Measuring Skills in Developing Countries

Rachid Laajaj* and Karen Macours**

September 2019

ABSTRACT

Measures of cognitive, noncognitive, and technical skills are increasingly used in developing country surveys, but have mostly been validated in high-income countries. We use a survey experiment in Western Kenya to test the reliability and validity of commonly used skills measures. Cognitive skills measures are found to be reliable and internally consistent, technical skills are very noisy, and measurement error in noncognitive skills is found to be non-classical. Addressing both random and systematic measurement error using common psychometric practices and repeated measures leads to some improvements and clearer predictions, though concerns remain. These findings hold for a replication in Colombia.

JEL codes: O12, O13, O15

Keywords: skills, measurement, agricultural productivity

* Rachid Laajaj is an associate professor at the Universidad de Los Andes. Corresponding author's email: r.laajaj@uniandes.edu.co

** Karen Macours is a professor at Paris School of Economics and INRA researcher[1]

The data and programs for full replication are available online at https://www.laajaj.com/research

[1] Jack Pfeiffer provided invaluable research assistance and programming skills, field support, tenacity and ideas throughout the project and we gratefully acknowledge all his contributions to this project. Manuel Camargo, Irene Clavijo, Katriel Friedman, Juan Diego Luksic, Freddy Felipe Parra Escobar, and Mattea Stein provided excellent research assistance. Data collection was organized through IPA Kenya and S.E.I. in Colombia and we acknowledge the excellent field teams for key contributions during piloting and translation and for all efforts during the data collection. We are indebted to Chris Soto for many suggestions and insights from the psychometric literature. We are grateful for inputs received from colleagues at PSE and seminar participants at Bristol, EUDN, IFPRI, Mannheim, Navarra, Nova, Oxford, Rosario, Trinity College Dublin, World Bank, the ISI World Statistics Congress, to Gero Carletto and Renos Vakis for continuous encouragements and support, to the World Bank's Living Standards Measurement Study (LSMS) team for funding of the survey experiment through a grant from DFID, and to DFID-ESRC and the Standing Panel on Impact Assessment (SPIA) of CGIAR under the grant 'Strengthening Impact Assessment in the CGIAR (SIAC) for funding of the follow-up data collection rounds in Kenya. All errors are our own.

doi:10.3368/jhr.56.4.1018-9805R1

I. Introduction

Cognitive and noncognitive skills are often considered key to understand economic decision-making. Empirical work with data from the US and Europe has made important advances in understanding both the causes and the consequences of skill formation (Heckman, 2007). Increasingly, cognitive, noncognitive and technical skills are the focus of analysis in development economics, with recent work on the determinants of skill formation (Attanasio et al, 2015), and the importance of skills for later life outcomes (Gertler et al, 2014). Development economists have also long worried about the role of many hard-to-observe skills as potential confounders in empirical analyses.

Low level of skills are often seen by policy makers as key constraints to reducing poverty, and large investments are made in training programs aimed at improving skills (1 billion US$ a year by the World Bank alone), even if there are many questions regarding their effectiveness (McKenzie and Woodruff, 2012; Blattman and Ralston, 2015). Low levels of skills among farmers in developing countries are thought to be one of the main drivers of productivity differences between sectors in the economy (Lagakos and Waugh, 2013). Young (2013) argues that sorting on skills explains the urban-rural productivity gaps observed in most developing countries and Gollin, Lagakos, and Waugh (2014) show that productivity differences become smaller when accounting for observed human capital differences. Understanding such potential selection at the micro-level arguably requires measures that go beyond years of schooling attained and more complex measures of skills are sometimes included in household surveys. Yet the measurement of skills in developing country field conditions poses substantial challenges, and the related measurement error and its implications for empirical work have received little attention.

This paper starts to address this gap with the results of a skill measurement experiment. It documents the measurement challenges and discuss potential solutions and implications. We designed and implemented a relatively large survey focused on measuring different types of skills, and use the data to shed light on the reliability and validity of a wide set of commonly used skill measures and on the predictive power of the measured skills. We refer to cognitive skills as those that capture hard skills such as abstract reasoning power, language and math skills; noncognitive skills capture soft or socio-emotional skills, including a wide set of personality traits and facets, such as self-esteem, tenacity, conscientiousness, locus-of-control, and attitudes-to-change. The measure of technical skills focuses on agricultural knowledge and know-how, given that the data come from poor rural individuals whose main occupation is in agriculture.[2]

There is a wide variety of existing questions, tests or modules to measure skills. Many instruments have been designed to assess skills in lab conditions, and some standardized instruments have been developed for inclusion in surveys in developed country settings. Increasingly, economists are also including measures of abilities and personality traits in household surveys conducted in developing countries. But little validation of survey instruments has occurred for such contexts.[3] Many questions can be raised about the applicability of some of the existing scales for poor rural populations, given the high level of abstraction of many questions, low levels of education in the respondent population, difficulties of standardization for enumerator-administered tests conducted in the field, and translation challenges.

This study aims to test the reliability and validity of several commonly used scales and tests, and highlights both random and systematic measurement error that needs to be accounted for when measuring skills in developing-country field conditions. The main analysis draws on a sample of 900 farmers in Western Kenya and first analyzes the test-retest reliability and the Cronbach's alpha to estimate internal consistency of various existing scales. We subsequently use exploratory factor analysis, correction for acquiescence bias, and item response theory (IRT) to reduce measurement error, and analyze validity and reliability of the improved constructs.[4] We then study the predictive validity of both the original scales and the improved constructs and analyze the extent to which the skills measured predict agricultural productivity and economic decisions related to the adoption of advanced agricultural practices.[5] This tests the potential role played by these skills in agricultural production, and shows that they might be important omitted variables when not included in the analysis of agricultural decision-making.

Almund et al. (2011) suggest that different skills and personality traits might help predict different outcomes, with cognitive ability being more important for complex tasks, while personality is shown to be more important for job performance. This study focuses on a specific population and consider outcomes that relate to their main occupation, farming. Even if the most highly skilled

[2] Hence we broadly follow the distinction of Heckman and Kautz (2012) who distinguish between cognitive abilities, personality traits, and other acquired skills. Jones and Kondylis (2018) is a recent example using a measure of acquired agricultural skills, in which measurement error is being discussed as potential reason for lack of impact.

[3] Measures of risk aversion and time preferences, which have a longer history of use in developing country surveys, have received more scrutiny. Chuang and Schechter (2015) review the evidence, including their stability over time.

[4] While explanatory factor analysis is used elsewhere in the economics literature on skills (Cunha and Heckman, 2010; Heckman, Stixrud and Urzua, 2006), we also build on insights from the psychometrics literature, such as for the corrections for acquiescence bias (the tendency to agree even when statements are contradictory, or "yea-saying").

[5] Following McKenzie (2012), in order to reduce the noise in the outcome variables, the measures of yield and practices are obtained from the average over the four seasons that followed the collection of skills data.

may have selected out of this occupation, understanding the importance of skills for agricultural decisions and productivity is important in its own right, given that the majority of the world's poor continue to live in rural areas, where agriculture remains the most important source of employment (World Bank, 2008). Differences in the willingness to exert effort are often considered key to understand heterogeneity in agricultural outcomes (de Janvry, Sadoulet, and Suri, 2016). More generally, farmers face many different tasks and decisions, some of which may depend more on knowledge, others on problem solving ability, and yet others on effort. In the predictive regressions, we consider a variety of outcomes to capture those potential differences, and analyze to what extent different skills explain a meaningful part of the variation in outcomes.

Our first set of results show that cognitive skills, using both standardized scales and tests developed for the specific context, can be measured with high levels of reliability and validity, similar indeed to those found in developed country settings. Cognitive skills also show good predictive validity, even after controlling for educational levels. On the other hand, we find that standard application of commonly used scales of noncognitive and technical skills suffer from large measurement error, resulting in low reliability and validity. For technical skills, factor analysis and item response theory results in a construct with higher predictive validity, even if large measurement error remains. Repeated measurement further helps to improve predictive power, and most of the measurement error in technical skills appears to be random measurement error. A possible explanation for this measurement error is the heterogeneity in optimal agricultural practices, which can be highly contextual and therefore perceived correct answers to technical skills questions may well vary between farmers and over time.

For noncognitive skills, the evidence suggests systematic measurement error. Combining questions according to pre-existing scales leads to low internal consistency and the latent noncognitive construct resulting from the factor analysis does not map in the personality domains typically found in developed countries. While the corrected noncognitive constructs are predictive of agricultural productivity, the estimates do not allow drawing conclusions about the relevance of specific noncognitive skills. Overall, the best predictions obtained after corrections of different sources of measurement error show that the three skill domains together explain up to 14 % of the variation in yield, with all three constructs being significant and with similar point estimates. Technical and noncognitive skills also help predict some agricultural practices and input use.

The last part of the paper analyzes the different challenges for measuring skills in household surveys in developing countries and discusses guidelines on how to address them. Building on the randomized allocation of enumerators to respondents, we show the potential key role of the interaction with enumerators to explain some of the identified measurement problems, a finding that, to the best of our knowledge, has not been quantified before. Other challenges include the respondents' ability to understand the questions, order effects, response biases, anchoring, different factor structures, and the idiosyncrasy of agricultural knowledge. Finally, we replicate key parts of the experiment and analysis with 804 farmers in Colombia and show that the results present similar patterns in this very different context and language.

The large amount of measurement error this study documents provides an important warning sign for studies trying to use similar measures in poor rural settings. Measurement error, when it is classical, could lead to important attenuation bias and lack of statistical power. This might well lead to an underestimation of the importance of skills for decision-making, or of the impact of external interventions on cognitive, noncognitive or technical skill formation. This can have important implications for sample size calculations and might point to usefulness of measuring individuals' skills at several points in time to reduce such error. [6] Yet, the evidence also suggests the measurement error in skills might well be non-classical, and could hence more broadly lead to erroneous conclusions. The results in this study – intended as a proof of concept - show that it can be particularly hard to distinguish different aspects of noncognitive skills at least in some rural developing contexts. This suggests that studies measuring only a subset of noncognitive skills need to be careful regarding the interpretation of which latent factor is being measured.

The measurement challenges identified in this paper relate to a wider literature in economics, which has highlighted measurement concerns for attitudinal, expectations or aspirations questions in both developed and developing countries (Bertrand and Mullainathan, 2001; Krueger and Schkade, 2009; Manski, 2004; Delavalande, Gine, and McKenzie, 2011; Bernard and Taffese, 2014; Bond and Long, 2018). The analysis of noncognitive skills also relates to the psychometrics literature on the validity of Big Five personality trait measures across cultures (Benet-Martinez and John, 1998; John, Naumann, and Soto, 2008). Using data from self-administered surveys collected mostly from college-students in high and middle-income settings, a number of papers argue that the Five Factor Model (FFM) is universal (McCrae and Costa, 1997; Piedmont et al, 2002; McCrae and Terracciano 2005), including in Africa (Schmitt, 2007). [7] Others claim, however, that no one scale can apply to all cultures (Cross & Markus 1999) and show, for instance, that an orally administered

[6] The observation that measurement error might be substantially higher for noncognitive than for cognitive skills, and the limitations of the use of Big Five questionnaires in large-scale surveys have also been pointed to by Borghans et al (2008) as potential reasons for underestimating the importance of noncognitive skills in developed country settings.

[7] McCrae and Terracciano (2005) administrated the survey to observers (asking one person about someone else's personality traits) but it was self-administrated, i.e. filled by the respondent rather than asked by an enumerator.

survey on a forager-horticulturalists indigenous population in the Bolivian Amazon does not find support for the FFM model (Gurven et al, 2013). In related work, we find strong evidence of lack of congruence for Big Five measures collected in household surveys across 23 low and middle income countries (Laajaj et al. 2019). These results confirm the concerns about the applicability of such measures in household surveys, raised in this paper. To the best of our knowledge, there are however no validation exercises using household survey measures of other non-cognitive skills, cognitive nor technical skills in developing countries even if they are commonly and increasingly used in the micro-development literature.

This paper also relates to a large literature on the importance of cognitive and noncognitive functioning in household economic decision-making. Cognitive ability has been shown to be an important predictor of socioeconomic success (Heckman, 1995, and Murnane, Willett, and Levy, 1995). Heckman, Stixrud, and Urzua (2006) and Heckman and Kautz (2012) argue that noncognitive abilities matter at least as much, despite the historically strong focus on cognitive ability. In developed countries, evidence shows noncognitive abilities and personality traits to be related to a large set of socio-economic outcomes such as wages, schooling, crime, and performance on achievement tests (Bowles, Gintis, and Osborne, 2001; Heckman, Stixrud, and Urzua, 2006; Cunha and Heckman, 2010; Almund et al. 2011). In the psychology literature, there is also substantial evidence on the predictive power of personality traits for socio-economic outcomes. Development economists are also increasingly including measurements and analysis of personality traits in empirical studies (Dal Bo, Finan, and Rossi, 2013; Callen et al., 2015).

