01	1.98	41.28	227.90	56.51	114	656
ELS:02	7,150	pen and paper	97.05	4.92	14.17	350.42	8.04	309	381

Note: Summary statistics are presented for the sample present in the full educational attainment model. The column marked "Perfect" indicates the percentage of students with item response rates of 100 percent. For NELS:88, some respondents were routed to additional questions based on answers to previous questions. A substantial portion of the optional questions are targeted at students whose parents are foreign-born or speak a language other than English. Item nonresponse to these questions is plausibly impacted by factors other than effort on the survey. We therefore excluded optional items on NELS:88 from our analysis. The number of observations in ELS:02 is rounded to the nearest ten per data-use license agreement.

Table 3: Summary Statistics for Years of Education and Labor-Market Outcomes

	Average Years of Education [standard deviation]	Average Employment Income ($) [standard deviation]		Percent Employed (%)	
		Males	Females	Males	Females
NLSY:79	12.92 [2.39]	25,364 [17,915]	17,891 [13,268]	82.93	71.86
HSB:80	13.19 [1.67]	n/a	n/a	n/a	n/a
NELS:88	14.24 [1.85]	30,979 [21,634]	22,897 [14,384]	97.28	94.91
Add Health	14.60 [2.12]	46,510 [58,601]	33,465 [35,278]	92.61	86.68
NLSY:97	13.52 [2.81]	35,261 [25,293]	27,877 [19,971]	73.92	66.64
ELS: 02	14.61 [1.96]	34,185 [26,424]	27,649 [20,929]	93.96	92.87

Note: In NELS:88 and ELS:02, years of education were imputed based on reports of highest degree completed. Dropouts were coded as 10 in NELS:88 and 11 in ELS:02, where baseline students were in the 8th grade and 10th grade, respectively. GED recipients and HS graduates were coded as 12, two-year college graduates as 14, four-year college graduates as 16, master's degree holders as 18, and higher graduate degree holders as 20. Summary statistics for employment income are restricted to panel participants who were employed. NLSY79 & 97 truncated employment income to the mean of the upper 2% for respondents with income at 98th percentile or higher for summary statistics.

Table 4: OLS Results for Years of Education

	(1)	(2)	(3)	(4)	(5)
NLSY:79 (N=8,230)					
Item Response Rate	0.134***		0.010		0.007
	(0.034)		(0.026)		(0.027)
Cognitive Ability		1.343***	1.342***	1.314***	1.313***
		(0.036)	(0.036)	(0.037)	(0.037)
Locus of Control				0.103***	0.103***
				(0.024)	(0.024)
R^2	0.290	0.482	0.482	0.483	0.483
HSB:80 (N = 6,073)					
Item Response Rate	0.291***		0.292***		0.269***
	(0.040)		(0.040)		(0.040)
Cognitive Ability		0.096**	0.096***	0.091**	0.092**
		(0.037)	(0.037)	(0.036)	(0.036)
Locus of Control				0.106***	0.097***
				(0.030)	(0.030)
Self-Esteem				0.107***	0.102***
				(0.027)	(0.029)
R^2	0.108	0.103	0.110	0.111	0.118
NELS:88 (N=9,989)					
Item Response Rate	0.107***		0.025		0.020
	(0.031)		(0.031)		(0.031)
Cognitive Ability		0.597***	0.594***	0.547***	0.545***
		(0.031)	(0.031)	(0.030)	(0.030)
Locus of Control				0.125***	0.125***
				(0.029)	(0.029)
Self-Concept				0.089***	0.089***
				(0.026)	(0.026)
R^2	0.332	0.402	0.402	0.411	0.411
Add Health (N=2,458)					
Item Response Rate	0.215***		0.144***		0.141***
	(0.042)		(0.043)		(0.043)
Cognitive Ability		0.519***	0.499***	0.528***	0.508***
		(0.059)	(0.059)	(0.059)	(0.059)
Self-esteem				0.126***	0.124***
				(0.045)	(0.044)
R^2	0.251	0.287	0.290	0.291	0.294
NLSY:97 (N=5,158)					
Item Response Rate	0.287***		0.139***		0.134***
	(0.048)		(0.045)		(0.045)
Cognitive Ability		1.444***	1.433***	1.353***	1.344***
		(0.039)	(0.039)	(0.045)	(0.045)
Coding Speed				0.177***	0.173***
				(0.043)	(0.043)
R^2	0.129	0.331	0.332	0.333	0.334

ELS:02 (N=7,150)					
Item Response Rate	0.325***		0.098*		0.033
	(0.057)		(0.053)		(0.054)
Cognitive Ability		0.726***	0.720***	0.642***	0.640***
		(0.029)	(0.029)	(0.030)	(0.030)
Control Expectations				0.116***	0.115***
				(0.036)	(0.036)
General Effort/ Persistence				0.170***	0.169***
				(0.036)	(0.036)
R^2	0.194	0.278	0.278	0.295	0.295

Notes: All independent variables are standardized. All models control for respondent's household and demographic characteristics. In NELS:88, ELS:02, Add Health, and HSB, years of education were imputed based upon highest degree attained. Such imputation may make the data left-censored and warrant Tobit regressions. However, results do not change whether one uses Tobit or OLS, so we report OLS estimates for simplicity. The number of observations in ELS:02 is rounded to the nearest ten per data-use license agreement.*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 5: Probit Results for Employment Status

	(1)	(2)	(3)	(4)	(5)
NLSY:79 (N=5,353)					
Item Response Rate	0.006		0.007		0.006
	(0.038)		(0.033)		(0.038)
Cognitive Ability		-0.007	-0.008	-0.007	-0.008
		(0.046)	(0.047)	(0.047)	(0.047)
Rotter Locus of Control				-0.007	-0.007
				(0.034)	(0.034)
Years of Education	0.031*	0.034*	0.033*	0.033*	0.032*
	(0.015)	(0.017)	(0.017)	(0.017)	(0.017)
R^2					
NELS:88 (N= 9,091)					
Item Response Rate	0.001		0.001		0.001
	(0.002)		(0.002)		(0.002)
Cognitive Ability		0.002	0.002	0.002	0.002
		(0.002)	(0.002)	(0.003)	(0.003)
Locus of Control				0.000	0.000
				(0.002)	(0.002)
Self-Concept				0.001	0.001
				(0.002)	(0.002)
Years of Education	0.006***	0.006***	0.006***	0.006***	0.006***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R^2					
Add Health (N = 2,395)					
Item Response Rate	0.004		0.003		0.003
	(0.007)		(0.007)		(0.007)
Cognitive Ability		0.008	0.008	0.008	0.008
		(0.008)	(0.008)	(0.008)	(0.008)
Self-Esteem				0.005	0.005
				(0.006)	(0.006)
Years of Education	0.010***	0.009***	0.009***	0.009***	0.009***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
NLSY:97(N=2,625)					
Item Response Rate	-0.017		-0.013		-0.013
	(0.045)		(0.045)		(0.045)
Cognitive Ability		-0.110*	-0.109*	-0.118*	-0.117*
		(0.059)	(0.059)	(0.063)	(0.054)
Coding Speed				0.016	0.016
				(0.054)	(0.054)
Years of Education	0.016	0.032*	0.033*	0.032*	0.032*
	(0.017)	(0.019)	(0.019)	(0.019)	(0.019)

ELS:02 (N= 6,200)					
Item Response Rate	-0.060 (0.074)		-0.070 (0.075)		-0.081 (0.076)
Cognitive Ability		0.032 (0.036)	0.037 (0.037)	0.047 (0.037)	0.052 (0.038)
Control Expectations				-0.124*** (0.041)	-0.123*** (0.041)
General Effort and Persistence				0.144*** (0.040)	0.147*** (0.040)
Years of Education	0.077*** (0.016)	0.072*** (0.017)	0.072*** (0.017)	0.070*** (0.017)	0.070*** (0.017)

Notes: All independent variables are standardized, except years of education, where the unit of measure is a single year of education completed. Coefficients are marginal effects holding all other variables at their mean. All models control for respondent's household and demographic characteristics. The number of observations in ELS:02 is rounded to the nearest ten per data-use license agreement. *** p<0.01, ** p<0.05, * p<0.10

Table 6: OLS Results for Log of Employment Income

	(1)	(2)	(3)	(4)	(5)
NLSY:79 (N= 4,280)					
Item Response Rate	0.035		0.028		0.028
	(0.023)		(0.024)		(0.024)
Cognitive Ability		0.126***	0.124***	0.123***	0.121***
		(0.025)	(0.025)	(0.025)	(0.025)
Rotter Locus of Control				-0.012	-0.011
				(0.015)	(0.015)
Years of Education	0.125***	0.102***	0.101***	0.101***	0.101***
	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)
R^2	0.348	0.355	0.356	0.355	0.356
NELS:88 (N=8,496)					
Item Response Rate	0.038***		0.035**		0.035**
	(0.014)		(0.014)		(0.014)
Cognitive Ability		0.023**	0.018*	0.013	0.008
		(0.022)	0.011	(0.012)	(0.011)
Locus of Control				0.031**	0.031***
				(0.012)	(0.012)
Self-Concept				0.020*	0.019*
				(0.011)	(0.011)
Years of Education	0.065***	0.063***	0.062***	0.059***	0.059***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
R^2	0.381	0.380	0.381	0.383	0.384
Add Health (N=2,098)					
Item Response Rate	-0.008		-0.012		-0.013
	(0.020)		(0.020)		(0.020)
Cognitive Ability		0.045	0.046	0.049	0.051
		(0.032)	(0.032)	(0.032)	(0.032)
Self-esteem				0.043**	0.043**
				(0.021)	(0.022)
Years of Education	0.111***	0.106***	0.106***	0.104***	0.104***
	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)
R^2	0.147	0.149	0.149	0.151	0.151
NLSY:97 (N=4,187)					
Item Response Rate	0.017		0.011		0.011
	(0.022)		(0.022)		(0.022)
Cognitive Ability		0.128***	0.128***	0.110***	0.110***
		(0.023)	(0.023)	(0.024)	(0.024)
Coding Speed				0.038*	0.038*
				(0.021)	(0.021)
Years of Education	0.099***	0.080***	0.080***	0.079***	0.079***

	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
R^2	0.159	0.166	0.166	0.167	0.167
ELS:02 (N= 6,420)					
Item Response Rate	-0.018 (0.039)		-0.048 (0.040)		-0.055 (0.040)
Cognitive Ability		0.107*** (0.023)	0.110*** (0.023)	0.103*** (0.023)	0.106*** (0.023)
Control Expectations				-0.008 (0.023)	-0.006 (0.023)
General Effort and Persistence				0.042** (0.019)	0.042** (0.019)
Years of Education	0.083*** (0.010)	0.067*** (0.010)	0.067*** (0.010)	0.064*** (0.010)	0.064*** (0.010)
R^2	0.119	0.124	0.124	0.125	0.126

Notes: All explanatory variables in the table are standardized, except years of education, where the unit of measure is a single year of education completed. Regressions restricted to panel participants who were employed. All models control for respondent's household and demographic characteristics. In ELS, years of education were imputed based upon highest degree attained. For NLSY79 the untruncated reported income was used. For the NSLY97 only the truncated income variable is available. We ran a tobit model to account for the truncation of the upper tail. The number of observations in ELS:02 is rounded to the nearest ten per data-use license agreement. Results were the same as those reported here for all practical purposes. *** $p<0.01$, ** $p<0.05$, * $p<0.10$

Chapter Two

Just Filling in the Bubbles:
Using Careless Answer Patterns on Surveys as a Proxy Measure of Noncognitive Skills

Collin Hitt
Department of Education Reform
University of Arkansas

Abstract

This paper develops a new and potentially important behavioral measure of noncognitive skills. I quantify the extent to which students provide unpredictable or "careless" answers on surveys. Specifically, I examine answer patterns on Likert-type items that comprise attitude scales. Apart from students' literal answers on these items, I examine the overall pattern of answers to determine whether students appear to be providing unpredictable or "careless" answers. Self-reported scales are fundamental tools for survey researchers and exist in hundreds of existing datasets. The methods I present can be used to create careless-answer variables in any such dataset. Using the National Educational Longitudinal Study of 1998 and the Educational Longitudinal Study of 2002, I test whether careless answer patterns from adolescent respondents are predictive of later educational attainment, independent of cognitive ability and other traditionally-measured noncognitive skills. An increase in careless-answering predicts lower later educational attainment. I posit that careless answers, as I have quantified them, proxy as a behavioral measure of conscientiousness.

Section 1: Introduction

Education researchers are examining a growing number of "noncognitive" outcomes. This is a promising break from past practice. Historically, the outcome measure of choice has been standardized tests. Standardized tests are not designed to measure noncognitive skills – the character traits and personality factors such as self-control and conscientiousness – that are now understood to be important determinants of educational attainment and labor market success. Education researchers are investigating programs that seek to impact these softer skills, but such research is encountering substantial challenges. Social scientists have struggled, and continue to struggle, to measure such noncognitive skills, especially in the context of education program evaluation.

These struggles stem partly from the assessment tools being used. Self-reported surveys are one of the main tools used in noncognitive skills research. Students are asked to report on their activities and beliefs, and those answers are used to form quantitative measures of noncognitive skills. Many factors can bias responses to these surveys. For example, the accuracy of the survey data obviously depends upon respondents' actually paying attention to the survey. This points to an irony when measuring conscientiousness. Student attentiveness and effort on surveys is often determined in part by noncognitive factors that the surveys are attempting to measure.

The noncognitive skills that receive the most attention from education researchers today are closely related directly to student discipline and the daily work of school. These include conscientiousness, grit, locus of control and mindfulness. When surveying students

about these skills, survey researchers are not only counting on the fact that respondents will be candid with them in their answers, they are counting on the fact that respondents are even taking the time to read the survey. Virtually by definition, students who lack noncognitive skills such as conscientiousness are less likely to focus on a survey that is dozens or hundreds of questions long.

Surveys can be long and boring. Conscientious effort (or skill) is required to complete a long survey. If respondents lose focus or become disengaged, their responses lose accuracy. This is a major measurement problem for education research.

I propose a solution. It is actually possible to assess whether students are providing meaningful answers to surveys, and it is possible to use that information as a proxy measure of noncognitive skills. Previous research has examined "straight-line" answer patterns and item nonresponse, as possible measures of survey disengagement (e.g. Barge and Gehlbach, 2012; Hitt, Trivitt and Cheng, 2016). In this study, I develop a novel method for detecting incoherent or unpredictable answer patterns from individuals. This method is based upon psychometric tools developed for the purpose of assessing the consistency of survey instruments. Using those tools, I attempt to assess the consistency, or unpredictability, of student answer strings.

When students provide careless answers - simply to satisfy the demands of the survey - they muddy the data. Their answers are inaccurate. And yet they may actually reveal something about their noncognitive skills. If it is possible to quantify the extent to which

students are engaged in surveys, it may be possible to use that information as a measure of noncognitive skills. This information, in essence, forms a behavioral measure.

This is why I seek to develop a novel method for detecting careless or inconsistent answer patterns. The logic of psychometric measures such as Cronbach's alpha is that, in an internally consistent and reliable scale, answers to different items should be correlated across the survey sample. It is logically equivalent to say that, in an internally consistent scale, the answer to each item should be reasonably well-predicted by answers to other items on the rest of the scale. That is, a student's answer to a given item should be predictable, given his or her answers to the other items on the same scale. In this article, I simply examine the extent to which student-respondents provide answers that are far different than what their previous responses would have suggested.

In particular, I examine the Likert-type items that comprise the attitudinal scales in the self-administered portion of the National Educational Longitudinal Study of 1988 (NELS:88) and the Educational Longitudinal Study of 2002 (ELS:02). For each item, I use regression analysis to estimate the relationship between that item and the average score on the other items on the scale, across all students. I then use regression estimates to calculate each student's predicted response, given his or her response to other items on the same scale. A student's regression-predicted response is based on "item-rest" regressions that are mathematically equivalent to the item-rest correlations used for other psychometric purposes. The residual to an item-rest regression represents the extent to which a student, literally, gave unpredictable responses on that item. Students with consistently large residuals

are students who, by definition, are providing unpredictable and relatively inconsistent responses.

For all students, I quantify the extent to which they provided unpredictable responses. I hypothesize that the unpredictability of their responses signals a noncognitive trait, which I call carelessness. I test whether answer-unpredictability, or a pattern of careless answers, is explained by cognitive ability. I find that it is not. Next, I test whether answer-unpredictability is strongly correlated with the self-reported noncognitive skills collected by NELS:88 and ELS:02. Again, I find that it is not. I then estimate whether answer-unpredictability is associated with later educational outcomes, measured on average at age 26. My prior expectation is that answer-unpredictability, as a measure of a detrimental behavior like carelessness, will be negatively correlated with educational attainment. Indeed, independent of cognitive ability, self-reported noncognitive ability and a rich set of demographic controls, an increase in the unpredictability (or carelessness) of a respondent's answers to Likert-type items is associated with a significant decrease in the number of years of schooling completed. In the NELS:88, this effect is driven mainly by a 1.7 percentage point decrease in the likelihood of graduating from high school. In ELS:02, whose baseline population largely graduated from high school and attended some college, the effect is driven mainly by a 2.0 percentage point decrease in the likelihood of completing a bachelor's degree. The effect sizes for careless-answers are similar in magnitude to noncognitive skills measured using self-reported scales.

