The Reliability and Validity of Skills Measurement in Rural Household Surveys

Rachid Laajaj and Karen Macours,

Universidad Los Andes, Paris School of Economics and INRA

Outline

- Motivation
- Research questions
- Reliability – Validity
- Improving measures
- Predicting agricultural production & decisions
- Further understanding measurement error
- Next steps
- Lessons learned

Why Measuring skills ?

- Having good measures of skills can
  - help better understand poor households' **decision** making (e.g. technology adoption dilemma)
  - Allow to **observe changes** (impact) on outcomes that themselves might trigger longer term results

  => can be key for dynamic understanding of poverty
  - Be useful to control for typical "**unobservables**"
- Skills are in themselves an outcome of interest: measure of **multidimensional poverty**
- Two main audiences: "multi-purpose surveys" vs "special purpose surveys"

Skills matter: Evidence from Developed Countries

- Good body of evidence in cognitive skills, mostly on developed countries.
  - Hanushek and Kimko (2000) use **math and science test scores**, and find it to **predict growth much better than years of education**.
  - Numerous studies establish that measured **cognitive** ability is a **strong predictor** of schooling attainment and wages, conditional on schooling (Cawley, Heckman, and Vytlacil 2001).

- Heckman, Stixrud & Urzua (2006) find **that non-cognitive skills** can be as important as cognitive skills to explain success in life (income, wages, criminal behavior, teenage pregnancy …)

Skills in developing countries

- Cognitive delays from early childhood and important **socio-economic gradients** in cognition
- Literature has documented:
  - Low levels of **aspiration**
  - High levels of **depression**
  - Lack of **Self-control**
- Lack of information or **"know how"** regarding agricultural practices

=> Can good measurement of adult skills allow us to better understand decision making?

Cross-cutting challenges for skill testing in household surveys

- Many existing measures are time consuming
- Initially designed for developed country settings, lab settings, self-administered surveys, etc
- Concept often more abstract – and/or more technical – translation and understanding can become a big issue
- Standardized application of tests
- Openness of adult respondents to test-taking
- Measurement error and imperfect proxies

Outline

- Motivation
- **Research questions**
- Reliability – Validity
- Improving measures
- Predicting agricultural production & decisions
- Further understanding measurement error
- Lessons learned

Research Questions

- Starting point: little to **no validation** of skills measures
- Can we measure skills in rural developing settings? Reliability? Validity?
- Which skills matter for agronomical decision making and agricultural productivity?
- Address specificities of rural developing context:
  - Are **scales** similar to the ones of developed countries?
  - Lower **education** (adapt questions)
  - Not filled by hand, but through **enumerators**

Methodology

- Designed an instrument with different alternative modules and approaches for each of the 3 skill types
- Conducted survey experiment in rural Kenya
  - Randomized survey instrument (finalized after extensive piloting)
  - Test-retest
  - Randomized field work implementation
- Use statistical analysis to analyze reliability and validity of the measurements

Which skills?

- **Cognitive skills (~ IQ)**
  - Memory (Forward and backward Digit Span), problem solving (Raven matrices)
  - "Class room" skills: Reading and math (achievement)
- **Non-cognitive skills (Socio-emotional skills)**
  - Big Five personality traits
  - Lower-order constructs: Locus of control, self-esteem, Self-control, perseverance, aspirations,
  - CESD (depression)
- **Technical skills**
  - Knowledge/Know-how. We worked with agronomists and soil scientists on targeted questions for main crops and practices

Measurement Non-Cognitive skills

- Traditional : 1-5 scale with statements about one-self
  - "On a scale from 1 to 5 - with 1 indicating you strongly disagree and 5 indicating you strongly agree : You see yourself as someone who tends to be lazy"
- 1-5 scales about causes of poverty
- Economic ladder
- Locus of control through "beans" (visual aid)
- CESD : E.g. "In the last 7 days, how many days were you hopeful about the future?"
- Standardized measures for risk aversion and time preference