The insights from this paper are relevant for various strands of the wider literature. In light of the large debate about whether the worldwide increase in schooling is leading to measurable and sustained gains in learning (Pritchett and Beatty, 2015), having widely comparable measures of cognitive abilities, that can be measured for adults, outside of the classroom, and in a variety of settings, is arguably key. Certain large data collection efforts covering wide and heterogeneous populations, such as the Young Lives Surveys or the World Bank STEPS surveys (Pierre et al., 2014), are now including measures for noncognitive abilities and personality traits. Increasingly skills measures are also included in impact evaluation surveys, specifically when interventions aim to change noncognitive traits (Bernard et al., 2014; Groh, McKenzie, and Vishwanath, 2015; Blattman, Jamison, and Sheridan, 2017; Ghosal et al., 2016; Adhvaryu, Kala, and Nyshadnam, 2016) but also when changes in noncognitive abilities are seen as potential mechanisms to explain changes in final outcomes (Blattman and Dercon, 2016). Moreover, there is a growing literature on learning, technology adoption and agricultural productivity in developing countries which often relies on measures of agricultural knowledge and learning (Jack, 2011; de Janvry, Sadoulet, and Suri, 2016). Finally, while this paper focuses on measures during adulthood, skills start to develop much earlier in life and poverty during early childhood can lead to very serious cognitive delays and affect socio-emotional development (Grantham McGregor et al., 2008). Growing evidence furthermore suggests a strong link between early childhood and adult outcomes, including recent work focusing on the long-term impact of external factors during childhood on noncognitive outcomes of adults (Leigh, Glewwe and Park, 2015; Krutikova and Lilleor, 2015). As such, a better measurement of adult skills can contribute to better understand the long-term returns to social policies targeting skill development in early childhood and beyond.

The paper is organized as follows: the next section provides more information about the context, the instrument and the implementation of the survey experiment. Section 3 provides the intuitive description of, and rational for the improved constructs, and discusses reliability and internal consistency. It also shows predictive validity results, using agricultural yield and practices as outcome variables, and comparing results with the naïve and the improved constructs. Section 4 presents additional analysis and derives lessons and practical recommendations for skills measurement. Section 5 then discusses the main findings for Colombia and section 6 concludes. Finally, the online appendix provides the technical details on psychometric concepts and methods used, documents the construction of the indices and the details of the questionnaire design. Together with the available code, it aims to provide guidance for other researchers. The appendix also contains the details on the Colombian replication and other secondary empirical results.

II. Setting, Sample and the Questionnaire Design

A. Setting

The survey experiment was conducted in Siaya province in Western Kenya targeting 960 farmers, spread across 96 villages and 16 sub-locations, of whom 937 were reached for the first measurement, and 918 for the second measurement. Half of the farmers were selected from a stratified random draw of farming households in each village, the other 50% were farmers nominated in village meetings for participating in agricultural trials. The village list was either pre-existing or drawn up by the village health worker. Given the individual nature of skills, the sample is a sample of individuals, identified as being the main farmer in the selected households. Farmers were surveyed twice with an interval of about three weeks between the test and retest.

Respondents have on average 6 years of education (substantially below the Kenyan average), a bit more than half of the respondents are female, 62% are head of household, and are on average 46 years old. Farms contain on average about 3 plots, and 65% of households own at least some cattle. Maize is the main staple crop, and is often intercropped or rotated with beans. Many farmers also have root crops and bananas.

B. Questionnaire design

The main instrument consists of 3 modules (cognitive, noncognitive, and technical agronomical skills) that were asked in random order. This section summarizes the content of each module as well as the rational for the choice of questions and tests. Appendix C provides a more comprehensive description of the questionnaire.

Many instruments have been designed to assess cognitive and non-cognitive skills in lab conditions, or among highly educated respondents in high-income settings. They have subsequently been integrated in survey instruments that are applied in field conditions, often without prior testing of their suitability. We therefore aim to test the validity of existing cognitive and non-cognitive scales administered in rural field conditions. An extensive desk review of papers allowed making an initial selection of questionnaire modules and questions that are similar to approaches used elsewhere in the literature. For technical skills, rather than starting from specific questions, we focus on different types of questions found in the literature.

1. *Cognitive skills*

With the objective of measuring different aspects of adult farmers' cognitive ability, we selected five cognitive tests: i) The Raven Colored Progressive matrices, measuring visual processing and analytical reasoning; ii) The digit span forwards and backwards, measuring short-term memory and executive functioning; iii) A written and timed test of basic math skills; iv) An oral 9-item test containing short math puzzles relevant for agriculture and increasing in difficulty level; and v) A reading comprehension test. Table A1.A in Appendix C provides detailed descriptions of each of the tests.

2. *Noncognitive skills*

The noncognitive part focuses on testing instruments derived from commonly used scales in noncognitive domains that the literature found to be predictive of success in life and that are potentially relevant for smallholder farmers. We use a subset of items from the 44-item BFI, a commonly used instrument for the Big Five personality traits. We also test commonly used instruments for lower-order constructs such as locus-of-control, self-esteem, perceptions about the causes of poverty, attitudes towards change, organization, tenacity, meta-cognitive ability, optimism, learning orientation, and self-control. The majority of these subscales are derived from a set of questions asking the respondent the level at which he agrees or disagrees with general statements about himself, with answers on a Likert scale from 1 to 5.[8]

In addition, we asked a set of locus-of-control questions with visual aids in which people are asked to attribute success to effort & good decisions, luck or endowments. We also included the CESD, a commonly used depression scale, validated in many developing countries, as it relates to some noncognitive domains captured in other scales (neuroticism and optimism). A standard risk aversion game and time preference questions were added, for comparison and completeness.

Table A1.B in Appendix C presents all items, and the first column indicates the sub-scale each of the items belongs to. As is the case in the original scales, some questions are positively-coded, indicating that a higher likelihood to agree with the statement indicates a higher score on the relevant noncognitive trait, while others are reverse-coded. The last column in Table A1.B shows which questions are reversed.[9] While the pilot revealed that reverse-coded questions were sometimes harder to understand (often due to negative phrasing), care was given to keep approximately equal number of positively and reverse-coded items, as they are key to detect acquiescence bias. For a few questions a binary choice was used instead of a Likert scale.

3. *Technical skills*

There are no standardized scales that measure technical skills, reflecting the fact that agricultural knowledge can be very specific to a geographical area, crop and type of inputs or practices. That said, different types of questions can be found in the literature, reflecting different underlying ideas about which knowledge could be the most relevant ones: questions on the timing at which inputs should be used, how to apply the inputs (quantity, location, etc.), knowledge of both basic and more

[8] The causes-of-poverty subscale does not ask directly about the respondents themselves but uses a Likert scale to ask about reasons for why poor people are poor.

[9] For neuroticism and CESD, we use reverse coding to refer to higher levels of neuroticism and stress, as lower neuroticism and stress should imply a higher noncognitive score.

complex practices (spacing, rotation, composting, conservation…), and general knowledge (the active ingredients in certain fertilizers). Based on this categorization, we then worked with local agronomists to design a module aimed at capturing agricultural knowledge relevant for the farmers in the study population, by covering production of the main crops and the use of the most common practices and inputs in Western Kenya. We use a mix of open questions and multiple-choice questions, some questions allow multiple answers, and a subset of questions had visual aids (e.g. pictures of inputs). The set of questions covered a relatively broad spectrum of practices, including a set of questions on maize, banana, soya, soil fertility practices, composting and mineral fertilizer. Table A1.C in Appendix C presents all questions, and the first column indicates the sub-scale each of the questions was grouped under.

4. *Piloting and questionnaire preparation*

We conducted extensive piloting of these modules and questions. Qualitative piloting allowed testing face validity, by asking qualitative follow-up questions regarding the understanding of the questions and meaning/reasoning of the answers. After qualitative piloting, an extended version of the skill questionnaire was piloted in November 2013 on 120 farmers in Siaya, close to the study area, and on farmers that had been selected in a similar way as those of the actual study population. A small subset of these farmers was also retested in December 2013 with the same survey instrument, in order to obtain retest statistics of the pilot. Based on this quantitative pilot, we eliminated questions with little variation.[10] As a consequence of these extensive piloting, the results presented in this paper should be interpreted as representative for studies with similar piloting, but an understatement of the amount of noise in the case of studies with limited effort to adapt the survey instruments and train enumerators.

We also removed questions that showed negative correlations with other variables meant to capture the same latent trait, and fine-tuned phrasing and translation of questions.[11] The final survey instrument took about 2.5 hours to complete. The vast majority of farmers in the sample (97%) were native Luo speakers (the local language) and were interviewed in Luo. The others were more comfortable in Swahili or English (Kenya’s two official languages), and hence interviewed in their

[10] For instance, experience with pesticides or irrigation is extremely limited in the population of study, so that any related questions did not provide variation.

[11] For the noncognitive module, a relatively large set of questions was identified with either very little variation (everybody agreed with a certain positive statement), or a bi-modal distribution, typically in the case of reverse-coded questions. In extreme cases this led to negative correlations between variables that should capture the same latent trait.

language of choice. The English-language survey therefore was translated in both Luo and Swahili. All versions were homogenized after independent back translation.[12]

C. Alternative measures of skills

Prior to the set of questions in the three main modules described above, respondents were asked their self-assessment for the same set of skills using a set of 14 questions, formulated to proxy the different subdomains captured by the questions in the main modules. And after answering all questions from the three main modules, each farmer was asked to assess the skill level of one of the other farmers from her village in the sample using similar proxy questions. This provides an independent (though clearly subjective and possibly mis-measured) assessment. A second proxy measure comes from asking the same questions to another household member (typically the spouse) also involved in farming. And a third independent measure was obtained prior to the survey from the village health worker, who was asked to classify each farmer according to his cognitive, noncognitive, and technical abilities, using a broad categorization (high, medium, low). The predictive power of these three proxy measures can be compared with the predictive power of the detailed skills measures, an issue we turn to in section 4.

D. Randomization of the survey instrument and fieldwork

To understand the drivers of measurement error, an important focus of the study was the extent to which the order of answers, of questions, and of modules, or any unobserved enumerator effects might affect answers. The data collection was done using mini laptops, and a program specifically designed to randomize the different components of the questionnaire. The order of the three main modules (cognitive, noncognitive and technical) was randomized, which allows to control and test for potential survey fatigue and to assess whether some tests tend to modify the responses of the following questions. The order of the questions within a module was randomized to control for a potential learning caused by the preceding questions. And in all multiple-choice questions, the order of the answers was also randomized. In order to test for enumerator effects, we also randomly assigned respondents to enumerators. For the re-test 40% of households was assigned to the same

[12] Back-translation initially revealed a substantial number of questions with translation problems, in particular in the noncognitive part. As noncognitive questions are often more abstract and use concepts that are not part of daily vocabulary, finding the appropriate translation can be a challenge. For modules, translations and back-translations were compared, and we worked together with native Luo and Swahili speakers to finalize translations, to assure that the original meanings of the questions were maintained (and hence to know which questions we are in fact testing). We suspect that similar translation issues affect other surveys trying to obtain answers related to more abstract concepts.

enumerator. Survey teams were allowed to deviate from the random assignment for logistical reasons. Overall compliance with the enumerator assignment was about 75%. Finally, we randomized the order in which the villages were surveyed to evaluate effects related to enumerators learning or changing how they administrate the survey over time.

E. Training and data collection

Prior to survey implementation, all field personnel participated in an intensive two-week training, with both classroom and field training and extensive practice to guarantee fluent and correct implementation of the different skill measurements. The first round of the survey started January 20th 2014, after the harvest of the 2013 agricultural season– and took approximately 3 weeks. The retest survey was conducted in the following 3 weeks. A small household and farm survey was implemented in parallel and provides the agricultural outcome variables. All survey activities, including tracking of harder to reach respondents were finished by the end of March. Almost all surveys were conducted before the start of the main agricultural season in 2014. Additional survey rounds were implemented at the end of the following four agricultural seasons with information on production outcomes and practices, and are used to investigate which skills best predict these economic outcomes.

III. Reliability and Validity of Different Skills Constructs

We aim to test the reliability and validity of the different skill measures. Reliability indicates the share of informational content (rather than noise) of a measure of a given skill and validity indicates whether it actually measures what it intends to measure. To do so we calculate for each measure the test-retest correlation, a pure reliability measure, and Cronbach's alpha, which is affected both by the noise and the extent to which items are measuring the same underlying construct (construct validity). We also test the predictive validity, by analyzing whether the skill measures predict different agricultural outcomes that they are theoretically expected to be correlated to. Appendix A and B provides a detailed methodological explanation of the different tests and methods used.

For each domain (cognitive, noncognitive and technical skills), we construct different measures. By comparing the reliability and validity of these different constructs, we demonstrate the importance of accounting for response patterns and latent factor structure. A "naïve" score aggregates the different questions using the existing sub-scales meant to measure certain abilities as they were included in the survey instrument. This has the advantage of simplicity and transparency, and mimics what is often done in practice. We also construct alternative “improved” aggregate measures, relying on different corrections to extract the most relevant information from the available items: we use exploratory factor analysis to determine the number of factors in each construct, item response theory to further improve the cognitive and technical constructs, and correct for acquiescence bias in the noncognitive construct.