The rest of the paper proceeds as follows. Section 2 provides a review of the literature on noncognitive skills. Section 3 describes the data available in the NELS:88 and ELS:02. Section 4 presents a brief overview of psychometric techniques used to assess the internal consistency and reliability of surveys. Section 5 presents a novel method for measuring survey answer-unpredictability, or careless-answers. Section 6 presents analyses of the association between answer-unpredictability and later attainment outcomes. Section 7 concludes.

Section 2: Literature Review

The growing field of noncognitive skills research includes contributions from economics, psychology and education policy. Its modern origins lie in the scholarship of James Heckman, whose groundbreaking work demonstrated that GED recipients possessed cognitive skills similar to high school graduates who never attended college, yet their lifetime outcomes were similar to those of high school dropouts (Heckman and Rubenstein, 2001). In other research, Heckman demonstrated that the lifelong, lasting effects of the Perry Preschool Project could not be explained by the cognitive impacts of the early childhood program (Heckman, Pinto and Savelyev, 2013).

Much of the foundational work of noncognitive skills research established that noncognitive skills were important, simply by showing that cognitive tests failed to measure important variations in educational attainment, health outcomes and labor market success. Heckman and Rubenstein (2001) initially referred to noncognitive skills as "dark matter," a powerful force that exists but goes unobserved (p. 149).

Methods for Measuring Noncognitive Skills

Personality psychologists have helped to better define noncognitive skills. The discipline has provided useful concepts for the behaviors and traits that make up noncognitive skills; the discipline is also the source for survey tools now being used to measure noncognitive skills in surveys and program evaluations. In large-sample datasets, skills are measured using self-reported scales. For example, Rotter's (1966) Locus of Control scale was a popular tool for decades. The Duckworth Grit Scale is a prominent, more recent tool (Duckworth and Quinn, 2009). Self-reported scales are by far the most popular tool used to measure noncognitive skills. Personality psychology has helped bring the "dark matter" of noncognitive skills into clearer view. As a result, the term noncognitive skills is in many places being replaced by the term "character" skills (e.g. Heckman et al., 2014). That said, as policy researchers and program evaluators are attempting to assess these skills in children, serious measurement challenges are becoming more apparent.

Self-reported surveys require that respondents accurately report their noncognitive skills. Some respondents simply do not provide credible or legitimate answers to questions asked (Robinson-Cimpian, 2014; Hitt, Trivitt and Cheng, 2016).

Self-reports also are limited by reference-group bias, where respondents differ in the standards by which they judge their own behavior (e.g. West et al. 2014). For example, two students who actually put forward similar effort on schoolwork may rate themselves differently as hard-workers, based on their individual understanding of the concept of hard work. Education researchers are beginning to use anchoring vignettes, in an attempt to

partially deal with reference group bias (e.g. Vonkova et al. 2015), but those efforts are nascent.

Given these problems, researchers have turned to behavioral tasks to measure student effort and engagement (as well as other noncognitive skills). For example, students can be timed on how long it takes them to abandon a difficult or impossible puzzle, in order to measure persistence (e.g. Egalite, Mills and Greene, 2014). Famously, Walter Mischel developed the marshmallow task, to measure self-control and delay of gratification (Mischel, Ebbeson and Raskoff Zeiss, 1972). These tasks can provide valuable information about behaviors related to conscientiousness and persistence (Duckworth and Yeager, 2015). But games and behavioral tasks also have limitations. Tasks can be complicated to administer. They can be costly to design and difficult to interpret. And perhaps most importantly, many performance tasks are only now being developed. Social science research depends heavily on longitudinal datasets that were begun years, even decades ago. It is impossible to travel back in time to administer new behavioral tasks to students in years past.

A promising solution to this problem may come from information inherent in surveys and standardized tests. They too can be viewed as tasks. The data collected from students not only includes literal answers to the questions, but also more subtle information about whether participants were engaged. For example, respondents frequently skip questions or plead ignorance. Hitt, Trivitt and Cheng (2016) show that the rate at which students skip questions is negatively predictive of later educational attainment and employment status, independent of cognitive ability. Borghans and Schils (2015) are able to

quantify diminishing effort at the end of tests, by examining scores at the beginning versus the end of a test whose question order was randomized, and show that diminished effort is predictive of later attainment.

This paper continues in the spirit of such research, while making an important advance. It is easy to count the extent to which students skip questions throughout a survey, as done in Hitt, Trivitt and Cheng (2016). But some students can also engage in what survey researchers call "satisficing," the process of technically completing a survey while not providing careful information (Krosnick, Narayan and Smith 1996). It has been an open question of whether "satisficing" can be identified with confidence.

Survey researchers traditionally view satisficing as a source of statistical noise. But for noncognitive skills research in education, satisficing has more serious implications. Skills such as conscientiousness, grit and self-control are conceptually related to completing assigned tasks, such as surveys. Self-reported assessments might ask students whether they remain focused on tasks, or whether they follow instructions well. Students who easily lose focus may simply provide careless answers to such questions without even reading the item.

In short, self-reported surveys rely on students who lack focus or motivation to actually stay focused and motivated long enough to answer questions about their focus and motivation. The problem here is obvious.

Due to this problem, it is possible that self-reported scales contain very little information about students who are truly low in skills such as conscientiousness, persistence or self-control. Yet it is precisely these low-skilled students that noncognitive-skills

interventions are supposed to help. If low-skilled students cannot be identified in the data, then it will be impossible to know which programs can make a difference in their lives. Psychometric techniques may help recover information about these students, by determining the extent to which the answers appear effortful or careless.

Methods for Detecting Careless Answers

On surveys, adolescent respondents sometimes provide dubious answers. That is, they provide answers that they know are untrue. Of respondents who give dubious answers, there are two relatively distinct groups of students. There are students who read and understand the questions and then intentionally provide mischievous answers. And then there are students who pay very little attention to the questions and just fill in the bubbles - or, to use the parlance of psychometrics, give "careless" answers.

Careless respondents are the focus of this paper, but I will briefly discuss the literature surrounding mischievous responders. When surveyed, students are almost always asked questions about their race and gender. Questions about their religion and parents' national origin are also frequently asked. Many surveys ask about life experiences and unhealthy behaviors, such as whether students suffer from disabilities, belong to a gang, have children, or use drugs. Mischievous responders are students who intentionally give extreme answers to questions about their lives. For example, students may identify as being blind *and* in a gang *and* having multiple children, perhaps because they find it humorous to say so (Robinson-Cimpian, 2014). Some students give response combinations that are too

improbable to be credible. Recent studies have found that mischievous responders can substantially bias analyses of underrepresented subgroups of students.

New methods have been developed to identify mischievous responders, with the purpose of removing those students from the data. While an interesting phenomenon, I do not examine mischievous responses in this paper.

Rather, the focus of this study is students who simply give careless answers. Simple psychometric techniques can be used to identify (at least some) cases where students appear to be just filling in the bubbles. The method I present in Section 4 builds upon, and in some ways synthesizes, several existing methods used by psychometricians to flag careless answers.

A brief description of existing methods follows. As with methods used to detect mischievous responses, the methods used to detect careless responders have heretofore been used to flag and remove students from data. My focus ultimately will be much different - to use careless answer metrics to gather information about students' noncognitive skills.

There are various *post hoc* methods for detecting careless answers, after data collection is complete (Meade and Craig, 2012; Maniaci and Rogge, 2014). Each builds upon the fact that surveys are comprised often of multi-item scales, and answers to questions within the same scale should correlate with one another. I discuss scale construction and inter-item correlation in greater detail in Section 4.

One simple approach for identifying careless answers uses "psychometric synonyms" and "antonyms." A pair of items whose answers are strongly and positively correlated can be

called psychometric synonyms. Psychometric antonyms are pairs of items whose answers are strongly and negatively correlated. If a student gives dissimilar answers to questions that are synonyms - or too-similar answers to questions that are antonyms - his response could be flagged as "careless."

This method has limitations. The synonym and antonym approach is somewhat a-theoretical. Whichever item pairs are found to be strongly and positively correlated can be called synonyms, and vice versa for antonyms. Also, the threshold for what constitutes a strong enough correlation to declare item-pairs as either a synonym or antonym is arbitrary. A common threshold is $r = 0.60$, so responses to items with a correlation of $r = 0.59$ would be ignored.

A simpler approach exists, grounded in *a priori* expectations about which items should be correlated with one another. Multi-item scales can be split into halves, with scores on either half compared to one another. This builds off of the psychometric test of Rulon's split-half reliability, which is used to assess the internal consistency of scales, not the credibility of respondents. Internal consistency tests are discussed in greater detail in the Section 4. For example, average scores on even and odd items can be compared to one another. Presuming that split-half averages are strongly correlated across the entire sample, a respondent with vastly different scores on even and odd items could be considered careless (Johnson, 2005; Meade and Craig, 2012). However, this approach too ignores certain information. For example, if a respondent on an eight-item scale gives zig-zagged answers scored 1-2-3-4-4-3-2-1, his even-odd (and first-half, last-half) averages would be equivalent.

Another way to measure careless answers is to focus on extreme outliers: answers that appear to be an outlying response, given that answers as a whole appear distributed around the population mean. One measure of extreme outliers is Mahalanobis distance, which calculates the "multivariate distance between [a] respondent's response vector and the vector of sample means" (Meade and Craig, 2012). Pyschometricians who have used Mahalanobis distance to detect careless answers have done so using scales with a large number of answer options, sometimes seven or more. The attitude scales employed in large-scale education datasets typically use items with a narrower answer range. For example, I examine answer patterns in the NELS:88, which uses a four-point Likert-type scale to measure self-reported Locus of Control and Self-Concept. Mahalnobis distance calculations may not function well in data with a truncated range.

The measure I outline in Section 4 synthesizes the information that would be captured by each of the measures above, while also capturing information that each of these measures ignore. It is more complex to calculate than psychometric synonyms or split-half differences. Other the other hand, it is more understandable and easier to calculate than Mahalanobis distance. The intuition behind these approaches and my own, however, is similar. Careless answers can be detected by identifying irregular patterns of responses.

A reasonable objection to these methods is that legitimate answers are being flagged as careless. Some students might be divergent thinkers, with unconventional combinations of views on the questions being asked.

Recent literature from psychology suggests strongly that careless-answering measures inattentiveness. Some surveys include "bogus items" - items that either instruct a respondent to answer a specific way, or ask a question for which there can only be one credible answer. Meade and Craig (2012) compared the measures of carelessness outlined above to the frequency of bogus answers, in a convenient sample of college students. Bogus items scores were strongly correlated with psychometric-synonym flags and split-half differences, and moderately correlated with Mahalanobis distance.

Summary

The measurement challenges with noncognitive skills are many. I have outlined only a few of those challenges in this section. No single solution - no single measurement tool or method - can overcome those challenges. Incremental improvements to measurement methods are needed. In the remainder of this paper, I present a new method of measuring noncognitive skills. It is intended, in a modest way, to bring the "dark matter" of noncognitive skills clearer into view.

Section 3: Data

The National Educational Longitudinal Study of 1988 (NELS:88) is a survey of more than 12,000 American students attending eighth grade in 1988. The survey panel continued until 2000. At baseline, students were assessed math and reading tests. They were also issued a self-administered, pen and paper, multiple choice survey that contained 320 items (or more for some students). Questions ranged in topic from parental occupation to perceptions of

school to participation in sports. Two well-established noncognitive skills scales were also included, as discussed below.

The Educational Longitudinal Study of 2002 (ELS:02) is a survey of more than 15,000 students attending tenth grade in 2002. The survey panel continued until 2012. As with the NELS:88, students were administered math and literacy tests at baseline, and were issued a lengthy pen-and-paper survey. Questions on the survey covered a wide range of topics about daily life. The survey also contained scales on certain noncognitive skills, using the common Likert-type items. Unlike NELS:88, the ELS:02 contained dozens of other Likert-type items on other topics as well. The inclusion of Likert-type questions that are not part of the noncognitive skills assessments allows me to conduct important robustness checks, as discussed in Section 7.

The answers-patterns within Likert-type items are the focus of my analysis. To illustrate the nature of these survey tools, I focus here on the NELS:88.

Figure 1 is an excerpt from the student questionnaire from the NELS:88 baseline survey. The questions shown comprise two attitude scales - the Locus of Control scale and Self Concept scale.[10] The survey items use a four-point Likert-type format, the only questions on the NELS:88 that used this format. Students are asked whether they strongly agree, agree, disagree or strongly agree with a number of statements. This question format is widely used in survey research, and especially in personality psychology.

[10] The Self Concept and Locus of Control Scales used in NELS:88 were based off of similar scales from the High School and Beyond survey of 1980 and the National Longitudinal Study of 1972. All scales were based off of instruments designed by Rosenberg (1965) and Rotter (1966).

<<Figure 1 Here>>

<<Tables 1A and 1B Here>>

Likert-type questions are popular because they allow individual items to be scored numerically. The items in Figure 1 are scored from 1 to 4, with strongly disagree scored a 1 and strongly agree scored a 4. For reverse coded items, the scores are reversed.

Tables 1A and 1B show the item-level summary statistics for each item in Figure 1. The items are grouped by scale. At the bottom of each table is a composite scale score, the simple average of the items above.[11]

In the estimates in Section 6, I used the following information measured at the baseline year: standardized (cognitive) test scores, noncognitive scale scores and student demographic information. In NELS:88, the self-reported noncognitive skills are Locus of Control and Self Concept. In ELS:02, the self-reported noncognitive skills are Effort (short for general effort and persistence) and Control Expectations. I also use information on educational attainment collected during the final year of the panel: the year 2000 for NELS:88 and the year 2012 for ELS:02.

Section 4: Reliability and Consistency

[11] This simple, composite scale score is calculated by me, and slightly different than composite scale scores reported in the NELS:88 dataset. Here, I report a simple average of raw item scores so that the reader can easily see how a scale score can be built. The main differences are as follows. The NELS:88 authors standardize item level answers, and then average those standardized scores. Again, I calculate a simple average of raw item scores. The NELS:88 pre-generated scores and the simple averages I report here are correlated at r = 0.999).

The field of psychometrics uses a standard set of procedures when creating composite scores from survey items. One of the most common procedures is a test for internal consistency called Cronbach's alpha, which reports the extent to which item-level answers co-vary. This is a popular test for a simple reason. Cronbach's alpha, and related statistics such as item-rest correlations, help to judge whether separate items are consistently measuring a similar construct.

A brief discussion of how survey scales are constructed will help illustrate the information contained in psychometric reliability statistics. Researchers, when creating a composite score, take individual answers to specific questions and then transform them into an abstract, composite value.[12] This is a potentially arbitrary process. Some questions are included in a composite score, others are not – sometimes these decisions are made after data is collected.

Within a particular scale, each item can be described as a different way of asking about the same underlying construct (or same set of constructs). In order for a scale to be deemed internally consistent, the answers to the component items should be correlated. This is what Cronbach's alpha is designed to test: the internal consistency and reliability of a multi-item scale.

[12] In creating a survey instrument, certain steps should typically be followed before the survey is deployed in the field. Researchers should have a strong theoretical reason, and some preliminary evidence, suggesting that a chosen set of questions can been combined to measure an underlying construct.

<<Tables 2A and 2B Here>>

Tables 2A and 2B report internal consistency and reliability statistics for the NELS:88 Locus of Control and Self Concept scales. The bottom right cell of the tables shows the Cronbach's alpha for the overall scale.

The item-level rows show individual item statistics. In the column 5, the Cronbach's alpha values represent what the overall scale alpha would be if that given item is removed. This statistic, when compared to the overall Cronbach's alpha for the scale, tells whether the scale can be made more reliable (or more internally consistent) by removing that particular item.

The values in column 5 are inversely related to the values in the three columns 2 through 4, which report the extent to which answers to an individual item are correlated with answers on the rest of the scale. For example, the item-rest correlation for item 44B is simply a Pearson's product-moment correlation coefficient. It reports the correlation (r = 0.406) between answers to item 44B and the simple average of the remaining items on the rest of the scale. Formally, within a given scale, item-rest correlations between student answers to item j and student answers to other items can be expressed as follows:

$$\text{corr}(x_j, \bar{x}_{i \neq j}) \qquad (1)$$

, where

$$\bar{x}_{i \neq j} = \frac{\sum_{i \neq j}^{n-1} x_i}{n-1} \qquad (2)$$

in a scale with *n* items. An item-rest correlation shows whether scores on an individual item are consistent with scores across the rest of the scale. A particularly weak item-rest correlation suggests that an item should perhaps be dropped from the composite calculation, since the item does not appear to be measuring the same construct as the other questions in the scale. In a scale considered highly reliable, individual item scores are moderately to highly correlated with scores on the remaining items.

I have presented a brief overview of these common psychometric tests because they perform a key role in my analysis. However, I propose to use these procedures – the item-rest correlations in particular – for an entirely different purpose. Rather than judge the reliability of a scale, I seek to quantify the unpredictability of respondents' answers.

Section 5: Identifying Unpredictable Answers

A problem in survey research is that respondents become disengaged, sometimes quickly. This is easy to imagine with respect to the NELS:88 or ELS:02. Eighth and Tenth graders respectively are given a low stakes, self-administered, pen-and-paper survey that is hundreds of items long. It's virtually certain that some students become disengaged. When they do, they might simply complete the survey by providing thoughtless or careless answers. That is, some students just fill in the bubbles. Such answers, when viewed together, can appear incoherent.

Most students dutifully fill out surveys. If this weren't so, survey data would be generally useless. This method identifies careless-answer patterns as those that are inconsistent with answer patterns across the entire population.

