DISCUSSION PAPER SERIES

DP13271

MEASURING SKILLS IN DEVELOPING COUNTRIES

Rachid Laajaj and Karen Macours

DEVELOPMENT ECONOMICS

ISSN 0265-8003

MEASURING SKILLS IN DEVELOPING COUNTRIES

Rachid Laajaj and Karen Macours

Discussion Paper DP13271
Published 21 October 2018
Submitted 19 October 2018

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

MEASURING SKILLS IN DEVELOPING COUNTRIES

Abstract

Measures of cognitive, noncognitive, and technical skills are increasingly used to analyze the determinants of skill formation or the role of skills in economic decisions in developing and developed countries. Yet in most cases, these measures have only been validated in high-income countries. This paper tests the reliability and validity of some of the most commonly used skills measures in a rural developing context. A survey experiment with a series of skills measurements was administered to more than 900 farmers in western Kenya, and the same questions were asked again after three weeks to test the reliability of the measures. To test predictive power, the study also collected information on agricultural practices and production during the four following seasons. 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 correlations between questions are driven in part by the answering patterns of the respondents and the phrasing of the questions. Addressing both random and systematic measurement error using common psychometric practices and repeated measures leads to improvements and clearer predictions, but does not address all concerns. We replicate the main parts of the analysis for farmers in Colombia, and obtain similar results. The paper provides a cautionary tale for naïve interpretations of skill measures. It also points to the importance of addressing measurement challenges to establish the relationship of different skills with economic outcomes. Based on these findings, the paper derives guidelines for skill measurement and interpretation in similar contexts.

JEL Classification: O12, O13, O15

Keywords: N/A

Rachid Laajaj - r.laajaj@uniandes.edu.co
Universidad de Los Andes

Karen Macours - karen.macours@psemail.eu
Paris School of Economics and INRA

Measuring Skills in Developing Countries

Rachid Laajaj* and Karen Macours**

October 2018

ABSTRACT

Measures of cognitive, noncognitive, and technical skills are increasingly used to analyze the determinants of skill formation or the role of skills in economic decisions in developing and developed countries. Yet in most cases, these measures have only been validated in high-income countries. This paper tests the reliability and validity of some of the most commonly used skills measures in a rural developing context. A survey experiment with a series of skills measurements was administered to more than 900 farmers in western Kenya, and the same questions were asked again after three weeks to test the reliability of the measures. To test predictive power, the study also collected information on agricultural practices and production during the four following seasons. 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 correlations between questions are driven in part by the answering patterns of the respondents and the phrasing of the questions. Addressing both random and systematic measurement error using common psychometric practices and repeated measures leads to improvements and clearer predictions, but does not address all concerns. We replicate the main parts of the analysis for farmers in Colombia, and obtain similar results. The paper provides a cautionary tale for naïve interpretations of skill measures. It also points to the importance of addressing measurement challenges to establish the relationship of different skills with economic outcomes. Based on these findings, the paper derives guidelines for skill measurement and interpretation in similar contexts.

JEL codes: O12, O13, O15

Keywords: skills, measurement, agricultural productivity

** Universidad de Los Andes*

***Paris School of Economics and INRA*[1]

[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, 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 the FAO and DFID-ESRC for funding of the follow-up data collection rounds in Kenya. All errors are our own. Contacts: r.laajaj@uniandes.edu.co; karen.macours@psemail.eu

1. 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 in developing countries. 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 in this paper 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.

---

[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 of article using a measure of acquired agricultural skills. The authors fail to detect any impact on agricultural knowledge, and attribute it to large measurement errors.

[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 the measures 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.

[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.

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. Beyond the advantage of focusing on decision making in a specific domain, 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 and combining questions according to pre-existing scales leads to low internal consistency. Related, 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 clear 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 types of skills together explain up to

17 % of the variation in yield, with all three skill constructs being significant and with similar point estimates. Technical and noncognitive skills also help predict agricultural practices and input use, though with varying degrees.

The last part of the paper analyzes the different challenges related to 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 as helping explain some of the identified measurement problem, a finding that, to the best of our knowledge, has not been quantified before. Other challenges include the respondent's ability to understand the questions, order effects, response biases, anchoring, different factor structures, and specific challenges related to 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. At the very least, the 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. If anything 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. As such they suggest that studies that only attempt to measure 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,

[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.

2014; Bond and Long, 2018), and general issues with data quality in household surveys in developing countries (Judge and Schechter, 2009). 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 is universal (McCrae and Costa, 1997; Piedmont et al, 2002; McCrae and Terracciano 2005), while others claim that no one scale can apply to all cultures (Cross & Markus 1999).[7] Specifically for Africa, Schmitt et al (2007) finds slightly lower reliability and congruence than in other regions, but generally concludes that the five-dimensional structure is robust. In a context that may be closer to the conditions of household surveys implemented in micro-development economics, Gurven et al (2013) do not find robust support for the FFM using data from an orally administered survey on a forager-horticulturalists indigenous population in the Bolivian Amazon. In related ongoing work, we find strong evidence of lack of congruence for Big Five measures collected in household surveys across 20 low and middle income countries (Laajaj et al. 2018). 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

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

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 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 for which having reliable measures of agricultural knowledge and learning is key (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. Indeed, it is now widely recognized that 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 description of, and rational for, the calculation of 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 related to measurement error and derives lessons and practical recommendations for skills measurement. Section 5 concludes and the appendix provides details on methodologies and data, as well as additional empirical results.

2. THE SETTING, THE SAMPLE, AND THE QUESTIONNAIRE DESIGN

2.1. 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 they 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.

2.2. 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 1 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.

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 provides detailed descriptions of each of the tests.

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 the appendix 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

[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.

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.

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 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 the appendix presents all questions, and the first column indicates the sub-scale each of the questions was grouped under.

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

[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.

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

2.3. 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 of his 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.

---

[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.

[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.

2.4 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 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.

2.5. 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 20^{th} 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 surveys 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.

3. RELIABILITY AND VALIDITY OF DIFFERENT SKILL 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 2 provides a detailed methodological explanation of the different tests and methods used.

For each domain (cognitive, noncognitive and technical skills), we construct different measures of which we test the reliability and validity. 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. We also construct alternative aggregate measures, using exploratory factor analysis, item response theory, and corrections of response biases recommended in the psychometric literature. By comparing the reliability and validity of the different constructs, we demonstrate the importance of accounting for response patterns and latent factor structure.

3.1. Construction of the Indexes

The "naïve score" is calculated as the simple average of items (questions) that belong to pre-determined sub-domains. This has the advantage of simplicity and transparency, and mimics what is often done in practice. For the "improved" construct we apply 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. This subsection describes the methods and the insights gained from the different steps. The following subsections compare results when using the different indexes.

Correcting noncognitive items for Acquiescence Bias

Acquiescence bias ("yea-saying") refers to the respondent's tendency to agree (more than disagree), even when statements are contradictory. We correct for Acquiescence Bias in all noncognitive questions answered on a Likert scale, following common practice in psychometrics (Soto et al. 2008; Rammstedt, Kemper, and Borg, 2013; and references therein). To obtain the acquiescence score of each individual, we subtract the mean of reverse-coded items (after reversing these items) to the mean of positively-coded items and divide the result by 2. Then to correct for the acquiescence bias, for positive questions, we subtract this acquiescence score from the Likert score given by the answer, and for reverse questions, we add the acquiescence bias to the Likert score.[13]

Exploratory factor analysis to determine the number of factors in each construct

We conduct exploratory factor analysis (EFA) separately for cognitive, noncognitive and the technical skills, and determine the number of factors that should be extracted from the data. To do so, we pool all data for each domain (pooling for instance all noncognitive questions together), instead of relying on pre-determined scales. Hence, we let the data indicate the potential factor structure and related latent traits, following an approach also used by Ledesma and Valero-Mora (2007), Cunha, Heckman and Schennach (2010), and Attanasio et al (2015).

For the cognitive skills, we use the score for each of the five tests as inputs in the EFA. For the noncognitive and technical skills, we use each of the questions separately. We determine the number of latent factors that can be extracted from all the measures, using four different criteria commonly used in the psychometric literature (see appendix for details). The results of the exploratory analysis indicate that the cognitive and technical skills can best be measured by one factor each, while the underling latent factors for noncognitive skills corrected for Acquiescence Bias are best captured by 6 factors (appendix Table A2).

More details on the factor analysis of the noncognitive skills

The factor analysis explicitly accounts for the fact that answers to items are imperfect proxies of the true underlying latent traits. Latent factor models estimate the joint distribution of the latent factors and help remove some of this measurement error. We estimate factor loadings, then rotate the factor loadings using a principal factor analysis with quartimin rotation to predict the resulting

---

[13] The logic is that if asking about the same skills with reversed question leads to a different answer on average, it can be attributed to the "yea-saying" and should be adjusted in a way that brings positive and reverse questions to the same average.

Some studies use instead a joint correction for acquiescence bias and extreme response bias (adjusting for individual variance in responses), referred to as ipsatizing. The value of correcting for extreme response patterns is debated in the psychology literature (Hicks, 1970; Fischer and Milfont, 2010). Implementing this alternative correction in our data significantly worsened the reliability and validity of the construct, and we therefore do not consider it further.

factors.[14] Table 1 presents the resulting factor loads of the acquiescence bias corrected items, sorted by dominant factor. Strikingly, the factor analysis does not result in a clear categorization of variables into the theoretical scales and subscales. With the exception of the first factor, most factors seem to have a mix of items from different sub-constructs (in theory meant to be measuring different latent skills).[15] CESD items are a clear exception. They uniquely load on two factors, which do not include other items, and separate negative from positive attitudes.[16] On the other hand the Big Five personality trait division typically found in the psychometrics literature is not confirmed by the factor structure, with the exception of conscientiousness related items, which mostly load on the second factor.[17] The fourth factor further raises some doubts as it is uniquely composed of the reverse questions from the "causes of poverty" sub-construct, while the positive ones load on other factors.[18] This may indicate that it is at least partially driven by a response pattern rather than the actual belief about the causes of poverty. Overall these results raise concerns about whether the scales actually measure what they intend to. Despite the mixing of items, we attempted to discern a dominant interpretation for each factor, indicated in the last column of Table 1.

We use the factor loadings to aggregate the different noncognitive skills. To obtain the predicted factors, and following Attanasio et al (2015), items are assigned to the factor for which they have the highest factor loadings, with factor loads of other items set to 0.[19] To analyze the test-retest, and to guarantee we are comparing similar constructs, we apply the factor loading obtained from the first survey round (the test) also to the variable values of the second survey round (the retest). When redoing the exploratory factor analysis on the retest data, the factor structure is broadly similar, justifying the use of the same factor loads for both test and retest data. To obtain the aggregated non-cognitive skills construct, we use the average of the 6 factors.

The Use of Item Response Theory for Cognitive and Technical skills

[14] The quartimin rotation re-weights the factor loadings so that each variable mostly loads on one factor. That said, some variables still load on multiple factors after rotation, and no further restrictions were imposed.

[15] 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.

[16] The original scale development paper for the CESD (Radloff, 1977) similarly identifies a positive subscale/factor.

[17] A similar result is found when restricting the EFA to items of the Big Five. Items meant to measure distinct personality traits are mixed into various factors (Appendix Table A3). We return to this lack of congruence in section 4.

[18] The sorting or reverse versus non reverse questions in the factors is substantially stronger when not correcting for acquiescence bias (Table A7).

[19] Setting factor loads to 0 for all loadings other than the highest one helps reducing correlation between factors. The results are qualitatively very similar when we do not apply this correction.

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. The question's parameter can include its difficulty, its discriminant (how much the probability depends on the latent factor) and the possibility of pseudo-guessing. IRT has become the standard tool for high stakes tests such as GRE or GMAT and is believed to provide greater precision than classical test theory. We apply it to cognitive skills and to technical skills to obtain the two "improved" constructs, in each case assuming uni-dimensionality given the result of the EFA.[20]

For the technical skills, we used IRT pooling all items together. Only 3 items were removed because they had a discriminant opposed to the expected one (meaning that respondents were more likely to have a correct answer if they had a lower predicted latent skill). Of the remaining 32 items, 28 had a significant discriminant parameter at the 5% level (and 24 items at the 1% level), indicating that most items contributed to the assessment of the latent trait.

For the cognitive skills, we applied a mixed method, given the format of the questions and the requirements of IRT. We first use IRT to calculate the subconstruct of the numeracy questions, the Raven test and reading test.[21] We then used factor analysis using these three indexes and the scores of the digit span, the reverse digit span, and the timed math test to obtain one latent factor.

3.2. Reliability and construct validity

Test-retest correlation

To test reliability, we calculate the correlation between the same construct measured twice over a period of three weeks. The test-retest correlation provides an estimate of the share of a measures' variance that is driven by the variance of the true ability it is intended to measure. This is equivalent to one minus the share of variance explained by pure measurement error.[22] Intuitively high measurement error means that the true score is an imprecise measure and leads to a low test-

[20] We do not use IRT for the noncognitive skills, as the "difficulty" of each question is less applicable to noncognitive questions, and because IRT can only be used on discrete measures (after subtracting the acquiescence score are not).

[21] IRT cannot be used on digit span, reverse digit span, and the timed math test given that its subcomponents are not independent from each other.

[22] If measurement error is classical, the test-retest correlation gives a good indication of the signal to total variance ratio. On the other hand, the test-retest correlation can under- or over-state the signal to total variance ratio in case of non-classical measurement error. If the errors in measurement are positively correlated over time, for instance because both measures suffer from persistent acquiescence bias, the test-retest correlation will overstate reliability.

retest correlation, hence a low reliability. A change in the true value of the skills between the test and the retest would also lower the statistic. This is a desirable property since measures that change too much in the short run would not be good for long terms outcomes. A threshold of minimum .7 test-retest correlation is often applied. All estimates are done using z-scores of the relevant constructs and subconstructs (i.e. after subtracting the mean and dividing by the standard deviation).

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. 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.[23]

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.[24]

[23] 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.

[24] The fact of being surveyed during the test may affect the answers in the retest, and hence the test-retest statistic. Table A4 in the appendix 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

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 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 2 for details. For the purpose of statistical analysis, a minimum threshold of .7 is often applied.