A. Exploratory factor analysis, acquiescence bias corrections, and item response theory

The results of the exploratory analysis indicate that the cognitive and technical skills can best be measured by one factor each, while the underlying latent factors for noncognitive skills corrected for acquiescence bias are best captured by 6 factors (Appendix Table A2). For the noncognitive skills, the factor analysis does not result in a clear categorization of variables into the theoretical scales and subscales (see Appendix B3 for details) and, most factors seem to have a mix of items from different sub-constructs (in theory meant to be measuring different latent skills).[13] Overall these results raise concerns about whether the scales actually measure what they intend to. We use the factor loadings to aggregate the different noncognitive skills and label each factor based on the type of questions that most frequently appear with high factor loads in the construct (Table 1). To obtain the aggregated non-cognitive skills construct, we use the average of the 6 factors.

Prior to the factor analysis, all variables were corrected for acquiescence bias (see Appendix B1) building on methods from psychometrics that use agreement to contradictory statements to measure the acquiescence bias. McCrae et al. (2001) find that not correcting for it prior to factor analysis often results in the emergence of a factor representing the response pattern.

For the cognitive and technical skills, Item Response Theory imposes further structure on a set of items to measure an underlying latent ability or trait. It assumes that the probability of getting the correct answer to a question (or a higher score on a given item) depends on the unobserved ability of the respondent and some parameters of the question, estimated simultaneously (see detailed explanation in Appendix section B5).

B. Reliability and construct validity

1. Test-retest correlation

[13] This means for example that a question that is expected to measure agreeableness and a locus of control question can better correlate together (and thus be assigned to the same underlying factor) than two locus of control questions.

To test reliability, we calculate the correlation between the same construct measured twice over a period of three weeks (see Appendix A3 for details). The first column of Table 2A shows the test-retest correlations of the "naïve" aggregate and of the sub-constructs by predefined subdomains. The results vary widely. The cognitive naïve construct reaches a test-retest correlation of .83 (with the test-retest for individual tests between .37 and .82, but only the digit span sub-constructs lower than .6) indicating a high degree of reliability, comparable to what is often obtained in lab or class-room conditions. By contrast, the noncognitive and technical test-retest correlations are .54 and .31 respectively, which is strikingly low given the large set of items used to compute them.[14] This probably points to a large role for guessing and possibly general uncertainty about the answers. Unsurprisingly given that the number of items reduces the noise, sub-constructs perform worse than the aggregate constructs. Among the noncognitive ones, test-retest statistics are slightly higher for locus of control, CESD and causes of poverty than for other sub constructs.[15]

The first column of Table 2B provides the test-retest correlations of the "improved" constructs and sub-constructs, calculated as described in section 3.1. Compared to the naïve constructs, the test-retest statistics are marginally higher for the cognitive and noncognitive skills and substantially higher for the technical construct (from .31 to .41). Hence the use of IRT, factor analysis and correction for acquiescence bias substantially improves the reliability of the constructs. That said, test-retest statistics remain below standard thresholds for the noncognitive sub-constructs and particularly low for the technical skill construct.[16]

2. Cronbach's Alpha

The Cronbach's alpha is one of the most widely used measures of internal consistency of a test. For a given number of items, it increases when the correlation between items increases. Hence it is

[14] For comparison, in meta-analysis of personality scales, mostly in developed countries, prior research has found average test-retest correlations between 0.73 and 0.82 (Schuerger et al. 1989, McCrae et al. 2011).

[15] For CESD, it is a priori not clear that answers should be stable over three weeks, as the reference period of the questions is the last week, and as mental health might be malleable in the short run. But in related work, Krueger and Schkade (2009) find that the test-retest reliability of a general life satisfaction question was no better than questions asking about effective experience on specific days, and attributed this to transient influences influencing the former.

[16] The fact of being surveyed during the test may affect the answers in the retest, and hence the test-retest statistic. Appendix Table A4 shows that indeed scores are slightly higher in the retest for all 3 skill constructs. To the extent that scores increase for all respondents this does not affect the test-retest statistics, as scores are standardized within survey round. Moreover, the standard deviations in the test and the retest for cognitive and noncognitive scores are very similar. They are however slightly lower for the technical scores in the retest than in the test, potentially indicating a learning effect by either the respondents, the enumerators, or both. Appendix Table A.12 shows that the item-level changes in correct answers for the technical items, further suggesting some learning. Results in section 4 further suggest that at least part of this learning is enumerator related. Such learning does not appear to explain the low test-retest statistics, however, as removing items that show a significant difference between average test and retest scores (9 items with 1% significant difference, or 16 items with 5% significant difference) leads to test-retest statistics between .32 and .35.

higher when the noise of each item is low (high reliability) and when they actually measure the same underlying factor (indicator of high validity). See Appendix A4 for details.

The second and third columns of Table 2A show the Cronbach's alpha of the naïve constructs of the test and retest, while Table 2B provides similar statistics for the improved constructs.[17] The conclusions for the aggregate constructs parallel those obtained from the test-retest correlations. The Cronbach's alpha is above the commonly used threshold of 0.7 for the cognitive skill construct, somewhat acceptable in the case of the noncognitive, and substantially below the acceptable threshold in the case of the technical skills.[18] For comparison, Schmitt et.al (2007) finds that the Cronbach's alpha of the Big Five in Africa for self-administered surveys among mostly college students were between 0.55 and 0.68, which was the lowest of all regions. For the same subconstructs, we find Cronbach's alphas between 0.31 and 0.51. The Cronbach's alphas do not differ much between the test and retest, which confirms that the retest is broadly comparable to the test. Cognitive sub-constructs with a large number of items reach very high Cronbach's alpha, as does the CESD. The alpha for the naïve causes-of-poverty sub-construct is also high, but recall that the factor analysis suggests this correlation may be driven by common response patterns rather than common meaning.

The Cronbach's alphas of the improved aggregate constructs are not higher than the ones of the naïve constructs, but the ones of the 6 noncognitive factors generally show large improvements compared to the naïve sub-constructs. These two observations are partly mechanical given that the factor analysis pools together items with higher correlations in the subconstructs, and the correlation between factors is minimized through the quartimin rotation.[19] The technical skills construct reaches a Cronbach's alpha of .54, which remains quite low given that it includes 32 items. This suggests that farmers' knowledge might be idiosyncratic (with different farmers having different pieces of knowledge), and therefore hard to aggregate in a knowledge score.

C. Predictive validity

In the psychometric literature, predictive validity consists in testing whether the skills constructs correlate with outcomes that should, in theory, be correlated with the skills. Hence in this section

[17] Interpretation of Cronbach's alpha should take into account that it tends to increase with the number of items.

[18] The high Cronbach's alpha of the cognitive construct is consistent with Table A1.A in Appendix C showing that scores of the 5 sub-components are highly correlated with each other. The correlations are highest among skills most clearly acquired in school (reading and the two math tests) and a bit lower with the more general cognitive tests (Raven and digit span). Correlations are also high with grades of education attained and self-assessed literacy.

[19] When we do not apply the new factors weights, but only correct scores for the acquiescence bias, neither the alpha's nor the test-retest systematically improve (Appendix Table A5).

we first look at correlations, then run regressions to test the extent to which skills predict agricultural productivity and practices. The estimates capture conditional correlations and are not meant to reflect particular causal relationships. Observed correlations may be driven by the fact that 1) the skills affect agricultural decisions and outcomes; 2) the agricultural outcomes are determinants of skills formation; or 3) some other variables are correlated with both skills and agricultural outcomes, making skills a potential confounder if not observed. Nonetheless, independently of which one of these factors drives the correlation, a high predictive power indicates that improving skills measures can contribute to a better understanding of agricultural productivity.

1. *Correlations with other variables*

Before turning to the regressions, Figure 1 shows unconditional correlations of the skill constructs as a first form of validation. Figure 1 and Appendix Table A1.A show a strong relationship between measures of cognitive skills and grades of education attained or self-assessed literacy. [20] The relationship between the number of years using mineral fertilizer and technical skills is also relatively strong. This provides some validation, but is also a reminder that the direction of causality is hard to infer. A farmer may know more about fertilizer because she has been using it for a while, or may have been using it exactly because he knew about it. Figure 1 also shows a relatively strong positive correlation between cognitive, technical and noncognitive skills. This points to the importance of studying the different skills together rather than independently, to avoid wrongly attributing to a skill the effect of other correlated skills.

2. *Yield predictions by construct*

The key outcome variable we use to test predictive validity is maize yield. Prior research has shown how skills or personality traits relate to job performance or wages in different contexts (Borghans et al. 2008). By contrast there is relatively less evidence about the relationship between skills and agricultural outcomes, perhaps because yield in rainfed agriculture is known to be a particularly noisy outcome variable, subject to many parameters other than skills, including weather conditions, soil characteristics, and availability of inputs or credit. Besides this, high skill individuals are likely to have transitioned out of agriculture, leaving a selected sample

[20] The cognitive construct is highly correlated with the respondent's reported education, with 59% of the variation in the cognitive score explained by grades attained and self-declared literacy. Respondents' education also explains a relatively large share of the variation in the noncognitive (19%) and technical (11%) skills, though less than for the cognitive skills.

(Lagakos & Waugh 2013).[21] And yet better understanding the relationship between skills and agricultural productivity deserves to be investigated, simply because agriculture is the primary activity of a large share of poor people worldwide.

Our study addresses these challenges through the following strategies: First, in order to limit the consequences of measurement errors and variability in the outcome variable, we use the average rank of yield over the four seasons following the completion of the skills data collection, (with ranks rescaled from 0 for the lowest yield to 100 for the highest).[22] Second, the comprehensive skills surveys and combination of the two surveys reduce the noise of the explanatory variables. Third, village fixed effects absorb a large share of spatial variation related to climatological, agro-ecological and economic conditions and imply that we focus on intra-village variation.[23] Fourth, we intentionally do not control for intra-village differences in soil characteristics, access to credit, nor any other input factors as they are (at least in part) a reflection of how skills (and related past actions) affect agronomic outcomes. Finally, virtually all farmers in the region produce maize, so maize yield is a useful common indicator of land productivity. Beyond yield, we also look at input use, without the prior that more input is necessarily better, but simply because their use should vary in function of the different skills. In all regressions, because random measurement error causes attenuation bias and a reduction of the R^2, an increase in the coefficients of the normalized skills measures and in the R^2 are signs of measures that are less noisy.[24]

We test how much of the variation in yield is explained by the measures of cognitive, noncognitive and technical skills, by regressing yield on the different skill constructs.[25] The first five columns in

[21] We find a significant correlation of 0.14 (p<0.0001) between a binary indicator variable of self-employment in agriculture as the main occupation of the household head and cognitive ability (but no significant correlation with non-cognitive skill and agricultural knowledge). This likely implies that cognitive ability in our sample has a lower mean and variance than in a representative sample of the entire population, making it more difficult to capture its relationship with yield.

[22] We use the rank because it is less sensitive to extreme values (Athey and Imbens, 2017). Appendix Tables A8-A11 shows similar regressions using the average of the log of the yield for the same seasons. The results are qualitatively similar but less precise.

[23] Variation in climatological and agro-ecological factors that differ between villages may well be correlated with both skills (by affecting selective outmigration) and yield.

[24] We intentionally put little structure. If skills matter one should observe significant correlations independently of the channels. Significant correlations of skills with yield would be consistent with skills mattering despite other constraints faced by the farmer, or because it helps release them. We leave the task of better understanding how these constraints interact with skills to other research and focus on improving the measures in order to facilitate such work.

[25] Our interpretations assume no correlation between measurement errors in of non-cognitive skills and measurement errors in the reporting of yields. If such correlation exists, then the explanatory power would likely be overestimated with respect to the prediction of true yields.

Table 3 do not include controls and demonstrate the share of variation explained by the three skill constructs (R^2). Columns 6 to 10 report results with controls and shows whether the skill measures remain significant after controlling for observed farmer, household and village characteristics.[26] Significant coefficients on skills in the later regression, with controls, point to the potential of skill measures to capture otherwise unobserved characteristics.

The results are presented for four different types of constructs: the naïve constructs, improved constructs, the naïve constructs averaged over test and retest, and the improved constructs averaged over test and retest. The comparison of estimates with the naïve constructs versus the improved constructs indicates how much gain in predictive power comes from aggregating the items in a way that better accounts for measurement errors. And the comparison of the improved constructs with the test-retest averages shows to what extent the improved constructs yield similar results as averaging over multiple waves, an alternative but costly method to reduce random measurement error. Finally, the test-retest average of the improved constructs provides our best estimate of the role of skills using all means available to get the most reliable constructs.

Results in Table 3 broadly show that the three types of skills matter, as all three coefficients are significant. Combined the measures explain a substantial share of the variation in yields. The R^2 of the naïve constructs without any control is 12.7 percent (column 1), compared to 14.4 percent when using the improved constructs (column 2). Interestingly this last figure is the same as the R^2 obtained when averaging the naïve scores of test and retest (column 3). Hence a better method to aggregate the information from the different items leads to as much improvement as the use of a second wave (which doubles the cost of data collection). The combination of both the improved method and averaging test and retest further raises the R^2 to 14.4 percent, providing our most reliable estimate of the contribution of the different skills to explaining variation in yields. This is likely still an underestimate of the explanatory power of skills, as we know from section 3.2 that the improved constructs remain fairly noisy.