As discussed above, item-rest correlations are used in psychometrics to assess survey items. The same tool could be used to flag inconsistent or unpredictable responses, at least on Likert-type items such as those that make up the attitude scales in the NELS:88 and ELS:02.

The logic behind item-rest correlations is that answers to a particular item should, in an internally-consistent scale, be correlated with the answers to the other items in the scale. A logically equivalent statement goes as follows. In a reliable scale, on average, a respondent's answers to item *j* should be reasonably well predicted based on his answers to the other scale items, as judged by the answers on item *j* given by other respondents who had responded similarly to him on the other items on the scale.

Consider the following bivariate regression equation:

$$Y_{jst} = B_0 + B_1 X_{jst} + \eta_{jst} \qquad (3)$$

Where Y_{jst} is the answer given to item *j* of scale *s* by student *t*, and X_{jst} is the average of items besides item *j* on scale *s* by student *t*. B_0 is a constant and η_{jst} is the error term. In a standardized bivariate regression, the constant drops out, and the standardized coefficient for B_1 is mathematically identical to a Pearson's correlation coefficient. That is, for a given scale, B_1 in a standardized version of equation 3 provides identical estimates as the item-rest correlation coefficient in equation 1. Thus I will refer to equation 3 as an "item-rest" regression.

<<Tables 3A and 3B Here>>

Let us turn to data from NELS:88, for illustrative purposes. Table 3A and 3B show estimates of "item-rest" bivariate regressions for every item in the NELS:88 Locus of Control and Self Concept scales. Column 5 shows the standardized coefficients for each regression, which are identical to the corresponding item-rest correlation coefficients in Tables 2A and 2B.

As discussed, psychometricians would traditionally be interested in the standardized coefficient B_1 to equation 3, as it is equivalent to the item-rest correlation coefficient. This is the estimate used, in part, to judge the appropriateness of an item and reliability of the scale. I, however, am interested in the error term η_{jst}, which is literally the degree to which student *t* provided an unpredictable answer to item *j*, according to the regression results.

In a highly reliable scale, by definition, the average student's answer to item *j* should be reasonably well predicted by the regression estimates. My focus in this study is respondents who provide careless or inconsistent answers, on scales that overall appear to be reliable. These may be respondents who simply answer in a straight line or who zig-zag across the page. These may be respondents who provide random answer patterns, with no meaningful pattern at all. The potential shapes and patterns that inconsistent answers can take on the written page are innumerable. By examining individual respondent-item residuals, I can plausibly capture many different "satisficing" behaviors at once.

<<Tables 4A and 4B Here>>

Tables 4A and 4B show the summary statistics of the absolute values of the residuals to each of the "item rest" regressions in Tables 3A and 3B. For example, the top row shows that the absolute difference between the predicted values and the actual values for item 44B in table 4A was on average 0.56 points. Keep in mind that this is on a scale ranging from 1 to 4. For Item 44B, the maximum absolute value of a regression residual was 2.77. This respondent had a score of 1 for the first item, and an average score of 4 for the remaining items – a dubious answer string.

For any given respondent, a large residual for *an individual item* could stem from a number of innocent factors. It could result by accidentally circling an unintended answer. It could result from coding error. It could result from confusion specific to that particular item. Respondents who are taking the survey seriously could end up with a peculiar item response in the survey record, occasionally. This is why I create a composite score of all item level residuals for each respondent, by averaging the absolute values of all item-level residuals from the "item rest" regressions. I'm interested mainly in respondents who provide incoherent or careless answers across the entire survey.

Respondents with relatively high item level residuals, on average, are respondents who consistently provide answers that appear at odds with one another, as judged by the answer patterns of other respondents. In the following sections, I discuss in greater detail what may drive patterns of unpredictable answers. For now, I treat careless-answers as a measure of noncognitive skills, and I test whether the measure performs as one would expect of a noncognitive measure.

Section 6: Validating "Careless-Answers" by Predicting Education and Income

I have proposed an unconventional but plausible measure of noncognitive skills. Student carelessness on surveys may capture a skill-deficit or trait that is related to academic work ethic. I hypothesize that relationship is negative: the more careless one is on a survey, the worse one will do in school.

The important question is whether careless-answers can be measured, and also whether that measure has worth in social science research. In order to actually validate any measure of noncognitive skills, it is important to submit the measure to two empirical tests. First, does the measure capture information independent of cognitive ability? Second, is it predictive of important outcomes, independent of cognitive ability?

The measure I have proposed must pass a second pair of tests as well, since I have argued that carelessness on surveys can capture new information not captured by self-reported measures of noncognitive skills. Thus I need to demonstrate that survey carelessness captures information that is independent of explicitly measured noncognitive skills, and also that the new measure is predictive of important outcomes, independent of explicitly measured noncognitive skills.

<<Tables 6A and 6B Here>>

Table 6A shows the pairwise correlations between cognitive test scores, Locus of Control, Self Concept and careless answers in NELS:88. The correlations between careless answers and the other variables are weak and negative. The correlation (r=-0.224) with

cognitive ability is negative but relatively weak. This is consistent with previous literature, which has found a moderate relationship between measured noncognitive and cognitive abilities (Almlund et al., 2011). Locus of Control (r=-0.325) and Self Concept (r=0.157) are correlated with cognitive ability as well.

Table 6B shows the pairwise correlations between cognitive test scores, Effort, Control Expectations and careless answers in ELS:02. Again the correlation of careless answers with cognitive ability is negative but weak (-0.201). The correlation of careless answers to the Effort and Control Expectations is virtually nil, and statistically insignificant.

The relatively weak correlation with cognitive ability demonstrates that the careless answers capture something other than cognitive ability. That of course could be random noise. Or it could be a completely unimportant behavioral trait, as far as educational attainment is concerned. Thus I turn to the question of whether careless-answering is predictive of later educational outcomes.

The NELS:88 and ELS:02 are longitudinal surveys. As discussed in the Section 3, all of the cognitive and noncognitive measures discussed thus far were measured during the baseline year, when respondents were in the eighth grade. Educational attainment information is available through the year 2000 for NELS:88 and 2012 for ELS:02.

I estimate the following two period model, to determine whether carelessness on surveys is predictive of educational attainment:

$$S_i = \beta_0 + \boldsymbol{\beta_1 X_i} + \boldsymbol{\beta_2 H_i} + \beta_3 C_i + \boldsymbol{\beta_4 N_i} + \beta_5 \eta_i + \epsilon_i \qquad (4)$$

Where S_i is the years of education completed by individual *i*. $\boldsymbol{X_i}$ is a vector of demographic and geographical control variables: gender, age and Census region. $\boldsymbol{H_i}$ is a vector of individual characteristics that influenced previously accumulated human capital: two-parent household, rage, mother's age at birth, and the highest grade completed by the head of the household. C_i is observed cognitive ability. $\boldsymbol{N_i}$ is a vector of self-reported noncognitive abilities: Locus of Control and Self Concept in NELS:88, Effort and Control Expectations in ELS:02. η_i is the average-answer-unpredictability, which I have otherwise referred to as the careless answering, a noncognitive trait. ϵ_i is a normally distributed error term. Tables 7 and 8 summarize the dependent variable, educational attainment.

<<Table 7 and Table 8 Here>>

Years of Education

Tables 9A and 9B contain the estimates of equation 4, where years of education is the dependent variable. Respectively for NELS:88 and ELS:02, with no cognitive controls, a one standard deviation increase in careless-answering is associated with a 0.179 and 0.154 decrease in the years of education completed, per column 2. The negative relationship is in the predicted direction, since the measures of careless answers theoretically capture a detrimental behavior. When cognitive controls are added, the negative relationship remains significant, although it does attenuate. Carelessness performs as one would expect of a noncognitive measure, that is, as a significant predictor independent of cognitive ability.

Column 5 presents the full model, which contains cognitive ability, self-reported noncognitive skills and unpredictable-answers. The unpredictable-answer measure of noncognitive skills remains negative and statistically significant. In NELS:88 the effect attenuates further when including additional noncognitive skills, whereas in ELS:02 the relationship becomes stronger in the full model. In NELS:88, a one standard deviation increase in unpredictable-answers is predictive of a 0.05 year decrease in the years of education completed; in ELS:02 the effect is a 0.10 year decrease.

The inclusion of the unpredictable-answer variable slightly improves the predictive power of the overall model in Tables 9A and 9B. The R-squared increases when average-absolute-residuals is included, as evidenced by comparisons of column 3 to column 1 and of column 5 to column 4. This provides additional evidence that the carelessness measure contains some truly new and independent information.

Attainment Levels

In the education attainment estimates above, I have treated attainment (years of education) as a continuous variable. However, these estimates may hide a more specific association between carelessness and attainment. Careless-answers may be differentially predictive of attainment at different rungs on the attainment ladder.

<<Tables 10A and 10B Here >>

Tables 10A and 10B examine the impact of careless-answers at four attainment thresholds: HS diploma or higher; some postsecondary education; completion of a

bachelor's degree or higher; and completion of a postgraduate degree. Column 1 contains all baseline participants; each column thereafter is limited to respondents who reached at least the previous level of attainment (e.g. Column 3 estimates the effects on Bachelor's degree completion, conditional on having at least enrolled in college at some point).

So, each column in Table 10 is a separate regression with samples that grow smaller as the attainment threshold goes higher. In each regression, the dependent variable is equal to one if a student reached that attainment level (conditional on reaching the previous level). Estimates are based on a linear regression of a dummy variable on the full set of regressors from equation 4, the educational attainment model. Estimates can be interpreted as probability estimates. Ordinary Least Squares estimates are shown for the sake of simplicity; probit and multinomial logit models provide qualitatively identical estimates.

In NELS:88, unpredictable-answers are associated with attainment levels at the lower end of the attainment distribution. Column 1 shows that a one standard deviation increase in average-answer-residuals is associated with a 1.7 percentage point decrease in the likelihood of earning at least a high school degree; put another way, a one standard deviation increase in average-answer-residuals is associated with a 1.7 percentage increase in the likelihood of dropping out of high school or earning only a GED. However, at higher levels of the attainment distribution, the predictive impact of carelessness dissipates entirely.

Interestingly, across Table 10A, the predictive power of careless-answers is strongest where that of the other noncognitive measures is weakest – at the lower end of the attainment distribution. Conversely in NELS:88, careless-answers loses power when

predicting postsecondary attainment, where the predictive power of self-reported noncognitive skills is strongest.

The pattern of findings is somewhat different in ELS:02, per Table 10B. It is worth noting again an important difference between the baseline populations of NELS:88 and ELS:02. The NELS:88 surveyed eighth graders, and thus was able to fairly accurately observe high school dropout patterns. The ELS:02, however, first surveyed students mid-way through the tenth grade. Many (or by some estimates most) of the students who drop out of high school leave high school within the first two years; such students are therefore not part of the ELS:02, which sampled students still in high school. A very high percentage of the ELS:02 sample also attended at least some college, as compared to NELS:88.

In ELS:02, careless-answers are predictive of attainment at the postsecondary level. Conditional on enrolling in at least some college, a one standard deviation increase in careless answers is associated with a 2.2 percentage point decrease in the likelihood of completing a bachelor's degree, independent of cognitive ability and self-reported noncognitive ability. Furthermore, conditional on receiving a four undergraduate degree, a one standard deviation increase in carless-answers is associated with a 3.0 percentage point decrease in the likelihood of completing a postgraduate degree.

Section 7: Discussion and Conclusion

Students who don't care to complete a survey do a poor job completing the survey. This is not a controversial claim amongst survey researchers. The question is whether careless answer patterns can be identified. In this paper, I have proposed a new method of detecting

careless answer patterns on the self-reported Likert-type scales that are so popular with noncognitive skills researchers. Furthermore, I hypothesize that detecting careless answer patterns may provide useful information about students' noncognitive skills and traits. Perhaps students who put little careful effort into completing a survey also put little careful effort into the paperwork that impacts future success, like homework or financial aid applications.

In order to detect careless answer patterns, I have used common psychometric methods for a new and different purpose. Commonly-known tests such as Cronbach's alpha and item-rest correlations are usually used to judge the consistency and reliability of survey instruments. I have instead used similar tools to identify unpredictable answers from students, examining responses to Likert-type items in the NELS:88 and ELS:02.

When unpredictable answers persist across many items for an individual student, I contend that this is an indicator of student disengagement, and not simply confusion or a lack of comprehension on the survey. Simple pairwise correlations show that careless answering in survey responses is largely independent of cognitive ability. Unpredictability in survey responses is also largely independent of explicitly measured noncognitive skills.

I test whether respondents' careless-answering is associated with later life outcomes. A defining feature of noncognitive skills research is that softer skills and personality traits are predictive of outcomes such as educational attainment. Independent of cognitive ability and traditionally measured noncognitive skills, a one standard deviation increase in careless answering is associated with between a 0.05 and 0.10 year decrease in years of schooling

completed. The sizes of these effects become more meaningful when examining particular attainment thresholds. In the NELS:88, a one standard deviation increase in careless answers is associated with a 1.7 percentage point decrease in the likelihood of completing high school. In ELS:02, a dataset with relatively few high school dropouts but relatively many college enrollees, a one standard deviation increase in careless answers is associated with a 2.2 percent point decrease on completing a bachelor's degree, conditional on having enrolled in college.

These effects are conservative estimates. In every regression model, I include a large number of variables (i.e. mother's year at birth, parental education, household income) that are correlated with noncognitive skills. The findings with respect to educational attainment are also robust to different estimation techniques. Educational attainment models could be estimated using probit, ordered probit or multinomial logit methods. Each of these methods produce attainment level findings that are thematically similar to those presented above.

The use of the word "careless" may make some readers uncomfortable. Throughout this paper, I have used the term carelessness to refer to the behavior of respondents who consistently provide unpredictable answers. This term is, of course, normative and conjectural. The true behaviors, skills or attitudes that underlie answer-unpredictability have yet to be determined. That should be the subject of a future study, one that is able to compare respondent answer patterns to independent information about their noncognitive skills. That said, I do not believe researchers who have conducted low stakes surveys of adolescent students will be upset with the term careless. It is virtually a given in education

research that some students don't put careful effort into the completion of surveys or standardized tests. The open question is whether we can quantify the extent to which students have exhibited low effort.

Surveys are a task. In an attempt to overcome the shortcomings of self reported scales, noncognitive skills researchers are developing behavioral tasks, measuring student engagement. Some tasks are indeed designed to measure student focus and persistence. Such tasks appear promising but will take time to develop and refine. I have argued that a survey is a task that in many ways resembles homework. If a survey is a proxy for a representative homework assignment, we would expect that students who fail to carefully complete it are likely to eventually do worse in school.

Beyond supplying a behavioral measure of noncognitive skills in future surveys, careless-answers also potentially provide information on noncognitive skills within existing surveys that did not initially attempt to measure such skills. I examine answer patterns on Likert-type items. In the ELS:02, Likert-type items are used to measure a wide array of student perceptions and attitudes, not just noncognitive skills.[13] As a robustness check, I created a careless-answer measure that uses only items not designed to measure noncognitive skills; when using a careless-answer measure based on this subset of items, the regression results to the attainment model are virtually identical to those above. That is to say, even if the ELS:02 had contained no items specifically covering noncognitive skills, my method of detecting careless-answers would have provided information about noncognitive skills.

[13] In the NELS:88, the only Likert-type items appears on scales used to measure noncognitive skills.

This paper is designed to advance rather than critique noncognitive skills research. Researchers in personality psychology, character skills and noncognitive skills have changed the conversation around education policy. Remarkable discoveries have been made using self-reported survey results. However, the limitations of self-reported data are real, and advancements in noncognitive skills research will depend heavily on overcoming these challenges. These measurement challenges are particularly acute in education program evaluation, where researchers need a bigger and better toolkit.

I have developed a new, behavioral measure that can add to information gathered through the typical survey process. Careless-answering captures information that other variables do not, which alone makes it important. As a proxy for noncognitive skills, it can be used to re-examine older, existing datasets - many of which have paltry measures of noncognitive skills. Even in rich data sets, it can be used alongside self-reported scores and other potential measures of student engagement (such as item nonresponse). And perhaps most importantly, it is convenient. As long as researchers are collecting survey data using Likert-type scales, they're collecting information on students' carelessness, which they're collecting information on students' noncognitive skills, even if they don't mean to do so.

Acknowledgements

I thank Marty West and Hunter Gelbach for their generous, detailed feedback on this work; Kieran Killeen and participants of the 2015 Annual Meeting of the Association of Education Finance Policy; and Joseph Robinson-Cimpian and participants University of Illinois Department of Educational Psychology QUERIES brown bag seminar.

References

Almlund, Mathilde, Angela Lee Duckworth, James J. Heckman, and Tim D. Kautz. 2011. *Personality Psychology and Economics*. National Bureau of Economic Research. http://www.nber.org/papers/w16822.

Barge, Scott, and Hunter Gehlbach. 2012. "Using the Theory of Satisficing to Evaluate the Quality of Survey Data." *Research in Higher Education* 53 (2): 182–200.http://link.springer.com/article/10.1007/s11162-011-9251-2.

Borghans, Lex, and Trudie Schils. 2015. "The Leaning Tower of Pisa." Working Paper. Accessed February 24. http://www.sole-jole.org/13260.pdf.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92 (6): 1087. http://psycnet.apa.org/psycinfo/2007-07951-009.