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.[25] The conclusions for the aggregate constructs parallel those obtained from the test-retest correlations. The Cronbach's alpha is above the bar 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.[26] 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 noncognitive 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.[27] The technical skills construct reaches a Cronbach's alpha of .54, which remains quite low given that it includes 32

scores (9 items with 1% significant difference, or 16 items with 5% significant difference) leads to test-retest statistics between .32 and .35.

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

[26] The high Cronbach's alpha of the cognitive is consistent with Table A1.A 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.

[27] 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).

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.

3.3. Predictive validity

To further investigate validity, we test to what extent the skills constructs predict real life outcomes. We analyze whether skills correlate with agricultural productivity and practices, and how much predictive power the measurements have for such outcomes. 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, independent 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.

Virtually all farmers in the region produce maize, hence getting higher yield is a common objective that higher skilled farmers could be expected to achieve. Because random measurement error causes attenuation bias and a reduction of the R^2, an increase in the coefficient (of the normalized skills measures) and in the R^2 are signs of measures that are less noisy.[28]

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.[29] 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 he's been using it for

[28] 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.

[29] 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.

a while, or may have been using it exactly because he knew about it. Finally, 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.

Yield predictions by construct

The key outcome variable we use to test predictive validity is maize yield. Maize is the main crop in the region of study, and the only crop that households in the sample have in common. As yield in rainfed agriculture is known to be a particularly noisy 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.[30]

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. The first five columns in Table 3 do not include controls and demonstrate the share of variation explained by the three skill constructs (R-squared). Columns 6 to 10 report results from a specification with controls and shows whether the skill measures remain significant after controlling for observed farmer, household and village characteristics.[31] Significant coefficients on skills in the later regression 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.

---

[30] We use the rank because it is less sensitive to extreme values (Athey and Imbens, 2016). The appendix shows similar regressions using the average of the log of the yield for the same seasons. The results are qualitatively similar but less precise.

[31] 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.

Results in Table 3 broadly show that the three types of skills matter, as all three coefficients are significant and combined the measures explain a substantial share of the variation in yields. The R-squared of the naïve constructs without any control is 12.1 percent (column 1), compared to 14.5 percent when using the improved constructs (column 2). Interestingly this last figure is practically the same as the R-squared 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-squared to 16.6 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 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.[32] 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

[32] 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.

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.[33]

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. All variables are measured as z-scores. The regressions are estimated with the controls, but the adjusted R-squared in absence of controls is added in the bottom of Table 4. Using sub-constructs increases the R-squared for the predictive model for yields from 12.1 percent with naïve constructs to 13.9 with the sub-constructs.

None of the cognitive tests on their own has a significant relationship with productivity or input use, and the F-test for joint significance is also low. The same finding 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.

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.

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

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.

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.[34] 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 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.

[34] 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.

4. 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.

4.1 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-squared 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 was sent to ask the questions, and this is up to 7 percent for technical skills and 9 percent for noncognitive skills.[35] This suggests that a large amount of noise is introduced by the 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.[36] Standard practice for test-retest correlations is to have the test administrated in similar conditions. However from a practical point of view, 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 of the latent skill. We find that assigning a different enumerator leads to a drop of .09, .06 and .07 in the test-retest correlation of the cognitive construct, the noncognitive one, and the technical one respectively. Hence enumerator effects

[35] 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.

[36] 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%.

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.

4.2 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 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.

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.[37]

4.3 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.[38]

Instead the test-retest correlation for noncognitive skills was highest, and indeed above the .7 threshold, when it comes last, and differences between different test-retest combinations are significant. In contrast, the test-retest 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, and any potential annoyance generated by such questions does not affect the other modules.

4.4 Response Biases

---

[37] 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.

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

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, and this happened more often for lower educated individuals.

Strikingly, the acquiescence score shows a strong negative correlation with yields (coefficient is -6.15 of the average rank of maize yield), significant at the 5%, indicating that 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. 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 can confuse respondents). While it may be tempting to instead drop reverse items, the benefits of being able to measure and correct acquiescence bias seem to clearly outweigh the costs.

Of course, acquiescence bias is only one of the possible response biases. Other biases include "extreme response bias" and "middle response bias". In this study, correcting for extreme response bias by standardizing the standard deviations of responses did not lead to improvements in validity or reliability, as has been found elsewhere. That said, the distribution of many of the positively-phrased noncognitive questions is highly skewed to the right, suggesting it remains a potential concern.[39] Finally, "social desirability bias" may lead a respondent to answer what he believes to give the best impression, or what he believes the enumerator wants to hear. The oral administration of a survey can reduce the impression of anonymity and increase the bias. In surveys related to impact evaluations, respondents may also believe their answers could affect the probability to receive benefits and attempt to answer strategically. Such biases are difficult to

[39] This is the case even after eliminating variables showing the least variation after the piloting.

avoid when using self-reports rather than observed outcomes and the extent to which they affect the measures and generate a non-random noise is difficult to assess. While this holds for many outcomes other than skills, such response patterns are likely to be more pronounced in questions that require a subjective assessment, and provide an additional challenge for the interpretation of the findings on skills, as the way of answering questions may be related to personality itself.

4.5 Anchoring and the use of other sources of information

The previous sub-section raises the possibility of different response biases affecting the answers. 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, not unlike the use of recommendation letters to evaluate skills in other settings. A priori, the random measurement of a person's skills should be noisier when asking somebody else, as the other person is likely to have asymmetric information about the true skill level. Yet if it helps to address the systematic bias, or if the introduced random measurement error is limited, this may constitute an alternative or complementary manner to measure skills. To test these 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).[40] 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 in the sample informed about 2 other people in the sample. 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 are good, but all other correlations are very low. Strikingly, for 7 out of 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 again points to the importance of systematic answering patterns, which appear as important as the actual skill differences between the two people about

[40] 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.

whom the proxy reports. Possibly, the relative influence of answering patterns is as high for proxy reports because information about another person is less salient than about oneself.

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 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. The results show that using proxy reports by a village informant we obtain broadly similar results as those obtained with direct reports, with all three proxies having predictive power for yield when no other controls are included, and the CHWs report on farmers' technical skills being particularly robust.

These results are striking, as they 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, clearly such proxy measures are not a good solution when one aims to obtain comparable skill measures across villages. More generally, we interpret these results as evidence of a large remaining amount of measurement error in the self-reported outcomes. Results with other proxy respondents are broadly similar but are less significant and robust, possibly because the response bias cannot be cancelled out. 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.

4.6 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 for details). We restrict the analysis to the 23 items from the Big Five included in our study, so as to compare constructs based on the same items. Table 12 presents the congruence with respect to the same items administrated in the United States where the BFI has been validated many times. It shows an average congruence across the five factors that is only .40. For comparison using the factor loads of the same items administrated in Spain, Holland and Germany results in congruence coefficients between .76 and .93.

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 that the lack of understanding of some items or response patterns is stronger than in high income contexts and does not allow detecting the same factor structure even if the true latent factors are similar. If the main problem was lack of understanding, we would expect the congruence with the US to be higher among individuals that are above the median of cognitive skills compared to the ones below the median, but this is not what we find.

The improved indexes used the factor analysis of items corrected for acquiescence bias. For comparison, the factor loads of the items without correction for acquiescence bias are presented in Appendix Table A7. This shows far less consistency in how the items are sorted, making it difficult to attribute a dominant interpretation to the factors. Instead, specific factors appear to be pooling questions with the same answer types and phrasing. The first, fifth and sixth factors are only pooling items that are not reversed, and second, third and fourth factors are pooling reversed items together. Almost all factors that are not in a one to five Likert-scale sorted themselves together in the sixth factor. In sum, without the correction of acquiescence bias, the share of the variance in the responses driven by acquiescence bias and other response patterns overwhelms variance in responses that is driven by the latent traits that are intended to be measured. A factor analysis that is driven by phrasing rather than actual content is arguably of little interest, hence prior correction for acquiescence bias is fundamental. The findings hence suggest it is advisable to correct for acquiescence bias first, and then systematically analyze the latent factor structure through exploratory factor analysis when using noncognitive skills data. Naïve interpretation of item aggregation following pre-existing constructs without such analysis is likely to lead to erroneous conclusions regarding noncognitive skills.

4.7 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.[41] In contrast, questions that had sufficient variance in the responses often were questions for which the right answer may depend on the context.[42] 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.

4.8 From Kenya to Colombia, evidence of similar results in a different context

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 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.

As for Kenya, the cognitive measure is the only one that shows high level of consistency both across tests and across time. By contrast the measure of technical skills is quite noisy, with a test retest that goes from .50 in the naïve measure to .62 when calculated with Item Response Theory. This is higher than in Kenya but remains quite low. Both the cognitive and technical indexes

[41] 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.

[42] 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.

become more precise with the improved measure. The results confirm that the use of Item Response Theory can bring substantial gains in dealing with the noise.

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 .65, in line with the Kenyan results. 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 squared of enumerators fixed effects is substantial. Both are signs that the interaction with enumerators is part of the challenge.

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. 2018, combining Big Five data covering more than 300,000 individuals from 30 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.

5. **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. 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.

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.[43]

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.

[43] 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.

REFERENCES

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

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* (forthcoming)

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.

Judge, G. and L. Schechter, 2009. “Detecting Problems in Survey Data using Benford’s Law”, *Journal of Human Resources*, 44(1): 1-24.

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., 2018. “How well do the Big Five measure personality in developing country surveys?”, mimeo.

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. 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.

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

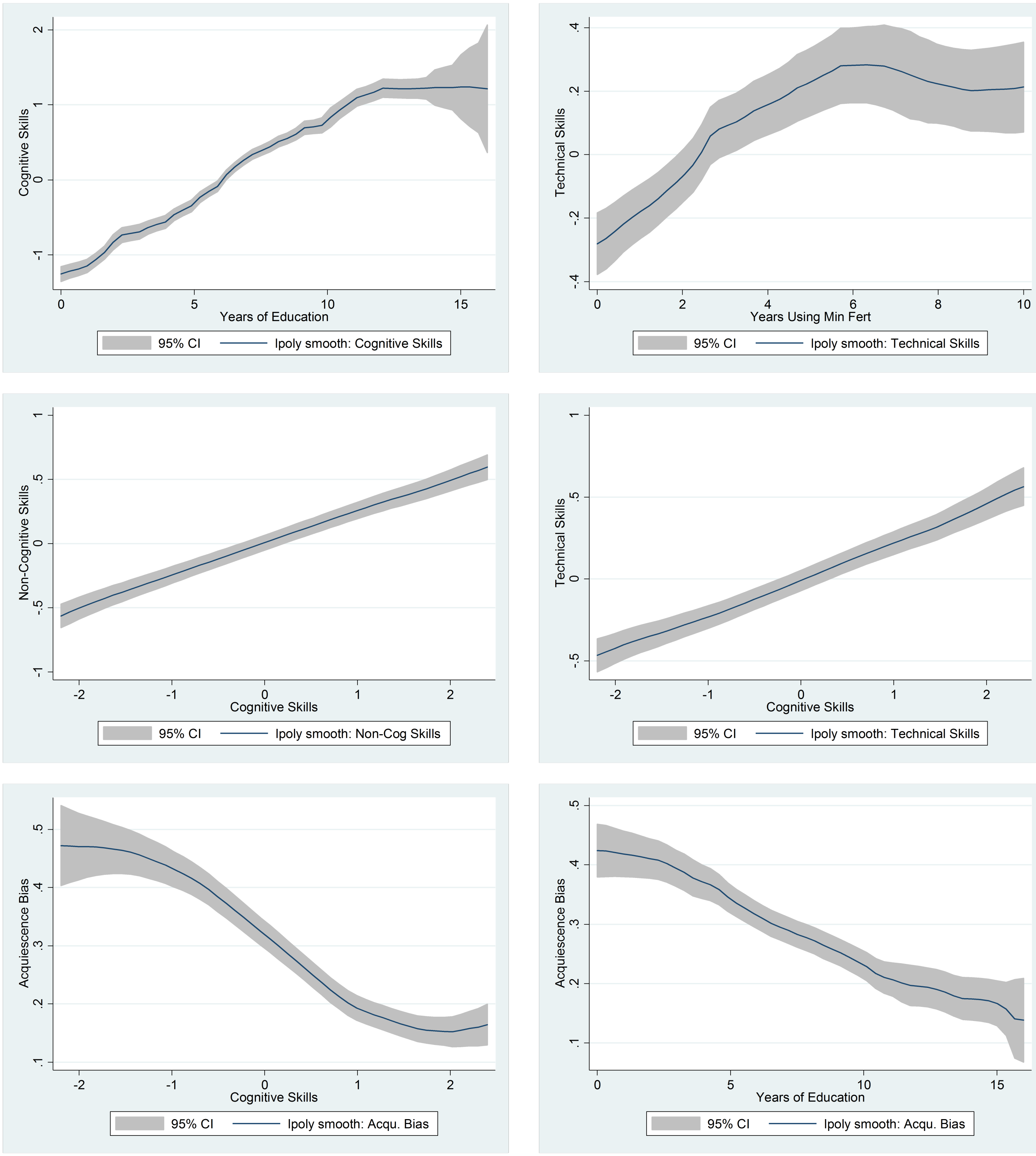


Table 1: Factor Loads of noncognitive items (corrected for acquiescence bias)