The estimations with controls (columns 6 to 9) show that these conclusions stand even after controlling for observables. Skills are jointly significant, and remarkably, cognitive skills remain significant even after controlling for education and literacy. Comparing across columns, the highest improvement in significance and size of the coefficients from cleaning up measurement error is seen in the technical construct. This is consistent with the fact that this was the noisiest construct

[26] Controls include education, literacy, gender, age and age squared of the farmer, land and cattle ownership, size and quality of the house, household size, whether the farmer is the household head, and household head's gender. We also include village and enumerator-assignment fixed effects. We use the randomly assigned enumerator as opposed to the actual enumerator, as only the former is exogenously determined.

according to test-retest and Cronbach's alpha. Column 4 further suggests that technical skills may be more important than other skills once measurement error is addressed, though this conclusion does not hold when adding controls (column 8). Hence more generally, the evidence shows that all three skills matter for agricultural productivity, but properly capturing this effect requires substantial effort, both in data collection and aggregation method.

Finally columns 5 and 10 use the most reliable constructs (improved average across test and retest) but do not include technical skills. This could be important because cognitive and noncognitive skills may have effects on yields that go through technical knowledge, possibly attenuating their coefficients when technical skills are controlled for. Indeed, both coefficients increase when the technical skill construct is removed.[27] This specification also assesses the relative importance of cognitive versus noncognitive skills, and suggests that both are equally important for productivity, a result that parallels results on the importance of skills in US labor markets in Heckman, Stixrud and Urzua (2006) and Heckman and Kautz (2012). The point estimates suggest that a one standard deviation increase in cognitive or non-cognitive skills increases the average rank of maize yield by 4 to 5 percentage points.[28]

3. Yield predictions by sub-construct

The level of aggregation used in Table 3, with one aggregate construct to measure each of the domains, is higher than what is often used in empirical work on skills. We hence also present predictions, separating out the different cognitive tests, the subscales for personality and lower order constructs, and subscales of technical skills by broad topic. Table 4 first presents estimates using the naïve sub-constructs with variables measured as z-scores. The regressions are estimated with controls, but the adjusted R^2 in absence of controls is added in the bottom of Table 4.

Using sub-constructs increases the adjusted R^2 for the predictive model for yields from 10.4 percent with naïve constructs to 12.5 with the sub-constructs. However, the F-test for joint significance of cognitive tests is also low and insignificant. The same holds for the technical skills. This is consistent with the earlier finding that test-retest and alphas are lower for the sub-constructs than for the aggregate constructs, indicating measurement error is introduced in the regressions with the sub-constructs, making it difficult to assess their relationship with yields.

[27] Comparing columns 1, 4 and 5 in Table 3, note that the cognitive skill construct loses explanatory power as the precision of the technical skill construct increases, but gains significance when it is removed. This suggests that the effect of cognitive skill on yield could be operating through its effect on technical knowledge.

[28] This corresponds to about 21 percent of a standard deviation in the average rank of yield.

In contrast, we find a few significant correlations with the noncognitive subscales. The 15 noncognitive sub-constructs are jointly significant. The few significant results suggest that causes of poverty, tenacity, agreeableness and CESD might have some predictive power for yields, but coefficients are only marginally significantly different from each other. To illustrate the risk of drawing erroneous conclusions from this type of regression, columns 2 to 6 present similar regressions where we only keep one sub-construct at a time, using each component of the Big Five. In four out of five cases the coefficient is significant. We only present the Big Five for conciseness, however 10 out of the 15 coefficients are significant when they are the only noncognitive variable in the estimate. When a noncognitive sub-construct is used as an explanatory variable without other measures of noncognitive skills, the latter ones are likely to be omitted variables, suggesting the observed effect should not be attributed to the sub-construct used in the regression.

Table 5 shows similar regressions, but now using the improved constructs and keeping the number of factors suggested by the EFA. Column 1 shows the cognitive construct becomes significant, as do the first and fourth noncognitive factor. The later mirrors the findings from Table 4, as the first factor is basically the CESD while the fourth factor is dominated by the reverse questions of the causes of poverty. Importantly for the interpretation, the coefficients of the noncognitive factors are not significantly different from each other, and columns 2 to 7 further illustrates that all but one are significant when they are included without the others. Hence even the estimates with the improved constructs do not allow to clearly discriminate between noncognitive skills. We hence conclude that while noncognitive skills matter for productivity, the data do not allow us to infer which of the noncognitive skills matter.

4. *Predictions of agricultural practices*

We complement the analysis with regressions on key agricultural practices, averaged over the four seasons. We analyze to what extent the different skill measures are predictive for the use of mineral fertilizer, manure, hybrid seeds, multiple time weeding and hiring labor.[29] The estimates with the improved constructs (Table 6) show that technical skills are positively correlated with a number of advanced farming practices, an encouraging sign for its validity. As for yield, the noncognitive construct is also strongly predictive. In contrast, the cognitive construct is not (if anything, there is a negative relationship with weeding), suggesting that the relationship between cognition and yield

[29] We focus on these practices as they show meaningful variation between households and across time, and can reasonably be expected to correlate to some of the domains we are trying to measure. We exclude other practices, such as row planting, which virtually all farmers in this context use.

is not driven by decisions regarding these practices. The overall predictive power of the skills varies widely between practices. Skills basically explain none of the variation in the use of manure, while they explain up to 11% of the use of hybrid seeds.

Table 7 presents results that separate out the different noncognitive constructs, and shows that for four out of the five practices, as for yield, the data do not allow to discriminate between the different noncognitive skills. The regression for weeding provides an interesting exception, as the factor that is dominated by conscientiousness is positively correlated with weeding while many of the other factors have small and negative coefficients. The F-statistic confirms the difference between the factors. Given the intuitive relationship between conscientiousness and efforts for weeding, this provides some validity to the improved noncognitive constructs.

IV. **Further Understanding Measurement Challenges**

Overall this set of results presents a mixed picture on the ability of the different tests and subscales to meaningfully measure the intended skills in the population studied. This section presents further evidence to understand the potential sources of measurement error and derives practical guidelines for the measurement of related skills in empirical work.

A. Interaction with enumerators

Most tests were initially designed to be self-administrated. Yet in a rural developing country setting, because many respondents are unable to read, the questions are typically asked by an enumerator. This may affect responses in multiple ways even after intensive efforts to harmonize practices during enumerator training. Drawing on the random assignment of enumerators to respondents, we therefore estimate to what extent answers are affected by enumerators. Table 8 shows the R^2 of a regression of the improved constructs on enumerator fixed effects. Ideally one would like these fixed effects to have no explanatory power. Yet 4 percent of the variance of the cognitive skills can be explained by which enumerator asked the questions (significant at 10% level), this is up to 7 percent for technical skills and 9 percent for noncognitive skills (both significant at 1% level).[30]

[30] As the regressions are based on the randomly assigned enumerator, and there were deviations from this assignment in 25% of interviews, these provide lower bound estimates of the variation explained by enumerator effects. We use randomized inference to calculate p-values (with 10.000 replications).

This suggests a large amount of noise introduced by enumerators, possibly due to the level of explanations they provide or other unintended nudges.

We also compare the test-retest statistics when the same enumerator was assigned to a given respondent for the test and the retest compared to when a different enumerator is sent each time.[31] Standard practice for test-retest correlations is to have the test administrated in similar conditions. Practically, however, test-retest correlations that are high with the same enumerator, but lower with different enumerators, could indicate the influence of the enumerator rather than the consistency of the measure. We find that assigning a different enumerator leads to a drop of .09 in the test-retest correlation of the cognitive construct (significant at 1%), .06 in the noncognitive one, and .07 in the technical one (significant at 10%). Hence enumerator effects substantially reduce the reliability, confirming the non-negligible role of enumerators for skill measurements.

This is further confirmed when analyzing effects of being surveyed at a later stage during the survey round, i.e. on days farther away from the training when standardization may be weakened. We use the random order in which the villages were surveyed, account for the imperfect compliance with the assignment through a 2SLS estimation, and find that technical scores are significantly higher for farmers surveyed on later dates during the test (Table A6).

These results point, first of all, to the importance of intensive training for standardized application of the different tests, and for the potential need of re-standardization during the survey rounds. This typically requires developing detailed scripts to be followed literally, and avoiding idiosyncratic interpretation or clarifications by enumerators. Attempts at standardization alone may however not be enough (as this study shows) and random assignment of enumerators to respondents is recommended, in order to properly account for any remaining enumerator effects. For impact evaluations with skills measures, ensuring balance of enumerators between control and treatment groups should also help avoiding bias. Moreover, when possible, it is worth considering introducing self-administration in at least part of the survey instrument.

B. Respondent's ability to understand the questions

Another difference between the population studied and the population for which most tests were designed is the low educational level of the respondents, which can affect respondents' ability to understand the questions. To assess this, Table 8 presents test-retest correlations, Cronbach's

[31] We assigned the same enumerator to test and retest in 40% of cases. As before, one would expect that the observed differences between same and different enumerator assigned would be greater if the compliance was 100%.

alphas, and the share of the variation explained by enumerator effects, comparing respondents for whom the aggregate cognitive index is below versus above the median. Differences between the two groups do not point at any clear direction for the cognitive and noncognitive constructs. For the aggregate technical construct there are relatively large differences in the indicators across the two groups, all pointing towards higher reliability in the group with higher cognitive skills. Hence respondents' difficulties in understanding the technical knowledge questions may help explain measurement error.[32]

These findings indicate the importance of extensive qualitative piloting to probe the understanding by different types of respondents in detail, to be done each time skill measures are used in a new context. They also suggest the need to adapt standardized questions taken from international scales to make them understandable, even if it weakens some of the international comparability. There may also be a trade-off between easing understanding by the respondent and introducing enumerator effects, as questions requiring more explanations and enumerator initiative, such as questions involving visual aids, are harder to standardize. The challenges resulting from the need to translate concepts to languages that may not have the relevant equivalents, and the complexity this introduces needs to be understood better.[33]

C. Order of the modules in the survey

Given the length of the survey, and of many other surveys in developing countries, one can hypothesize that the duration of the survey and the order of questions play a role in explaining measurement error. We randomly assigned the order of the cognitive, noncognitive and technical modules in both the test and the retest and use this to assess the order effect. Table 9 shows that for the cognitive and noncognitive skills the order in which the module appeared in the test and retest does indeed significantly affect their test-retest correlations. But contrary to our prior, there is no clear evidence of survey fatigue, as there is no systematic degradation of the reliability when a module comes later in the survey.[34]

Instead, the test-retest correlation for noncognitive skills was highest when it comes last, and differences between different test-retest combinations are significant. In contrast, the test-retest

[32] We also analyzed heterogeneity with respect to age, education and gender. Splitting the sample by education level leads to similar conclusions as the split by cognitive ability. Results by age and gender do not allow pointing to subgroup of the population that clearly outperforms the other on the different measures. The psychometric indicators of noncognitive and technical skills do not pass the bar for any subsample of the population that we examined.

[33] The noncognitive questions posed the largest challenges for translation, and understanding concepts such as "active imagination", or "generating enthusiasm" were difficult even for the (university level trained) enumerators.

[34] Analysis of the random order of the questions within modules leads to a similar conclusion.

correlation for technical skills is highest when it comes first. This matches well with observations from the field that noncognitive questions, which are more abstract, tend to raise eyebrows when the survey starts with them, whereas discussion about farming practices allowed a smoother start. Hence careful attention to the order of different modules when designing a survey instrument can reduce measurement error, while survey duration and fatigue may not be that important. Good practice may be to start with questions on topics that the respondents are comfortable talking about, and ask more abstract noncognitive questions towards the end of the survey, when respondents are more at ease.

D. Acquiescence bias: implications and corrections

Acquiescence bias may be more likely in populations with lower levels of education or cognition than those for which Big Five questionnaires and lower-order noncognitive subscales were originally designed. The bottom panel in Figure 1, showing a strong negative correlation between the acquiescence score and the cognitive index (left) or the educational level (right), is suggestive in this regard. The gradients are steep, and the acquiescence score is twice as large for somebody with no education compared to somebody with 10 years of education. This is consistent with qualitative observations during piloting: "yea-saying" was more likely when respondents did not fully understand a question, which happened more often for lower educated individuals.

Strikingly, the acquiescence score shows a strong negative correlation with yields as well. The coefficient of correlation with the average rank of maize yield is -0.15, with a p-value < 0.0001. Hence farmers with a higher propensity of agreeing with different statements have lower yields on average. The importance of acquiescence bias and its high correlation with both cognitive skills and outcomes of interest imply that the effects of the noncognitive skills may be confounded with response patterns when the later are not properly dealt with.