Duckworth, Angela L., and David Scott Yeager, 2015. "Measurement Matters Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes." *Educational Researcher* 44, no. 4 (2015): 237-251.

Duckworth, Angela Lee, and Patrick D. Quinn. "Development and validation of the Short Grit Scale (GRIT–S)." *Journal of personality assessment* 91, no. 2 (2009): 166-174.

Egalite, Anna J., Jonathan N. Mills, and Jay P. Greene. 2014. *The Softer Side of Learning: Measuring Students' Non-Cognitive Skills*. EDRE Working Paper. http://www.uaedreform.org/site-der/wp-content/uploads/EDRE-WP-2014-03.pdf.

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program."*American Economic Review*, 145–49.http://www.jstor.org/stable/2677749.

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*, *103*(6), 2052–2086.

Heckman, James J., John Eric Humphries, and Tim Kautz, eds. *The myth of achievement tests: The GED and the role of character in American life*. University of Chicago Press, 2014.

Hitt, Collin, Julie Trivitt, and Albert Cheng. 2016. "When You Say Nothing at All: The Surprisingly Predictive Power of Student Effort on Surveys," forthcoming, *Economics of Education Review*.

Johnson, John A. 2005. "Ascertaining the Validity of Individual Protocols from Web-Based Personality Inventories." *Journal of Research in Personality* 39, no. 1 (2005): 103–29.

Krosnick, Jon A., Sowmya Narayan, and Wendy R. Smith. 1996. "Satisficing in Surveys: Initial Evidence." *New Directions for Evaluation* 1996 (70): 29–44.

Maniaci, Michael R., and Ronald D. Rogge. 2014. "Caring about Carelessness: Participant Inattention and Its Effects on Research." *Journal of Research in Personality* 48 (February 2014): 61–83. doi:10.1016/j.jrp.2013.09.008.

Meade, Adam W., and S. Bartholomew Craig. 2012. "Identifying Careless Responses in Survey Data." *Psychological Methods* 17, no. 3 (2012): 437.

Mischel, Walter, Ebbe B. Ebbesen, and Antonette Raskoff Zeiss, 1972. "Cognitive and attentional mechanisms in delay of gratification." *Journal of personality and social psychology* 21, no. 2 (1972): 204.

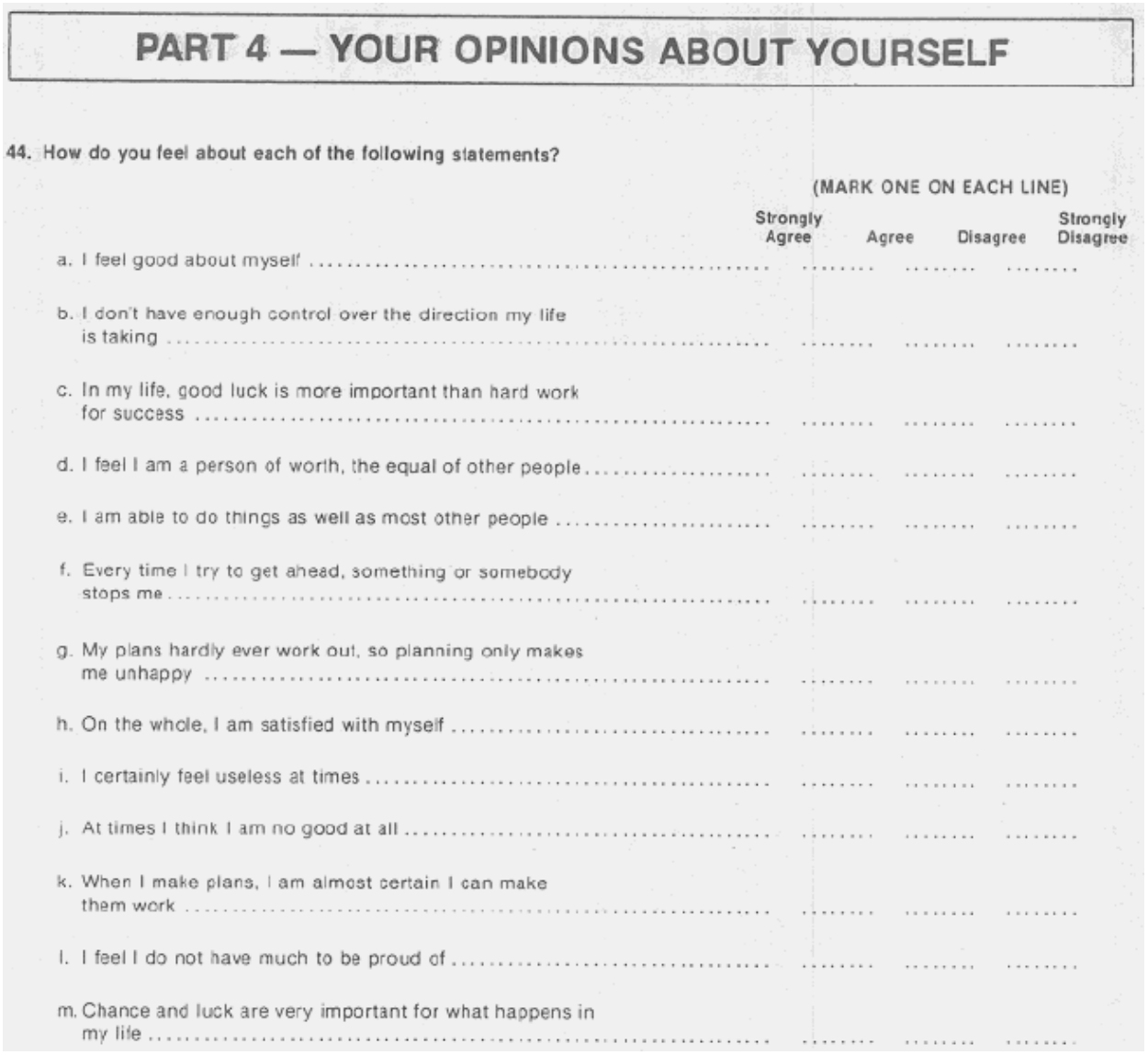
Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80*(1), 1-28.

Vonkova, Hanka, Gema Zamarro, Vera DeBerg, & Collin Hitt, 2015. Comparisons of Student Perceptions of Teacher's Performance in the Classroom: Using Parametric Anchoring Vignette Methods for Improving Comparability. (EDRE WP 2015-01). http://www.uaedreform.org/comparisons-of-student-perceptions-of-teachers-performance-in-the-classroom-using-parametric-anchoring-vignette-methods-for-improving-comparability/

West, Martin R., Matthew A. Kraft, Amy S. Finn, Rebecca Martin, Angela L. Duckworth, Christopher FO Gabrieli, and John DE Gabrieli, 2014. "Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling." In *CESifo Area Conference on Economics of Education Munich: CESifo (September)*. 2014.

Figure 1: The Locus of Control and Self-Concept Scales, from the NELS:88 Baseline Year Student Survey



PART 4 — YOUR OPINIONS ABOUT YOURSELF

44. How do you feel about each of the following statements?

(MARK ONE ON EACH LINE)

	Strongly Agree	Agree	Disagree	Strongly Disagree
a. I feel good about myself				
b. I don't have enough control over the direction my life is taking				
c. In my life, good luck is more important than hard work for success				
d. I feel I am a person of worth, the equal of other people				
e. I am able to do things as well as most other people				
f. Every time I try to get ahead, something or somebody stops me				
g. My plans hardly ever work out, so planning only makes me unhappy				
h. On the whole, I am satisfied with myself				
i. I certainly feel useless at times				
j. At times I think I am no good at all				
k. When I make plans, I am almost certain I can make them work				
l. I feel I do not have much to be proud of				
m. Chance and luck are very important for what happens in my life				

Note: Items B, C, F, G, K and M make up the Locus of Control Scale. Items A, D, E, H, I, J and L make up the Self Concept Scale.

Table 1A: Locus of Control Scale, Item and Composite Score Summary Statistics

Item	N	Mean	Std. Dev.	Min	Max
44B	11,269	3.09	0.80	1	4
44C	11,243	3.29	0.72	1	4
44F	11,248	2.85	0.76	1	4
44G	11,251	3.05	0.78	1	4
44K	11,227	2.98	0.68	1	4
44M	11,254	2.75	0.89	1	4
Composite	11,315	3.00	0.48	1	4

Source: NELS88, Student Baseline Year Questionnaire
Note: Item K is reverse coded.

Table 1B: Self Concept Scale, Item and Composite Score Summary Statistics

Item	N	Mean	Std. Dev.	Min	Max
44A	11,291	3.27	0.61	1	4
44D	11,163	3.32	0.65	1	4
44E	11,213	3.31	0.64	1	4
44H	11,201	3.21	0.68	1	4
44I	11,192	2.54	0.83	1	4
44J	11,199	2.75	0.91	1	4
44L	11,226	3.28	0.78	1	4
Composite	11,320	3.10	0.48	1	4

Source: NELS88, Student Baseline Year Questionnaire
Note: Items A, D, E and H are reversed coded.

Table 2A: Locus of Control Scale, Internal Consistency and Reliability

	(1)	(2)	(3)	(4)	(5)
Item	N	item-test correlation	item-rest correlation	average interitem covariance	alpha
44B	11,269	0.627	0.406	0.153	0.634
44C	11,243	0.596	0.393	0.162	0.639
44F	11,248	0.655	0.459	0.148	0.616
44G	11,251	0.708	0.524	0.135	0.591
44K	11,227	0.499	0.288	0.182	0.669
44M	11,254	0.622	0.369	0.153	0.651
Test scale				0.155	0.676

Note: Item K is reverse coded before calculations conducted.

Table 2B: Self Concept Scale, Internal Consistency and Reliability

	(1)	(2)	(3)	(4)	(5)
Item	N	item-test correlation	item-rest correlation	average interitem covariance	Alpha
44A	11,291	0.676	0.555	0.185	0.744
44D	11,163	0.614	0.470	0.192	0.758
44E	11,213	0.567	0.416	0.199	0.767
44H	11,201	0.690	0.558	0.179	0.742
44I	11,192	0.679	0.506	0.173	0.752
44J	11,199	0.729	0.556	0.159	0.742
44L	11,226	0.655	0.489	0.179	0.755
Test scale				0.181	0.779

Note: Items A, D, E and H are reversed coded before calculations conducted.

Table 3A: "Item-Rest" Regressions, Locus of Control Scale

	(1)	(2)	(3)	(4)	(5)
	Coef.	Std. Err.	T	P>t	Beta
44B	0.664	0.014	47.11	0.00	0.406
Constant	1.112	0.043	26.09	0.00	.
44C	0.565	0.012	45.30	0.00	0.393
Constant	1.628	0.037	43.70	0.00	.
44F	0.710	0.013	54.82	0.00	0.459
Constant	0.695	0.040	17.48	0.00	.
44G	0.856	0.013	65.33	0.00	0.524
Constant	0.495	0.040	12.47	0.00	.
44K	0.376	0.012	31.82	0.00	0.288
Constant	1.849	0.036	51.34	0.00	.
44M	0.676	0.016	42.12	0.00	0.369
Constant	0.682	0.050	13.74	0.00	.

Table 3B: "Item-Rest" Regressions, Self Concept Scale

	(1)	(2)	(3)	(4)	(5)
	Coef.	Std. Err.	T	P>t	Beta
44A	0.680	0.010	70.78	0.00	0.555
Constant	1.178	0.030	39.43	0.00	.
44D	0.605	0.011	56.31	0.00	0.470
Constant	1.465	0.033	43.96	0.00	.
44E	0.518	0.011	48.38	0.00	0.416
Constant	1.722	0.033	51.87	0.00	.
44H	0.772	0.011	71.20	0.00	0.558
Constant	0.829	0.034	24.52	0.00	.
44I	0.876	0.014	62.10	0.00	0.506
Constant	-0.248	0.045	-5.45	0.00	.
44J	1.084	0.015	70.85	0.00	0.556
Constant	-0.667	0.049	-13.68	0.00	.
44L	0.778	0.013	59.33	0.00	0.489
Constant	0.900	0.041	22.12	0.00	.

Table 4A: Absolute Values of Residuals to "Item-Rest" Regressions, Locus of Control Scale, Summary Statistics

Item	N	Mean	Std. Dev.	Min	Max
44B	11,266	0.56	0.47	0.00	2.77
44C	11,242	0.53	0.41	0.02	2.89
44F	11,246	0.51	0.44	0.00	2.53
44G	11,251	0.50	0.43	0.01	2.92
44K	11,226	0.47	0.46	0.02	2.35
44M	11,253	0.66	0.49	0.02	2.39

Table 4B: Absolute Values of Residuals to "Item-Rest" Regressions, Locus of Control Scale, Summary Statistics

Item	N	Mean	Std. Dev.	Min	Max
44A	11,282	0.40	0.31	0.01	2.90
44D	11,163	0.45	0.35	0.02	2.89
44E	11,213	0.47	0.34	0.02	2.79
44H	11,201	0.43	0.37	0.01	2.92
44I	11,191	0.58	0.42	0.02	2.93
44J	11,199	0.62	0.43	0.02	3.22
44L	11,226	0.50	0.46	0.01	3.01

Table 5: Careless Answers, Summary Statistics

Absolute Values of Residuals to Item-Regressions, Averaged Across Scales

	N	Mean	Std. Dev.	Min	Max
NELS:88	11,313	0.51	0.19	0.15	1.96
ELS:02	14,343	0.50	0.14	0.10	1.69

Note: The row "Total" provides the summary statistics for the Careless-Answer variable in Tables 6 through 9.

Table 6A: Correlations between Cognitive and Noncognitive Variables, NELS:88

	Unpredictable -Answers	Cognitive Ability	Locus of Control	Self Concept
Careless-Answers	1			
Cognitive Ability	-0.2239	1		
Locus of Control	-0.2426	0.325	1	
Self Concept	-0.0904	0.1567	0.5357	1

Note: All correlations are significant at $p<0.001$

Table 6B: Correlations between Cognitive and Noncognitive Variables, ELS:02

	Unpredictable-Answers	Cognitive Ability	Effort	Control Expectations
Careless-Answers	1			
Cognitive Ability	-0.2006	1		
Effort	0.001	0.2241	1	
Control Expectations	0.017	0.3218	0.7239	1

Note: The correlation of Unpredictable-Answers to Effort and Control-Expectations are not significant. All other correlations are significant at $p<0.001$

Table 7: Summary Statistics for Years of Education

	Mean	SD	Minimum	Maximum	Outcome Year
NELS:88	14.24	1.85	10	20	2000
ELS:02	14.61	1.96	11	20	2012

Note: In NELS:88 and ELS:02, years of education were imputed based on reports of highest degree completed. Dropouts were coded as 10 in NELS:88 and 11 in ELS:02, where baseline students were in the 8th grade and 10th grade, respectively. GED recipients and HS graduates were coded as 12, two-year college graduates as 14, four-year college graduates as 16, master's degree holders as 18, and higher graduate degree holders as 20.

Table 8: Summary Statistics for Highest Educational Attainment Level

	Less than High School	GED	High School Diploma	Some Postsecondary Education	Bachelor's Degree	Postgraduate Degree
NELS:88	6.45	3.35	12.61	44.82	29.12	3.65
ELS:02	1.69	1.11	6.95	47.07	32.82	10.35

Note: All numbers are percentages.

Table 9A: OLS Estimates for Years of Education, NELS:88

	(1)	(2)	(3)	(4)	(5)
Cognitive Ability	0.618***		0.600***	0.557***	0.550***
	0.033		0.033	0.030	0.030
Careless-Answers		-0.179***	-0.086***		-0.05**
		0.023	0.023		0.026
Locus of Control				0.153***	0.144***
				0.042	0.044
Self Concept				0.094***	0.095***
				0.026	0.026
N	10,015	10,208	9,991	9,992	9,990
R^2	0.3848	0.3207	0.3864	0.3961	0.3967

Note: All control variables standardized at mean zero, σ of one. *** = p<0.01; ** = p<0.05; * = p<0.10

Table 9B: OLS Estimates for Years of Education, ELS:02

	(1)	(2)	(3)	(4)	(5)
Cognitive Ability	0.678***		0.673***	0.599***	0.586***
	0.024		0.024	0.030	0.031
Careless-Answers		-0.154***	-0.080***		-0.101***
		0.022	0.020		0.026
Effort				0.170***	0.166***
				0.034	0.034
Control Expectations				0.116***	0.125***
				0.035	0.036
Observations	12,125	11,729	11,729	9,801	9,801
R^2	0.2887	0.2025	0.2931	0.2946	0.2968

Note: All control variables standardized at mean zero, σ of one. *** = p<0.01; ** = p<0.05; * = p<0.10

Table 10A: OLS Estimates by Attainment Level, NELS:88

	(1)	(2)	(3)	(4)
	HS Diploma or Higher	Some Postsecondary	Bachelor's Degree or Higher	Postgraduate Degree
Cognitive Ability	0.041***	0.048***	0.136***	0.041***
	0.005	0.008	0.009	0.008
Careless-Answers	-0.017***	-0.001	-0.009	-0.001
	0.005	0.005	0.007	0.008
Locus of Control	0.012*	0.023***	0.014*	0.006
	0.006	0.007	0.008	0.009
Self Concept	0.010*	0.010	0.014*	0.006
	0.006	0.006	0.008	0.008
N	9,987	9,424	8,291	4,501
R^2	0.2221	0.1314	0.2617	0.0379

Note: All control variables standardized at mean zero, σ of one. *** = p<0.01; ** = p<0.05; * = p<0.10

Table 10B: OLS Estimates by Attainment Level, ELS:02

	(1)	(2)	(3)	(4)
	HS Diploma or Higher	Some Postsecondary	Bachelor's Degree or Higher	Postgraduate Degree
Cognitive Ability	0.028***	0.054***	0.142***	0.054***
	0.004	0.006	0.009	0.012
Careless-Answers	-0.001	-0.004	-0.023**	-0.030**
	0.003	0.005	0.009	0.012
Control Expectations	-0.002	0.013*	0.037***	0.039***
	0.005	0.007	0.010	0.014
Effort	0.012***	0.008	0.034***	0.005
	0.004	0.006	0.009	0.014
N	9,801	9,601	9,104	5,740
R^2	0.0625	0.1003	0.2359	0.0603

Note: All control variables standardized at mean zero, σ of one. *** = p<0.01; ** = p<0.05; * = p<0.10

Chapter Three

When Students Don't Care:

Reexamining International Differences in Achievement and Noncognitive Skills, Using Novel Measures of Student Effort on Surveys and Tests

Gema Zamarro, University of Arkansas

Collin Hitt, University of Arkansas

Idelfonso Mendez, University of Murcia

Abstract

Policy debates in education are often framed by using international test scores, such as the Programme for International Student Assessment (PISA). The obvious presumption is that observed differences in test scores within and across countries reflect differences in cognitive skills and general content knowledge, the things which achievement tests are designed to measure. We challenge this presumption, by demonstrating that a substantial amount of the within-country and between-country variation in PISA test scores is associated with student effort on the tests, rather than true academic content knowledge. Drawing heavily on recent literature, we posit that our measures of effort are actually proxy measures of noncognitive abilities such as conscientiousness and self-control.