Question's short name	Fact Load 1	Fact Load 2	Fact Load 3	Fact Load 4	Fact Load 5	Fact Load 6	Dominant interpretation
cesd17	**0.60**	-0.03	0.03	0.03	-0.02	0.04	CESD
cesd6	**0.60**	0.01	-0.04	0.01	0.08	0.05	
cesd15	**0.59**	0.02	0.10	0.00	-0.10	-0.06	
cesd14	**0.59**	-0.03	-0.02	-0.02	0.05	0.02	
cesd10	**0.58**	0.00	-0.01	-0.04	-0.01	-0.06	
cesd1	**0.58**	-0.01	-0.10	0.03	0.06	-0.03	
cesd3	**0.57**	0.03	0.07	-0.01	-0.07	0.01	
cesd18	**0.55**	0.01	0.02	-0.02	-0.06	-0.06	
cesd11	**0.54**	0.01	-0.02	0.02	0.01	0.08	
cesd19	**0.53**	-0.02	0.09	0.00	-0.03	-0.02	
cesd2	**0.51**	-0.08	-0.02	0.00	0.12	0.07	
cesd20	**0.48**	-0.04	0.00	0.09	0.00	0.02	
cesd13	**0.47**	0.03	-0.03	-0.03	-0.08	-0.02	
cesd9	**0.41**	0.13	-0.20	0.01	0.15	0.09	
cesd5	**0.39**	0.02	-0.15	0.03	0.04	-0.09	
cesd12	**0.35**	-0.04	0.07	-0.07	-0.06	0.26	
cesd7	**0.27**	0.08	-0.18	0.07	0.06	-0.10	
patience1	**0.11**	0.04	-0.07	-0.01	0.11	-0.03	
BF_C1	0.03	**0.51**	0.00	0.20	-0.03	-0.01	Conscientiousness Tenacity
BF_C5	0.03	**0.49**	0.17	0.19	0.07	-0.05	
BF_C3	-0.01	**0.48**	0.00	-0.12	0.16	0.06	
BF_C4	-0.03	**0.45**	0.03	-0.11	0.05	0.12	
BF_A4	0.10	**0.45**	0.01	-0.20	0.00	-0.18	
BF_N1	0.14	**0.41**	0.08	-0.08	0.02	0.13	
BF_C6	0.12	**0.39**	0.23	0.15	0.06	-0.03	
tenac2	-0.06	**0.39**	-0.13	0.11	-0.06	0.07	
tenac1	-0.06	**0.36**	0.09	0.06	0.10	0.06	
BF_O1	-0.04	**0.31**	0.06	0.27	-0.05	0.21	
BF_N2	0.14	**0.30**	0.00	0.10	0.01	0.23	
tenac3	0.10	**0.30**	0.16	0.13	0.07	0.08	
BF_E4	0.03	**0.28**	0.26	0.17	0.10	-0.09	
metacog2	0.09	**0.27**	0.16	0.20	0.15	0.06	
LOC1	-0.09	**0.24**	-0.05	-0.04	0.20	0.04	
selfesteem1	0.08	**0.23**	0.18	0.10	0.06	0.14	
tenac4	-0.03	**0.18**	-0.02	0.08	0.08	0.11	
riskav3	-0.08	**0.03**	-0.07	0.02	-0.02	0.00	
LOC4	0.01	0.05	**0.46**	-0.03	-0.04	0.02	Locus of Control Metacognitive Openness
LOC3	0.10	-0.10	**0.43**	0.15	0.17	0.05	
BF_A1	0.02	0.30	**0.40**	-0.26	-0.18	-0.02	
LOC2	0.10	0.14	**0.38**	0.10	0.05	0.10	
BF_O4	0.02	0.16	**0.37**	0.18	0.02	0.14	
metacog3	0.08	0.22	**0.37**	0.12	0.18	-0.03	
metacog1	0.09	0.24	**0.35**	0.04	0.17	-0.09	
BF_E2	0.10	0.09	**0.34**	0.17	0.02	0.13	
BF_O2	0.05	0.18	**0.34**	0.19	0.06	0.07	
BF_A2	0.03	0.26	**0.33**	0.12	-0.04	-0.13	
causepov2	0.03	-0.02	**0.33**	0.26	0.08	0.04	
LOC7	0.06	-0.01	**0.32**	0.18	0.18	-0.07	
BF_A3	-0.08	0.27	**0.32**	0.25	0.00	-0.04	
optim3	-0.03	-0.08	**0.30**	0.04	0.19	0.17	

BF_O5	0.02	0.05	**0.28**	-0.24	-0.07	0.03	
tenac6	-0.07	-0.08	**0.27**	-0.04	0.07	0.01	
LOC5	-0.06	0.04	**0.26**	0.01	0.10	0.11	
BF_N3	-0.07	0.18	**0.24**	0.07	0.12	0.12	
LOC6	0.13	-0.15	**0.15**	0.10	0.11	0.13	
causepov6	-0.04	0.03	0.03	**0.62**	0.10	-0.05	Causes of poverty (all reversed items)
causepov5	0.06	0.06	-0.07	**0.56**	-0.06	-0.01	
causepov7	0.09	0.02	0.08	**0.55**	0.07	-0.12	
causepov9	0.06	0.00	0.07	**0.55**	0.09	-0.02	
causepov8	-0.02	-0.01	0.18	**0.51**	0.07	0.00	
optim1	-0.03	0.13	-0.41	**0.21**	0.08	0.06	
att_change4	0.03	-0.02	-0.04	0.08	**0.56**	0.10	Attitude toward Change Locus of Control with visual aid
att_change5	-0.01	0.04	0.01	0.07	**0.54**	0.12	
BF_E1	0.09	0.10	-0.02	-0.11	**0.45**	-0.23	
att_change2	0.06	-0.16	0.09	0.13	**0.43**	0.02	
BF_C2	0.08	0.14	0.09	-0.11	**0.43**	-0.28	
BF_E3	0.05	0.17	-0.17	-0.17	**0.41**	-0.09	
att_change3	-0.04	-0.20	0.19	0.15	**0.40**	0.03	
selfesteem4	-0.02	0.21	-0.02	-0.01	**0.39**	0.20	
BF_N4	0.08	0.11	0.04	-0.15	**0.32**	0.07	
BF_O3	-0.04	-0.33	0.22	-0.09	**0.31**	0.01	
causepov4	-0.04	0.01	0.11	-0.36	**0.20**	0.03	
LOC_va2	0.07	-0.03	0.07	0.14	**0.17**	0.02	
LOC_va1	0.13	-0.09	0.07	0.13	**0.16**	-0.02	
causepov3	0.05	-0.04	0.08	-0.55	**0.15**	-0.03	
LOC_va3	0.07	-0.15	0.03	0.07	**0.13**	-0.02	
causepov1	-0.07	0.03	0.01	-0.57	**0.07**	-0.10	
cesd4	0.15	-0.03	0.10	-0.13	0.02	**0.40**	CESD positive Self-esteem Risk aversion
selfesteem2	-0.05	0.24	-0.26	-0.02	0.15	**0.39**	
cesd21	0.12	-0.05	0.10	-0.09	0.02	**0.38**	
selfesteem3	0.10	0.07	-0.14	0.21	0.04	**0.38**	
cesd16	0.22	-0.09	0.06	-0.09	0.05	**0.38**	
att_change1	-0.04	0.00	0.10	-0.18	-0.06	**0.37**	
cesd8	0.06	-0.05	0.17	0.01	-0.02	**0.36**	
optim2	-0.03	0.18	-0.01	-0.02	0.12	**0.33**	
tenac5	0.02	0.25	0.07	-0.09	-0.04	**0.32**	
riskav2	-0.08	-0.06	-0.03	-0.06	-0.13	**0.20**	
riskav1	-0.09	-0.15	0.07	-0.01	0.04	**0.13**	

Note: items with possible acquiescence bias are demeaned by subtracting the person-specific acquiescence score.

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)	Same as panel 2A		
	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	2.75***	2.27***	2.23***	1.60**	3.74***	2.47**	2.54**	3.27***	3.30***	4.17***
	(0.754)	(0.735)	(0.815)	(0.802)	(0.768)	(1.150)	(1.167)	(1.099)	(1.157)	(1.118)
Noncognitive skills	4.10***	3.90***	4.87***	4.40***	5.40***	3.88***	3.97***	4.31***	4.23***	4.68***
	(0.716)	(0.701)	(0.854)	(0.847)	(0.895)	(0.754)	(0.727)	(0.943)	(0.908)	(0.922)
Technical skills	3.41***	4.40***	5.50***	6.41***		0.61	1.41	2.38**	3.04***	
	(0.828)	(0.830)	(0.989)	(0.951)		(0.924)	(0.866)	(1.058)	(1.050)	
Observations	903	893	903	893	893	903	893	903	893	893
R-squared	0.121	0.145	0.145	0.169	0.122	0.431	0.438	0.441	0.449	0.441
Controls	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
R^2 Adj. (w/o controls)	0.118	0.142	0.143	0.166	0.120	0.331	0.338	0.342	0.351	0.343
F Test	0	0	0	0	0	2.08e-07	1.43e-09	2.98e-09	0	0
F Test Diff.	0.523	0.237	0.0890	0.00727	0.256	0.0225	0.105	0.399	0.702	0.767

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	-0.40	0.13	0.25	0.20	0.28	0.22
	(1.067)	(1.111)	(1.138)	(1.126)	(1.115)	(1.123)
Reading	1.17	1.46	1.41	1.50	1.46	1.45
	(1.102)	(1.101)	(1.080)	(1.081)	(1.065)	(1.090)
Raven	0.60	0.72	0.66	0.55	0.64	0.62
	(1.054)	(1.098)	(1.098)	(1.112)	(1.080)	(1.112)
Digit Span	0.48	0.39	0.45	0.45	0.55	0.52
	(0.752)	(0.743)	(0.752)	(0.751)	(0.731)	(0.745)
Math (timed)	1.62	2.16**	2.11*	2.22**	2.05*	2.13**
	(1.141)	(1.077)	(1.080)	(1.085)	(1.072)	(1.071)
CESD	2.04***					
	(0.763)					
Locus of Control (LOC)	-0.11					
	(0.975)					
Self-esteem	-0.06					
	(0.697)					
Causes of poverty	2.24***					
	(0.841)					
Attitude towards change	0.03					
	(0.670)					
Tenacity / Organization	2.22***					
	(0.770)					
Metacognitive	0.45					
	(0.742)					
Optimism	0.67					
	(0.689)					
Risk aversion	-0.70					
	(0.671)					
Big 5 Agreeableness	1.03	1.99**				
	(0.790)	(0.768)				
Big 5 Extraversion	0.33		1.20*			
	(0.733)		(0.713)			
Big 5 Conscientiousness	-1.12			1.39**		
	(0.871)			(0.658)		
Big 5 Neuroticism	0.90				2.18***	
	(0.680)				(0.605)	
Big 5 Openness	-0.08					0.97
	(0.764)					(0.763)
Other noncognitive	0.04					
	(0.695)					
Intercrop /Compost	-0.03	-0.20	-0.19	-0.20	-0.24	-0.18
	(0.745)	(0.744)	(0.737)	(0.743)	(0.737)	(0.745)
Maize	0.51	1.04	1.16	1.12	0.94	1.13
	(0.770)	(0.741)	(0.753)	(0.747)	(0.758)	(0.762)
Banana	0.58	0.68	0.47	0.46	0.44	0.53
	(0.814)	(0.796)	(0.794)	(0.796)	(0.779)	(0.796)
Soya	0.11	-0.06	0.02	-0.01	-0.03	0.00
	(0.761)	(0.767)	(0.772)	(0.770)	(0.761)	(0.769)
Fertilizer	0.35	0.87	0.86	0.90	0.91	0.89
	(0.801)	(0.783)	(0.786)	(0.786)	(0.778)	(0.792)
Observations	900	902	902	902	902	902
R-squared	0.454	0.426	0.422	0.422	0.427	0.421
R2 Adj.	0.339	0.318	0.313	0.313	0.319	0.312
R2 Adj. (w/o controls)	0.139	0.112	0.118	0.115	0.124	0.114
F Test (Cog)	0.250					
F Test (Noncog)	0.0006					
F Test (Tech)	0.919					
Test NC diff.	0.0636					

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)	2.14*	3.12***	3.69***	3.17**	2.45**	3.62***	3.55***
	(1.247)	(1.185)	(1.148)	(1.206)	(1.190)	(1.153)	(1.132)
NC Factor 1	1.90**	2.52***					
(CESD)	(0.726)	(0.713)					
NC Factor 2	0.09		1.73**				
(Conscientiousness/Tenacity)	(0.813)		(0.665)				
NC Factor 3	0.38			1.71**			
(LOC / Metacog / Openness)	(0.721)			(0.691)			
NC Factor 4	2.59**				3.62***		
(Causes of poverty, negative items)	(0.989)				(0.687)		
NC Factor 5	0.10					0.77	
(Attitude towards change / LOC_va)	(0.700)					(0.690)	
NC Factor 6	1.09						2.24***
(CESD positive / Self-esteem / Risk av.)	(0.662)						(0.583)
Technical skills (IRT)	1.20	1.68*	1.94**	1.85**	1.40	1.92**	1.77*
	(0.877)	(0.894)	(0.861)	(0.876)	(0.857)	(0.882)	(0.895)
Observations	893	893	893	893	893	893	893
R-squared	0.443	0.429	0.424	0.423	0.435	0.419	0.427
R2 Adj.	0.339	0.327	0.321	0.320	0.334	0.316	0.324
R2 Adj. (w/o controls)	0.140	0.126	0.126	0.124	0.137	0.123	0.123
F Test (Cog)	0.089						
F Test (Noncog)	0.000						
F Test (Tech)	0.174						
Test NC diff.	0.157						

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-squared	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
R2 Adj.	0.435	0.200	0.260	0.176	0.136	0.448	0.197	0.271	0.182	0.146
R2 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-squared	0.522	0.325	0.377	0.315	0.274
Mean	0.679	0.606	0.460	0.576	0.564
R2 Adj.	0.432	0.199	0.261	0.187	0.139
R2 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 Enum 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	0.67	0.68	0.54	0.48	0.04	0.05	0.08
Noncognitive	0.60	0.54	0.50	0.46	0.63	0.67	0.09	0.16	0.11
Technical	0.45	0.38	0.25	0.44	0.46	0.55	0.07	0.15	0.12

Note: R2 of enumerator FE is the R2 of a regression of the improved construct on (randomly) assigned enumerator fixed effects.