Because acquiescence bias leads to observable contradictions in the responses of reversed and non-reversed items, it can be corrected for, as we do in this paper.[35] For this it is preferable to balance reverse and non-reversed items in all scales, a practice commonly used in psychology but often ignored by economists. As reverse items can be harder to understand or translate, they may require more adaptation (e.g. avoiding double negations that confuse respondents). While it may be

[35] Other sources of bias include "extreme response bias" and "middle response bias", which we did not find to improve estimations when corrected for, and "social desirability bias", which may lead a respondent to answer what he believes to give the best impression, or what he believes the enumerator wants to hear.

tempting to instead drop reverse items, the benefits of being able to measure and correct acquiescence bias seem to outweigh the costs.

E. Anchoring and the use of other sources of information

Another important response pattern comes from the fact that each respondent may interpret the Likert Scale differently and use different thresholds to decide between answer categories. One way of breaking the relationship between answer patterns and skills is to ask another person about the skills of the person of interest, which is why correlations between self-reports and other observer ratings are often used as evidence of validity (McCrae et al 2011). A priori, the other person is likely to have asymmetric information about the true skill level. Yet if it helps addressing the systematic bias, this may also provide an alternative or complementary manner to measure skills.

To test the validity and trade-offs, we collected proxy information from a number of different sources. First we asked the community health worker (CHW), a person well informed about different village members, based on her regular home-visits, to classify households according to their cognitive, noncognitive and technical skills (3 questions).[36] In addition, we ask another household member, as well as two other village members (one at test and one at retest) to assess 14 specific skills of the respondent, answered on a Likert scale. Each person was also asked the same 14 questions about herself.

Table 10 shows the correlations of the proxy measures with the relevant scales or subscales. Correlations between observable and objective skills (language and math) and proxy measures by random village members or other household member are good, but all other correlations are very low. Strikingly, for 7 out of the 14 measures, the correlation between proxy measures of skills of the same person by two different people is smaller than the correlation between proxy measures of skills of two different people by the same respondent. This points to systematic answering patterns, which appear as important as the actual skill differences between the two people about whom the proxy reports.

Table 11 shows results for the predictive power of the proxy report by the CHW. Asking the same person about the skills of several persons presents the advantage of ensuring that the same anchoring is used, making the resulting measure more comparable within this group. As each CHW was asked about the 10 sample farmers of the village, we include village fixed effects to take out

[36] Because the CHW's responsibilities require her to regularly visit villagers, we expect her to be well informed about the skills of others. Picking a random person in the village may not yield the same results.

any systematic CHW effect. Column 1 shows the variation explained by only the fixed effects, column 2 show the additional variation explained by the three skill proxies obtained from the CHW, and column 3, shows the specification with the full set of controls. Using proxy reports by a village informant gives results that are relatively close to those obtained with self-reports. with in particular the cognitive and technical skills proxies predictive of yield in the specification without controls. Moreover the predictive power of the CHWs report on farmers' technical skills remains robust to adding controls, and the R^2 is similar as in the model with self-reports.

These results suggest that some of the first order results can be obtained by asking 3 simple questions to a well-informed key-informant, instead of asking 2.5 hours of targeted skill questions and tests to each respondent. That said, such proxy measures are not a good solution when one aims to obtain comparable skill measures across villages. Hence for proxy information, there appears to be a benefit of asking one well-informed and connected person about many people in the village, rather than using several proxy respondents.

F. Differences in the factor structure

The factor analysis of the noncognitive skills raised concerns because it often did not pool items that were expected to belong to the same sub-constructs into the same factors. To formalize this finding, a congruence test considers the degree of correlation of the factor loads of similar items obtained in other contexts (see appendix B6). We restrict the analysis to the 23 items from the Big Five included in our study. Table 12 presents the congruence with respect to the same items administrated in the United States where the BFI has been validated many times. In Kenya, it shows an average congruence of .55 across the five factors. For comparison using the factor loads of the same items administrated in Spain, Holland and Germany results in congruence coefficients between .76 and .93. The average congruence coefficient for the subsample with higher cognitive ability is higher, but still substantially below the three developed country ones.

This finding could indicate that the underlying factor structure is different for this population than for populations on which it was previously validated, perhaps for cultural reasons. But it could also be due to the lack of understanding of some items or systematic response patterns that may be stronger than in high- income contexts and therefore do not allow detecting the same factor structure (even when the true latent factors are similar).

G. Explaining the noise in the measure of technical skill

As the technical skills questions attempt to measure knowledge, one would expect them to be less affected by systematic response biases. They require respondents to choose between a series of answers that do not have a clear ranking, or to give open answers, and while respondents certainly can guess, systematic bias of all questions in a given direction is less likely. The results indicate however that random measurement error is much more important for technical than for cognitive skills. This can be inferred from the low test-retest statistics, low Cronbach's alpha, and the gains in precision and predictive power when using means of test and retest, and factor analysis.

As discussed above, respondents' difficulties in understanding the technical knowledge questions may partly explain the measurement error. Insights from the qualitative field work provide additional insights for why the technical measure is noisy. To effectively assess a skill, a question needs to have only one correct answer and have enough variation in the responses to be informative of the respondents' knowledge. However, after working with agronomists to identify the most fundamental knowledge that can affect the farmer's productivity and piloting the questions, we found that most of them fell into one of the two following categories. Questions with unambiguous correct answers were answered correctly by the vast majority of farmers.[37] In contrast, questions that had sufficient variance in the responses often were questions for which the right answer may depend on the context.[38] Informative questions with one correct answer were difficult to find, precisely because the difficulty to make the right decisions in farming often comes from the difficulty to adapt to each context rather than applying a "one size fits all" solution. Obtaining better measures of technical skills may require the development of new techniques assessing whether the different micro-decisions taken by farmers fit their environment.

V. From Kenya to Colombia and Beyond: Similar Results in Different Contexts

Do the findings of this paper apply to rural developing contexts other than Luo speaking Western Kenya? To address this question we replicated key parts of the experiment in the department of Sucre in Colombia in 2017, applying a similar skills measurement survey twice to 804 farmers also with 3 weeks interval. Items were translated into Spanish but were kept as similar as possible in order to allow comparison. The technical knowledge module followed the same structure, but was adapted, working again with local agronomists to identify key farming knowledge required for good practices. Like in Kenya, we first piloted the questionnaire in order to make the adjustments

[37] This may be different when farmers have recently been exposed to new information (for instance through an extension intervention) as differences in exposure and internalization of the new messages may create more empirical variation in knowledge of this new information.

[38] For instance, the optimal number of seeds in a hole at planting can depend on the quality of the seeds and the spacing between seeds, and when farmers answer this question, their benchmark quality and spacing might be different than those of the agronomist. And their answers may change over time if answers reflect their most recent experiences.

necessary for the questionnaire to be well understood by the population. We apply the same calculations to compute the naïve and improved indexes and present the summary statistics in Table 13. Besides this, we replicate the results of figure 1, table 1, 2, 3 and 10 in figure D1 and tables D1, D2, D3 and D4, respectively.

As in the case of Kenya, both the cognitive and technical indexes become more precise with the improved measures. The results confirm that the use of Item Response Theory can bring substantial gains in dealing with the noise. Again, the cognitive measure is the only one that shows high level of consistency both across tests and across time, although it requires the use of improved measure. The measure of technical skills remains quite noisy, but less so than in the case of Kenya, with a test retest that goes from .50 in the naïve measure to .62 when calculated with Item Response Theory.

For Colombia, the noncognitive construct reaches a test-retest and Cronbach's alpha that are just around .7. The test-retest for the sub-constructs varies however between .21 and .66, and the Cronbach alpha's between .21 and .71, in line with the Kenyan results (table D2). Moreover the aggregate indicators of consistency are lower for the improved score. This loss after correcting for acquiescence bias and sorting items by factor suggests that the correlations were partly driven by response patterns. The acquiescence bias is of the same order of magnitude as in Kenya (.37). For noncognitive and technical measures, we also find that having the same enumerator assigned leads to greater test-retest correlations, and that the R^2 of enumerators fixed effects is substantial and is significant for noncognitive and technical skills (table D3), although in the case of Colombia, the R^2 of enumerators is much higher for technical skills (0.21) than for noncognitive skills (0.04). And as for Kenya, the difference in the test-retest correlated between having the same enumerator or a different enumerator assigned is significant for the cognitive and technical skills. Both are signs that the interaction with enumerators is part of the challenge. Finally, as shown in table D4, we find that in all four questions asked to proxies in the same village, the correlation between proxy measures of skills of the same person by two different people is smaller than the correlation between proxy measures of skills of two different people by the same respondent. This provides additional evidence the variance from systematic answering patterns may in fact exceed that variance from the true skills that one aims to capture.

In sum, the replication in a very different context and language lead us to strikingly similar conclusions, both in the validation of cognitive skills and regarding measurement error in technical and noncognitive skills. Along the same lines, results in Laajaj et al. (2019), combining Big Five data covering more than 300,000 individuals from 23 countries establishes lack of congruence with the expected five-factor model in face-to-face surveys (section 4.6), further suggesting issues found in Kenya are not specific to that context.

VI. Conclusions

Cognitive, noncognitive and technical skills are thought to play a key role for many economic decisions and outcomes in developing countries and are increasingly incorporated in empirical analyses. The measures of skills through enumerators in a developing setting brings differences that are such that it requires its own validation, and yet little is known about the validity or reliability of commonly used skill measures in surveys conducted in developing countries. This study is the first to investigate the reliability, validity, and the predictive power of a large range of skill measures on a poor rural adult population in developing country settings. We do so using data from a survey experiment, specifically designed for this purpose, and a variety of statistical tools and methodologies. The results show the cognitive skills measures are reliable and internally consistent, while technical skills are difficult to capture and very noisy. The evidence further suggests that measurement error in noncognitive skills is non-classical, as correlation between questions are driven in part by the answering patterns of the respondents and the phrasing of the questions.

These results first of all imply that collecting information on cognitive skills in large household surveys in field conditions is feasible. The high correlations between cognitive measures further suggest that it may not be necessary to ask the full battery of tests and shorter survey modules can hence be used to obtain a reliable proxy of cognitive skills. Further validation of such measures in other contexts will be important to establish whether these conclusions hold in different settings than Kenya and Colombia and the extent to which such measures allow comparing cognitive skills across countries, or across regions and groups within countries. The study further shows how specifically accounting for measurement error through factor analysis and item response theory can help increase the validity, reliability and predictive power of the technical skill measures. It also highlights that obtaining a good aggregate and stable measure of agricultural knowledge is challenging, as the “right” answer to many agricultural questions is context-specific, so that it can differ both between respondents, and even for the same respondent over time. Nevertheless, once the measurement error is reduced, the technical skills seem to lead to coherent predictions. The paper also shows the weaknesses of instruments designed to capture noncognitive outcomes when applied through enumerator-administered surveys in poor rural settings. It highlights the importance of using factor analysis and corrections for response patterns to obtain more reliable and valid measures, and warns against naïve interpretation of existing scales. This result further suggests that a narrow focus on a subset of questions from a few scales can lead to misleading conclusions. Instead, collecting an extensive battery of questions in a pilot stage, and using factor analysis and IRT on those pilot data, could help determine the subset of questions to keep in a larger survey, and their relevant interpretation.

Finally we find that skill measures can explain meaningful variation in agricultural productivity and practices. When using our best estimates to address measurement errors, we find that the three skills constructs contribute about equally to explaining yield. However, the presence of systematic measurement error in non-cognitive construct raises concern about the possible interpretation of the relationship observed in these regressions. One step towards a more cautious use of such measures would be to require evidence of their coherence before presenting the related results. Demonstrating a proper sorting of the items into coherent factors hence ought to be a pre-condition for separately interpreting the subconstructs.

Overall, this study provides a proof of concept that in some developing country contexts, technical and non-cognitive skill measures may not properly measure what they intend to measure. The evidence from Kenya and Colombia, combined with the absence of validation of such skill measures for similar contexts, raises obvious concerns. Even if the methods applied in this paper helped reduce some of the measurement error, a large amount of measurement error remained after corrections in both the noncognitive and the technical constructs. The evidence further suggests that having a relatively large set of items, and repeated measures, is important to reduce the measurement error.

The results also flag the large variation in answers due to variation across enumerators, pointing to the importance of carefully accounting for such enumerator effects in the data collection design. The use of enumerators and oral questions distinguishes developing country data from data obtained and validated in mostly developed settings, and this study provides clear evidence that they can have major implications for measurement, and ought to be accounted for. Finally, while the purpose of this study was to explicitly test for measurement error with existing scales, the sobering results arguably suggest the need for noncognitive and technical skill measurement instruments that are more adapted to a poor rural study population, and subsequently validated.[39]

[39] Decision based measures or incentivized real effort tasks are alternatives that can be explored, even though they are costly and face their own challenges (Duckworth and Kern 2011, Forstmeier et al. 2011; Alan, Benova and Ertac, 2016). In this study we purposefully limited the skill measures and scales to those that are commonly used (and relatively easily to incorporate) in large sample surveys. For both reasons, we did not include task-based measures, but highlight the need for exploration and validation of such measures as possible alternatives to current practice.