Measures of student effort yield information that is much more relevant than just whether a student was paying attention during some low-stakes test. Students may actually reveal something about their conscientiousness and self-control in the amount of effort they show on tests and surveys. Our previous work, and that of others validates this claim (e.g.

Borghans and Schils, 2012; Hitt, Trivitt and Cheng, 2016; Hitt, 2016). This study pilots and refines several such behavioral measures of student effort, studying student answer patterns on tests and surveys. For example, we examine the frequency with which students skip questions on surveys and tests, give careless answers, and show diminishing effort over the course of the test.

Our results show that measures of test and survey effort help explain between 33 and 40 percent of the observed variation in test scores across countries, while explaining only a minor share of the observed variation within countries.

"U.S. 15-year-olds made no progress on recent international achievement exams and fell further in the rankings, reviving a debate about America's ability to compete in a global economy."

- ***The Wall Street Journal,*** December 3, 2012

"Finland's schools owe their newfound fame primarily to one study: the PISA survey, conducted every three years by the Organization for Economic Co-operation and Development (OECD)."

- ***The Atlantic Monthly,*** December 29, 2011

1. Introduction

Since their introduction, large scale international assessments have been used to make sweeping statements about the quality of countries' schools and the cognitive skills of their students. The Programme for International Student Assessment (PISA), the Third International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRSL), have become important sources, and for some countries the only sources, of information on student performance in key subjects such as math, reading and science.

The tests ostensibly measure student content knowledge or, more generally, cognitive ability. However, in reality, student performance is driven by more than just cognitive ability and content knowledge. Some students put forward less effort than others during exams. This is a commonsense observation. Test scores cannot tell us about the math, reading or science ability of students who don't pay attention or put forward effort while taking low-stakes tests. So, observed differences in test scores within and across countries might reflect differences with student effort on the tests, rather than just true academic content knowledge.

In this paper, we contend that it is possible to measure student effort on tests. Doing so allows us to adjust observed test scores for estimated student effort: students who show low effort on tests probably possess greater math and reading ability than their test scores indicate. We examine student effort and test scores on the 2009 wave of PISA.

Survey and test data available from PISA allow us to build a number of possible measures of student effort. For example, the random ordering of questions in different test booklets and the random assignment of booklets to students in PISA is a key feature of this data that we exploit in order to estimate measures of student effort on each item, as affected by its order.

Our motivation for conducting this study goes beyond the simple question of whether students try hard on tests, however. We are interested in within-country and cross-country differences in noncognitive skills. Previous research (see e.g. Borghans and Schils, 2012; Hitt, Trivitt and Cheng, 2016; Hitt, 2016) has shown that measures of student effort, based on students' response patterns to surveys and tests, are predictive of important life outcomes, independent of measured cognitive ability. Drawing upon this literature, we argue that our measures of student effort derived from PISA can be understood as meaningful measures of student's noncognitive skills. Students plausibly tell us something about their character - their conscientiousness, self-control, or persistence - through their effort on tests and surveys. Perhaps they tell us about how they approach the routines of schooling. Tests and surveys are tasks that resemble everyday schoolwork. By measuring effort on those

tasks, we not only correct PISA test scores for student effort, we believe we also identify potential indicators of international differences in noncognitive skills.

In particular, this paper aims to respond to the following four research questions.

1. Can student effort on tests and surveys be measured?
2. Does effort on tests and surveys vary across countries?
3. Does varying student effort on tests impact our understanding of cross-country differences in test scores?

The rest of the paper goes as follows: Section 2 presents a simple conceptual framework for understanding the role that effort plays on the assessment of content knowledge and cognitive skills. Section 3 describes the PISA study and the data used in this paper. Section 4 explains our proposed measures of student's effort in tests and surveys and describes our approach for obtaining corrected measures of PISA test performance. Section 5 describes our results. Section 6 discusses our findings, reviews previous literature that provides persuasive evidence that student effort on surveys and tests is driven partly by noncognitive ability, and concludes by highlighting implications and limitations of our findings.

2. Theoretical Framework

In order to understand the role that effort plays in the testing and survey process, it is useful to think about the assessment process. Below we briefly outline some of the very basic elements of standardized tests and student surveys.

What is a Standardized Test?

A standardized test is an instrument designed to measure student content knowledge with respect to specific subjects. The PISA 2009, for example, contained questions on reading, math and science. The test is largely multiple choice. Items vary in difficulty. The tests are often long: the average PISA 2009 booklet contained approximately 60 questions, and was expected to take two hours. And importantly, the tests are low-stakes: student scores on PISA are anonymous and have no effect on the students themselves.

What is a Survey?

A survey is an instrument designed to gather information and opinions from students. It should be easily readable: surveys are typically constructed to be readable even to students whose reading ability is 3 to 5 years below the grade level. Surveys are often long. The PISA 2009 Student Survey contains approximately 170 items, almost all of which are multiple choice. The surveys are confidential and low stakes: the answers that students give have no effect on the students themselves.

How does effort relate to Test Scores?

Figure 1 presents a simple theoretical framework. A well-designed test should measure student cognitive ability or content knowledge. Realistically, we cannot observe actual cognitive ability or true content knowledge, especially not for students participating in PISA. What we observe is how well students perform on a test. Therefore, in order to conclude that test scores accurately measure true cognitive ability or content knowledge, one must assume that nothing moderates (or interferes with) the relationship between true ability and the performance on the test. We know this assumption is untrue.

<<Figure 1 Here>>

Student effort is a moderator between cognitive ability and test scores. Some students, simply put, don't try very hard on tests. This leads to an underestimate of those students' cognitive abilities. Previous literature has modeled student effort as a product of incentives (e.g. Kautz, 2015). And indeed incentives have been shown to alter student performance on tests. But on PISA, as on most standardized tests, the explicit incentives are the same for all students; the 2009 PISA is a low stakes test. Therefore, if student effort differs across students taking PISA tests, it does so for a reason other than individual incentives.

This is not to say that students in all parts of the world view PISA tests identically. Some national or regional educational authorities attach great importance to their students' performance on PISA. In Spain, for instance, PISA tests are the sole measure of educational achievement in Spanish states. PISA tests serve a similar role to the National Assessment of Educational Progress in the United States. One might expect that, in Spain, the competition between Spanish states may lead regional authorities to prepare specifically for PISA tests. This instance is likely an exception to the rule.

We argue that effort on PISA, and on standardized tests more generally, is driven by student noncognitive skills. Skills such as conscientiousness, persistence and self-control are, practically by definition, needed to complete long, mundane, low-stakes tasks.

In Figure 1, we show a simple conceptual model where student effort is driven by such noncognitive skills. Effort is a mediator of the relationship between noncognitive skills

and cognitive test scores. Put another way, noncognitive skills impact test scores because students who possess high levels of self-control or conscientiousness try harder on low-stakes tests.

In making the case that noncognitive skills drive student effort on tests, we rely largely on recent literature, which we describe in Section 6 after presenting our empirical results. We cannot actually test this case, however, relying solely on data available from PISA. Just as with cognitive ability, we cannot actually observe the true noncognitive skills of students participating in PISA. But we *can observe* test-taking and survey-taking behaviors of students in PISA datasets, and we can point to recent literature that shows that test-taking and survey-taking effort are linked to later life outcomes, independent of test scores. This literature suggests strongly that student effort during assessments is not some behavior that is idiosyncratic to tests and surveys. Instead it suggests that student effort is indicative of other noncognitive traits that have a broader and long-lasting impact.

Again, strictly speaking, our analysis of PISA data is limited to the relationship between our measures of effort and student test performance. So in this narrow sense, we hope to eventually correct PISA scores for student effort, allowing for more valid comparisons of student content knowledge and cognitive ability. But more generally, we believe that our research also produces information that will allow researchers to compare student noncognitive skills across and within countries, using effort on PISA as a proxy measure.

3. Data Source

Our data are from publicly available data sets published online by the Organization for Economic Co-operation and Development (OECD), the sponsoring agency of PISA. In particular, we focus on data from 2009. As we have mentioned, our study builds partly on research by Borghans and Schils (2012) that examines earlier waves of PISA We focus on 2009 and not the more recent 2012 wave of PISA due to the fact that, as of 2012, PISA no longer published detailed information on question ordering within each test booklet.

In 2009, seventy-four countries and regional "economies" participated in PISA. In total, tests were administered to 515,958 students, comprising representative samples within their home countries. The 2009 PISA test was a standardized test of math, reading and science ability.[14]

Each student took a test of approximately 60 items. Within each country, each participating student was randomly assigned one of several test booklets. Each booklet was comprised of test items drawn from an item bank of several hundred entries. Through 2006, all countries participating in PISA were issued the same set of thirteen booklets. That changed in 2009, according to the PISA 2009 Technical Manual:

[14] Scores were calculated separately for each content area, using an IRT framework that produces five "plausible values" for each student's abilities. Put plainly, each plausible value takes into account that a student's test score misestimates the student's true ability. Students of differing ability can receive the same raw score. Or, put differently, a student of a given ability could randomly receive any number of scores within a given range, due to error. Therefore, plausible values are, "random numbers drawn from the distribution of scores that could be reasonably assigned to each individual – that is, the marginal posterior distribution," according to the PISA 2009 Technical Manual. We use the variables PV1MATH, PV1READ and PV1SCIE in this analysis, which we still consider exploratory.

> "*In PISA 2009 some countries were offered the option of administering an easier set of booklets. The offer was made to countries that had achieved a mean scale score in reading of 450 or less in PISA 2006, and to new countries that were expected – judging by their results on the PISA 2009 field trial conducted in 2008 – to gain a mean result at a similar level. The purpose of this strategy was to obtain better descriptive information about what students at the lower end of the ability spectrum know, understand and can do as readers. A further reason for including easier items was to make the experience of the test more satisfying for individual students with very low levels of reading proficiency*."

Our current analysis is limited to the 44 countries who took the standard, harder set of booklets. Within those countries, we also exclude a relatively small group of students who received a booklet specially developed for schools that primarily serve students with disabilities. Our total sample is 311,484 students.

The PISA testing session lasted two hours.[15] Students were given an accompanying survey about learning environment, home factors, and student attitudes. The surveys were administered immediately after the completion of the test, and we expected to take one hour to complete.

PISA provides each student's full test and survey record. This includes details of each student's response to each question. We use patterns of student item-level responses to build a number of plausible measures of student effort, which we discuss in the following section.

4. Measuring Student Effort

On both the survey and the test itself, students actually provide indicators of their overall effort during the assessment process. We explore three potential measures of student effort

[15] A one-hour test developed for schools serving special needs students. Students taking the one-hour test are excluded from our analysis.

in PISA. The first of the measures is derived from student answer patterns on the test form, the other two are derived from answer patterns on the subsequent survey form. On the test form, we explore the rate of performance decline over the course of the test. On the survey form, which as mentioned above was administered immediately after the test, we explore: item nonresponse rates and a measure of careless answer patterns.

We will now describe each measure of effort in greater detail. Descriptive statistics for each measure can be found in Table 1.

<<Table 1 Here>>

PISA Test: Declining Effort

Across PISA tests, performance has been found to decline on average as students move from the beginning to end of the test (e.g. Borghans and Schils, 2012). Figure 2 presents the average performance on each question of the test as the test progresses, for a selected group of countries, using data from PISA 2009. Performance declines as the test goes on. As can also be seen in this figure, the rate of performance-decline over the course of the test varies across countries. Even in countries with relatively high PISA test scores, such as Korea and Finland, a decline is observed. For some countries the decline in performance can be dramatic. That is the case of Greece as it is observed in Figure 2. This doesn't have anything to do with the content of the final items on the test. Because question order is randomized across students as part of the PISA test, this suggests strongly that the observed decline is a matter of "test motivation," rather than a difference in question difficulty at the beginning versus the end of the test.

<<Figure 2 Here>>

In 2009, the order and assortment of test questions was randomized across PISA test booklets. Test booklets are then randomly assigned to students. So, across students, a given item varies in its position on the test. Some students begin with difficult questions, some with easier questions. The independence of question difficulty and question ordering allows us to calculate the effect that "order" has on the probability that a student answers a question correctly. Students who show no decline in motivation should have an equal probability of answering a given question correctly regardless of whether it appears at the beginning or the end of the test.

Our measure of test effort expands on the work by Borghans and Schils (2012), who examined data from the 2006 wave of PISA. Within each country, they examine the relationship between question position and the probability that it is answered correctly. Their approach generates country-level estimates of the decline in performance over the course of the test. The effects of order vary by country, suggesting motivation varies by country. They found that cross-country differences in motivation explained 19 percent of the variance in PISA scores between countries.

We seek to advance beyond the methods used by Borghans and Schils (2012), whose analysis of PISA identified country-level estimates of the relationship between question order and student performance. In particular, we explore a variety of approaches to produce student-level estimates of decline in performance over the course of the test.

One approach would simply be to subtract performance on the end of the test from performance at the beginning of the test. Table 1 shows the average number of items correct on the first ten and last ten items of the test. Across the sample, average performance declines from 5.85 items correct on the first ten items to 4.46 items correct on the final ten items. That said, for some students the decline in performance may reflect the fact that their booklets randomly contained relatively difficult items toward the end of the exam. Indeed, ANOVA estimates find that 16.3 percent of the variation in decline from the first ten to last ten items is explained by booklet number.

In order to account for the effects of booklet on decline in student performance, we simply estimate decline in performance, regression-adjusted for booklet number. Per Table 1, the adjusted rate of decline in performance from the first ten to the last ten items is 1.37 points.

This is a somewhat simplistic approach. By averaging performance over the first ten and last ten items, and then comparing those averages to one another, we are assuming that the rate of decline is fairly steady across the course of the test. However, if in fact student performance actually drops within the first ten items and remains steady thereafter, this measure will fail to fully capture decline in performance. The plots shown in Figure 2 do not point to such a pattern. The rate of decline in performance on average appears to take place steadily over the course of the test.

The approach described above is our primary means of quantifying decline in performance on tests, used throughout the following sections of the paper. However, below,

we also propose a more sophisticated approach to generate student-level estimates of performance decline. This approach requires a large amount of computing power. Given computing constraints, the only feasible way to estimate the following model is by repeatedly running the analysis on subsamples of students. We are currently exploring two sampling strategies.

The first, which we would use in order to actually produce internationally comparable estimates of effort for every student, is to repeatedly take random samples of 10,000 students from across the global sample, with replacement, and to repeat the analytical procedure until we have developed estimates for all students. We are currently in the process of developing this sampling strategy.

Another strategy, given the computational challenges, is to estimate the following model within each country. As explained below, this method provides valid country-level estimates of effort, but the student-level estimates are only valid for making within-country comparisons (e.g. Mendez et al. 2015). This is the approach we've pursued thus far.

In our more sophisticated approach, we use a linear random coefficient model as the base of our estimates of test effort as follows:

$$y_{ij} = \alpha_0 + \alpha_0^i + \beta_1 O_{ij} + \beta_1^i O_{ij} + \gamma_j + \varepsilon_{ij} \quad (2)$$

Where the dependent variable y_{ij} takes value 1 if the answer of student *i* to question *j* was correct, value 0.5 if they got half credit for that question, and 0 if the answer was wrong. The independent variable of interest O_{ij} is the sequence order of the test question, rescaled such

as the first question is numbered as 0 and the last question as 1. The constant α_0 then represents the average performance of students in the very first question on the test. By introducing a random intercept in the model (α_0^i) we allow for different students to deviate from the average performance in the first question. The intercept coefficient β_1 presents the average decline in test performance from the first to the last question of the test. By introducing a random slope in the model (β_1^i) we allow for different students to deviate from the average decline in performance. The introduction of this random intercept (α_0^i) and random slope component (β_1^i) has the advantage of better taking into account the structure of the data and allow for estimations of how individual students differ from the average observed pattern. This is the main difference between our model specification and that of Borghans and Schils (2012) who excluded this components and estimated just an average constant and slope. γ_j are question fixed effects to control for the difficulty level or nature of each question (e.g. Multiple choice or open question).