Measurement Technical skills

- Basic **knowledge required to perform a task**: very field specific by definition.
- Use proxies or try to obtain actual tests of relevant knowledge?
  - Self-assessment
  - Past years of experience
  - Knowledge tests => Which type of question?
    - Recognition techniques/practices, timing, knowing how to implement, ”scientific” understanding,
      - work with agronomists and soil scientists on targeted questions for main crops and practices
    - Attempt to have “unambiguous” questions with varying difficulty, mostly multiple choice, visual aids

Examples of Technical Skills Q.

- When planting hybrid maize in rows, how many seeds per hole should be applied?
- When planting bananas what is the optimal distance between banana trees?
  - 1. 1mx1m
  - 2. 2m x 2m
  - 3. 2m x 3m
  - 4. 3m x 3m

Measurement Exercise and Context

- Context:
  - 960 (918) farmers in 96 villages in Siaya - Western Kenya
  - Mainly maize & other annual crops. Most also have livestock
  - About 50/50 men-women
  - On average 6 years of education
- Implemented:
  - Review of instruments & work with local agronomists
  - Extensive piloting (qual. and quant. with 120 hh)
  - Test of Skills Measurement
  - Retest of Skills measurement
  - Other hh member survey
  - Typical household survey on agricultural practices

CAPI Questionnaire design

- 3 main sections
  - cognitive, non-cognitive, technical
- Randomization of:
  - Order of modules
  - Order questions in modules
  - Order of answer options

  => Allows analysis of survey fatigue & order effects
- Random assignment to enumerators
  - 2x same enumerator in 40% of cases
- Followed by a household survey

Outline

- Motivation
- Research questions
- **Reliability – Validity**
- Improving measures
- Predicting agricultural production & decisions
- Further understanding measurement error
- Lessons learned

Measures of Reliability

- Reliability: the share of variance not due to noise
- Indicators:
  - Consistency across time (pure reliability): High **Test-Retest Correlation** if you replicate the measure within a period short enough that it should not have changed.
  - Consistency across items: High correlation among items that intend to measure the same skill: **Cronbach's Alpha** (also validity)
  - Results not subject to the conditions i.e. enumerator, order of questions or responses, mood of the day.

Second Criteria: Validity

- Are you measuring what you intend to measure?
- **Indicators**:
  - Face validity: use of **Validated** (in other context) Psychometric scales & Piloting experience
  - Correlation with other measures (same round) **Cronbach's Alpha** & factorial analysis
  - Should **predict** well related behaviors: regressions on agronomical decisions and outcomes

Reliability of "Naïve Scores"

- We first look at reliability on "Naïve" Score (unweighted addition of points)"
- Typical norm in psychometrics: test-retest of 0.70 is considered good and usable:
  - Cognitive is above the norm, not Noncog or Technical.

Naïve Scores			
	Test-Retest	**Chronbach's Alpha**	**Nb of Indexes**
Cognitive	0.83	0.84	5
Non-Cog	0.53	0.76	14
Technical	0.31	0.43	6

Cognitive skills: Test-retest and Internal Reliability

Indicator	Test-retest	Cronbach's Alpha
All Cog	**0.83**	**0.84**
Raven	0.63	0.88
Numeracy Q.	0.60	0.70
Math sheet	0.68	
Reading	0.82	0.92
Digit Span	0.52	

Non-cognitive Reliability Test

Indicator	Test-retest	Cronbach's Alpha
All Non-Cog	**0.53**	0.76
Locus of Control	0.42	0.55
Causes of Pov	0.40	0.35
Attit. Change	0.43	0.46
Risk Aversion	0.14	
BF_Extrav.	0.24	0.21
BF_Agree	0.26	0.40
BF_Conscious	0.33	0.51
BF_Neurotic	0.26	0.46
BF_Open	0.17	0.23
CESD	0.42	0.83