The policy literature analyzing whether and how skills can be moved with policy interventions or connecting skills acquisition to improved outcomes asks a question that is related to, but not the same as, the one examined in this paper. While we focus on the cross-sectional variation in the level of different skills measures, we leave it to future research to investigate how the validity and reliability concerns affect findings when skills, and their malleability, are measured as outcomes of external interventions (as one would typically do for impact evaluation purposes).

Obtaining good measures of adult skills is a prerequisite for empirical work on the importance of skills for economic decision-making in developing countries and can be key to fully analyze the optimal design and potential benefits of a wide range of policies. For the rural sector, a better understanding of adult skills is particularly pertinent, given the often-hypothesized selection of higher skilled individuals into the non-agricultural occupations. Further improvements in skill measurement are needed to better understand the importance of this selection, and more generally to analyze the role of skills, and their interactions with other factors, for economic and social outcomes.

REFERENCES

Adhvaryu, A., N. Kala and A. Nyshadnam, 2016. "Soft Skills to Pay the Bills: Evidence from Female Garment Workers", Unpublished, University of Michigan

Alan, S., T. Benova and S. Ertac, 2016. "Ever Failled, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit", *HCEO Working Paper*.

Almund, M. A. Duckworth, J. Heckman, and T. Kautz, 2011. "Personality Psychology and Economics," (with). In E. Hanushek, S. Machin, and L. Woessman, eds., *Handbook of the Economics of Education* , Amsterdam: Elsevier. pp. 1-181.

Ashton, M.C., Lee, K., Perugini, M., Szarota, P., De Vries, R.E., Di Blas, L., Boies, K. and De Raad, B., 2004. "A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages". *Journal of personality and social psychology*, 86(2), p.356.

Athey, S., and G.W. Imbens, 2016. "The Econometrics of Randomized Experiments", Banerjee, A. and E. Duflo, (eds.), *Handbook of Economic Field Experiments*. Volume 1. Elsevier.

Attanasio, O., S. Cattan, E Fitzsimons, C. Meghir, and M. Rubio-Codina, 2015. "Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia", *NBER Working Paper* No. 20965

Benet-Martinez, V. and O.P. John, 1998. "Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analysis of the Big Five in Spanish and English", *Journal of Personality and Social Psychology* (75): 729-250.

Bernard, T. and A. Taffese, 2014. "Aspirations: An Approach to Measurement with Validation Using Ethiopian Data" *Journal of African Economies* (2014) 23 (2): 189-224

Bernard, T., S. Dercon, K. Orkin, and A. Taffese, 2014. "The Future in Mind: Aspirations and Forward-Looking Behaviour in Rural Ethiopia", *CEPR Discussion Papers* 10244.

Bertrand, M. and S. Mullainathan, 2001. "Do People Mean What They Say? Implications for Subjective Survey Data", *American Economic Review*, 91 (2): 67–72.

Blattman, C. and L. Ralston, 2015. "Generating Employment in Poor and Fragile States: Evidence from Labor Market and Entrepreneurship Programs", Available at SSRN 2622220.

Blattman, C., J. Jamison and M. Sheridan, 2017. "Reducing crime and violence: Experimental evidence on adult noncognitive investments in Liberia", *American Economic Review*, 107(4): 1165–1206

Blattman, C., and S. Dercon, 2016. "Occupational Choice in Early Industrializing Societies: Experimental Evidence on the Income and Health Effects of Industrial and Entrepreneurial Work", *IZA discussion Paper* 10255.

Bond, Timothy N., and Kevin Lang, 2018. "The Sad Truth about Happiness Scales" *Journal of Political Economy* 127, no. 4

Borghans, L., A. L. Duckworth, J.J. Heckman, and B. ter Weel, 2008. "The Economics and Psychology of Personality Traits," *Journal of Human Resources*, 43(4): 972-1059.

Bowles, S., H. Gintis, and M. Osborne 2001. "The determinants of earnings: a behavioral approach." *Journal of Economic Literature*, 39(4), 1137–76.

Callen, M., S. Gulzar, A. Hasanain, Y. Khan, and A. Rezaee, 2015. "Personalities and Public Sector Performance: Evidence from a Health Experiment in Pakistan." *NBER working paper* 21180

Chuang, Y. and L. Schechter, 2015. “Stability of Social, Risk and Time Preferences over Multiple Years”, *Journal of Development Economics*. 117: 151-170.

Cunha, F. and J. Heckman, 2010. “Investing in our Young People.” *NBER Working Paper* No. 16201.

Cunha, F., Heckman, J., and Schennach, S., 2010. “Estimating the technology of cognitive and non-cognitive skill formation.” *Econometrica* 78(3):883-931.

Dal Bo, E., F. Finan & M.A. Rossi, 2013. "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service," The *Quarterly Journal of Economics*, vol. 128(3), pages 1169-1218

de Janvry, A., E. Sadoulet and T. Suri, 2016. “Field Experiments in Developing Country Agriculture”, Banerjee, A. and E. Duflo, (eds.), *Handbook of Economic Field Experiments*. Volume 1. Elsevier.

Delavalande, A., X. Gine´ and D. Mckenzie, 2011. “Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence”, *Journal of Development Economics*, 94 (2): 151–63.

Duckworth, A.L. and Kern, M.L., 2011. “A meta-analysis of the convergent validity of self-control measures.” *Journal of Research in Personality*, 45(3), pp.259-268.

Fischer R., and T. L. Milfont, 2010. “Standardization in psychological research.” *International Journal of Psychological Research*. 3, 88–96.

Forstmeier, S., Drobetz, R. and Maercker, A., 2011. “The delay of gratification test for adults: Validating a behavioral measure of self-motivation in a sample of older people”. *Motivation and Emotion*, 35(2), pp.118-134.

Gertler, P., J. Heckman, R. Pinto, A. Zanolini, C.Vermeersch, S. Walker, S. M. Chang, S. Grantham-McGregor, 2014. “Labor market returns to an early childhood stimulation intervention in Jamaica”, *Science* 30: (344): 6187 pp. 998-1001.

Ghosal, S., S. Jana, A. Mani, S. Mitra and S. Roy, 2016. “Sex Workers, Stigma and Self-Belief: Evidence from Kolkata brothels”, *Working Paper 302* Department of Economics, Warwick University.

Gollin, D., D. Lagakos, and M. E Waugh, 2014. “The Agricultural Productivity Gap," *Quarterly Journal of Economics*, 129 (2), 939-993.

Grantham-McGregor, S., YB. Cheung, S. Cueto, P. Glewwe, L. Richter, and B. Strupp, 2007. Developmental Potential in the First 5 Years for Children in Developing Countries. *The Lancet* 369(9555): 60–70.

Groh, M., D. McKenzie, and T. Vishwanath, 2015. “Reducing Information Asymmetries in the Youth Labor Market of Jordan with Psychometrics and Skill Based Tests”, *World Bank Economic Review* 29(suppl 1): S106-S117.

Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H. and Lero Vie, M., 2013. “How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon”. *Journal of personality and social psychology,* 104(2), p.354.

Heckman J.J. and T. Kautz, 2012 “Hard Evidence on Soft Skills”, *Labour Economics,* 19(4):451–464.

Heckman, J. J. 2007. “The Economics, Technology and Neuroscience of Human Capital Formation.” *Proceedings of the National Academy of Sciences* 104(33): 13250–255.

Heckman, J. J., 1995. “Lessons from The Bell Curve”. *Journal of Political Economy,* 103 (5), 1091-1120.

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

Hicks, Loue E., 1970. “Some properties of ipsative, normative and forced-choice normative measures.” *Psychological Bulletin,* 74, 167-184.

Jack, B.K., 2011. “Market inefficiencies and the adoption of agricultural technologies in developing countries” White paper prepared for the Agricultural Technology Adoption Initiative, JPAL (MIT) / CEGA (Berkeley)

John, O.P; L.P. Naumann, and C.J. Soto, 2008. “Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In O.P. John, R.W.robins & L.A. Pervin (Eds.), *Handbook of personality: Theory and research*, pp114-158. New York, NY: Guilford Press.

Jones, M. and Kondylis, F., 2018. “Does feedback matter? Evidence from agricultural services”. *Journal of Development Economics*, 131, pp.28-41.

Krueger, A.B, and D.A Schkade, 2008. “The Reliability of Subjective Wellbeing Measures.” *Journal of Public Economics* 92(89):1833-1845.

Krutikova, S. and H.B. Lilleor, 2015. “Fetal Origins of Personality: Effects of early life circumstances on adult personality traits”, *CSAE working paper* 2015-03.

Laajaj, R., Macours K., Arias O., Gosling S., Pinzón Hernández, D.A, Potter, J., Rubio-Codina M., Vakis, R., 2019. “How well do the Big Five measure personality in developing country surveys?”, *Science Advances* 5: eaaw5226

Lagakos, D. and M. Waught, 2013. “Selection, Agriculture, and Cross-Country Productivity Differences”, *American Economic Review*. 103 (2): p. 948-980

Ledesma, R.D., and P. Valero-Mora, 2007. "Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis." *Practical assessment, research & evaluation* 12.2: 1-11.

Leight, J., P. Glewwe, and A. Park, 2015. “The Impact of Early Childhood Shocks on the Evolution of Cognitive and Non-Cognitive skills”, *Gansu Survey of Children and Families Papers*. Paper 51

Manski, C. F., 2004. “Measuring Expectations”, *Econometrica,* 72 (5): 1329–76.

McCrae, R.R. and Costa, P.T., Jr, 1997. “Personality trait structure as a human universal”, *American Psychologist,* 52, 509-516.

McCrae, R.R., Kurtz, J.E., Yamagata, S. and Terracciano, A., 2011. “Internal consistency, retest reliability, and their implications for personality scale validity.” *Personality and social psychology review*, 15(1), 28-50.

McCrae, R.R. and Terracciano, A., 2005. “Universal features of personality traits from the observer's perspective: data from 50 cultures”. *Journal of personality and social* *psychology*, 88(3), p.547.

McKenzie, D., 2012. “Beyond baseline and follow-up: The case for more T in experiments”, *Journal of Development Economics*, 99(2): 210-221.

McKenzie, D. and C. Woodruff, 2014. “What Are We Learning from Business Training and Entrepreneurship Evaluations around the Developing World?” *World Bank Research Observer*, 29(1): 48-82.

Murnane, R. J., J. B. Willett, and F. Levy, 1995. “The growing importance of cognitive skills in wage determination.” *Review of Economics and Statistics*, 77 (2), 251-266.

Piedmont, R.L., Bain, E., McCrae, R.R. and Costa, P.T., 2002. “The applicability of the Five-Factor Model in a sub-Saharan culture. In *The five-factor model of personality across cultures* (pp. 155-173). Springer, Boston, MA.

Pierre, G., ML. Sanchez Puerta, A. Valerio, and T. Rajadel, 2014. “STEP Skills Measurement Surveys - Innovative Tools for Assessing Skills." *Social protection and labor discussion paper* 1421. World Bank Group, Washington, DC.

Pritchett, L. and A. Beatty, 2015 “Slow down, you’re going too fast: Matching curricula to student skill levels” *International Journal of Educational Development: 40C : 276-288.*

Radloff, L.S., 1977. “The CES-D Scale: A Self-Report Depression Scale for Research in the General Population”, *Applied Psychological Measurement* 1(3): 385-401.

Rammstedt, B., C.J. Kemper and I. Borg, 2013. “ Correcting Big Five Personality Measurements for Acquiescence: An 18-Country Cross-Cultural Study”, *European Journal of Personality*, 27(1): 71-81.

Schmitt, D.P., Allik, J., McCrae, R.R. and Benet-Martínez, V., 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), pp.173-212.

Schuerger, J.M., Zarrella, K.L. and Hotz, A.S., 1989. “Factors that influence the temporal stability of personality by questionnaire.” *Journal of Personality and Social Psychology*, 56(5), p.777.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J., 2008. “The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20.” *Journal of Personality and Social Psychology, 94,* 718-737.

World Bank, 2008. *Agriculture for Development. World Development Report*. World Bank. Washington DC.

Young, A., 2013. “Inequality, the Urban-Rural Gap and Migration,” *Quarterly Journal of Economics*, 128(4), 1727–1785.