The model presented in (2) is then estimated for each country separately using Maximum Likelihood methods allowing for the random constant (α_0^i) and random intercept (β_1^i) components to be correlated. This process provides us with estimates of the country average performance in the first question (α_0), country average decline in test performance (β_1) and estimated question dummies effects. The model does not directly estimate the random effects but obtains estimates of their standard deviation and their estimated covariance. With this information, however, one could obtain best linear unbiased predictors of the random effects (α_0^i) and (β_1^i) to recover the individual performance of a student in the

first question ($\alpha_0+\alpha_0^i$) and decline in test performance ($\beta_1+\beta_1^i$) as compared to a reference group of students.[16]

The second estimate of declining performance is based on the random coefficient models presented in (2). As stated (see footnote 2), this method yields county-level estimates of dimished likelihood of answering the questions correctly, as students move from the beginning to the end of the test.. The mean of β_1 in is 0.1196 (which can be found in Table 4, discussed in greater detail later in the text). That is, at the country level, the average estimated effect of moving an item from being the very first question on the test to being the very last item on test would be a 11.96 percentage point decline in the probability of answering the item correctly. This estimate is similar to our variable "decline," described above, which shows a 13.7 percentage point decrease in the percent of items answered correctly from the first ten to the last ten items on the test.

Student Survey: Item Nonresponse

The item nonresponse rate on a survey is the rate at which students skip questions, or answer "I don't know." For decades survey methods researchers have seen survey item nonresponse as a measure of disengagement in the survey process, presuming of course the survey is well -designed.

[16] As of this writing, we are still working to obtain these student-specific measures using best linear unbiased predictors.

In PISA surveys, "I don't know" is virtually never offered as an answer choice. Survey item nonresponse rates are then measured as the rate at which students skip questions. Per Table 1, the average survey item nonresponse rate is 3 percent. The standard deviation of survey item nonresponse is 5 percent within country and 1 percent between countries.

Boe, May and Boruch (2002) examined this question more than a decade ago, in an unpublished working paper. They examine item response rates on a student survey given as part of the TIMSS. Item response rates are used to form a measure of; the authors call it, "student task performance." More than 50 percent of the cross-country variation in test scores was explained by survey item nonresponse. Our analysis of PISA 2009 follows an approach similar to Boe, May and Boruch (2002) and Hitt, Trivitt and Cheng (2016).

Student Survey: Careless Answer Patterns

The measure of student effort on surveys that we outline above - item nonresponse rate - is simple to calculate. We count the frequency with which students skip questions. Identifying careless answer patterns is more complicated. Commonsense intuition tells us that some students don't skip questions at all, but instead just fill in the "bubbles". We attempt to identify this type of behavior. We term "careless" answers as a series of answers on the student survey that appear inconsistent with one another.

We use a novel method developed by Hitt (2016), in order to distinguish between legitimate answers and answers that appear to have been entered carelessly. We exploit the fact that a large number of items on the PISA Student Survey are part of larger multi-item

scales that use a Likert-type response format. For example, as part of a scale to assess "attitude toward school" students are asked the extent to which they agree with a number of statements. The first item is, "School has done little to prepare me for adult life when I leave school." A subsequent item is, "School has taught me things which could be useful in a job." A priori, one would think, students who agree with the first statement should be unlikely to agree with the later statement.

When inquiring about the "attitude toward school" or some other concept, survey administrators ask multiple, similar questions for a simple reason. Asking multiple, simple questions about a related concept yields more reliable information than asking only a single question. In a well-constructed scale, answers to each of the questions should be reasonably well correlated with one another. If they weren't, one could hardly argue that the questions were actually measuring the same concept. Standard psychometric tests such as Cronbach's alpha and item-rest correlations are used to report whether items within a scale are in fact correlated.

In a scale deemed consistent and reliable, in psychometric terms, item-answers within a given scale are correlated with one another. That is to say, answers to any given item ought to be predicted reasonably well by answers to the other items on the scale. We examine the frequency with which students give answers that appear inconsistent, or more specifically, unpredictable, given their answers on the other related questions that are part of the scale.

Following Hitt (2016), we conduct a separate bivariate regression for every Likert-type item on the PISA Student Survey. In total, we examine 84 items across 12 scales. Every

item is regressed on the average of answers given to the remaining items on the same scale. For example, student responses to the first item of the "attitude towards school" scale are regressed on average score of the remaining items of that same scale.

Consider the following "item-rest" bivariate regression equation, adapted from Hitt (2016):

$$Y_{jst} = B_0 + B_1 X_{jst} + \eta_{jst} \quad (1)$$

Where Y_{jst} is the answer to item *j* within scale *s* provided by student *t*. The coefficient of interest is X_{jst}, the average of the rest of the items (all items not *j)* within the same scale (*s),* by student *t*. B_0 is a constant, which drops out when the regression is standardized, and η_{jst} is the error term. These bivariate regressions are mathematically equivalent to the item-rest correlations used in psychometric evaluations of scales (Hitt 2016).[17]

We store the estimated student-level residuals η_{jst} to each regression. Each residual literally measures the extent to which a given student gave an unpredictable answer, as judged by the regression model (which is based on the answer patterns of all students) and that student's answers to other items on the scale.

We then standardize the absolute value of each residual, with a mean of zero and a standard deviation of one. The average of these standardized scores is combined into a composite "careless answer" score. Displayed in Table 1, the unit of change for careless answer score does not have a conversational interpretation. A lower score signifies that on

[17] When the regression is standardized, the coefficient B_1 becomes mathematically identical to the coefficient of a Pearson product-moment correlation.

average a student's individual answers were well predicted by their other answers. A higher score signifies that the student consistently gave answers that did not appear consistent. The mean careless score is zero, with a standard deviation of 0.24 within country and 0.07 between countries.

5. Results

5.1 Can student effort be measured?

We have laid out a number of plausible measures of student effort on tests and surveys. We now examine the extent to which they are related to student test scores, and the extent to which these variables are related to one another.

Table 2A displays pairwise student-level correlations between PISA 2009 math, reading and science scores and our measures of student effort. All correlation coefficients shown are statistically significant ($p<0.01$). Each of our effort variables are constructed such that a higher value signifies lower effort (i.e. higher detrimental behavior). All correlations are negative, as expected.

Of the survey-based effort measures, item nonresponse and careless-answer patterns are all negatively related to test scores. On the PISA math score, the correlations are -0.27 and -0.08, respectively. The magnitudes on reading and science tests are similar, an interesting fact we will discuss momentarily.

The rate of performance decline is also negatively related to total score, an unsurprising fact. The correlation coefficient is -0.09.

<<Tables 2A and 2B Here>>

The pattern of results is noticeably similar across test subjects. One critique of our measures of student effort is that cognitive ability could be the real driver of student engagement on tests. If this was true, one would expect that reading ability above all else would be a driver of nonresponse and careless answers. Students who cannot read at all cannot read surveys and tests. And yet the correlations of our effort variables are hardly higher with reading than with other topics. This finding indicates that student effort impacts each test score similarly - something that would be true if our measures captured student effort, and likely would not be true if our measures were driven by reading limitations.

The correlations between our effort measures are also interesting. Neither survey item nonresponse nor careless-answering is strongly correlated with decline-in-performance on the test. Again, the variable "decline" is based on comparisons of performance at the beginning of the test versus the end of the test. The survey is administered immediately after the test. One might argue that cognitive fatigue causes students to decline in performance after the test. If this was the case, then one would expect that fatigue to impact student effort on the subsequent hour-long survey - and therefore one would expect "decline" to be correlated with survey item nonresponse and careless answering. Yet the results tell a different story.

Decline in performance on the test is very weakly related to survey item nonresponse and careless answers. This is consistent with the notion that survey effort in fact signals a lack of effort throughout the entire assessment process. "Decline" captures diminished effort

over the course of the test. Some students, however, never display much effort in the first place - their performance starts off low, stays low, and they show little effort on the survey. A measure that identifies such students would not be correlated with decline, but would be correlated with overall test score. That is the case for our survey-based measures of effort.

These results again are consistent with the fact that students who skip questions or give careless answers do so from the very beginning of the test onward. We mentioned this possibility above.

Careless answer patterns are not strongly correlated with survey item nonresponse. This again is unsurprising. Within a given question, giving a careless answer and not responding at all are mutually exclusive options. Over the course of the assessment, it's possible that different students take different approaches. Some just skip questions frequently, while others complete every question but do so with little care - few switch back and forth between skipping items and answering carelessly.

While the student-level correlations between effort measures are weak, the correlations at the country level are much stronger. Table 2B displays correlations of test scores and effort measures at the country level. While students who decline in performance are not the same students who skip items or who give careless answers, such students are concentrated together within countries. We delve further into the country-level concentrations of student effort in the following section.

<<Tables 3A, 3B and 3C Here>>

Tables 3A, 3B and 3C are regression estimates, where student-level PISA test score is regressed on each of our effort measures. All results are standardized, and significant, at $p<0.01$. The first three columns are standardized bivariate regressions, with a single regressor. The coefficients across the first seven columns are identical to the corresponding correlation coefficients in Table 2A.

Of primary interest in Tables 3A, 3B and 3C, is the estimated R-squared. No individual measure of effort explains more than a minor share of the individual variation in PISA test scores. Nevertheless, when used in combination, our measures of effort explain a substantially greater share of the overall variance than any standalone variable.

The eighth column of Tables 3A, 3B and 3C contains all measures of student effort in a single regression. For math, reading and science scores 9.8, 11.1 and 10.5 percent of the respective student-level variation in PISA scores is explained by our measures of effort.

Our first research question asked whether student effort on tests could be quantified. We have explored a number of quantifiable measures. We will now explore whether these indicators of effort affect our understanding of international comparisons of PISA performance.

5.2 International Comparisons of Student Effort Measures

We now turn to our second research question: Does effort on tests and surveys vary across countries? As shown in the descriptive statistics of the previous section, the variance between countries is smaller than the variance within countries. Nonetheless, there is a measurable between-country difference in each effort measure. To test the significance of

the between-country variance, we conducted a one-way ANOVA of each effort measure, with country as the independent grouping variable. The model F-statistic is statistically significant in every case, with between 1 and 7 percent of the overall variation explained by country dummy variables.

<<Tables 4A and 4B Here>>

<<Figure 3 Here>>

We'll now focus at some length on decline in performance over the course of the test. Tables 4A and 4B present the results of the random coefficient model described in (2) and that we used to obtain measures of decline in test performance. Figure 3 displays the estimates of these regressions for a selected group of countries.

As can be seen in this table and figure, and as it was anticipated in the descriptive averages presented in Figure 2, we observe a considerable amount of heterogeneity across countries not only in initial performance in the test but also on our country average estimate of the rate of decline in performance as the test progresses. Some high performing countries like South Korea start at a high performance level and remain at a higher level as the test progresses. Other high performing countries like Finland do not start at especially high levels in the response of the very first question of the test but present low rates of decline as the test progresses, which makes them end up at a very good final position in performance by the end.

Interestingly, countries like Spain or Greece have an average performance on the first question of the test that is above average and above that of the high performing country of Finland. However, their higher rate of decline in performance as the test progresses quickly drags down their cumulative scores. This is especially dramatic for the case of Greece, the country in our sample that presented the highest estimated rate of decline. It is important to stress here that this was also the country that presented the highest rate of decline in performance in PISA 2006 according to the estimates presented in Borghans and Schils (2012). This is reassuring as it suggests that our estimates of the country-level average rates of decline on test performance are capturing permanent country-specific noncognitive skills and are not the result of just one specific year of the PISA study.

Finally, the last column of Table 4A presents the estimated correlation between the individual specific random intercept and random slope components of the model. It is interesting to observe that, although overall the correlation seems to be small if we obtain the average for all countries together, these estimated correlations vary substantially across countries. This indicates that in countries with lower estimated correlations (e.g. Korea, Japan, Sweden or U.S) both high performing and low performing students present rates of decline in test performance that are similar. In other countries we observe bigger positive correlations (see e.g. Spain, Greece, Singapore) indicating that lower performing students present much higher rates of decline in test performance than higher performing students.

5.3 International Comparisons of Student's Performance Accounting for Effort

We now turn to our third research question: to what extent does the international variation in student effort on PISA tests explain international variation in PISA test scores? To answer this question we conduct a random-effect multilevel analysis of our PISA data. Our empirical model is as follows:

$$y_{it} = \beta_1 \boldsymbol{X}_i + (\alpha_{ct} + \varepsilon_{it})$$

where y_{it} is the PISA score for student *i* on test *t*, $\boldsymbol{X}_i$ is an array of measures of effort, α_{ct} is a country level random effect and ε_{it} is a normally distributed error term. This allows us to estimate the relationship between effort and test scores: across the overall sample, within country, and across countries. Tables 5A, 5B and 5C display the estimates of the within-country, between-country and overall variance in PISA scores explained by our measures of effort.

<<Tables 5A, 5B and 5C Here>>

Within the top three rows are estimates, by column, of the proportion of the variance of PISA scores in a given subject area explained by each measure of effort. The overall R-squared in each model corresponds with the R-squared numbers in Tables 3A, 3B and 3C. As discussed, for every variable (other than test item nonresponse), the overall R-squared is modest. The between-country estimates of our multi-level model tell a very different story.

Altogether, our measures of effort explain a substantial portion of the variation in between-country test scores. The first column displays results for decline in performance: whereas only 0.8 percent of the variation in student-level test scores is explained by decline in performance, 29.6 percent of the country-level variation in math test scores is explained by decline-in-performance. Decline-in-performance explains 23.0 percent of the international variation in reading scores, and 32.3 percent of international variation in science scores. These estimates are slightly larger than the 19 percent estimate of Borghans and Schils (2012) in their analysis of PISA 2006.

Survey item nonresponse is an even stronger predictor of international variation in test scores. In standalone models, survey item nonresponse explains 41.3, 33.0 and 37.8 percent of the international variation in PISA math, reading and science scores, respectively. These estimates are largely consistent with the findings of Boe, May and Baruch (2002), who examined TIMSS scores and found that 53 percent of the international variation in math scores was attributable to item response rates on a corresponding survey.

Careless answers on the survey are by far the weakest predictor of test scores, within and across countries. In all subjects, careless answering explains only about 2 percent of the international variation in test scores, among the countries in our analytical sample.

The final column in tables 5A, 5B and 5C displays estimates when all measures of student effort are included in the random effects model. Of the between country variation in PISA test scores, our combined measures of student effort on the test explain 39.6 percent

of the variation in math, 33.4 percent of the variation in reading, 38.6 percent of the variation in science.

Given the popular use of PISA scores to make international comparisons, it is useful to examine how the international distribution of test scores changes once adjusting for student effort on surveys. The simplest approach for calculating adjusted scores would be to use the estimates of the models presented in tables 5A, 5B and 5C and obtain adjusted scores as the estimated residuals from these regressions. This is the approach we use, taking the residuals to unstandardized versions of these regressions, adding in the constant.[18]

Table 6A displays the summary statistics of the raw and adjusted scores at the student level. We aggregate those results to the country level. Table 6B shows the summary statistics for country level raw and adjusted scores. As at the individual level, the overall distribution of country-level test scores tightens. The standard deviation in math scores, for example, shrinks from 38.0 in the raw scores to 33.9 in the adjusted scores; the range shrinks from 227.3 points to 202.4. As shown in these simple descriptive statistics, the gap between the highest and lowest performing countries in our sample is driven partly by student effort.

[18] We could take a different approach, using the estimates from our random coefficient models. We could adjust country average performance for differential student's effort in the test and survey, per Borghans and Schils (2015). We could simply use the estimated country average performance in the very first question in the PISA test estimated from our random coefficient model specification in (2) as a measure of performance purged of decline in test performance effects. However, although one could argue that this measure is not affected by fatigue in the test, it can be affected by different rates of nonresponse or other measures of test effort. Some students show low effort throughout the test, from the very onset. We have argued that our survey-based measures help identify such students. Therefore we prefer the approach outlined in the main text, which takes into account all of the information we've collected on student effort.

<<Tables 6A and 6B Here>>

6. Conclusion and Discussion

We have examined measures of student effort on PISA tests. We have shown that these measures differ by country, and have shown that the distribution of international test scores can change substantially once adjusting the effort that students put forward on the test. However, the information contained in PISA datasets does not allow us to directly test one final question: does effort on tests and surveys provide a proxy measure of student noncognitive skills?

Using only data available from PISA, we can only posit that these effort-based measures of effort are proxies for noncognitive skills, such as conscientiousness and persistence. However, previous research provides compelling evidence that our measures of effort actually capture student noncognitive skills.

Beyond their analysis of PISA scores, Borghans and Schils (2012) also examined student motivation on tests that were administered as part of a longitudinal study of British youth. At the baseline year, when respondents were 16 years old, a math test was given that had similar psychometric properties to PISA. Borghans and Schils (2012) found that the estimated decline in performance on this test was predictive of later labor market outcomes, including employment and wages, independently of final scores on the test.

The fact that decline in performance contains independent information that is predictive of objective measures of well-being shows that student motivation on tests is not

idiosyncratic to the testing session. It suggests strongly that decline in performance captures noncognitive skills.

