Comments on "Non-Cognitive Skills as Human Capital"

David J. Deming

December 2016

Shelly Lundberg has written an important paper about the rapidly growing study of "non-cognitive" skills in economics. This paper should be required reading for social scientists who seek to use measures of non-cognitive skills in schools and other educational settings to make important policy decisions. I largely agree with her conclusions about the state of the literature, which I summarize crudely as follows. Although the evidence is overwhelming that so-called "non-cognitive" skills are important predictors of many important life outcomes, we do not really agree on what they are (and importantly, what they are not). Thus we have very little idea of how to measure non-cognitive skills well, and even less idea of how to use measures of non-cognitive skills to make high-stakes policy decisions.

In my view, *measurement* is the fundamental challenge for social scientists who want to study non-cognitive skills. I would characterize existing measures of non-cognitive skills as having one of two problems. First, self-assessment measures such as personality inventories (e.g. the "Big 5") are arguably valid but unreliable across contexts, often in ways that make them difficult to use for any practical purpose. On the other hand, administrative records of behavior are reliable (in a statistical sense) and predictive - but possibly invalid - measures of the underlying skill. All of the measures of non-cognitive skills that I have seen used in research have – to varying degrees - one of these two problems.

While no measure is perfect, cognitive skills are much better measured than non-cognitive skills in terms of both validity and reliability. One might conclude from this that the construct of cognitive skill is *inherently* more valid. However, this ignores the history of measurement. Psychologists – and the testing industry – spent several decades and millions of dollars systematically improving and refining the

measurement of cognitive skills. I conclude by advocating for an equally careful and rigorous approach to the theoretical refinement and measurement of non-cognitive skills.

Reliability and self-assessment

As Lundberg (2016) points out, most self-assessment measures ask individuals to answer questions that indicate "what I am like" or "this is what I believe". An example is the Big 5 personality inventory, a rigorously developed psychological model that distills human personality into five factors – extraversion, conscientiousness, agreeableness, neuroticism and openness to experience.

The five factors were originally derived from a statistical factor analysis of a much larger number of potential personality traits (see John and Srivastava (1999) for an overview and history of the Big 5.) Thus in a sense they are statistical rather than theoretical constructs, chosen because they are distinct and orthogonal to one another rather than for higher-minded reasons. Still, the existence of these five distinct and mostly comprehensive set of personality factors has been replicated by psychologists in many other settings spanning geography, culture and time. Agreement is hardly unanimous and criticisms of the Big 5 abound, yet it probably represents the best case scenario for "non-cognitive" skill measures that are based on self-assessment. Moreover, Big 5 personality measures – especially conscientiousness – are strongly positively correlated with educational attainment, labor market earnings and other important life outcomes (e.g. Heckman and Kautz 2012).

Are self-assessments such as the Big 5 *reliable*? That depends on what you mean by reliable. The most basic definition is *test-retest* reliability, where one administers the same assessment to the same person under the same conditions over a very short period of time, and estimates the correlation between assessments. Of course a perfect replication of the test environment is never possible, but under ideal conditions the test-retest reliability of the Big 5 is extremely high. The correlation between assessments ranges from 0.8 to 0.9, depending on the length of the test instrument and the specific

factor being studied (John and Srivastava 1999). This is very similar to the test-retest reliability for IQ, for example.

However, policy is not made in a lab, and the evidence for the reliability of self-assessments in the field and across contexts is much less reassuring. Schmidt et al (2007) administer the Big 5 personality questionnaire in a number of OECD countries and show that the correlation between conscientiousness (the tendency to work hard and be persistent) and average hours worked is *negative.* This is particularly striking in the case of respondents in France and South Korea. South Koreans report working nearly 2500 hours per year, compared to around 1500 hours for their French counterparts. Yet France places 4th and South Korea places 25th out of 26th when respondents are asked to self-assess their conscientiousness. West et al (2014) find that students who are randomly assigned to a set of schools known for their emphasis on character-building and hard work (so-called "No Excuses" charter schools) self-report *lower* levels of conscientiousness, self-control, and "grit". In both cases respondents are comparing themselves to those around them. This makes it difficult or impossible to compare measures of "non-cognitive" skills across very different contexts.