Figure 1: Relationships between constructs and other indicators

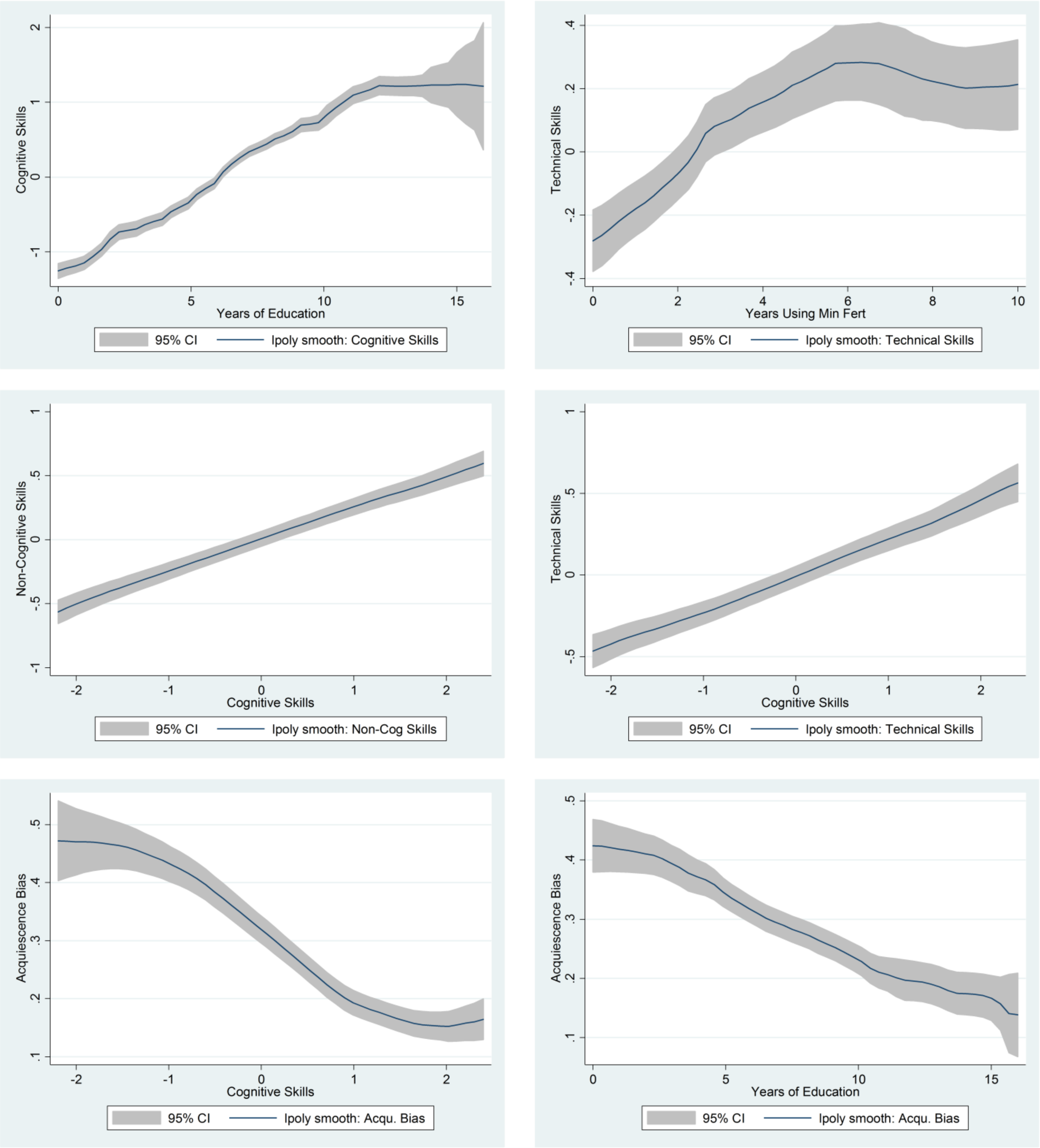

Table 1: Factor Loads of noncognitive items (corrected for acquiescence bias)

Factor	**Label**	**Items with factor Loadings higher than 0.3**
Factor 1	CESD	16 negatively phrased CESD
Factor 2	Conscientiousness	10 Big Five personality questions (5 conscientiousness; 2 neuroticism; 1 agreeableness; 1 openness)
	Tenacity	3 Tenacity questions
Factor 3	Locus of control	4 Locus of control questions
	Metacognitive	2 Metacognitive questions
	Other	6 Big Five personality questions (3 agreeableness; 2 openness; 1 extraversion)
		1 causes of poverty; 1 optimism
Factor 4	Causes of poverty	5 reverse items of causes of poverty scale
Factor 5	Attitudes towards change	4 attitude towards change
	Other	5 Big Five personality questions (2 extraversion; 1 conscientiousness; 1 neuroticism; 1 openness)
		Locus of control with visual aid
Factor 6	CESD positive	4 CESD positively phrased items
	Self-esteem	2 self-esteem
	Other	Risk aversion, 1 attitude to change, 1 optimism, 1 tenacity

Table 2: Measures of reliability and Internal Consistency

		2A - Naïve Score			
	Construct	Test-retest correlation	Cronbach's alpha of test	Cronbach's alpha of retest	Nb of items
	Cognitive	0.83	0.82	0.81	6
	Noncognitive	0.54	0.76	0.76	15
	Technical	0.31	0.45	0.48	5
Decomposition by sub-construct:					
Cog	Oral math questions	0.60	0.70	0.73	9
	Reading	0.82	0.77	0.77	12
	Raven	0.64	0.88	0.88	36
	Math (timed)	0.69	0.99	0.99	139
	Digit Span	0.37	NA	NA	1
	Digit Span Backwards	0.46	NA	NA	1
Non Cog	Locus of Control	0.49	0.56	0.62	9
	Self-esteem	0.32	0.28	0.36	4
	Causes of poverty	0.40	0.82	0.86	9
	Attitude towards change	0.37	0.37	0.43	5
	Organization/tenacity/self-control	0.26	0.42	0.48	6
	Metacognitive ability	0.19	0.46	0.54	4
	Optimism	0.22	0.17	0.26	3
	Risk aversion	0.12	0.21	0.03	2
	Patience	0.27	NA	NA	1
	Big 5 Agreeableness	0.25	0.39	0.31	4
	Big 5 Extraversion	0.23	0.33	0.37	4
	Big 5 Conscientiousness	0.33	0.51	0.26	6
	Big 5 Neuroticism	0.26	0.31	0.33	4
	Big 5 Openness	0.15	0.37	0.43	5
	CESD	0.41	0.82	0.85	21
Tech	Intercrop/Compost	0.21	0.18	0.15	7
	Maize	0.26	0.29	0.24	7
	Banana	0.17	0.19	0.17	6
	Soya	0.13	0.13	0.11	4
	Fertilizer	0.29	0.44	0.50	11

		2B - Factor / IRT Method		
	Construct	Test-retest correlation	Cronbach's alpha	Nb of items
	Cognitive (IRT)	0.86	0.82	6
	Noncognitive (Factor)	0.56	0.70	6
	Technical (IRT)	0.41	NA	1
Decomposition by sub-construct:				
Cog using IRT	Oral math questions	0.65	Same as panel 2A	
	Reading	0.80		
	Raven	0.61		
	Math (timed)			
	Digit Span			
	Digit Span Backwards			
Noncog 6 factors	CESD	0.43	0.84	18
	Conscientiousness/Tenacity	0.28	0.75	17
	LOC/Metacog/Openness	0.32	0.71	19
	Causes of poverty (all negative)	0.53	0.62	6
	Attitudes towards change/Beans	0.38	0.60	14
	CESD positive/Confidence/Risk aversion	0.30	0.56	11
Tech using IRT	Technical	0.41	0.54	32

Note: NA: Not Applicable, Cronbach's alpha cannot be calculated when there is only one item. In table 2B, noncognitive variables have been demeaned to correct for the Acquiescence Bias.

Table 3: Regressions of the average rank of maize yield across seasons on skill constructs

	SKILLS CONSRTUCTS USED AS REGRESSORS:									
	Naïve Score	Improved Index	Mean Naïve Score	Mean improved Index	Mean improved Index	Naïve Score	Improved Index	Mean Naïve Score	Mean improved Index	Mean improved Index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cognitive skills	2.80***	2.38***	2.40***	1.83**	3.88***	2.85**	3.06**	3.85***	3.97***	4.72***
	(0.802)	(0.787)	(0.869)	(0.859)	(0.832)	(1.267)	(1.267)	(1.230)	(1.267)	(1.229)
Noncognitive skills	3.78***	3.54***	4.60***	4.03***	5.00***	3.72***	3.72***	4.27***	3.95***	4.32***
	(0.742)	(0.737)	(0.902)	(0.892)	(0.924)	(0.782)	(0.787)	(1.015)	(0.975)	(0.978)
Technical skills	3.58***	4.38***	5.25***	6.11***		0.63	1.40	1.94	2.64**	
	(0.911)	(0.884)	(1.100)	(1.017)		(1.013)	(0.964)	(1.203)	(1.183)	
Observations	900	890	900	890	890	900	890	900	890	890
R^2	0.107	0.127	0.127	0.144	0.106	0.378	0.386	0.387	0.395	0.390
Controls	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
R^2 Adj. (w/o controls)	0.104	0.124	0.124	0.141	0.104	0.269	0.277	0.280	0.287	0.282
F Test	0	0	0	0	0	1.16e-06	2.69e-08	3.27e-08	3.96e-10	1.95e-10
F Test Diff.	0.751	0.357	0.210	0.0299	0.470	0.0490	0.213	0.329	0.681	0.831

Note: Dependent variable is the average rank of maize yields calculated over the 4 seasons (short rain 14 to long rain 16). Controls include education, literacy, gender, age and age squared of the farmer, land and cattle ownership, size and quality of the house, household size, whether the farmer is the household head, household head's gender, village fixed effects and enumerator-assignment fixed effects. Standard errors clustered at village level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 4: Regressions of the average rank of maize yield on naïve skill sub-constructs

	(1)	(2)	(3)	(4)	(5)	(6)
Oral math questions	-0.73	-0.14	-0.03	-0.06	0.01	-0.04
	(1.137)	(1.186)	(1.208)	(1.200)	(1.191)	(1.193)
Reading	1.67	1.91	1.86	1.94	1.91	1.91
	(1.200)	(1.207)	(1.186)	(1.192)	(1.181)	(1.203)
Raven	0.97	1.14	1.07	0.98	1.06	1.04
	(1.131)	(1.172)	(1.168)	(1.173)	(1.153)	(1.182)
Digit Span	0.60	0.54	0.59	0.59	0.68	0.65
	(0.779)	(0.763)	(0.773)	(0.771)	(0.751)	(0.768)
Math (timed)	1.58	2.17*	2.13*	2.22*	2.07*	2.15*
	(1.221)	(1.148)	(1.148)	(1.155)	(1.146)	(1.145)
CESD	2.34***					
	(0.820)					
Locus of Control (LOC)	-0.03					
	(1.094)					
Self-esteem	-0.36					
	(0.810)					
Causes of poverty	2.46**					
	(0.940)					
Attitude towards change	-0.53					
	(0.764)					
Tenacity / Organization	2.73***					
	(0.833)					
Metacognitive	0.24					
	(0.782)					
Optimism	0.93					
	(0.802)					
Risk aversion	-0.72					
	(0.714)					
Big 5 Agreeableness	0.95	1.86**				
	(0.888)	(0.834)				
Big 5 Extraversion	0.35		1.14			
	(0.799)		(0.752)			
Big 5 Conscientiousness	-1.23			1.19*		
	(0.959)			(0.712)		
Big 5 Neuroticism	0.59				1.87***	
	(0.723)				(0.643)	
Big 5 Openness	-0.13					0.75
	(0.852)					(0.825)
Other noncognitive	0.32					
	(0.763)					
Intercrop /Compost	0.06	-0.21	-0.21	-0.22	-0.26	-0.21
	(0.834)	(0.841)	(0.840)	(0.848)	(0.839)	(0.846)
Maize	0.66	1.09	1.21	1.16	1.01	1.18
	(0.853)	(0.830)	(0.839)	(0.834)	(0.847)	(0.845)
Banana	0.74	0.85	0.64	0.63	0.61	0.70
	(0.913)	(0.885)	(0.875)	(0.880)	(0.865)	(0.879)
Soya	-0.06	-0.26	-0.19	-0.21	-0.23	-0.21
	(0.820)	(0.822)	(0.827)	(0.825)	(0.815)	(0.824)
Fertilizer	0.20	0.71	0.70	0.74	0.75	0.74
	(0.852)	(0.848)	(0.846)	(0.845)	(0.842)	(0.845)
Observations						
R^2	897	899	899	899	899	899
R^2 Adj.	0.408	0.377	0.373	0.373	0.377	0.372
R^2 Adj. (w/o controls)	0.125	0.101	0.106	0.102	0.108	0.102
F Test (Cog)	0.149					
F Test (Noncog)	0.0017					
F Test (Tech)	0.878					
Test NC diff.	0.0917					

Note: Dependent variable is the average rank of maize yields calculated over the 4 seasons (short rain 14 to long rain 16). Controls include education, literacy, gender, age and age squared of the farmer, land and cattle ownership, size and quality of the house, household size, whether the farmer is the household head, household head's gender, village fixed effects and enumerator-assignment fixed effects. Standard errors clustered at village level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5: Regressions of the average rank of maize yield on improved skill sub-constructs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Cognitive skills (IRT)	2.58*	3.53***	4.13***	3.63***	2.95**	4.09***	4.00***
	(1.343)	(1.271)	(1.239)	(1.290)	(1.284)	(1.244)	(1.220)
NC Factor 1	1.98**	2.55***					
(CESD)	(0.770)	(0.769)					
NC Factor 2	-0.17		1.45**				
(Conscientiousness/Tenacity)	(0.900)		(0.722)				
NC Factor 3	0.40			1.67**			
(LOC / Metacog / Openness)	(0.773)			(0.725)			
NC Factor 4	2.59**				3.48***		
(Causes of poverty, negative items)	(1.088)				(0.721)		
NC Factor 5	-0.06					0.55	
(Attitude towards change / LOC_va)	(0.741)					(0.734)	
NC Factor 6	1.08						2.19***
(CESD positive / Self-esteem / Risk av.)	(0.723)						(0.637)
Technical skills (IRT)	1.17	1.68*	1.88*	1.79*	1.34	1.88*	1.70*
	(0.969)	(0.985)	(0.968)	(0.983)	(0.962)	(0.985)	(0.995)
Observations							
R^2	890	890	890	890	890	890	890
R^2 Adj.	0.391	0.380	0.374	0.374	0.384	0.371	0.377
R^2 Adj. (w/o controls)	0.279	0.270	0.262	0.263	0.274	0.259	0.267
F Test (Cog)	0.121	0.113	0.111	0.110	0.119	0.109	0.109
F Test (Noncog)	0.0575						
F Test (Tech)	1.55e-05						
Test NC diff.	0.230						