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 corresponding subconstruct					Test-retest correlation	
Question	Corresponding subconstruct	Self-assessment	Other HH member	Other village member	Retest 2nd village member	Average 2 village members	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	0.04	0.11	0.06	**0.08**
How well can you read and write?	Read	0.55	0.48	0.39	0.34	0.43	**0.23**	0.13
How good are you at math?	Math (timed)	0.31	0.37	0.27	0.25	0.33	**0.16**	0.14
How much does your life depend on your own action?	Locus of Control	0.13	0.00	0.02	0.07	0.06	**0.08**	0.07
How self-confident are you?	Self-esteem	0.13	0.03	0.06	0.05	0.08	0.11	**0.11**
How open to change are you?	Attitude towards change	0.22	0.06	0.05	0.07	0.07	**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	0.11	0.06	**0.12**
How hard working are you?	Organization/tenacity/ self-control	0.10	-0.02	0.04	0.03	0.05	**0.11**	0.08
How optimistic are you?	Optimism	0.11	0.05	0.02	0.05	0.04	0.08	**0.10**
How patient are you?	Patience	-0.01	0.00	0.08	0.00	0.07	**0.14**	0.10
How outgoing and social are you?	Big 5 Extraversion	0.12	0.05	0.07	0.02	0.07	**0.12**	0.08
How kind and sensitive are you?	Big 5 Agreeableness	0.15	-0.02	0.08	0.02	0.09	0.06	**0.15**
How easily do you get stressed?	Big 5 Neuroticism	-0.03	0.01	-0.03	0.03	0.02	0.10	**0.14**
How knowledgeable are you about farming techniques?	Technical skills	0.00	0.07	0.07	0.04	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		5.08***	1.47
				(1.240)	(1.496)
Noncognitive	0.22	Active/Motivated		3.45*	3.00
				(1.773)	(1.924)
Technical	0.15	Agricultural knowledge		5.44***	4.85***
				(1.487)	(1.505)
		Controls	Vil. FE	Vil. FE	All
		Observations	887	887	887
		R-squared	0.274	0.355	0.422
		F Test		0	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	0.42	0.56	0.43	0.27	0.34	**0.40**
	Cog above median	0.49	0.44	0.47	0.30	0.28	**0.40**
	Cog below median	0.60	0.50	0.44	0.35	0.27	**0.43**
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.

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)		R2 of Enum FE	Acquiescence Bias
				Same	Different		
Cognitive	0.84	0.65	5	0.84	0.83	0.05	-
Noncognitive	0.70	0.70	15	0.73	0.66	0.05	0.37
Technical	0.50	0.33	7	0.51	0.50	0.14	-

Improved Score	Test-retest correlation	Cronbach's alpha of test	Nb of items	Enumerator Assigned (Test-retest correlation)		R2 of Enum FE
				Same	Different	
Cog (IRT and factor)	0.94	0.81	6	0.94	0.93	0.06
Noncognitive (factor)	0.66	0.54	7	0.68	0.65	0.22
Technical (IRT)	0.62	-	1	0.66	0.58	0.04

APPENDIX TABLES

Table A1.A: List of tests of the cognitive skill module

Oral Math questions	An oral 9-item math test containing short math puzzles and increasing in level of difficulty. Each puzzle contains one or two phrases including the question. Answers can be open or multiple choice, some questions are framed to mimic math problems farmers might need to solve in daily live but they never require actual farming knowledge.
Reading	A reading comprehension test. Farmers were given 3 small texts of 5 to 7 lines (2 in Swahili and 1 in English or vice versa). In each exercise, they were asked to read the text and then asked 4 questions about the text (2 multiple choice and 2 open). They were allowed to consult the text while answering. The texts present (fictitious) information regarding agricultural inputs and were inspired by guidelines found on input packages. No time limit was imposed.
Raven	The 36 item Raven Colored Progressive matrices, measuring visual processing and analytical reasoning and executive functioning. People are asked to select the missing puzzle piece from a bigger image, with patterns becoming increasingly difficult.
Math (timed)	A timed math tests with 160 basic additions, subtractions and multiplications. Respondents are given a sheet of paper with the questions, and get 3 minutes to complete as many as they can.
Digit Span	The digit span forwards and backwards (measuring short-term memory and executive functioning). People are asked (without visual aid) to repeat a series of numbers that the enumerator reads to them. It starts with a 3 number series with the series becoming progressively longer as long as the respondent manages to repeat the series correctly. Afterwards, they are asked to repeat different series backwards (starting from a 2 number series and again gradually increasing the length of the series).

Correlations between cognitive measures, education and literacy

	Oral Math questions	Reading	Raven	Math (timed)	Digit Span	Literacy dummy	Years of Education
Oral Math questions	1						
Reading	0.61	1					
Raven	0.52	0.51	1				
Math (timed)	0.57	0.62	0.46	1			
Digit Span	0.46	0.48	0.42	0.46	1		
Literacy dummy	0.41	0.57	0.36	0.55	0.40	1	
Years of Education	0.58	0.70	0.47	0.65	0.49	0.65	1

The first 3 sub-constructs are calculated using item response theory

Table A1.B: List of items of the noncognitive skill module

Subscale / Naïve Category	Question's short name	Question	Positive or Reversed	Answer scale
Locus of Control	LOC1	It's not always wise for you to plan too far ahead because many things turn out to be a matter of good or bad fortune	R	1 to 5
	LOC2	Your life is determined by your own actions	P	1 to 5
	LOC3	When you get what you want, it's usually the result of actions	P	1 to 5
	LOC4	You feel like what happens in your life is mostly determined by others	R	1 to 5
	LOC5	Getting what you want requires pleasing the influential people	R	1 to 5
	LOC6	Please tell me which of the two propositions you most agree with 1. Each person is primarily responsible for his/her own success or failure in life 2. One's success or failure is a matter of his/her destiny	NA	Choose 1 answer
	LOC7	Only those who inherited large farms become successful farmers	R	1 to 5
Beans (locus-of-control with visual aid)	LOC_va1	What do you think explains why some people have more ECONOMIC SUCCESS than others? The enumerator records number of beans allocated to "Effort or Decisions" (Compared to "luck or birth") [with additional explanations and visual aid]	P	Allocate 10 beans
	LOC_va2	What do you think explains why some people are more PRODUCTIVE IN AGRICULTURE than others? The enumerator records number of beans allocated to Effort or Decisions (Compared to luck or birth) [with additional explanations and visual aid]	P	Allocate 10 beans
	LOC_va3	Between effort and good decision-making, how much do you think that each one matters for being productive in agriculture [with additional explanations, visual aid and the respondent allocating beans to the possible options]	NA	Allocate 10 beans
Self-esteem	selfesteem1	You feel that you have many good qualities	P	1 to 5
	selfesteem2	All in all, you are inclined to feel that you are a failure	R	1 to 5
	selfesteem3	On the whole, you are satisfied with yourself	P	1 to 5
	selfesteem4	You certainly feel useless at times	R	1 to 5
Causes of Poverty	causepov1	Poor people are poor because they lack the ability to manage money	P	1 to 5
	causepov2	Poor people are poor no matter what they do	R	1 to 5
	causepov3	Poor people are poor because they waste their money on inappropriate items.	P	1 to 5
	causepov4	Poor people are poor because they do not actively seek to improve their lives.	P	1 to 5
	causepov5	Poor people are poor because they are exploited by rich people.	R	1 to 5
	causepov6	Poor people are poor because the distribution of land between poor and rich people	R	1 to 5
	causepov7	Poor people are poor because they lack opportunities because they come from poor	R	1 to 5
	causepov8	Poor people are poor because they lack luck	R	1 to 5
	causepov9	Poor people are poor because they are born with less talent	R	1 to 5
Attitude toward Change	att_change1	When you learn about a new farming technique, compared to most of your neighbours: 1. You are more willing to try first 2. You let others try it first	NA	Choose 1 answer
	att_change2	On the farm: 1. You prefer doing routine things 2. You prefer doing something new	NA	Choose 1 answer
	att_change3	Choose one of the following 2 options: 1. You generally prefer leaving things the way they are 2. You generally prefer changing things	NA	Choose 1 answer
	att_change4	You often go to the plots of fellow farmers to observe what they do	P	1 to 5
	att_change5	You have tried to experiment on your own plot some of the techniques learned from fellow farmers.	P	1 to 5
Tenac/Organiz	tenac1	You can think of many times when you persisted with work when others quit	P	1 to 5

/Self-cont.	tenac2	You normally don't rest until the job is fully completed	P	1 to 5
	tenac3	Your family and friends would say you are a very organized person	P	1 to 5
	tenac4	You are much happier if everything is planned well ahead of time	P	1 to 5
	tenac5	You often spend money and regret later that you spent it	R	1 to 5
	tenac6	When you see something you like, you buy it right away, rather than waiting to see how you feel about it later	R	1 to 5
Metacognitive	metacog1	You think a lot about something before taking a decision about it	P	1 to 5
	metacog2	You set goals for yourself in order to direct your activities	P	1 to 5
	metacog3	You spend a lot of time reflecting on your mistakes in order to improve your farming practices	P	1 to 5
Optimism	optim1	In uncertain times you usually expect the best.	P	1 to 5
	optim2	Things go wrong for me most of the time.	R	1 to 5
	optim3	You talk more about solutions than problems.	P	1 to 5
CESD	cesd1	During the last 7 days, how many days … were you bothered by things that usually don't bother you?	R	0 to 7
	cesd2	… did you not feel like eating? (your appetite was poor)	R	0 to 7
	cesd3	… did you feel that you could not shake off the blues even with help from your family and friends?	R	0 to 7
	cesd4	… did you feel that you were just as good as other people?	P	0 to 7
	cesd5	… did you have trouble keeping your mind on what you were doing?	R	0 to 7
	cesd6	… did you feel depressed?	R	0 to 7
	cesd7	… did you feel that everything you did was an effort?	R	0 to 7
	cesd8	… were you hopeful about the future?	P	0 to 7
	cesd9	… did you think your life had been a failure?	R	0 to 7
	cesd10	… did you feel fearful?	R	0 to 7
	cesd11	… was your sleep restless?	R	0 to 7
	cesd12	… were you happy?	P	0 to 7
	cesd13	… did you talk less than usual?	R	0 to 7
	cesd14	… did you feel lonely?	R	0 to 7
	cesd15	… people were unfriendly?	R	0 to 7
	cesd16	… did you enjoy life?	P	0 to 7
	cesd17	… did you have crying spells?	R	0 to 7
	cesd18	… did you feel sad?	R	0 to 7
	cesd19	… did you feel that people disliked you?	R	0 to 7
	cesd20	... could you not get 'going'?	R	0 to 7
	cesd21	… did you feel that you are moving forward in life?	P	0 to 7
Big 5 Agreeableness	BF_A1	You see yourself as someone who tends to find fault with others	R	1 to 5
	BF_A2	You see yourself as someone who has a forgiving nature	P	1 to 5
	BF_A3	You see yourself as someone who is generally trusting	P	1 to 5
	BF_A4	You see yourself as someone who is sometimes rude to others	R	1 to 5
Big 5 Conscientious ness	BF_C1	You see yourself as someone who does things carefully and completely	P	1 to 5
	BF_C2	You see yourself as someone who can be somewhat careless	R	1 to 5
	BF_C3	You see yourself as someone who tends to be disorganized	R	1 to 5
	BF_C4	You see yourself as someone who tends to be lazy	R	1 to 5

	BF_C5	You see yourself as someone who does things efficiently (quickly and correctly)	P	1 to 5
	BF_C6	You see yourself as someone who makes plans and sticks to them	P	1 to 5
Big 5 Extraversion	BF_E1	You see yourself as someone who is reserved; keeps thoughts and feelings to self	R	1 to 5
	BF_E2	You see yourself as someone who generates a lot of enthusiasm	P	1 to 5
	BF_E3	You see yourself as someone who tends to be quiet	R	1 to 5
	BF_E4	You see yourself as someone who is outgoing, sociable	P	1 to 5
Big 5 Neuroticism	BF_N1	You see yourself as someone who is depressed, or gets blue	R	1 to 5
	BF_N2	You see yourself as someone who is relaxed, handles stress well	P	1 to 5
	BF_N3	You see yourself as someone who doesn't get easily upset, and is emotionally stable	P	1 to 5
	BF_N4	You see yourself as someone who gets nervous easily	R	1 to 5
Big 5 Openness	BF_O1	You see yourself as someone who is clever, thinks a lot	P	1 to 5
	BF_O2	You see yourself as someone who has an active imagination	P	1 to 5
	BF_O3	You see yourself as someone who likes work that is the same every time (routine)	R	1 to 5
	BF_O4	You see yourself as someone who likes to think and play with ideas	P	1 to 5
	BF_O5	You see yourself as someone who doesn't like artistic things (plays, music)	R	1 to 5
Risk Aversion	riskav1	You never try anything you are not sure of	R	1 to 5
	riskav2	A person can get rich by taking risks	P	1 to 5
	riskav3	Imagine that you can chose between 5 games in which you will flip a coin. First I am going to explain you 5 games, and then I am going to ask you which one you would prefer to play. In the 1st game, you get 2500 Ksh if you get head, and 2500 Ksh if you get tail 2nd game 2000Ksh vs 4000 Ksh 3rd game 1500 Ksh vs 5500 Ksh 4th game 1000 Ksh vs 7000 Ksh 5th game 0 Ksh vs 10000 Ksh Which game would you pick? The question includes more explanations and a table to visualize the choices.	P	Choose 1 answer. An index is calculated based on the response.
Patience	patience1	People often make decisions that involve trading off something soon for something else later. For example, people sometimes have to choose between having some money soon, or having more money later. The next set of questions asks how you make such decisions. There are no right or wrong answers. For each pair of options please indicate which you prefer between option (1) and option (2). Would you prefer: (1) 1000 Ksh now, or (2) 900 Ksh in one month? [following questions asked as long as the respondent picks (1)]: (1) 1000 Ksh now, or (2) 1100 Ksh in one month? (1) 1000 Ksh now, or (2) 1300 Ksh in one month? (1) 1000 Ksh now, or (2) 1500 Ksh in one month? (1) 1000 Ksh now, or (2) 2000 Ksh in one month? (1) 1000 Ksh now, or (2) 2500 Ksh in one month?	NA	Choose 1 answer per question. With Visual representation. An index is calculated based on the responses

