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

Collin Hitt
cehitt@uark.edu
Department of Education Reform
University of Arkansas

EDRE Working Paper No. 2015-06
Last Updated July 9, 2015
Please Do Not Cite Without Permission

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 used in attitude scales. Apart from students' actual answers on these scales, I examine the overall pattern of answers to determine whether students appear to be providing careful or careless answers. Self-reported scales are fundamental tools for survey researchers and exist in hundreds of existing datasets. Using two national longitudinal datasets, I show that careless answer patterns from adolescent respondents are negatively predictive of later educational attainment, independent of cognitive ability and other traditionally-measured noncognitive skills. I posit that careless answers, as I have quantified them, proxy as a behavioral measure of a negative noncognitive trait.

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 grit and conscientiousness – that are now understood to be important determinants of educational attainment and labor market success. Education researchers are investigating programs that seek impact these softer skills, but such research is encountering substantial challenges. Social scientists have struggled, and continue to struggle, to measure such non-cognitive 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. Ironically perhaps, 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, 2015). 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 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 calculate each student's predicted response, given his or her responses 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 non-cognitive skills research includes contributions from economics, psychology and education policy. Its modern origins lie in the work 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).

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, serous 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, 2015).

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 difficult to administer, and perhaps most importantly, many task-based measures are new. Social science research depends heavily on longitudinal surveys that were begun years, even decades ago. It is impossible to travel back in time to administer 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 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 (2015) show that the frequency with which students skip questions is 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 (2015). 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. 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 depend upon students who lack focus or motivation to stay focused and motivated enough to answer questions about their focus and motivation in school – 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 or persistence. Yet it is precisely these low-skilled students that noncognitive-skills interventions are supposed to help.

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 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. 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.

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.[1]

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

The field of psychometrics uses a set of standard 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.[2] This is a potentially arbitrary process. Some questions

[1] 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 composite for the sake of simplicity in interpretation. The main differences are as follows. The NELS:88 authors create an average of standardized item level answers. 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).

[2] 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.

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.

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 table 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 to the rest of the test. 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 item j and other items can be expressed as follows:

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

, where

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

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 than at 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 not particularly difficult to envision 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. When students become disengaged, 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 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.

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 NELS88 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 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 straight line or zig-zag across the page. These may be respondents who provide random answer patterns, with no meaningful effort 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 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 a 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 accidently 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. I average the absolute values of all item-level

residuals from the "item rest" regressions. I am 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 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 from 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 noncognitve skills.

Table 6A shows the pairwise correlations between cognitive test scores, Locus of Control, Self Concept and answer-unpredictability in NELS:88. The correlations between answer-unpredictability 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 answer-unpredictability in ELS:02. Again the correlation of answer-unpredictability with cognitive ability is negative but weak (-0.201). The correlation of

answer-unpredictability to the Effort and Control Expectations is virtually nil, and statistically insignificant.

The relatively weak correlation with cognitive ability demonstrates that the answer-unpredictability captures something other than cognitive ability. That of course could be random noise. Or it could a completely unimportant behavioral trait, as far educational attainment is concerned. Thus I turn to the question of whether careless-answers 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 carelessness, a noncognitive trait. ϵ_i is a normally distributed error term.

Years of Education

Tables 7A and 7B 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 in average-answer-unpredictability 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 answer-unpredictability conceptually captures a detrimental behavior, which I have called carelessness. 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 non-cognitive skills and unpredictable-answers. The unpredictable-answer measure of noncognitive skills remains negative and statistically significant. In NELS:88 the effect attenuates slightly when including additional noncognitive skills, where is 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 7A and 7B. 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 8A and 8B examines 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 8 is a separate regression with samples that grow smaller as the attainment threshold goes higher. And in each regression, the dependent is equal to one if a student reached that attainment level conditional (conditional on reaching the previous level). Estimates are based 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 level at the lower end of the attainment distribution. Column 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.

In NELS:88 the predictive effects of carelessness carry into college, somewhat. A one standard deviation increase in average-answer-residuals is associated with a 1.2 percentage point decrease in the likelihood of completing at least one year of postsecondary education. However, at higher levels of the attainment distribution, the predictive impact of carelessness dissipates entirely.

Interestingly, across Table 8A, 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 non-cognitive skills is strongest.

The pattern of findings is somewhat different in ELS:02, per Table 8B. 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.

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 non-cognitive ability. Furthermore, conditional on receiving a four undergraduate degree, a one standard deviation increasing 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 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 unpredictability 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' answer-unpredictability 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 answer-unpredictability is associated with between a 0.05 and 0.10 year decrease in years of schooling completed. The size 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, 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.

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.[3] 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.

This paper is designed to advance rather than critique noncognitve skills research. Noncognitive skills researchers 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

[3] In the NELS:88, the only Likert-type items appears on scales used to measure non-cognitive 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 more typical means. It can be used to reanalyze older, existing datasets. 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' noncognitive skills, whether or not they even mean to do so.