Technical Agricultural Knowledge Reliability Test

Indicator	Test-retest	Cronbach's Alpha
All Tech	**0.31**	**0.43**
Intercrop & Rotat.	0.16	0.04
Maize	0.23	0.30
Banana	0.20	0.22
Soybean	0.13	0.13
Composting	0.25	0.19
Min. Fertilizer Use	0.28	0.40

Issues with Technical Skills

- Requires a lot of preparation work
- Issue faced since piloting: 2 types of questions
  - Too easy: no variation in responses, everyone knows the answer
  - More complicated because right answer depends on context (and even agronomists often disagree among them)
- Narrow set of questions that fits between these 2 categories

Outline

- Motivation
- Research questions
- Reliability – Validity
- **Improving measures**
- Predicting agricultural production & decisions
- Further understanding measurement error
- Lessons learned

Towards less Naive Measures of the Skills

We Apply some corrections used in psychometrics:

- Item Response Theory for cognitive and technical tests.
- Factorial Analysis to group questions and weigh them
- Correct Acquiescence and extreme response bias in Non-cog questions.

  ~ “yay saying”

tendency to say yes, even to contradictory questions

Towards less Naive Measures of the Skills (2)

- Item Response Theory improves test-retest correlation of Cog from 0.83 to 0.85, but Tech went slightly down from 0.32 to 0.31.

  => Marginal improvements when well behaved

- Factorial analysis of NonCog gives worrying results:
  - Pools Items in non coherent groups (except CESD)
  - Chronbach's alpha average per group increase from 0.42 to 0.75 (~mechanical). Test-retest from .53 to .33.
  - 1st Factor is the Acquiescence Bias (if not corrected)

Correcting for response patterns

- Even after selection of items in pilots:
  - positively phrased questions often skewed to the right
  - reverse-coded questions have a bi-modal distribution

⇒ Borrow from psychometrics and ipsatize

- Calculate acquiescence score : averaging between the mean of the positively-coded items and the mean of reverse-coded items
  - Subtracted from all answers
- Correct for extreme response bias by dividing by s.d. of person's responses

Noncognitive construct after ipsatizing

- Factor loadings mostly still pool across scales
  - Don’t confirm Big Five, or other lower-level constructs (self esteem, locus-of-control,…)
  - CESD items do load together (separating positive from negative feelings, as the original scale)
- Acquiescence score itself becomes predictive

Outline

- Motivation
- Research questions
- Reliability – Validity
- Improving measures
- **Predicting agricultural production & decisions**
- Further understanding measurement error
- Lessons learned

Predictions of Productivity and Agricultural Decisions

- Clearly only correlations, no causality! (measurement exercise)
- Correlated for 3 possible reasons (all useful):
  - Skills affect agr. (decisions or productivity)
  - Agr. can affect skills
  - Omitted variables can affect both skills and agr.
- Also run regressions with large set of controls (including education, assets) additional predictive power ? Do we capture "unobservables"?

Predicting Agricultural Productivity

VARIABLES	Outcome variable: Log of MAIZE YIELD					
Estimation:	Naïve score		IRT / Factor		Mean Naïve Score	
Cog IRT	0.07**	-0.03	0.10***	-0.00	0.07*	-0.01
	(0.033)	(0.053)	(0.032)	(0.054)	(0.039)	(0.058)
NonCog	0.17***	0.12***	0.06	0.05	0.17***	0.11**
	(0.034)	(0.036)	(0.040)	(0.044)	(0.042)	(0.044)
Tech	0.06	0.01	0.11***	0.07**	0.10**	0.07
	(0.038)	(0.044)	(0.032)	(0.034)	(0.045)	(0.052)
Controls	N	Y	N	Y	N	Y
Observations	903	881	894	872	928	885
R-squared	0.053	0.344	0.039	0.336	0.056	0.337
Ftest	3.12e-10	0.00755	1.37e-07	0.194	5.99E-11	0.0243