Table A1.C: List of items of the technical skill module

Subscale	Question	Listed Answers (when not an open question)
Maize	If one wants to cover the soil with maize stalk, should you apply or leave maize stalks:	1. Between the lines 2. On the lines, as close as possible to the next crop
	When planting hybrid maize in rows, how many seeds per hole should be applied?	
	What quantity of planting fertilizer should you apply per seed of maize :	1. Less than half of a Teaspoon 2. Half of a Teaspoon 3. A full Teaspoon 4. Two Teaspoons
	Where should you apply commercial planting fertilizer for maize: [distances shown with ruler]	1. In the same hole mixed up with the soil 2. In the same hole in contact with the seed 3. 5 cm from the hole 4. 15 cm from the hole
	Imagine a maize field is inclined like this [show]. If on such a field you need to put top dressing on the ground, where do you put the fertilizer?	1. Uphill 2. Downhill 3. On the side 4. Same hole
	How many weeks after planting should you apply commercial top dressing to maize?	
	Where should you apply commercial top-dressing fertilizer for maize:	1. In contact with the plant 2. Spread closely around the plant 3. At 15 cm from the plant 4. Apply through broadcasting
Banana	When cultivating bananas, how many adult trees should be left per banana mat?	
	(for banana) How many of the youngest trees (suckers) should you leave on a mat?	
	When do you need to prune the leaves of banana trees:	1. Never 2. When the leaves start turning yellow 3. When the leaves are completely dry 4. Prune only the green leaves
	When planting bananas, what is the optimal distance between banana trees:	1. 1m x 1m 2. 2m x 2m 3. 2m x 3m 4. 3m x 3m
	If you want to keep only one of two healthy suckers, which one should you leave:	1. The one facing the sunrise 2. The one facing the sunset 3. The youngest one
	What can you do to prevent the Cigar-end disease: [show image]	1. Remove the male part 10 days after bunch formation 2. Remove the male part 10 days before bunch formation 3. Make sure that the male part does not fall 4. Increase the water provided to the tree
Soya	When planting soybean in rows, how many seeds per hole should be applied?	
	When planting soybean, what is the optimal distance between seeds:	1. 10cm x 30cm 2. 20cm x 30cm 3. 30cm x 30cm 4. 50cm x 30cm 5. 50cm x 5cm
	How is powder biofertilizer used when planting soybeans:	1. It is applied directly to the soil and then soybean is planted 2. The biofertilizer is mixed with the seed and a sticky solution if needed 3. Put the soybean first and then put biofertilizer on top of it 4. Fill a bucket of water, pour the biofertilizer in, and then the soybean is soaked in it
	How much time should be left between mixing the seeds with powder biofertilizer and planting the seeds:	1. 5 min 2. 4 hours 3. 8 hours 4. 24 hours
intercrop/ compost	Imagine that someone intercrops beans and maize in the same field. In which order should he plant:	1. Plant the maize first and then the beans 2. Plant the beans first and then the maize 3. Plant both at the same time 4. He should not intercrop maize and beans
	Among the following crop rotations, which one is best for long term soil fertility:	1. Rotate soya with soya 2. Rotate soya with maize 3. Rotate maize with millet 4. Rotate beans with soya

	How can you use Nepia grass and Desmodium to control maize stalk borer: [Answers come with corresponding images]	1. Plant Desmodium with the maize and put Nepia grass around the parcel 2. Plant Nepia grass with the maize, and Desmodium around the parcel 3. Intercrop both Desmodium and Nepia grass with the maize 4. Rotate Maize with Desmodium and Nepia grass
	Imaging you are making compost. While it is maturing, where should it be stored:	1. In a uncovered pit 2. In an uncovered heap 3. In a covered heap 4. Inside of the house
	Is it better to apply compost when it is humid or when it is dry?	1. Humid. 2. Dry
	If you want to use the waste from your own cattle to improve the fertility of the soil, is it better to:	1. Apply some manure everyday in part of the field2. Keep it covered and then apply it all at once3. Keep it uncovered and then apply it all at once
	Please tell me all the different ways you can you use to check whether the compost is ready to be applied to the field?	10 possible components were listed and multiple answers were allowed.
Fertilizer	In the cultivation of banana, which fertilizer should be applied at planting?	4 pictures of fertilizers are shown Multiple answers allowed
	In the cultivation of banana, which fertilizer should be applied at the vegetative stage?	4 pictures of fertilizers are shown Multiple answers allowed
	In the cultivation of banana, which fertilizer should be applied at flowering?	4 pictures of fertilizers are shown Multiple answers allowed
	[A picture of a fertilizer is shown] Do you think it is:	1. Planting Fertilizer 2. Top Dressing 3. Both
	[A picture of a fertilizer is shown] Do you think it is:	1. Planting Fertilizer 2. Top Dressing 3. Both
	[A picture of a fertilizer is shown] Do you think it is:	1. Planting Fertilizer 2. Top Dressing 3. Both
	[A picture of a fertilizer is shown] Do you think it is:	1. Planting Fertilizer 2. Top Dressing 3. Both
	Which ones of these fertilizers should be used on Sweet Potatoes?	4 pictures of fertilizers are shown Multiple answers allowed
	Which ones of these fertilizers provide Nitrogen?	4 pictures of fertilizers are shown Multiple answers allowed
	Which ones of these fertilizers provide Phosphorous?	4 pictures of fertilizers are shown Multiple answers allowed
	Which ones of these fertilizers provide Potassium?	4 pictures of fertilizers are shown Multiple answers allowed

Table A2: Number of factors to be retained according to different methods

	Number of factors recommended according to the following methods:				
	Kaiser's eigenvalue rule	Cattell's scree plot	Velicer's MAP rule	Horn's parallel analysis (p95)	Retained for analysis
Cog	1	1	1	1	1
Tech	1	1 or 3	1	8	1
Noncog naïve	7	3 or 7 or 9	4	9	7
Noncog demeaned	22	6	3	10	6
Big 5 demeaned	1	1 or 5	1	3	5

Table A3: Factor Loads of Big 5 personality traits

Question's short name	Fact Load 1	Fact Load 2	Fact Load 3	Fact Load 4	Fact Load 5
BF_C7	**0.57**	0.08	0.00	0.05	-0.01
BF_C1	**0.54**	0.00	-0.01	-0.05	0.08
BF_C8	**0.53**	0.05	0.02	0.07	-0.03
BF_E8	**0.37**	-0.08	0.25	0.11	-0.08
BF_A5	**0.29**	0.10	0.19	-0.03	-0.06
BF_C4	**0.22**	0.13	0.06	0.09	0.18
BF_N2	0.07	**0.46**	-0.12	0.04	0.05
BF_O4	0.15	**0.34**	0.08	0.06	-0.16
BF_N1	0.07	**0.32**	0.09	0.06	0.10
BF_N5	-0.09	**0.31**	0.27	0.08	0.04
BF_C5	0.12	**0.30**	0.00	0.02	0.14
BF_O3	0.18	**0.27**	0.12	-0.12	-0.04
BF_E4	0.15	**0.26**	0.12	0.03	-0.23
BF_A4	0.13	-0.01	**0.41**	0.06	0.03
BF_A1	0.07	0.05	**0.31**	-0.05	0.21
BF_O8	0.17	0.20	**0.20**	0.03	-0.12
BF_O9	-0.04	0.09	**0.12**	-0.03	-0.04
BF_E2	-0.01	-0.02	0.03	**0.39**	0.01
BF_E5	0.07	-0.07	-0.10	**0.34**	0.06
BF_C2	0.04	0.09	-0.02	**0.34**	0.02
BF_N8	0.07	0.10	-0.07	**0.23**	0.05
BF_O7	-0.25	0.05	0.11	**0.18**	-0.19
BF_A8	0.04	0.09	0.15	0.09	**0.37**

All items were demeaned to correct for acquiescence bias

Table A4: Comparison of naïve scores in test and retest

	Average Naïve score				
	Test		Retest		p-value of difference
	Average Naïve score	Standard Deviation	Average Naïve score	Standard Deviation	
Cognitive	0.425	0.167	0.454	0.169	0.000
Noncognitive	3.419	0.282	3.458	0.281	0.000
Technical	0.409	0.107	0.431	0.108	0.000

Only observations available for both test and retest are kept

Table A5: Measures of reliability and validity for noncognitive measures corrected for acquiescence bias

		2A - Naïve Score			
	Construct	Test retest correlation	Chronbach's alpha of test	Chronbach's alpha of retest	Nb of items
	Noncog DE-MEANED	0.53	0.78	0.79	15
Decomposition by subconstruct:					
Noncog	Locus of Control	0.45	0.50	0.51	9
	Self-esteem	0.32	0.37	0.41	4
	Causes of poverty	0.34	0.69	0.74	9
	Attitude towards change	0.41	0.39	0.46	5
	Organization/tenacity/self-control	0.29	0.37	0.32	6
	Metacognitive ability	0.31	0.44	0.55	4
	Optimism	0.23	0.04	0.09	3
	Risk aversion	0.12	0.03	0.14	2
	Big 5 Agreeableness	0.25	0.43	0.38	4
	Big 5 Extraversion	0.23	0.32	0.26	4
	Big 5 Conscientiousness	0.33	0.57	0.57	6
	Big 5 Neuroticism	0.26	0.41	0.36	4
	Big 5 Openness	0.19	0.32	0.34	5
	CESD	0.41	0.82	0.85	21

Table A6: The effect of time (instrumented date of survey) on scores

	Test			Retest		
	Cognitive	Noncog	Technical	Cognitive	Noncog	Technical
I. First Stage						
Assigned Order	1.363***	1.363***	1.364***	1.718***	1.704***	1.701***
	(0.0410)	(0.0411)	(0.0410)	(0.109)	(0.109)	(0.108)
II. Second Stage						
Day of survey	0.000442	0.00233	0.00137**	-0.000373	-6.48e-05	-1.09e-05
	(0.000851)	(0.00144)	(0.000536)	(0.000698)	(0.00119)	(0.000449)
Observations	922	919	921	895	884	893

Note: First stage instruments the date (day since start of survey) with the randomly assigned order of survey assigned in planning. All regressions include randomly assigned enumerator fixed effects. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table A7: Factor Loads of noncognitive items
(items taken as they are, without correction for acquiescence bias)

Question's short name	Fact Load 1	Fact Load 2	Fact Load 3	Fact Load 4	Fact Load 5	Fact Load 6	Fact Load 7	Positive or Reverse	Dominant interpretation
BF_C5	**0.62**	-0.01	0.00	0.09	-0.04	-0.05	0.07	P	Acquiescence Bias: 100% Positive Items
BF_C6	**0.56**	0.06	-0.03	0.08	-0.02	-0.02	0.03	P	
BF_C1	**0.54**	-0.02	-0.01	-0.01	-0.02	-0.08	0.15	P	
BF_A3	**0.53**	-0.08	0.08	-0.05	-0.05	-0.03	-0.08	P	
BF_O2	**0.53**	0.02	0.07	-0.11	0.01	0.12	-0.07	P	
metacog3	**0.52**	0.03	-0.04	0.08	0.05	0.06	-0.14	P	
BF_O4	**0.49**	-0.01	0.07	-0.09	0.10	0.05	-0.11	P	
BF_O1	**0.49**	-0.07	0.10	-0.14	0.05	0.00	0.15	P	
tenac3	**0.48**	0.05	-0.02	-0.01	0.04	0.02	0.03	P	
BF_E4	**0.47**	0.00	0.00	0.07	-0.04	-0.02	-0.06	P	
metacog1	**0.46**	0.05	-0.10	0.13	0.00	0.07	-0.12	P	
BF_E2	**0.46**	0.07	0.05	-0.11	0.03	0.06	-0.04	P	
metacog2	**0.46**	0.03	0.03	0.04	0.02	0.06	0.06	P	
LOC2	**0.45**	0.04	-0.02	0.03	0.13	-0.02	-0.12	P	
BF_A2	**0.44**	0.01	-0.01	0.05	-0.05	-0.10	-0.11	P	
selfesteem1	**0.43**	0.04	-0.01	-0.03	0.05	0.06	0.04	P	
BF_N3	**0.43**	-0.06	-0.03	-0.03	0.01	0.10	-0.01	P	
tenac1	**0.41**	-0.08	-0.10	0.08	-0.01	0.01	0.07	P	
BF_N2	**0.41**	0.07	-0.05	-0.06	0.07	0.03	0.17	P	
tenac2	**0.38**	-0.06	-0.08	-0.10	-0.02	-0.03	0.15	P	
LOC3	**0.33**	0.07	0.03	-0.09	0.05	0.15	-0.21	NA	
optim3	**0.27**	-0.03	-0.04	-0.05	0.05	0.16	-0.10	P	
tenac4	**0.27**	-0.05	-0.09	-0.06	0.06	0.04	0.06	P	
riskav2	**0.13**	-0.07	-0.18	-0.25	0.06	0.01	0.05	P	
cesd15	0.05	**0.61**	0.02	0.02	0.00	-0.15	-0.14	R	100 % reversed CESD
cesd10	-0.02	**0.60**	-0.04	0.03	-0.06	-0.01	-0.01	R	
cesd17	0.01	**0.59**	0.05	-0.01	0.09	-0.03	-0.02	R	
cesd6	0.04	**0.58**	0.02	-0.02	0.02	0.16	0.07	R	
cesd3	0.04	**0.57**	0.02	0.04	0.04	-0.08	-0.02	R	
cesd18	0.00	**0.57**	-0.01	0.00	-0.02	-0.09	-0.08	R	
cesd14	-0.04	**0.57**	0.00	0.01	0.03	0.10	0.03	R	
cesd1	-0.03	**0.55**	0.03	-0.01	-0.01	0.08	0.14	R	
cesd19	0.02	**0.54**	0.02	0.02	0.04	-0.07	-0.15	R	
cesd11	0.02	**0.51**	0.04	0.01	0.09	0.03	0.09	R	
cesd13	0.02	**0.48**	-0.05	-0.03	-0.01	-0.06	0.01	R	
cesd2	-0.04	**0.47**	0.02	-0.02	0.08	0.20	0.04	R	
cesd20	-0.01	**0.45**	0.13	-0.05	0.06	0.02	0.06	R	
cesd5	-0.11	**0.37**	0.00	0.06	-0.09	0.00	0.17	R	
cesd7	0.01	**0.28**	0.05	0.00	-0.22	0.10	0.24	R	
causepov6	0.01	-0.04	**0.69**	0.02	-0.03	0.05	0.01	R	100% reversed with