Similarly, recent research examines whether item nonresponse is a proxy measure for noncognitive skills such as conscientiousness (e.g. Hedengren and Strattman, 2012). The most robust examination of survey item nonresponse as an indicator of noncognitive skills can be found in Hitt, Trivitt and Cheng (2016). Within six longitudinal surveys of adolescents from the United States, the frequency with which students skip questions or answer "don't know" is found to be predictive of later educational attainment or labor market outcomes, independent of controls for cognitive ability. The fact that, after adjusting for cognitive ability, survey item nonresponse rates are still associated with later outcomes suggests strongly that item nonresponse is tied to relevant noncognitive abilities.

Careless answering, similar to item nonresponse, has been explored as a proxy measure of noncognitive skills in the literature. In a pair of longitudinal datasets that follow American adolescents into adulthood, Hitt (2016) finds that careless answer patterns are predictive of educational attainment, independent of cognitive ability. As with survey item response rates, careless-answer patterns appear not to be only a measure of effort on a survey, but an indicator of other student behavioral traits that impact later life outcomes.

In total, this research combined with our findings suggests strongly that international differences in test scores are driven a great deal by international differences in noncognitive skills. Our analysis produces country-level estimates of student effort and persistence, separate from adjusted PISA scores. We are hopeful that the effort-based measures we

develop can be used in future research of noncognitive skills. There is growing interest in international comparisons of noncognitive skills. Our results can be used to inform this research. Noncognitive skills research relies heavily on student self-reported scales. These scale scores provide valuable but imperfect information about student noncognitive skills; self-reported scales are prone to a number of biases, and of course are affected by differences in student effort on surveys.

Our results suggest that standardized test scores reflect more than student learning, they reflect the character traits of students taking the tests. As designed, test scores provide valuable but imperfect information on student cognitive abilities. But testing data can also contain information about the effort that each student put forward on the test. As researchers seek to examine international differences in noncognitive skills, they may be able to exploit the measures of effort we have laid out here.

In summary, we calculate international and regional differences in test-effort and survey-effort, using our new measures, which we argue proxy as measures of noncognitive skills. We then decompose international differences in test scores based upon our novel measures of noncognitive skills, finding that between 33 and 40 percent of the between country variation in PISA scores is driven by our measures of effort.

Importantly, our analysis is presently limited to the 44 countries that used the primary set of PISA testing booklets. We expect that our results will be unchanged by the inclusion of the remaining countries that used an easier set of PISA test booklets. However, the interpretation of these findings becomes more complicated once including these countries.

Our variable "decline" is adjusted for booklet number, which within our sample is not related with student or country academic abilities. However, the new set of easier test booklets is necessarily correlated with country-level academic abilities, which would bias the variable "decline" derived from those booklets.[19]

Another important caveat about expanding the analysis into the countries we exclude: the use of careless answer variables is sensitive to the actual differences across countries in the internal reliability of scales. It is possible that the words and concepts tested by Likert-type items do not translate well into many of the small countries that are now using easier test booklets. In unreliable scales, individual answers do not readily predict other answers (because answers to items are largely uncorrelated). Given that careless answers are quantified as being unpredictable answers, it makes sense only to use this measure of effort in settings where scales are internally consistent (which is the case in 44 countries used in our analysis).[20] In any case, careless answer patterns on surveys are a relatively weak driver of our main findings, so we do not believe this problem presents a challenge to our results.

---

[19] Including countries that use the easier set of booklets would create problems, if we were to use the more complicated random-coefficients method (laid out above) to estimate decline in performance over the course of the test. If we add the countries that use easier booklets, we would likely be adding countries that are vastly different in test performance from countries that use the standard booklets. Random coefficient estimates are all in reference to the average performance; this then would make it difficult to capture, for instance, that Greece does worse than other developed countries. The differences across countries using standard booklets become too small in comparison with the new average performance and the only differences observed are between countries using standard versus easier booklets. It is important to note that practically all of the countries that elected to use easier booklets are developing countries.

[20] For example, in OECD member countries, almost all of whom used the standard set of booklets, the median reliability for the "attitude towards school" scale was a Cronbach's alpha of 0.70, according to the 2009 PISA Technical Manual. However, for example in

The policy implications of international and regional gaps in test scores are based in large part on what test scores are seen to represent. Our work examines the extent to which these differences in test scores are really driven by differences in math, science and literacy skills, rather than by differences of another sort – differences in how students approach the routine tasks of school and work. Our analysis synthesizes methods from previous research and applies them to a new sample of students, those participating in the 2009 wave of PISA. The finding is remarkably consistent across time, using each approach. A substantial portion of the international variation in test scores is driven by student effort on the test itself.

Albania and Peru (two countries that used the alternate set of booklets) the Cronbach's alpha was 0.46 and 0.53.

References

Boe, Erling E., Henry May, and Robert F. Boruch. 2002. "Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Acheievement Differences at the National, Classroom, and Student Levels." *Center for Research and Evaluation in Social Policy, CRESP-RR-2002-TIMSS1*. http://eric.ed.gov/?id=ED478493.

Borghans, Lex, and Trudie Schils. 2015. "The Leaning Tower of Pisa." Working Paper. Accessed February 24. http://www.sole-jole.org/13260.pdf.

Hedengren, David, and Thomas Stratmann. 2012. "The Dog That Didn't Bark: What Item Nonresponse Shows about Cognitive and Non-Cognitive Ability." *Available at SSRN 2194373*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2194373.

Hitt, Collin, Julie Trivitt and Albert Cheng. 2016. "When You Say Nothing At All: The Predictive Power of Student Effort on Surveys," accepted manuscript, in press, *Economics of Education Review* (doi: 10.1016/j.econedurev.2016.02.001)

Hitt, Collin. 2016. "Just Filling in the Bubbles: Using Careless Answers Patterns on Surveys as a Proxy Measure of Noncognitive Skills," EDRE Working Paper 2015-6. http://www.uaedreform.org/just-filling-in-the-bubbles-using-careless-answer-patterns-on-surveys-as-a-proxy-measure-of-noncognitive-skills/

Kautz, Tim, James J. Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans. 2014. "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success." National Bureau of Economic Research. http://www.nber.org/papers/w20749

Méndez, Ildefonso, Gema Zamarro, José García Clavel, and Collin Hitt. 2015. *Habilidades No Cognitivas Y Diferencias de Rendimiento En PISA 2009 Entre Las Comunidades Autónomas Españolas*. Ministerio de Educación. https://books.google.com/books?hl=en&lr=&id=rcOaCgAAQBAJ&oi=fnd&pg=PA53&dq=mendez+zamarro+hitt&ots=6-ejo1t9t_&sig=xzWgrJuWJynseBM2c18xU8PW558 .

Figure 1: Theoretical Framework

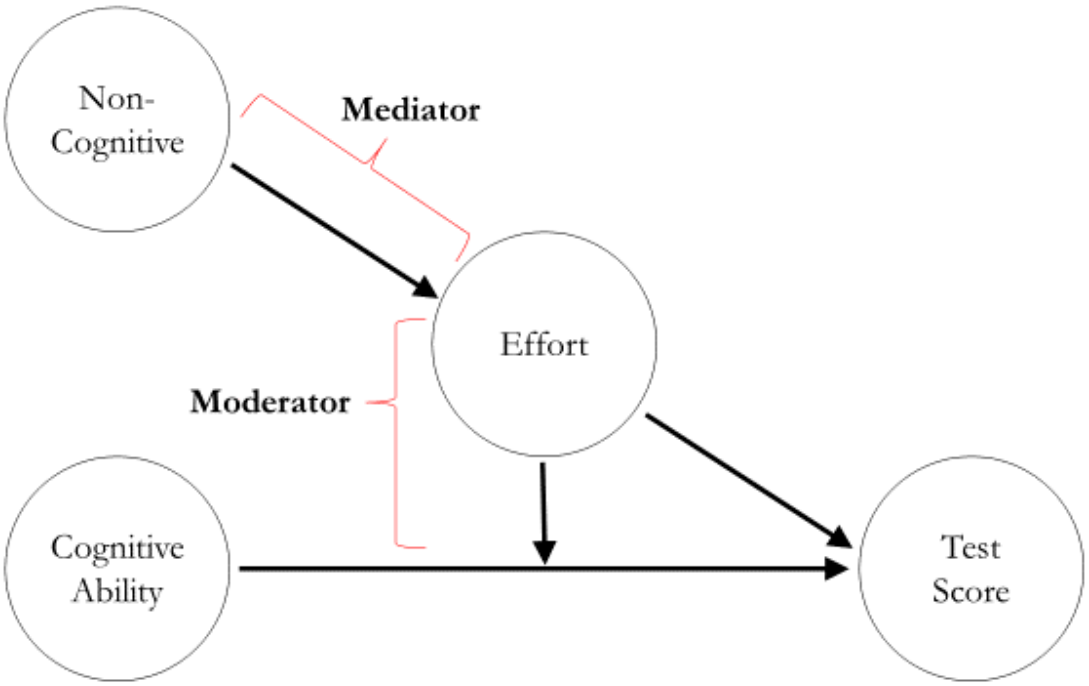


Figure 2: Average Performance by Question Position in Selected Countries

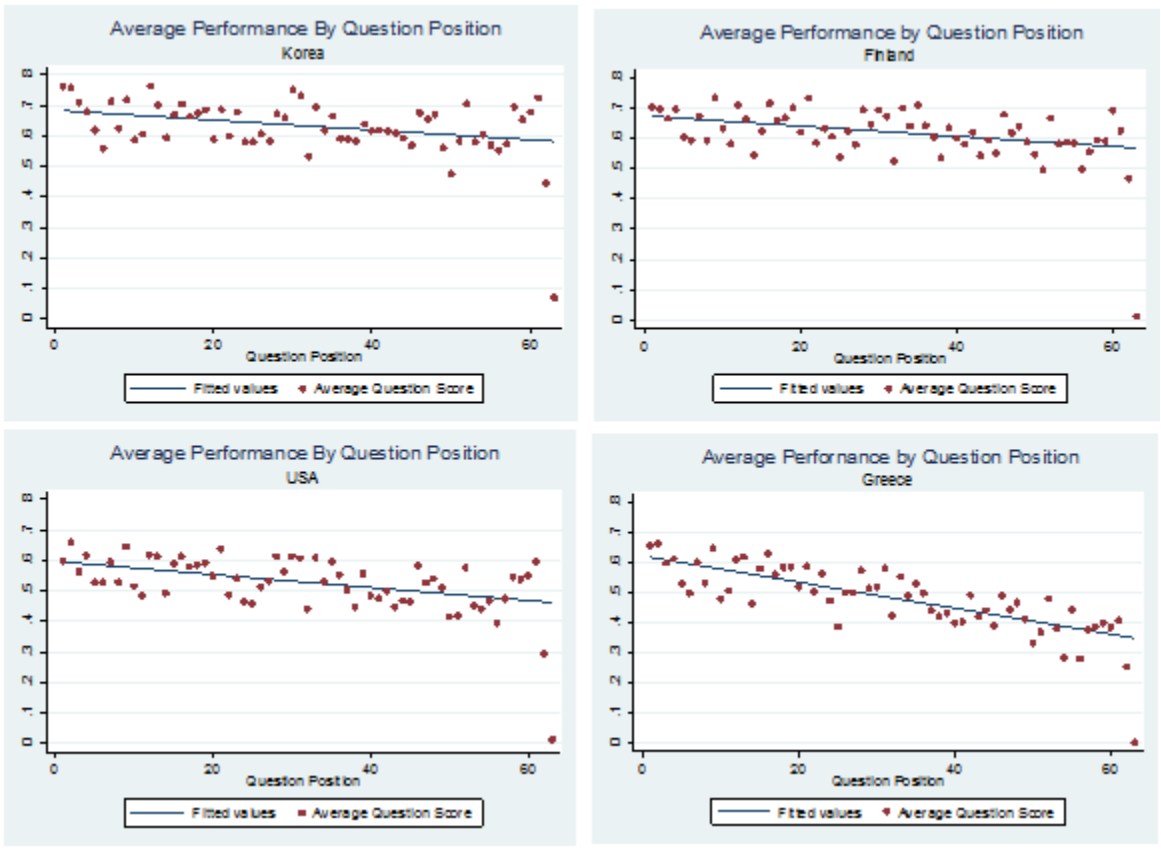
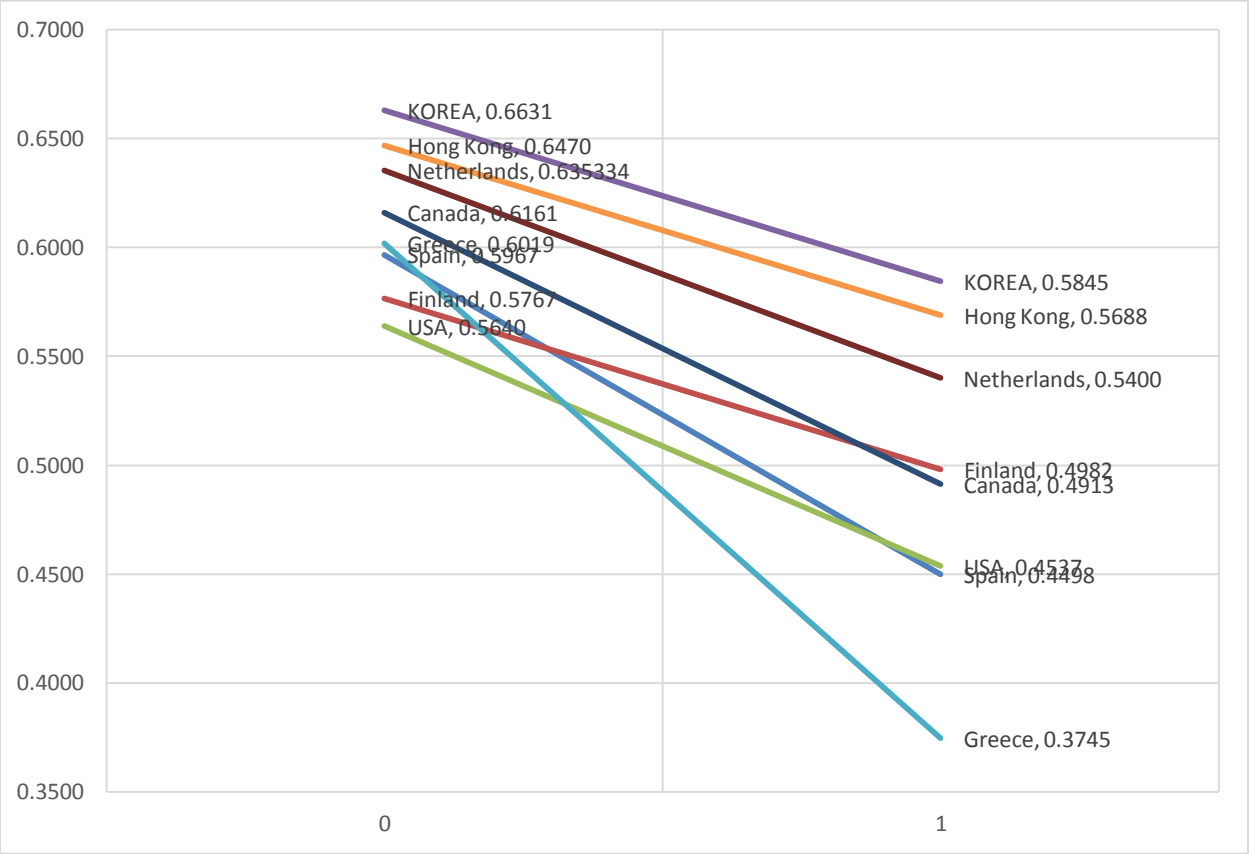


Figure 3: The Estimated Decline in Performance during the PISA test, by Country



Note: Estimates above were obtained using random coefficient regression estimates by country including a random constant and slope and shows just the performance at the beginning and at the end of the test. Details of this model are explained in Section 4.