Predicting Agricultural Decisions

	FARMING PRACTICES (DUMMIES)			
VARIABLES	Used Fertilizer	Used Hybrd Seed	Used Manure or Compost	Hired Labor
Cog IRT	0.03*	0.03*	0.01	0.05**
	(0.015)	(0.018)	(0.020)	(0.020)
NonCog Fact	0.02	0.04**	0.02	-0.02
	(0.017)	(0.019)	(0.017)	(0.016)
Tech IRT	0.07***	0.02	0.03	0.00
	(0.018)	(0.019)	(0.021)	(0.017)
Observations	903	813	830	900
R-squared	0.047	0.019	0.008	0.009
Ftest	6.39e-07	0.00154	0.136	0.0752

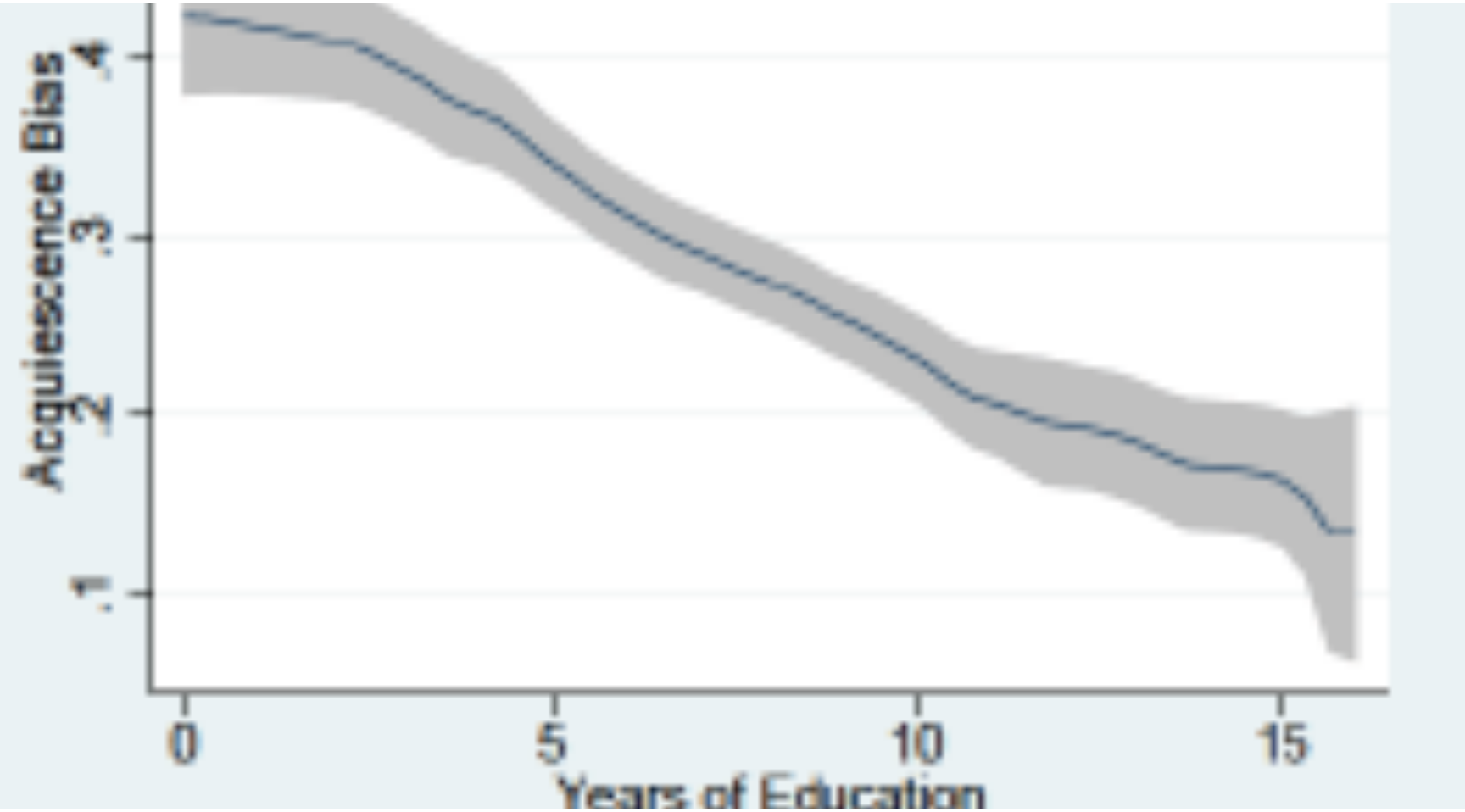
Predictive Power of Skills

- **Cognition** overall has good predictive power, but not when controlling for education (correlation of .74)
- **Technical** skills has some predictive power for production and agricultural decisions (coherent when disaggregating e.g. fertilizer and compost practices), but requires reducing the noise
- **Non Cog** looses predictability when corrected for Acquiescence bias. Less coherent groups. When disaggregated CESD more consistently significant.

Outline

- Motivation
- Research questions
- Reliability – Validity
- Improving measures
- Predicting agricultural production & decisions
- **Further understanding measurement error**
- Lessons learned

Acquiescence Bias and Cog Skills



Answer patterns predictive…

VARIABLES	Outcome variable: Log of MAIZE YIELD					
Estimation:	log(Maize Yield)		Used fertilizer		Hired labor	
Cog IRT	0.07**	-0.03	0.02	-0.02	0.03	0.04
	(0.033)	(0.053)	(0.016)	(0.022)	(0.021)	(0.032)
NonCog	0.05	0.04	0.02	0.03	-0.01	-0.01
	(0.041)	(0.044)	(0.017)	(0.018)	(0.016)	(0.021)
Tech	0.11***	0.07**	0.07***	0.05**	0.00	0.01
	(0.031)	(0.034)	(0.018)	(0.018)	(0.017)	(0.022)
Acquiescencse score	-0.23**	-0.17	-0.08	-0.08	-0.09	-0.12