Non-cognitive skill measures that are sensitive to context are particularly problematic in high-stakes settings. Put bluntly, personality assessments can be easily gamed if one knows what the "right answer" is supposed to be. For example, personality tests are often administered by large retail companies as part of the job applicant screening process. A cursory web search for "job application test answers" reveals that there is a robust market in teaching people how to successfully game personality assessments.

Notably, gaming is possible even without access to specific test items. Conscientiousness is among the best predictors of job performance, and so employers would like to screen for this personality trait. Big 5 question items that measure conscientiousness include Likert scale items (1 to 5 numerical

responses that range from strongly disagree to strongly agree) such as "I see myself as someone who does a thorough job" or "I am always prepared." It is not hard to foresee that placing a high weight on conscientiousness in hiring will lead to a sudden and dramatic increase in the self-reported persistence and diligence of the average applicant!

Using behaviors to measure "non-cognitive" skills

Kautz et al (2015) discuss this problem of "reference bias" in self-assessed measures. They propose using behaviors as alternative measures of skills:

> "…all tasks or behaviors can be used to infer a skill as long as the measurement accounts for other skills and aspects of a situation…Self-reported scales should not be assumed to be more reliable than behaviors, although personality psychologists often assume so. The question is which measurements are most predictive and which can be implemented in practice. The literature suggests that there are objective measurements of non-cognitive skills that are not plagued by reference bias." (Kautz et al 2015, p.17-18).

In other words, behavioral measures of non-cognitive skills might be better than self-assessments if they are predictive and reliable across contexts (e.g. not plagued by reference bias).

Lundberg (2016) points out that using observed behaviors to measure skills is potentially problematic if behavior also depends on social context. Using the AddHealth data, she shows that 1) self-reported impulsivity is correlated with school suspensions and with crime; 2) African-Americans are much more likely to be suspended from school; 3) there are no racial differences in self-reported impulsivity. Thus it is problematic to use school suspensions as a behavioral measure of impulsivity, since suspensions are also determined by school context, racial discrimination and other unknown factors.

I think this critique is extremely important, and it points out deeper issues with the measurement of non-cognitive skills. Sometimes measures are too predictive – or alternatively, they are predictive *because* the underlying construct is invalid. School suspensions capture some measure of the student's impulsivity, but also what type of school they attend, their gender and race, and many other things. In these situations, one's confidence in the ability to use the behavior as a proxy for skills hinges on one's ability to control for everything else that is important. This is a classic omitted variables bias problem – the behavior (school suspensions) captures the underlying skill, but also many other things.

I must note that Kautz et al (2015) includes a very careful discussion of the pros and cons of these issues, and Borghans et al (2011) go into even more detail on identification issues in the use of behavior measures. So the authors are not unaware of these concerns.

Nonetheless, I think the issue of construct validity is mostly underappreciated in the literature. There is simply no substitute for careful development of a theoretically sound underlying construct. We will never be able to measure "non-cognitive" skills well if we do not understand what we are measuring.

The way forward

One pessimistic response is that we will never be able to measure non-cognitive skills well because non-cognitive skills do not exist. The most reductive view, which we have all seen from time to time, is that IQ is everything. While it is easy to reject this extreme form of the argument, it is not so easy to reject a weaker form that cognitive skills are more important predictors of life outcomes than non-cognitive skills. This argument starts with the observation that cognitive tests are both more predictive and more reliable than non-cognitive measures, whether self-assessed or behavioral.

However, if measurement error in skills is classical, then the coefficient on skills in a regression with an outcome such as log wages will be attenuated toward zero, with the degree of attenuation

decreasing in the reliability of the measure. Thus, if “non-cognitive” skills are measured more poorly than cognitive skills, we will tend to underestimate their importance.

More broadly, we must recognize that measures of cognitive and non-cognitive skills do not just appear from nowhere. Rather, they are developed over many years and by many different researchers, often for an initially narrow purpose. The modern IQ test was created as a means to diagnose mental retardation in school children, with lower scores simply indicating that children were unable to perform tasks that were “typical” for their same-age peers. The later reification of “g” as general intelligence was based on the observation that childrens’ grades and test scores can be statistically best explained by a single common factor.

All this is to say that the scholarly consensus about the importance of different human capacities is often driven by how well these capacities can be measured. For example, if we could develop reliable and context-invariant tests of important “non-cognitive” capacities such as self-control and social intelligence, I would not be surprised if they ended up being better predictors than IQ of labor market outcomes.