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. 2015. "When You Say Nothing at All: The Surprisingly Predictive Power of Student Effort on Surveys." EDRE Working Paper. http://www.uaedreform.org/downloads/2014/10/edre-wp-2014-10.pdf.

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

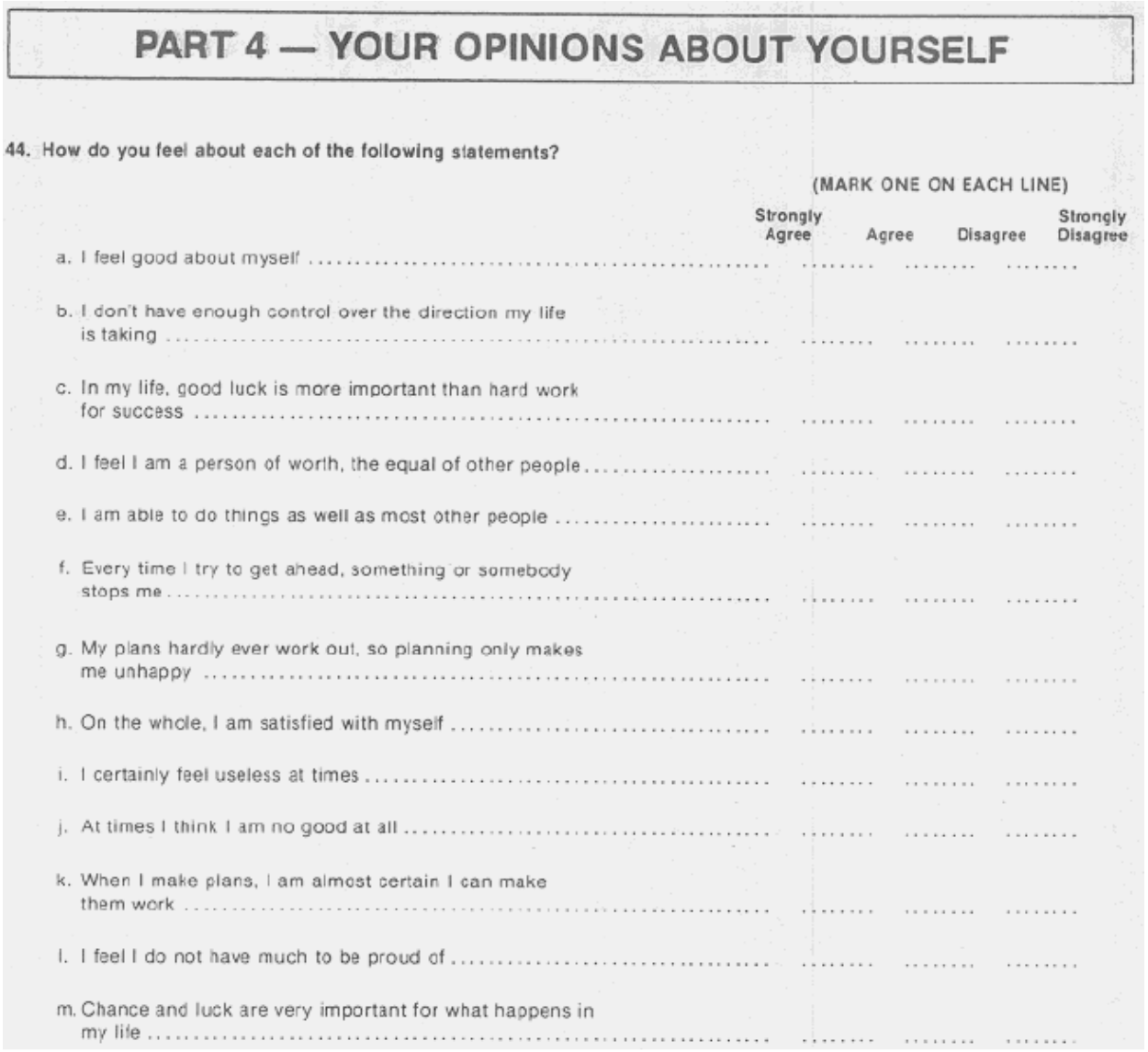
Mischel, Walter, Ebbe B. Ebbesen, and Antonette Raskoff Zeiss, 1972. "Cognitive and attentional mechanisms in delay of gratification." *Journal of personality and social psychology* 21, no. 2 (1972): 204.

Rotter, J. B. (1966). Generalized Expectancies for Internal versus External Control of Reinforcement. *Psychological Monographs, 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: Average-Unpredictability: 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 Unpredicitibility 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
Unpredictable-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
Unpredictable-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 7A: 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
Unpredictable-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 7B: 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
Unpredictable-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 8A: 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
Unpredictable-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 8B: 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
Unpredictable-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