causepov7	0.01	0.07	**0.63**	0.07	-0.05	0.01	-0.03	R	1 NA
causepov8	0.07	0.00	**0.62**	0.01	-0.03	0.09	-0.08	R	
causepov9	0.01	0.03	**0.62**	0.03	0.01	0.06	-0.01	R	
causepov5	-0.02	0.03	**0.58**	0.03	-0.02	-0.08	0.09	R	
causepov2	0.02	0.03	**0.42**	0.17	0.07	0.05	-0.17	R	
LOC7	0.00	0.05	**0.34**	0.27	0.03	0.05	-0.17	R	
LOC6	-0.09	0.10	**0.21**	0.03	0.14	0.13	-0.12	NA	
riskav1	-0.13	-0.04	**0.14**	0.02	0.06	0.08	-0.07	R	
BF_C3	0.14	0.03	0.07	**0.45**	-0.03	0.02	0.09	R	100 % reversed with 1 NA
selfesteem4	-0.01	-0.04	0.15	**0.44**	0.14	0.11	0.11	R	
LOC5	-0.12	-0.06	0.19	**0.42**	0.10	-0.10	-0.04	R	
BF_E1	-0.12	0.09	0.05	**0.38**	-0.08	0.17	-0.04	R	
BF_C2	-0.03	0.11	0.07	**0.38**	-0.11	0.20	-0.13	R	
LOC1	-0.06	-0.05	0.12	**0.38**	0.01	-0.01	0.05	R	
BF_E3	-0.12	0.06	-0.02	**0.37**	-0.03	0.16	0.03	R	
LOC4	-0.01	0.02	0.19	**0.36**	0.07	-0.13	-0.17	R	
BF_A4	0.13	0.13	-0.04	**0.36**	-0.10	-0.08	-0.01	R	
BF_A1	0.18	0.09	0.00	**0.35**	-0.01	-0.15	-0.21	R	
BF_N1	0.20	0.14	0.12	**0.35**	0.02	-0.02	0.08	R	
BF_N4	-0.05	0.08	0.04	**0.35**	0.05	0.13	-0.02	R	
BF_C4	0.22	0.03	0.09	**0.34**	-0.02	0.00	0.08	R	
tenac5	0.04	0.03	0.10	**0.33**	0.12	-0.10	0.08	R	
selfesteem2	-0.05	-0.06	0.11	**0.31**	0.15	-0.01	0.29	R	
optim2	-0.01	-0.02	0.17	**0.31**	0.11	0.04	0.13	R	
tenac6	-0.09	-0.01	0.15	**0.21**	-0.01	0.02	-0.13	R	
BF_O5	-0.01	0.07	-0.01	**0.21**	-0.02	0.00	-0.11	R	
patience1	-0.04	0.09	-0.04	**0.13**	0.01	0.05	0.06	NA	
cesd4	0.01	0.04	-0.09	0.04	**0.56**	0.00	0.00	P	100 % Positive or NA, and dominated by CESD and other non 1 to 5 formats
cesd21	-0.02	0.01	-0.04	0.03	**0.54**	0.00	0.01	P	
cesd16	-0.03	0.09	-0.04	-0.01	**0.52**	0.08	0.06	P	
cesd8	0.06	-0.03	0.09	-0.05	**0.52**	-0.04	-0.08	P	
cesd12	0.02	0.27	-0.04	-0.01	**0.38**	-0.07	0.02	P	
att_change1	0.01	0.01	-0.02	0.03	**0.16**	0.07	-0.04	NA	
att_change4	0.18	-0.03	-0.04	0.03	0.08	**0.42**	0.08	NA	
att_change5	0.25	-0.06	-0.05	0.08	0.08	**0.40**	0.08	NA	
att_change2	0.05	0.02	0.09	0.07	0.03	**0.27**	-0.04	NA	
LOC_va2	-0.01	0.04	0.21	0.00	0.02	**0.26**	-0.11	P	
LOC_va1	-0.05	0.12	0.19	-0.03	-0.02	**0.26**	-0.15	P	
att_change3	0.01	-0.06	0.14	0.08	0.04	**0.25**	-0.12	NA	
LOC_va3	-0.09	0.07	0.12	-0.09	-0.03	**0.24**	-0.09	P	
BF_O3	-0.24	0.00	0.13	0.18	0.04	**0.18**	-0.18	R	
causepov3	0.02	0.04	-0.57	0.04	0.02	**0.13**	-0.11	P	
causepov4	0.06	-0.04	-0.39	0.09	0.06	**0.10**	-0.07	P	
causepov1	0.03	-0.04	-0.62	0.03	-0.06	**0.07**	-0.05	P	

cesd9	0.00	0.34	-0.01	0.18	0.01	0.10	**0.37**	R	Mixed
optim1	0.09	-0.07	-0.03	-0.14	0.03	-0.04	**0.24**	P	
selfesteem3	0.19	0.02	0.06	-0.13	0.17	0.01	**0.21**	R	
riskav3	0.01	-0.07	0.01	-0.03	-0.04	-0.03	**0.06**	P	

Table A8: Regressions of the average of log maize yield across seasons on skill constructs

	SKILLS CONSTUCTS USED AS REGRESSORS:									
VARIABLES	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	0.10**	0.10**	0.08*	0.06	0.15***	0.14**	0.18***	0.18**	0.20***	0.23***
	(0.041)	(0.041)	(0.043)	(0.044)	(0.042)	(0.068)	(0.067)	(0.071)	(0.072)	(0.070)
Noncognitive skills	0.13***	0.11***	0.16***	0.13**	0.17***	0.12**	0.12**	0.15***	0.13**	0.15**
	(0.040)	(0.041)	(0.046)	(0.048)	(0.050)	(0.047)	(0.053)	(0.057)	(0.061)	(0.061)
Technical skills	0.16***	0.16***	0.25***	0.27***		0.03	0.04	0.10*	0.13**	
	(0.045)	(0.043)	(0.050)	(0.048)		(0.052)	(0.048)	(0.063)	(0.061)	
Observations	900	890	900	890	890	900	890	900	890	890
R-squared	0.058	0.066	0.078	0.086	0.055	0.294	0.305	0.306	0.315	0.310
Controls	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
R2 Adj. (w/o controls)	0.0553	0.0631	0.0753	0.0825	0.0526	0.170	0.181	0.183	0.192	0.188
F Test	4.39e-09	7.80e-11	0	0	1.97e-09	0.00556	0.000910	0.000141	2.18e-05	2.01e-05
F Test Diff.	0.690	0.578	0.0815	0.0174	0.794	0.302	0.278	0.720	0.777	0.437

Note: Dependent variable is the average log 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. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A9: Regressions of the average log maize yield on naïve skill sub-constructs

	(1)	(2)	(3)	(4)	(5)	(6)
Oral math questions	0.02	0.04	0.05	0.04	0.05	0.05
	(0.064)	(0.064)	(0.065)	(0.065)	(0.065)	(0.064)
Reading	0.06	0.06	0.06	0.06	0.06	0.06
	(0.062)	(0.061)	(0.060)	(0.061)	(0.060)	(0.061)
Raven	0.04	0.05	0.04	0.04	0.04	0.04
	(0.059)	(0.058)	(0.058)	(0.058)	(0.058)	(0.059)
Digit Span	0.00	0.00	0.00	0.00	0.01	0.01
	(0.041)	(0.040)	(0.040)	(0.040)	(0.039)	(0.040)
Math (timed)	0.09	0.11*	0.11*	0.11*	0.10*	0.11*
	(0.061)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)
CESD	0.08*					
	(0.046)					
Locus of Control	0.02					
	(0.058)					
Self-esteem	-0.06					
	(0.045)					
Causes of poverty	0.07					
	(0.059)					
Attitude towards change	-0.08**					
	(0.039)					
Tenacity / Organization	0.14***					
	(0.044)					
Metacognitive	0.02					
	(0.044)					
Optimism	0.06					
	(0.043)					
Risk aversion	-0.03					
	(0.032)					
Big 5 Agreeableness	-0.02	0.04				
	(0.046)	(0.043)				
Big 5 Extraversion	0.02		0.04			
	(0.037)		(0.035)			
Big 5 Conscientiousness	-0.02			0.05		
	(0.055)			(0.045)		
Big 5 Neuroticism	0.01				0.05	
	(0.044)				(0.040)	
Big 5 Openness	0.01					0.04
	(0.044)					(0.040)
Other noncognitive	0.02					
	(0.041)					
Intercrop /Compost	0.02	0.00	0.00	0.00	0.00	0.00
	(0.044)	(0.043)	(0.043)	(0.043)	(0.042)	(0.043)
Maize	0.04	0.05	0.05	0.05	0.04	0.05
	(0.048)	(0.048)	(0.046)	(0.046)	(0.048)	(0.047)
Banana	0.04	0.05	0.04	0.04	0.04	0.04
	(0.046)	(0.045)	(0.044)	(0.044)	(0.044)	(0.044)
Soya	0.00	-0.01	-0.01	-0.01	-0.01	-0.01
	(0.038)	(0.037)	(0.038)	(0.038)	(0.037)	(0.037)
Fertilizer	-0.02	0.00	-0.00	-0.00	0.00	-0.00
	(0.047)	(0.048)	(0.047)	(0.047)	(0.047)	(0.047)

Observations	897	899	899	899	899	899
R-squared	0.325	0.299	0.299	0.299	0.299	0.299
R2 Adj. (w/o controls)	0.0673	0.0528	0.0560	0.0551	0.0558	0.0538
F Test (Cog)	0.307					
F Test (Noncog)	0.0437					
F Test (Tech)	0.822					
Test NC diff.	0.0299					

Note: Dependent variable is the average of log 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. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A10: Regressions of the average log of maize yield on improved skill sub-constructs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Cognitive skills	0.17**	0.19***	0.21***	0.19***	0.18***	0.21***	0.20***
	(0.068)	(0.067)	(0.066)	(0.067)	(0.066)	(0.066)	(0.064)
Factor 1	0.05	0.07*					
(CESD)	(0.041)	(0.043)					
Factor 2	0.02		0.06				
(Conscientiousness/Tenacity)	(0.051)		(0.047)				
Factor 3	0.03			0.05			
(LOC/Metacog/Openness)	(0.047)			(0.042)			
Factor 4	0.05				0.10***		
(Causes of poverty, negative items)	(0.052)				(0.038)		
Factor 5	-0.03					-0.00	
(Attitude towards change/Beans)	(0.043)					(0.045)	
Factor 6	0.06						0.09**
(CESD positive/Confidence/Risk aversion)	(0.040)						(0.037)
Technical skills	0.04	0.05	0.05	0.05	0.04	0.06	0.05
	(0.048)	(0.050)	(0.048)	(0.049)	(0.048)	(0.049)	(0.050)
Observations	890	890	890	890	890	890	890
R-squared	0.308	0.302	0.301	0.301	0.304	0.299	0.304
R2 Adj. (w/o controls)	0.0581	0.0577	0.0604	0.0575	0.0602	0.0569	0.0577
F Test (Cog)	0.0127						
F Test (Noncog)	0.0599						
F Test (Tech)	0.456						
Test NC diff.	0.589						

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. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A11: Skills asked to a village informant: prediction of average log of maize yield

Corresponding Skill Index	Explanatory variables: Question asked to village informant	Regressions with average log of maize yield as dependent variable		
Cognitive	Level of education		0.14**	0.03
			(0.068)	(0.079)
Non-cognitive	Active/ Motivated		0.13	0.09
			(0.119)	(0.120)
Technical	Agricultural knowledge		0.36***	0.33***
			(0.097)	(0.102)
	Controls	Vil. FE	Vil. FE	All
	Observations	883	883	883
	R-squared	0.186	0.244	0.299
	F Test		5.83e-08	0.0003

Note: Skill proxies obtained through village informant (CHW), scored on scale from 1 (low) to 3 (high). 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. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix 1: questionnaire design and sources

A review of studies and questionnaires using different approaches to measure cognitive, noncognitive and technical skills of adults preceded the initial questionnaire design.[44] This appendix discusses the choices made to design the final instruments, including methods used to reduce the number of questions and hence the overall duration of the questionnaire. It also, provides information about the source of the different scales and tests used, and references other papers that use them.

Cognitive module

In most empirical work in development economics, a household's skill level is proxied by the education level of the household head, the maximum number of years of education in the household, or an individual's self-assessment of his literacy level. However, education is not always a strong correlate of productivity differences in agriculture. Existing literature reviews (Lockheed, Jamison and Lau, 1980; Phillips, 1994) indicate that the production increase resulting from four years of additional schooling is typically 7% to 8%. While the correlation is most often positive, in many papers it is not statistically significant. It may well be that grades attained or self-assessed literacy are not good measures of the farmers' active knowledge of reading or math. Farmers reading skills might matter e.g. for the processing of information regarding input use, and his math skills might be crucial to make optimal cost-benefit analysis. A priori it is also quite possible that it is a farmer's broader cognitive skills (such as memory, processing ability, or analytical thinking) rather than his classroom knowledge (such as reading or math) that help him adapt to the varying conditions of climate or soil.[45]

There are many tests that are designed specifically to measure cognitive skills but many are hard to apply as part of a large household survey. Not only do these tests typically require a level of standardization and quality control that goes beyond the usual training and supervision of household survey enumerators, they can also be very time-consuming, might require a standardized test-taking environment and/or specialized professional test administrators (such as licensed psychologists), might have content that is inappropriate for developing country settings, or require the use of test material that is unpractical in field circumstances. Moreover, any language-based tests are likely to suffer from lack of comparability across countries – and often also within countries - and lack of standardization

[44] We do not consider the literature on measuring skills for children and teenagers, as most instruments would not necessarily be relevant for adults See, for instance, Cueto and Leon (2012) for psychometric analysis of the skills measures in the Young Lives surveys.

[45] A useful distinction can be between *fluid intelligence* (the ability to solve novel problems) and *crystallized intelligence* (knowledge and developed skills) – Cattell (1987).

upon translation. Existing short and non-language based tests (often based on visual aids) that do not suffer from these limitations are sometimes used as alternative for inclusion in household surveys.