Here I am optimistic that we can more fruitfully exploit comparative advantage between psychologists and economists. Psychologists have carefully developed measures that map cleanly to underlying constructs, but they have (for the most part) not subjected these measures to rigorous testing in a variety of field settings. Economists, on the other hand, have gleefully used convenient, off-the-shelf measures of questionable validity (NB I am as guilty as anyone in this regard) to make broad generalizations about the importance of non-cognitive skills, with an exact definition of these skills TBD. When it comes to “non-cognitive” skills, we economists are the proverbial drunk searching under the street lamp for his keys, because that is where the light is located.

I will close with a specific example of this possible complementarity across disciplines. The Reading the Mind in the Eyes Test (RMET) is a test of emotion recognition or social sensitivity developed by Simon Baron-Cohen and colleagues (e.g. Baron-Cohen et al 2001). The RMET was originally created for a narrow purpose – to diagnose so-called "theory of mind" deficits such as Asperger Syndrome and high-functioning autism in otherwise well-functioning adults. However, much like IQ, psychologists have discovered that the RMET has predictive power for a wide variety of outcomes within a general population. Woolley et al (2010) randomly assign participants to teams and find that the team's average score on the RMET predicts task performance after controlling for group average IQ.

While the RMET is not perfect, it is superior to many other measures of "non-cognitive" skills in at least two respects. First, the RMET overcomes some of the limitations of self-assessment because there is a correct answer to the question items. This prevents reference group bias as well as strategic responses in high-stakes settings. Second, there is a well-grounded theory of how the underlying capacity (theory of mind) relates to task performance (emotion recognition in human faces), and in turn how task performance relates to outcomes (see Deming 2017 for a more thorough discussion of the connection between social skills and labor market success). This helps with the concern that a poorly defined construct measures "too much".

There are many studies in psychology journals that probe the validity and reliability of the RMET and other measures of social and emotional intelligence across settings, samples and cultures. A recent meta-analysis finds a modest positive correlation of about 0.25 between IQ and the RMET (Baker et al 2014). Most of the studies in this meta-analysis rely on small convenience samples.

What we do not have – and what I am hoping economists can provide – is a sense of how the RMET or other measures of social intelligence vary in a broader population. What is the correlation between social intelligence and measures of SES such as income and parental education? Does the RMET predict

life outcomes at all, and is it differentially predictive for key subgroups? These are only initial questions in what I hope is an emerging paradigm – improving the theory and measurement of "non-cognitive" skills.

References

Baker, C. A., Peterson, E., Pulos, S. and Kirkland, R. A.: 2014, Eyes and iq: A meta-analysis of the relationship between intelligence and reading the mind in the eyes, *Intelligence* **44**, 78–92.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. and Plumb, I.: 2001, The reading the mind in the eyes test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism, *Journal of Child Psychology and Psychiatry* **42**(2), 241–251.

Borghans, Lex, Bart HH Golsteyn, James Heckman, and John Eric Humphries. "Identification problems in personality psychology." Personality and Individual Differences 51, no. 3 (2011): 315-320.
Heckman, James J. & Kautz, Tim, 2012. "Hard evidence on soft skills," Labour Economics, Elsevier, vol. 19(4), pages 451-464.

Deming, David J. 2017. "The Growing Importance of Social Skills in the Labor Market," *Quarterly Journal of Economics,* 132(4): 1593-1640.

John, Oliver P., and Sanjay Srivastava. "The Big Five trait taxonomy: History, measurement, and theoretical perspectives." *Handbook of personality: Theory and research* 2.1999 (1999): 102-138.

Kautz, Tim, James J. Heckman, Ron Diris, Bas ter Weel, Lex Borghans. Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success. OECD (2015).

Schmitt, D. P., J. Allik, R. R. McCrae, and V. Benet-Mart´ınez (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. Journal of Cross-Cultural Psychology 38 (2), 173–212.

West, Martin R., et al. "Promise and Paradox Measuring Students' Non-Cognitive Skills and the Impact of Schooling." *Educational Evaluation and Policy Analysis* (2015): 0162373715597298.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. and Malone, T. W.: 2010, Evidence for a collective intelligence factor in the performance of human groups, *Science* **330**(6004), 686–688.