Note: Dependent variable is the average rank of maize yields calculated over the 4 seasons (short rain 14 to long rain 16). Controls include education, literacy, gender, age and age squared of the farmer, land and cattle ownership, size and quality of the house, household size, whether the farmer is the household head, household head's gender, village fixed effects and enumerator-assignment fixed effects. Standard errors clustered at village level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 6: Regressions of agricultural practices on skill constructs

	Naïve scores used as regressors					Mean improved indexes used as regressors				
SKILLS	Mineral Fertilizer	Manure	Hybrid Seeds	Multiple Weeding	Hiring Labor	Mineral Fertilizer	Manure	Hybrid Seeds	Multiple Weeding	Hiring Labor
Cognitive skills	0.00	0.02	-0.00	-0.04*	0.02	0.01	0.01	-0.01	-0.07***	0.01
	(0.017)	(0.017)	(0.020)	(0.022)	(0.022)	(0.019)	(0.018)	(0.025)	(0.023)	(0.028)
Noncognitive skills	0.03**	-0.01	0.03**	0.03**	0.03*	0.04***	-0.01	0.05**	0.05***	0.04**
	(0.013)	(0.015)	(0.014)	(0.015)	(0.015)	(0.015)	(0.019)	(0.019)	(0.017)	(0.019)
Technical skills	0.01	0.02	0.05***	0.00	0.03*	0.05***	0.02	0.07***	0.02	0.06***
	(0.013)	(0.018)	(0.015)	(0.017)	(0.016)	(0.016)	(0.019)	(0.019)	(0.016)	(0.022)
Observations	900	900	900	900	900	890	890	890	890	890
R^2	0.520	0.320	0.371	0.299	0.266	0.532	0.319	0.381	0.306	0.276
Mean	0.679	0.606	0.460	0.576	0.564	0.679	0.606	0.460	0.576	0.564
R^2 Adj.	0.435	0.200	0.260	0.176	0.136	0.448	0.197	0.271	0.182	0.146
R^2 Adj. (w/o controls)	0.039	-0.001	0.080	0.005	0.015	0.075	-0.001	0.108	0.019	0.023
F Test	0.133	0.385	0.001	0.066	0.019	0.000	0.705	0.000	0.002	0.000
F Test Diff.	0.359	0.246	0.171	0.029	0.915	0.336	0.676	0.125	0.001	0.367

Note: Dependent variables are the averages of binary variables calculated over the 4 seasons (short rain 14 to long rain 16). Controls include education, literacy, gender, age and age squared of the farmer, land and cattle ownership, size and quality of the house, household size, whether the farmer is the household head, household head's gender, village fixed effects and enumerator-assignment fixed effects. Standard errors clustered at village level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7: Regressions of agricultural practices on improved skill sub-constructs

SKILLS	Mineral Fertilizer	Manure	Hybrid Seeds	Multiple Weeding	Hiring Labor
Cognitive skills (IRT)	0.01	0.03	-0.00	-0.04*	0.01
	(0.019)	(0.018)	(0.021)	(0.023)	(0.024)
NC Factor 1	0.01	-0.00	0.01	-0.01	0.00
(CESD)	(0.012)	(0.011)	(0.012)	(0.013)	(0.016)
NC Factor 2	0.00	0.02*	0.01	0.04***	-0.02
(Conscientiousness/Tenacity)	(0.014)	(0.013)	(0.015)	(0.015)	(0.015)
NC Factor 3	0.01	0.00	-0.01	-0.01	0.02
(LOC / Metacog / Openness)	(0.014)	(0.018)	(0.014)	(0.015)	(0.019)
NC Factor 4	0.00	-0.05**	0.01	-0.00	0.02
(Causes of poverty, negative items)	(0.014)	(0.019)	(0.020)	(0.021)	(0.019)
NC Factor 5	0.01	-0.00	0.03**	0.02	0.03**
(Attitude towards change / LOC_va)	(0.013)	(0.016)	(0.014)	(0.014)	(0.015)
NC Factor 6	0.02	0.02**	-0.00	0.02	-0.01
(CESD positive / Self-esteem / Risk av.)	(0.014)	(0.012)	(0.013)	(0.014)	(0.015)
Technical skills (IRT)	0.02	0.01	0.04***	0.00	0.04**
	(0.013)	(0.017)	(0.015)	(0.015)	(0.017)
Observations	890	890	890	890	890
R^2	0.522	0.325	0.377	0.315	0.274
Mean	0.679	0.606	0.460	0.576	0.564
R^2 Adj.	0.432	0.199	0.261	0.187	0.139
R^2 Adj. (w/o controls)	0.054	-0.001	0.090	0.023	0.019
F Test (Cog)	0.703	0.154	0.844	0.085	0.763
F Test (Noncog)	0.548	0.127	0.120	0.004	0.123
F Test (Tech)	0.154	0.529	0.006	0.984	0.034
Test NC diff.	0.958	0.092	0.338	0.035	0.163

Note: Dependent variables are the averages of binary variables calculated over the 4 seasons (short rain 14 to long rain 16). Controls include education, literacy, gender, age and age squared of the farmer, land and cattle ownership, size and quality of the house, household size, whether the farmer is the household head, household head's gender, village fixed effects and enumerator-assignment fixed effects. Standard errors clustered at village level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 8: Test-retest correlations, Cronbach's alpha and influence of enumerators by subgroups

	Test-retest correlation				Cronbach's alpha		R2 of enumerator FE		
Sample split:	Enumerator assigned for test and retest		By Cognitive Skill		By Cognitive Skill		By Cognitive Skill		
	Same	Different	Below Median	Above Median	Below Median	Above Median	All	Below Median	Above Median
Cognitive	0.91	0.82 (0.000)	0.67	0.68	0.54	0.48	0.04 [0.066]	0.05 [0.822]	0.08 [0.407]
Noncognitive	0.60	0.54 (0.116)	0.50	0.46	0.63	0.67	0.09 [0.000]	0.16 [0.000]	0.11 [0.029]
Technical	0.45	0.38 (0.096)	0.25	0.44	0.46	0.55	0.07 [0.000]	0.15 [0.000]	0.12 [0.008]

Note: The p-values between parentheses indicate significance of the difference between same and different enumerator assigned for test and retest. R2 of Enumerator FE is the R2 of a regression of the improved construct on (randomly) assigned enumerator fixed effects. The p-values between brackets provide the significance levels of the R2s. All p-values were obtained by Randomization inference (10,000 repetitions).

Table 9: Test-retest correlations as a function of order of the module in the survey instrument

Cognitive		Order in Retest			p-val all coef equal
		1	2	3	
Order in Test	1	0.87	0.90	0.87	0.049
	2	0.80	0.91	0.83	
	3	0.84	0.87	0.83	

Noncognitive		Order in Retest			p-val all coef equal
		1	2	3	
Order in Test	1	0.60	0.49	0.32	0.008
	2	0.57	0.62	0.57	
	3	0.50	0.54	0.73	

Technical		Order in Retest			p-val all coef equal
		1	2	3	
Order in Test	1	0.52	0.37	0.36	0.505
	2	0.51	0.30	0.47	
	3	0.37	0.42	0.38	

Table 10: Correlation of different skill proxy measures with subscales measuring same domain

		Correlation with subconstruct			Test-retest correlation	
Question	Corresponding subconstruct	Self-assessment	Other HH member	Other vlg member	Asking different person about same person	Asking same person about different person
How smart are you, how quickly do you understand things?	Raven	0.10	0.16	0.11	0.06	**0.08**
How well can you read and write?	Read	0.55	0.48	0.39	**0.23**	0.13
How good are you at math?	Math (timed)	0.31	0.37	0.27	**0.16**	0.14
How much does your life depend on your own action?	Locus of Control	0.13	0.00	0.02	**0.08**	0.07
How self-confident are you?	Self-esteem	0.13	0.03	0.06	0.11	**0.11**
How open to change are you?	Attitude towards change	0.22	0.06	0.05	**0.12**	0.11
How much do you think that you are someone who is organized?	Big 5 Conscientiousness	0.18	0.01	0.12	0.06	**0.12**
How hard working are you?	Organization/tenacity/ self-control	0.10	-0.02	0.04	**0.11**	0.08
How optimistic are you?	Optimism	0.11	0.05	0.02	0.08	**0.10**
How patient are you?	Patience	-0.01	0.00	0.08	**0.14**	0.10
How outgoing and social are you?	Big 5 Extraversion	0.12	0.05	0.07	**0.12**	0.08
How kind and sensitive are you?	Big 5 Agreeableness	0.15	-0.02	0.08	0.06	**0.15**
How easily do you get stressed?	Big 5 Neuroticism	-0.03	0.01	-0.03	0.10	**0.14**
How knowledgeable are you about farming techniques?	Technical skills	0.00	0.07	0.07	0.09	**0.16**

Note: table reports correlations between the 14 summary questions and the subconstruct most closely corresponding to each question. We use the demeaned measures of non-cognitive subconstructs, and the improved indexes of cognitive subconstructs and technical skills.

Table 11: Skills asked to a village informant: correlation with skills index and prediction of average rank of maize yield

Corresponding Skill Index	Correlation with corresponding skill index	Explanatory variables: Question asked to village	Regressions with the average rank of maize yield as dependent variable		
Cognitive	0.43	Level of education		4.73***	1.36
				(1.30)	(1.55)
Noncognitive	0.22	Active/Motivated		2.34	1.73
				(2.06)	(2.16)
Technical	0.15	Agricultural knowledge		6.43***	5.89***
				(1.72)	(1.74)
		Controls	Vil. FE	Vil. FE	All
		Observations	883	883	883
		R^2	0.239	0.310	0.372
		F Test		0.000	0.000

Note: Skill proxies obtained through village informant (CHW), scored on scale from 1 (low) to 3 (high). The right side of the table presents the correlation of the three questions asked to the village informant with the improved index of the corresponding skill, which the question intended to proxy. In the regressions, the dependent variable is the average rank of maize yields calculated over the 4 seasons (short rain 14 to long rain 16). Controls include education, literacy, gender, age and age squared of the farmer, land and cattle ownership, size and quality of the house, household size, whether the farmer is the household head, household head's gender, village fixed effects and enumerator-assignment fixed effects. Standard errors clustered at village level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 12: Congruence of Big 5 factors compared to United States, comparisons with other countries

		Extraversion	Agreeableness	Conscientious-ness	Neuroticism	Openness	**Average**
Kenya Sample	All	0.64	0.23	0.46	0.63	0.81	**0.55**
	Cog above median	0.72	0.27	0.51	0.70	0.76	**0.59**
	Cog below median	0.52	0.46	0.07	0.48	0.55	**0.42**
Colombia Sample	All	0.57	0.60	0.03	0.74	0.60	**0.51**
	Cog above median	0.69	0.64	0.49	0.76	0.73	**0.66**
	Cog below median	0.54	0.56	0.32	0.65	0.48	**0.51**
Other countries	Spain	0.95	0.93	0.95	0.96	0.86	**0.93**
	Dutch	0.94	0.91	0.94	0.94	0.64	**0.87**
	German	0.93	0.60	0.96	0.94	0.39	**0.76**

Note: To have comparable results for different countries, the same subset of questions was kept, using Big 5 data from other countries, and the factorial analysis uses varimax rotation. The absolute value of correlations between factor loads is used to calculate congruence. To maintain comparability the items are not corrected for acquiescence bias.

Table 13: Analysis of reliability and validity applying a similar survey in Colombia

Naive Score	Test-retest correlation	Cronbach's alpha of test	Nb of items	Enumerator Assigned (Test-retest correlation)		R^2 of Enum FE	Acquiescence Bias
				Same	Different		
Cognitive	0.78	0.62	6	0.75	0.75	0.04	-
Noncognitive	0.70	0.70	15	0.73	0.67	0.04	0.37
Technical	0.50	0.33	7	0.51	0.50	0.13	-

Improved Score	Test-retest correlation	Cronbach's alpha of test	Nb of items	Enumerator Assigned (Test-retest correlation)		R^2 of Enum FE
				Same	Different	
Cog (IRT and factor)	0.94	0.81	6	0.94	0.93	0.05
Noncognitive (factor)	0.64	0.69	7	0.66	0.61	0.04
Technical (IRT)	0.62	-	1	0.66	0.58	0.21