With the objective of measuring different aspects of adult farmers' cognitive ability, we selected five different cognitive tests: i) The 36 item Raven Colored Progressive matrices; ii) The digit span forwards and backwards; iii) A timed math test with 160 basic additions, subtractions and multiplications; iv) An oral 9-item math questions tests containing short math puzzles and increasing in level of difficulty; and v) A reading comprehension test. Table A1A provides a detailed description for each of these tests.

Versions of the Raven and the digit span are very frequently used in surveys in developing countries (de Mel, Mckenzie and Woodruff, 2009a, 2010; Beaman and Magruder, 2012; Dupas and Robinson, 2013; Giné and Mansuri, 2014; Djankov et al., 2015); and the timed math tests has also been used before (Barham, Macours, Maluccio, 2017). Jamison and Moock (1984) used numeracy questions, literacy tests and raven tests. The specific math puzzles and the reading comprehension tests used in this paper were designed for the purpose of this experiment. The outcomes of these test give us an observed outcome of the farmers' cognitive skills.

Noncognitive module

The choice of subscales was based on comparisons with the seminal papers in the literature on noncognitive skills, complemented with scales used in the literature on small business development in developing countries.[46] We also added measures used in the small but growing empirical literature on aspirations and locus of control in developing countries.

For each of the scales, we selected a subset of items to be included in the final survey instrument after piloting. We followed standard practices in psychology and broader insights from the psychometric literature, regarding selection of questions, question type and mode of analysis. In particular, questions of the different subscales as well as all 44 questions of the BFI were incorporated in the pilot version of the questionnaire. During piloting, a relatively large set of questions was identified with either very little variation (because 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. Qualitative field observations allowed to

[46] While most of these scales were originally designed for self-administration (i.e. respondents directly filling in answers), in Kenya they were asked by the enumerators to the respondent, reflecting how they are typically used in large household or individual surveys.

interpret the underlying answering pattern, with people either understanding the reverse question, in which case they often disagreed, or not understanding the reverse question, in which case they reverted to agreeing, as a fallback. Hence these distributions of the individual variables suggested a relatively high level of acquiescence bias or "ya-saying" in the study population. Variables were eliminated if they showed very little correlations with other variables belonging to the same construct, or showed very little variation.

For the Big Five personality traits, we use a version of the Big Five Inventory (BFI) written for a population with 5 years of education. The BFI is a commonly used instrument for the Big Five factor model, a higher module assumed to encompass the most generally important personality traits (John, Donahue, and Kentle, 1991; John, Naumann and Soto, 2008). The BFI has been used in the development economics literature by Dal Bo, Finan, and Rossi (2013); Callen et al. (2015). The BFI instrument has 44 items, and all 44 items were included in the pilot version of the instrument. After piloting, the number of BFI items was reduced to 23, keeping at least 3 questions for each personality trait, and a balance between the positive and reverse-coded items. The 23 items include the 10 items of the shorter BFI-10 scale (Rammstedt and John, 2007).

For Locus-of-Control we use a subset of the Levenson's (1981) "Internality, Powerful Others and Chance scales". This scale is used, for instance, by Acharya et al. (2007) and Bernard et al (2014). We also use a subset of items from the Attributions for Poverty scale (Feagin, 1972, 1975), similarly used in Bernard et al (2014).

We also added a series of locus-of-control questions with visual aids. The respondent was asked to allocate 10 beans between different answer options for three locus of control questions. The questions followed the general concepts of standardized locus-of-control instruments, and the visual aid aimed at increasing engagement and understanding of the respondents. For example a question asked the respondent to allocate 10 beans to three possible reasons for why some individuals have more economic success than others from: 1) Efforts and Decisions, 2) Luck, and 3) Birth.

The literature on the formation and predictive power of noncognitive skills in the US often uses measures of both locus-of-control and self-esteem (Heckman, Stixrud and Urzua, 2006; Heckman and Kautz, 2012; among many others). Following this literature, the questions of self-esteem used come from the Rosenberg (1965) scale. In development economics, similar measures are used in Blattman, Jamison and Sheridan (2016); Blattman and Dercon (2016); Adhvaryu, Kala and Nyshadham (2016).

A number of additional subscales were included because of their frequent use in the literature on small business development in developing countries. While this literature typically focuses on non-agricultural businesses, it seems plausible that some of the same characteristics may affect success in farming business. In particular, we followed studies analyzing what distinguishes entrepreneurs from others in developing contexts that have used measures of optimism, attitudes towards change, tenacity&organization, self-control, meta-cognitive activity, risk aversion and patience, similar to the ones we test, in addition to the BFI, internal locus of control, and self-esteem (Krauss et al. 2005; De Mel et al. 2009b, 2010; Giné and Mansuri, 2014; Djankov et al., 2015). Some of these subscales are conceptually closely related to one of the Big Five personality traits (with tenacity&organization, for instance, related to conscientiousness).

Most items were asked using a Likert scale, following the original scales. A few items were however changed to a binary scale, and some others we adapted to fit the agricultural context (see details of items in Table A1.B).

Finally, the CESD was added as it is often used in studies in developing countries, including in national representative panel surveys such as the Indonesian Family Life Survey and the Mexican Family Life Survey. It is arguably related to the Neuroticism personality trait. It consists of a set of 20 questions asking about the respondents' emotions in the last 7 days. As such it is more direct and arguably less abstract than the Likert scale questions. One additional question was added to capture perceptions of upward mobility using the same 7 day format.

Technical module

Technical skills and agricultural knowledge required for farming are likely to differ a lot from context to context. For this reason rather than replicating specific questions from prior survey instruments, we reviewed the literature to categorize existing approaches, and then designed a questionnaire that uses similar categories, but with specific questions adapted to the study population.

The majority of studies measuring technical knowledge do it with the intention of evaluating learning from a training provided to farmers, for instance through Farmer Field Schools (Godtland et al. 2004; Feder et al 2004; Maureci et al. 2007; David 2007; Buck and Alwang 2011). Consequently, they tend to provide an assessment of the practices that are taught by the intervention in order to track progresses in related knowledge. A few studies apply a broader technical knowledge assessment, including Goldstein and Udry (1999), Hanna, Mullainathan, and Schwartzstein (2014), Kondylis, Mueller and Zhu (2015) and the Ethiopia Rural Household Surveys (ERHS) of 1999.

Based on a review of questionnaires used in those studies, one can classify questions based on form, (i.e. how they evaluate technical skills), the technology or practices referred to when assessing skills, and the type of knowledge asked about when assessing skills.

With regard to the form, knowledge tests with a series of multiple-choice or open-ended questions are used in a number of studies, and the skill measures used in this paper are constructed from such questions.[47]

When assessing skills, the technology or practices referred to have to be a function of the crops in the region of study. Hence for a general knowledge tests, initial fieldwork and local knowledge is important to first identify which crops, technologies and practices are most common in the region and can best help distinguish farmers with best practices from less knowledgeable ones. The particular crops, technologies and practices referred to in our survey instrument were based on qualitative fieldwork prior to questionnaire design and knowledge of local agronomists.

The review of the literature further revealed a relatively large commonality regarding the types of knowledge that are being assessed (even if it were applied to different crops and practices). For example, questions often ask for mode of application, quantity and timing of inputs, all common practical issues faced by farmers with important consequences for yield. The ability to recognize deficiencies, pests etc. are also common. A potential challenge for such questions is that the optimal practice may depend on a number of other factors, making it hard to evaluate whether the answer provided by a farmer is "correct" or not. A different type of question asks for theoretical knowledge such as, for instance, the type of nutrients included in certain fertilizers. These have the advantage of having unambiguous correct answers, but one can wonder whether they capture the type of practical knowledge that matters for productivity (in case farmers, for instance, know which fertilizer to use and when, but do not know its composition). Whether such theoretical questions are good predictors of practices and productivity is part of the questions of interest for this study.

Based on the categorizing of existing questions in the literature, we designed an instrument that covered the different types of knowledge for the practices and technologies relevant in the region of study. Extensive fieldwork in cooperation with local agronomists was required to design, test and adapt the questions. As for the noncognitive module, only questions showing sufficient variation in

[47] A relatively large number of questionnaires also ask farmers to self-assess their level of knowledge. Alternatively, farmers are sometimes asked what they actually do rather than what they know. The former is likely prone to subjectivity, while the later measures the combination of many other constraints (budget, time, etc.) in addition to differences in technical skills. Given these concerns, this study focuses agricultural knowledge tests.

answers during piloting were kept. This led to exclusion of certain practices (such as pest management or irrigation) as knowledge about them was extremely limited in the region.

Figure: Classification technical skills questions

Form of Evaluation	Technology or Practices	Type of knowledge
Test: Multiple choice or open questions about best practices, assessing whether the respondent finds the right answer.	Seeds	How to apply an input: - Where to apply it - What quantity - Timing of application - Other decisions (spacing…)
Self-assessment: subjective assessment or "do you know…"	Fertilizer (mineral / biofertilizer)	Recognizing: pests, plant deficiencies, better seeds… to decide what inputs or practices to apply.
What the farmer does: use of practices or technology	Herbicide, Pesticide or Integrated Pest Management	How to use a complex practice (composting, fertilizer mix...)
Sources of information: training received, extension, etc.	Irrigation	Theoretical knowledge (e.g. name of nutrients in mineral fertilizer)
	Soil management practices: - Manure, compost, use of stalk - Rotation, intercropping - Tillage	
	Planting practices (number of seeds, spacing, gapping, etc.)	
	Storage / usage / commercialization	

Appendix 2: Brief introduction to psychometrics concepts and methods used

Reliability

Reliability is the overall consistency of a measure. A measure is said to have a high reliability if it produces similar results under similar conditions. A measure that is very noisy is a measure with low reliability.

In classical theory, it is assumed that a person's observed or obtained score on a test is the sum of a true score (T) and a Measurement Error (E):

$$X = T + E$$

Hence the variance of X is given by:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

In this setting, reliability is the ratio of variability in item X due to variation of the true score T:

$$Reliability = \frac{\sigma_T^2}{\sigma_X^2}$$

It can thus be interpreted as the ratio of the variance of a given measure that is driven by the true variance of the score across the population, or equivalently 1 minus the share of variance explained by pure measurement error.

An estimation of the reliability can be obtained with the test-retest correlation (consistency across time).

If measurement error is classical, the test-retest correlation gives a good indication of the signal to total variance ratio. On the other hand, the test-retest correlation can under or over-state the signal to total variance ratio in case of non-classical measurement error. If the errors in measurement are positively correlated over time, for instance because both measures suffer from persistent acquiescence bias, the test-retest correlation will overstate the reliability of the data.

The Cronbach's alpha is also an indicator of reliability, and provides a measure of consistency across items expected to measure the same latent construct. As such the Cronbach's alpha is, however, also an indicator of validity.

Validity

Test validity is the extent to which a test measures what it is supposed to measure.

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.

Among the key indicators of validity are the following ones:

- Face validity assesses the extent to which a test is subjectively viewed as covering the concept it purports to measure. For example a question about self-confidence should seem to ask about self-confidence (hence a question with high correlation with related measures but seemingly asking something very different cannot be considered valid)
- Content validity refers to the extent to which a measure represents all facets of a given construct.
- Piloting experience and use of psychometric scales validated in other contexts
- Construct validity: Correlation with other measures intending to measure the same construct
- Predictive validity: it should predict well related behaviors that are theoretically expected to be correlated with the measure

A more detailed explanation of reliability and validity can be found in American Educational Research Association et al. (1999).

<u>Test-retest correlation</u>

Test-retest correlation is the correlation between measures using the same instrument, measured twice, on the same person, in similar conditions within a relatively short period of time. Temporal stability provides an assessment of the reliability of a measure. Typically a similar test is applied twice to the same population within a period short enough that that the traits that the researcher intends to measure should not have changed, but long enough that respondents do not remember their original responses. A standard period between test and retest goes from two weeks to one month.

Under classical theory assumptions, the correlation between the test and the retest can be interpreted as a direct measure of reliability as defined above ($\frac{\sigma_T^2}{\sigma_X^2}$) hence the correlation can directly be interpreted as the share of variance of the measure explained by the variance of the true score.

Crocker and Algina (2006) and Nunally, and Bernstein (1994) provide a broader explanation of classical test theory and test-retest correlation.

Cronbach's alpha

The Cronbach's alpha (Chronbach 1951) is one of the most widely used measures of internal consistency of a test.

Cronbach's alpha is mathematically equivalent to the expected value of the split-half reliability. Split-half reliability is obtained by 1) randomly splitting the items into two sets of items of equal size, 2) calculating the average of each set of items, and 3) calculating the correlations between these two sets of items. Although not calculated this way, the Cronbach's alpha is equal to the average of the correlations obtained through all the possible combinations of split-half reliability. It provides an indicator of how well the items correlate among them (although it also increases with the number of items).

Assume that we have a measure X made of k items: $X = Y_1 + Y_2 + \cdots + Y_k$

Its Cronbach's alpha is given by:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

Where $\sigma_{Y_i}^2$ is the variance of item i and σ_X^2 is the variance of the measure X.

The Cronbach's alpha provides an assessment of both the construct's validity and its reliability. It is said to provide a lower bound on the reliability of a test, because for the case where all items are measuring exactly the same construct, the Cronbach's alpha would only be affected by the measurement error of each item and is a pure measure of reliability. When it is not the case, then the Cronbach's alpha is also affected by the extent to which items are measuring the same latent construct. Hence a low Cronbach's alpha indicates that either the items are measuring very different latent constructs (the validity is poor since the items are usually pooled with the intention to measure one latent construct) or they are measuring the same latent construct but with a lot of noise, hence the reliability is low.

A Cronbach's alpha of 0.9 tends to be required when individual decisions will be made based on a specific test (for example student's admissions, Nunnally and Bernstein, 1994; Kline, 2013), but an alpha of .7 is often considered acceptable for the purpose of statistical analysis.