Table 1: Measures of Student Motivation during PISA Assessment, Summary Statistics

Variable		Mean	SD	Min	Max	Observations
Test: First Ten Score	overall	5.85	2.44	0.00	10.00	N = 311,484
	between		0.59	3.51	7.05	n = 44
	within		2.39	-1.20	12.33	T-bar = 7,079.18
Test: Last Ten Score	overall	4.46	2.70	0.00	10.00	N = 311,484
	between		0.74	2.07	6.23	n = 44
	within		2.62	-1.77	11.50	T-bar = 7,079.18
Test: Adjusted "Decline"	overall	1.37	2.36	-11.12	10.45	N = 311,484
	between		0.27	-0.57	0.86	n = 44
	within		2.34	-10.67	10.55	T-bar = 7,079.18
Survey: Item Nonresponse	overall	0.03	0.05	0.00	0.95	N = 311,484
	between		0.01	0.01	0.05	n = 44
	within		0.05	-0.02	0.96	T-bar = 7,079.18
Survey: Careless Answers	overall	-0.03	0.25	-1.04	5.19	N = 309,425
	between		0.07	-0.15	0.13	n = 44
	within		0.24	-1.04	5.16	T-bar = 7,032.39

Table 2A: Student-level Correlations Between Test Scores and Measures of Motivation

	1	**2**	**3**	**4**	**5**	**6**
1. Math Score	1.00					
2. Reading Score	0.82	1.00				
3. Science Score	0.88	0.88	1.00			
4. Test: Decline	-0.09	-0.11	-0.11	1.00		
5. Survey: Items Missing	-0.27	-0.29	-0.28	0.03	1.00	
6. Survey: Careless Answers	-0.08	-0.08	-0.08	-0.01	0.06	1.00

Note: All coefficients significant at $p<0.001$

Table 2B: Country-level Correlations Between Test Scores and Measures of Motivation

	1	**2**	**3**	**4**	**5**	**6**
1. Math Score	1.00					
2. Reading Score	0.90	1.00				
3. Science Score	0.95	0.93	1.00			
4. Test: Decline	-0.54	-0.48	-0.57	1.00		
5. Survey: Items Missing	-0.64	-0.57	-0.61	0.36	1.00	
6. Survey: Careless Answers	-0.14	-0.14	-0.14	0.24	0.25	1.00

Note: All coefficients significant at $p<0.01$, except for all coefficients in Row 6, which are not statistically significant at $p<0.10$

Table 3A: Regression of PISA Math Score on Effort

Test: Decline	-0.087			-0.078
Survey: Items Missing		-0.274		-0.289
Survey: Careless Answers			-0.085	-0.069
R-squared	0.008	0.075	0.007	0.098
N	311,484	311,484	309,425	309,425

Note: All coefficients significant at $p<0.001$

Table 3B: Regression of PISA Reading Score on Effort

Test: Decline	-0.114			-0.105
Survey: Items Missing		-0.291		-0.304
Survey: Careless Answers			-0.078	-0.062
R-squared	0.013	0.085	0.006	0.111
N	311,484	311,484	309,425	309,425

Note: All coefficients significant at $p<0.001$

Table 3C: Regression of PISA Science Score on Effort

Test: Decline	-0.106			-0.098
Survey: Items Missing		-0.283		-0.297
Survey: Careless Answers			-0.077	-0.062
R-squared	0.011	0.080	0.001	0.105
N	311,484	311,484	309,425	309,425

Note: All coefficients significant at $p<0.001$

Table 4A*:* Results, Random Coefficients Estimates of Item Order on Performance

	Country	α_0	β_1	$SD(\alpha_0^i)$	$SD(\beta_1^i)$	$Corr(\alpha_0^i, \beta_1^i)$
1	JPN	-0.1210	0.6804	0.1677	0.1787	-0.0666
2	KOR	-0.0786	0.6631	0.1242	0.1449	0.0078
3	HKG	-0.0781	0.6470	0.1443	0.1702	-0.0886
4	DEU	-0.0966	0.6468	0.1640	0.1901	-0.1790
5	SWE	-0.0781	0.6393	0.1554	0.1878	-0.0578
6	NLD	-0.0953	0.6353	0.1289	0.1701	-0.0453
7	LTU	-0.1231	0.6343	0.1813	0.1755	-0.2339
8	PRT	-0.0692	0.6329	0.1159	0.1552	-0.0885
9	MAC	-0.1302	0.6256	0.2007	0.1745	-0.4444
10	CZE	-0.0974	0.6252	0.1636	0.1893	-0.1638
11	BEL	-0.0998	0.6230	0.1594	0.1914	-0.1478
12	IDN	-0.1133	0.6228	0.1853	0.1904	-0.2311
13	TAP	-0.1493	0.6218	0.1950	0.1937	-0.2397
14	NZL	-0.1493	0.6216	0.1862	0.1815	-0.2462
15	DNK	-0.1090	0.6166	0.1602	0.1867	-0.1994
16	CAN	-0.1248	0.6161	0.1802	0.1813	-0.1917
17	ITA	-0.1562	0.6153	0.2032	0.1932	-0.3211
18	NOR	-0.1067	0.6119	0.1617	0.1845	0.0061
19	QCN	-0.0849	0.6038	0.1493	0.1836	-0.1971
20	GRC	-0.2274	0.6019	0.2272	0.1972	-0.4489
21	FRA	-0.1473	0.5977	0.1990	0.2024	-0.2013
22	CHE	-0.1222	0.5968	0.1589	0.1808	-0.1650
23	ESP	-0.1469	0.5967	0.1952	0.1903	-0.3127
24	LVA	-0.1438	0.5962	0.1823	0.2061	-0.2377
25	AUT	-0.0922	0.5945	0.1562	0.1983	-0.2185
26	AUS	-0.1166	0.5932	0.1615	0.1903	-0.0586
27	SGP	-0.1915	0.5875	0.2320	0.1996	-0.4541
28	EST	-0.0867	0.5867	0.1552	0.1707	-0.2696
29	POL	-0.1161	0.5836	0.1765	0.1810	-0.2897
30	RUS	-0.1526	0.5825	0.1969	0.1910	-0.3984
31	FIN	-0.0786	0.5767	0.1546	0.1657	-0.1747
32	SVK	-0.1232	0.5752	0.1641	0.1935	-0.3230
33	USA	-0.1103	0.5640	0.1481	0.1862	-0.0977
34	HRV	-0.0999	0.5500	0.1476	0.1845	-0.3115
35	GBR	-0.1093	0.5445	0.1425	0.1842	-0.0774
36	LIE	-0.1176	0.5435	0.1722	0.1733	-0.3633
37	ISR	-0.1764	0.5404	0.2198	0.2190	-0.3407
38	ISL	-0.1108	0.5367	0.1663	0.1862	-0.1772
39	SVN	-0.1185	0.5309	0.1439	0.1847	-0.1202
40	LUX	-0.1040	0.5152	0.1520	0.1857	-0.3180
41	HUN	-0.0962	0.5134	0.1393	0.1731	-0.1908
42	TUR	-0.1267	0.4535	0.1641	0.1804	-0.3563
43	THA	-0.1330	0.4127	0.1592	0.1771	-0.4133
44	IRL	-0.1556	0.3802	0.1777	0.1498	-0.5418

Table 4B: Random Coefficient Models Estimates of Test Decline

Overall	$\boldsymbol{\alpha_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{SD(\alpha_0^i)}$	$\boldsymbol{SD(\beta_1^i)}$	$\boldsymbol{Corr(\alpha_0^i, \beta_1^i)}$
Mean	0.5849	-0.1196	0.1835	0.1686	-0.2270
SD	0.0607	0.0321	0.0138	0.0257	0.1318

Table 5A: Variation in PISA Math Scores Explained by Student Motivation, Multilevel Model.

	Test: Decline	Survey: Item Nonresponse	Survey: Careless Answers	Combined
within country	0.0051	0.0664	0.0063	0.0861
between countries	0.2964	0.4126	0.0182	0.3958
Overall	0.0075	0.0749	0.0072	0.0983
Country n	44	44	44	44
Student n	311,484	311,484	309,425	309,425

Table 5B: Variation in PISA Reading Scores Explained by Student Motivation, Multilevel Model.

	Test: Decline	Survey: Item Nonresponse	Survey: Careless Answers	Combined
within country	0.0112	0.0788	0.0051	0.1033
between countries	0.2300	0.3301	0.0201	0.3337
Overall	0.0130	0.0845	0.0061	0.1114
Country n	44	44	44	44
Student n	311,484	311,484	309,425	309,425

Table 5C: Variation in PISA Science Scores Explained by Student Motivation, Multilevel Model.

	Test: Decline	Survey: Item Nonresponse	Survey: Careless Answers	Combined
within country	0.0087	0.0727	0.0047	0.0943
between countries	0.3226	0.3777	0.0183	0.3857
Overall	0.0113	0.0800	0.006	0.1053
Country n	44	44	44	44
Student n	311,484	311,484	309,425	309,425

Table 6A: Student-level PISA Scores, Raw and Adjusted for Student Motivation

Variable	Mean	Std. Dev.	Min	Max
Math	501.4	97.0	48.1	1022.2
Math, Adjusted	520.5	92.1	103.0	1129.8
Reading	495.2	93.9	6.7	871.1
Reading, Adjusted	514.7	88.5	89.0	1103.7
Science	504.5	96.1	0.8	883.8
Science, Adjusted	524.0	90.9	80.8	1135.4

Note: N = 309,425

Table 6B: Country-level PISA Scores, Raw and Adjusted for Student Motivation

Variable	Mean	Std. Dev.	Min	Max
Math	501.5	38.0	372.8	600.1
Math, Adjusted	520.5	33.9	404.2	606.5
Reading	494.4	27.1	402.4	556.0
Reading, Adjusted	513.9	23.7	434.2	562.3
Science	504.1	32.6	383.1	575.2
Science, Adjusted	523.6	28.7	414.8	581.5

Note: N = 44

Conclusion

This dissertation has explored new measures of noncognitive skills. The most popular method for measuring noncognitive (or "character") skills in students is through self-reported surveys. Yet some students do not provide reliable self-reports. Along with my co-authors I examine not only what respondents say on these surveys, but what they do. Do they frequently skip questions? Do they just fill in the bubbles? Do they trail off in performance over the course of the test?

The amount of effort that students show on surveys is predictive of later life outcomes, my co-authors and I have demonstrated. In Chapter 1, Julie Trivitt, Albert Cheng and I examine the rate at which students skipped questions or answered "unsure." In Chapter 2, I develop a novel method for detecting careless answer patterns. In both chapters, we follow the same process that one would follow in validating a new scale or performance task. We present a new measure, and then test whether over time it is independently predictive of important, objective outcomes.

Item nonresponse and careless answers both perform as valid measures of noncognitive skills. Measured in adolescence during a single survey session, each measure is independently predictive of later life outcomes. This discovery could lead to important developments in education research.

For example, in Chapter 3, Gema Zamarro, Ildefonso Mendez and I demonstrate that international differences in reading, math and science scores might also reflect noncognitive skills. We use survey item nonresponse and careless answer patterns, as well as

the decline in student effort over the course of a standardized test, to demonstrate this fact. We also address an important critique of these measures. A potential concern with our measures of effort is that they could all actually be capturing cognitive ability, albeit in a messy way. If this was the case, one would expect the correlation between our measures and test scores to be at least as strong within countries as across countries. We find the opposite: the correlation between our noncognitive measures and test scores is much stronger across countries than within countries.

In the previous chapters, I have outlined several limitations to measures of noncognitive skills that are based on self-reports. Using behavioral measures of noncognitive skills can help address those limitations. But those limitations aside, another weakness of existing datasets is that self-reports are typically collected on a very limited basis. In the case of the 1997 National Longitudinal Study of Youth, for example, self-reports of noncognitive skills weren't collected at all at baseline. In such datasets, behavioral measures of survey and test effort can help fill voids. In datasets with richer amounts of self-reported data, survey effort can still be used to supplement and strengthen existing data.

After a given noncognitive (or cognitive) skill is validated, another important question then follows: what can impact the skills being measured? Researchers are already exploring these questions using self-reported data. For example, using the same value-added models that show teacher impacts on test scores (see: Koedel, Mihaly and Rockoff, 2015), Matthew Kraft and colleagues have found in multiple studies that teachers can impact self-reported noncognitive skills (Blazar and Kraft, 2015; Kraft and Grace, 2016). Interestingly, the same

paradox I've written about earlier becomes apparent again: the teachers who have positive impacts on test scores are not the same teachers who have impacts on self-reported noncognitve skills. The teacher *impacts* on achievement appear largely uncorrelated with the *impacts* on noncognitive skills.

Our new measures of survey effort can be used to extend this research, which Albert Cheng and colleagues are doing in forthcoming work (Cheng and Zamarro, 2016; Cheng, 2016). They are exploring whether teachers can impact student conscientiousness on surveys. Moreover, they are exploring whether student conscientiousness on surveys mirrors that of their teachers. Survey item nonresponse and careless answers are the rare noncognitive measures in that they are often collected (inadvertently) on students, teachers, parents and peers alike. This is a promising and exciting direction for future research, but it also points to the largest limitation to measures of noncognitive skills - they are not fit for use in high stakes accountability policies.

Research is increasingly showing that character skills are important and that educators can impact them. In turn, presumably, policymakers will seek to push schools to improve these skills. Under the current paradigm of accountability in public schools, this could mean aligning incentives and pressures to impact those behaviors. The problem is that accountability policy is rarely designed to impact skills but rather is designed to force improvement on *a specific metric* of skill.

In social science, there is a principle called Campbell's Law: "The more any quantitative social indicator (or even some qualitative indicator) is used for social decision-

making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." (Campbell, 1976). This problem is well known in standardized testing. Test preparation does not improve student math and reading skills, but can impact math and reading scores. Yet under intense pressure to raise test scores, it is rational for schools and teachers to engage in test preparation (Witte et al. 2014).

This problem would be far more problematic for the measurement of noncognitive skills. Both self-reports and survey effort are easily coached - whereas the underlying behaviors that are supposed to be captured by these measures are much harder to impact.

If accountability policies are created to pressure educators to improve metrics of noncognitive skills, those metrics will instantly become contaminated. For example, students taking the Duckworth Grit Scale are asked to rate themselves on statements like "I am a hard worker" and "I am diligent." These items are one quarter of the eight-item Grit Scale. Unlike standardized tests that are drawn from banks of hundreds of items, all designed to measure math or reading skills, there is no Grit item bank. The items that comprise the Duckworth Grit Scale are the same for every student.

Students can easily be coached to say that they are hard workers and diligent. This of course doesn't make those students hard working and diligent. The same goes for item nonresponse. If educators were held accountable for the frequency with which students failed to respond to questions, it would be a solid bet that item nonresponse would fall to zero. This doesn't mean that students suddenly became more conscientious.

The other measures of noncognitive skills that I have explored are more difficult but not impossible to game. Careless answers can likely be minimized somewhat by coaching and cajoling students to take surveys seriously.

Put plainly, the measures I have explored above are useful for research and research only. Perhaps technology and tools will emerge to address the problems of coaching survey responses. This is not yet the case.

In the Introduction to this dissertation, I outlined a paradox in education policy, made apparent only recently by leading research. Impacts on test scores do not routinely equate to impacts on later outcomes. Nor do impacts on test scores equate to impacts on noncognitive skills. Policies that reward only test score impacts will fail to reward, or even punish, educators who are having other important impacts on students. This isn't simply a challenge to priorities in accountability policy. In other words, it is not simply a question of, should schools focus on reading skills or work ethic or math skills or self-control? This is a challenge to the entire K-12 policy paradigm in the United States.

Current policy rests on two major assumptions: student skills of all sorts can be reliably measured; and the impacts that educators have on those skills can be identified in the data. The former fails to be true for noncognitive skills once high stakes are attached, so sensitive are the measures to corruption. Once the first assumption fails, so does the second. If policymakers decide that character skills should be a top priority of schools, a move away from current accountability models will be absolutely necessary.

On the other hand, if policymakers decide that noncognitive skills are not the proper focus of public education, or if they decide that educators should be held accountable for things that can be measured in high stakes setting, then perhaps current policy should remain in place. In this case, programs that focus on improving noncognitive skills but not test scores will continue to be ignored or punished.

Education policy research is entering an exciting and crucial era. For the past fifteen years, education policy in the United States has been focused intensely on improving test scores. Leading education research in turn has focused on test score gains. This agenda was a response to a generation of research that had demonstrated an unconscionable gap in test scores between white and minority students. Research on noncognitive skills is challenging the primacy of achievement tests. In doing so, this body of research challenges the foundation of many education reforms over the past fifteen years.

None of this is to say that the problems spotlighted by the achievement gap should be forgotten. The need to improve opportunities for disadvantaged children remains as salient as ever, and is indeed what motivates the noncognitive skills research program. For example, the legendary economist James Heckman and others now argue that the best way to help disadvantaged children is through programs that focus on noncognitive skills (Heckman, Humphries and Kautz, 2014).

However, if public education in America is to be oriented toward something other than test scores, parents and policymakers will need a guiding concept more specific than "noncognitive" skills. Measures of character skills need to be developed, improved and

validated - in order to show what these skills are and how they matter. That is what this dissertation has attempted to accomplish, in some small part.

References: Conclusion

Blazar, David, & Kraft, Matthew A. (2015). "Teacher and teaching effects on students' academic behaviors and mindsets (Working Paper No. 41). Cambridge, MA: Mathematica Policy Research.

Campbell, Donald T. (1976). "Assessing the Impact of Planned Social Change," Paper #8, Occasional Paper Series, December 1976, The Public Affairs Center, Dartmouth College.

Cheng, A., & Zamarro, G. (2016). Measuring Teacher Conscientiousness and its Impact on Students: Insight from the Measures of Effective Teaching Longitudinal Database (EDRE WP No. 2016-04). University of Arkansas: Fayetteville, AR.

Cheng, A. (2016). Like Teacher, Like Student: Teachers and the Development of Student Noncognitive Skills (EDRE WP No. 2015-02). University of Arkansas: Fayetteville, AR.

Heckman, James J., John Eric Humphries, and Tim Kautz. (2014) *The Myth of Achievement Tests: The GED and the Role of Character in American Life*. University of Chicago Press, 2014.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. (2015) "Value-added modeling: A review." *Economics of Education Review* 47 (2015): 180-195.

Kraft, Matthew A. and Sarah Grace. (2016). Teaching for tomorrow's economy? Teacher effects on complex cognitive skills and social-emotional competencies (Working Paper). Brown University: Providence, RI

Witte, John F., Patrick J. Wolf, Joshua M. Cowen, Deven Carlson, and David F. Fleming. (2014) "High Stakes Choice: Achievement and Accountability in the Nation's Oldest Urban Voucher Program," *Education Evaluation and Policy Analysis*, 36(4), December 2014: 437-456.