Exploratory Factor Analysis and deciding the number of factors

Here we provide a very brief description of four methods that we used and that are commonly used to determine the number of factors to be retained. Valero-Mora (2007) provides a more detailed explanation of the methods and their advantages and caveats.

1) Kaiser's (1958) criterion only keeps factors with an eigenvalue higher than one;

2) Visual inspection of the Catell (1966) scree plot. Cattell's rule is such that the number of factors should be equal to the number of eigenvalues before which the smooth decrease of eigenvalues appears to level off on to the right of the plot;

3) Velicer's (1976) Minimum Average Partial minimizes the unexplained partial correlation;

4) Horn 's (1965) Parallel Analysis keeps the factors as long as they explain more than the 95^{th} percentile of eigenvalues from randomly generated data (Cota, Longman, Holden, Fekken, & Xinaris, 1993; Glorfeld, 1995).

Valero-Mora (2007) argues that the methods of Velicer and Horn are more reliable. Given that different methods do not always lead to the same conclusions, we opted for the number of factors most commonly suggested by the methods, putting more emphasis on the last two.

Item Response Theory

Item Response Theory offers a structural way of using a set of items to measure a latent ability or trait. It is based on the idea that the probability of a correct/keyed response to an item is a mathematical function of person and item parameters. For example, in the case of binary items, it considers that the probability of getting the correct answer to each item is a logarithmic function of the difficulty of the item and the latent ability of the respondent. IRT simultaneously estimates the difficulty of each item and the ability of each respondent, such that it maximizes the likelihood of the responses observed in the data.

IRT has become the standard tool for high stakes tests such as GRE or GMAT because it is believed to provide a greater precision than Classical Test Theory.

IRT requires the following assumptions:

1) Uni-dimensionality; assessed through factor analysis;

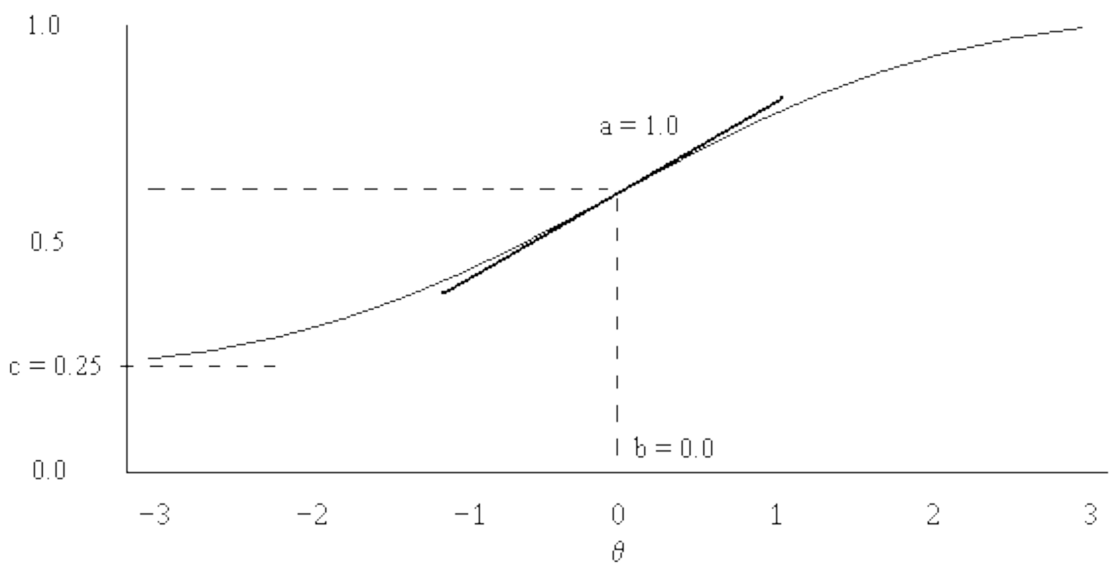
2) Local independence (which is violated for timed tests, or for tests for which one question affects answers to the following one(s));

3) Monotonicity (item characteristic curve well behaved – see below).

The graphs below represents the Item Characteristic Curve (ICC) which is the probability of getting a correct answer to a given item, conditional on the respondent's underlying ability. In the one parameter model, also called Rasch Model, the difficulty of each item is the parameter estimated (where b-value is 0 in the graph). The two-parameter model also estimates the discriminant, which is the slope (a-value) at difficulty and can be interpreted as the effect of the underlying ability on the respondent's probability to answer the question correctly. The three-parameter model adds a pseudo guessing parameter (c-value), which estimates the probability of a respondent with lowest level of ability to obtain a correct answer.

Using these three parameters, the conditional probability of getting a correct answer to item i for an individual with underlying ability $\theta$is given by:

$$P_i(\theta) = c_i + (1 - c_i)\,\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$



The Graded Response Model applies a similar logic, with multiple difficulty parameters in order to deal with ordered polytomous variables. More about GRM can be found in Van der Linden and Hambleton (2013).

IRT also allows hybrid models that combine the different types of model.

For the technical skills we combine the Graded Response Model with the Two-Parameter Model. We do so, because we have two types of questions. The vast majority of questions are multiple choice questions where the respondent can choose only one possible answer. Based on this answer, we created a binary variable for whether the answer was correct or not. In some questions, however, it was possible to select multiple answers, in which case we created a count variable indicating the number of correct answers, but penalizing for wrong answers selected.

In the cognitive sub-constructs, we used a hybrid of Three-Parameter Model and Two-Parameter Model, because for some questions the guessing parameter was found to be zero (in which case there is no gain from the three parameter model).

For a general introduction to IRT, see Hambleton and Swaminathan (2013).

Tucker's Congruence Coefficient

The Tucker's congruence coefficient (or simply congruence coefficient) is an index that assesses the similarity between factor structures of the same set of items applied to two different populations. One first applies a factor analysis to the two populations. In order to assess the similarity between a factor x and a factor y, after applying factor analysis to two different population, one calculates the correlation coefficient (by item) of the two vectors of factor loadings.

$$\varphi(x,y) = \frac{\sum_{i,} x_i y_i}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$$

Where $x_{i,j}$ and $y_{i,j}$ are the loadings of item i on factors x and y, respectively (each one extracted from applying the factor analysis of the same items to a different population).

$\varphi(x,y)$ can be interpreted as a standardized measure of proportionality of elements in both vectors. A coefficient that is equal to 1 corresponds to a perfectly identical factor structure between the two populations, while a coefficient equal to 0 corresponds to a factorial that is completely orthogonal.

For an order of magnitude, Lorenzo-Seva and Ten Berge (2006) indicate that a congruence coefficient over .95 implies a good similarity, and a range of [.85 - .94] shows fair similarity.

More about Tucker's congruence coefficient can be found in Abdi (2007).

Additional References from the appendices

Abdi, H., 2007. "RV coefficient and congruence coefficient." *Encyclopedia of measurement and statistics*, pp.849-853.

Acharya, V., A. Rajan, and A. Schoar, 2007. "What determines entrepreneurial success? A psychometric study of rural entrepreneurs in India." *mimeo*, London Business School, IFMR, MIT.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US), 1999. "Standards for educational and psychological testing." *Amer Educational Research Assn.*

Barham, T., K. Macours, and J. A. Maluccio, 2017. "Are Conditional Cash Transfers Fulfilling Their Promise? Schooling, Learning, and Earnings After 10 Years." *CEPR Discussion Paper* 11937

Beaman, L. and J. Magruder, 2012. "Who Gets the Job Referral? Evidence from a Social Networks Experiment." *American Economic Review*, 102(7): 3574-93.

Buck, S. and J. Alwang, 2011. "Agricultural extension, trust, and learning: results from economic experiments in Ecuador." *Agricultural Economics*, 42(6): 685-699.

Cattell, R.B., 1966. "The scree test for the number of factors." *Multivariate Behavioral Research* 1: 629-637.

Cattell, R.B., 1987. *Intelligence: Its Structure, Growth and Action.* Amsterdam: North-Holland.

Cota, A.A., Longman, R.S., Holden, R.R., Fekken, G.C., & Xinaris, S., 1993. "Interpolating 95th percentile eigenvalues from random data: An empirical example." *Educational & Psychological Measurement, 53:* 585-596.

Crocker, L. and Algina, J., 1986. *Introduction to classical and modern test theory.* Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Cronbach, L.J., 1951. "Coefficient alpha and the internal structure of tests." *Psychometrika*, 16(3), pp.297-334.

Cueto, S and J. Leon, 2012. "Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 3 of Young Lives", *Young Lives technical note* 25.

David, Soniia, 2007. "Learning to Think for Ourselves: Knowledge Improvement and Social Benefits among Farmer Field School Participants in Cameroon", *Journal of International Agricultural and Extension Education*, 14(2): 35-49 de Mel, S. D. McKenzie, C. Woodruff, 2010. "Who are the Microenterprise Owners? Evidence from Sri Lanka on Tokman versus De Soto", in: Josh Lerner and Antoinette Schoar (eds.): *International Differences in Entrepreneurship*, University of Chicago Press. Chicago.

de Mel, S., D. McKenzie and C. Woodruff, 2009a. "Are Women More Credit Constrained? Experimental Evidence on Gender and Microenterprise Returns." *American Economic Journal: Applied Economics*,1(3): 1-32.

De Mel, S., D. McKenzie, and C. Woodruff, 2009b. "Innovative firms or innovative owners? Determinants of innovation in micro, small, and medium enterprises." *IZA Discussion Paper* No. 3962

Djankov, S., E. Miguel, Y. Qian, T. Roland, and E. Zhuravskaya, 2005. "Who are Russia's entrepreneurs?" *Journal of the European Economic Association*, *3*(2-3): 587-597.

Dupas, P. and J. Robinson, 2013. "Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya" *American Economic Journal: Applied Economics* 2013, 5(1): 163–192

Feagin, J. R., 1972. "Poverty: We Still Believe That God Helps Those Who Help Themselves." *Psychology Today 1*, 101–29.

Feagin, J. R., 1975. *Subordinating the Poor: Welfare and American Beliefs*. Englewood Cliffs, NJ: Prentice Hall

Feder, G., R. Murgai, and J.B. Quizon, 2004. "The acquisition and diffusion of knowledge: The case of pest management training in farmer field schools, Indonesia." *Journal of agricultural economics*, 55(2): 221-243.

Giné, X. and G. Mansuri, 2014. "Money or ideas? A field experiment on constraints to entrepreneurship in rural Pakistan." *World Bank Policy Research Working Paper* 6959, World Bank, Washington DC.

Glorfeld, L.W., 1995. "An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain." *Educational and Psychological Measurement, 55*: 377-393.

Godtland, E.M., Sadoulet, E., De Janvry, A., Murgai, R. and Ortiz, O., 2004. "The impact of farmer field schools on knowledge and productivity: A study of potato farmers in the Peruvian Andes." *Economic Development and Cultural Change*, 53(1): 63-92.

Goldstein, M. and C. Udry, 1999. "Agricultural innovation and resource management in Ghana." *Final Report to IFPRI under MP17*, Mimeo. Yale University.

Hambleton, R.K. and Swaminathan, H., 2013. *Item response theory: Principles and applications*. Springer Science & Business Media.

Hanna, R., S. Mullainathan, and J. Schwartzstein, 2014. "Learning through noticing: theory and experimental evidence in farming", *Quarterly Journal of Economics*, 1311–1353.

Horn, J. L., 1965. "A rationale and test for the number of factors in factor analysis." *Psychometrika, 30:* 179-185.

Jamison, D.T. and P.R. Moock, 1984. "Farmer education and farm efficiency in Nepal: The role of schooling, extension services, and cognitive skills." *World Development*, *12*(1): 67-86.

John, O.P., E.M. Donahue, R.L. Kentle, 1991. *The Big Five Inventory – Versions 4a and 54*. Berkeley CA: University of California, Berkeley, Institute of Personality and Social Research.

Kaiser, H., 1960. "The application of electronic computers to factor analysis." *Educational and Psychological Measurement* 20: 141-151.

Kaiser, H.F., 1958. "The varimax criterion for analytic rotation in factor analysis." *Psychometrika,* 23(3), pp.187-200.

Kline, P., 2013. *Handbook of psychological testing*. Routledge.

Kondylis, F., V. Mueller, and S. Zhu, 2015. "Measuring agricultural knowledge and adoption." *Agricultural Economics*, *46*(3): 449-462.

Krauss, S.I., M. Frese, C. Friedrich, and J.M. Unger, 2005. "Entrepreneurial orientation: A psychological model of success among southern African small business owners" *European Journal of Work and Organizational Psychology*, *14*(3): 315-344.

Levenson, H., 1981. "Differentiating among Internality, Powerful Others, and Chance." In H. M. Lefcourt (Ed.), *Research with the Locus of Control Construct*, pp. 15–63. New York: Academic Press.

Lockheed, M. E., T. Jamison, and L.J. Lau, 1980. "Farmer Education and Farm Efficiency: A Survey." *Economic Development and Cultural Change*, 29 (1): 37-76.

Lorenzo-Seva, U. and Ten Berge, J.M., 2006. "Tucker's congruence coefficient as a meaningful index of factor similarity." *Methodology*, 2(2), pp.57-64.

Mauceri, M., J. Alwang, G. Norton, and V. Barrera, 2007. "Effectiveness of integrated pest management dissemination techniques: a case study of potato farmers in Carchi, Ecuador." *Journal of Agricultural and Applied Economics*, 39(03): 765-780.

Nunnally, J.C. and Bernstein, I.H., 1994. *Psychometric theory* (3rd ed) New York: McGraw-Hill.

Phillips, J.M., 1994. "Farmer Education and Farmer Efficiency." *Economic Development and Cultural Change*, Vol. 43(1): 149-166.

Rammsted, B. and O.P. John, 2007. "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German." *Journal of Research in Personality*, *41*: 203– 212

Rauch, A. and M. Frese, 2000. "Psychological approaches to entrepreneurial success: A general model and an overview of findings." *International review of industrial and organizational psychology*, 15: 101-142.

Rosenberg, M., 1965. *Society and the Adolescent Self-Image.* Princeton, NJ: Princeton University Press.

Van der Linden, W.J. and Hambleton, R.K. eds., 2013. *Handbook of modern item response theory.* Springer Science & Business Media.