	(0.101)	(0.112)	(0.054)	(0.047)	(0.061)	(0.073)
Controls	N	Y	N	Y	N	Y
R-squared	0.047	0.333	0.051	0.466	0.011	0.206
Ftest	3.91e-08	0.0885	1.82e-06	0.0120	0.0853	0.183

Enumerators Matter

- Enumerator fixed effects **explain up to 15% of variation** especially for non-cog and Tech
  - Questions with visual aids, open questions, more difficult questions seem to be more sensitive
- Test-retest reliability largely affected by changing enumerators (randomized)
- Only enumerator fixed-effects would not solve the problem
- Important to balance enumerators

	TEST-RETEST CORRELATIONS			
	All Tests	**Same enumerator**	**Different enumerator**	**All, with enum. FE**
Cognitive	0.83	0.88	0.81	0.84
Non-Cog	0.53	0.63	0.49	0.51
Technical	0.31	0.45	0.26	0.32

Some further qualitative insights

- Cognitive outcomes are ‘observed’
- Technical ‘objective’, but noisy
- Non-cognitive outcomes
  - Simple sentence structure matters
  - More abstract questions might be harder
    - “In the last 7 days, how many days did you feel depressed ?“

    versus
    - “On a scale from 1-5, you see yourself as somebody who is depressed or gets blue”
  - Negatively phrased questions are difficult
    - But important to have reverse coded questions to correct for response patterns!
  - Translation makes all of this harder

Where do we go from here?

- Towards more general lessons
  - 2nd survey experiment in Nicaragua
  - Possible analysis of internal consistency of non-cognitive outcomes in other datasets/countries?
- How to obtain more valid&reliable non-cognitive measures?
  - First isolate translation issues
    - Keeping intent of questions through translation is hard
  - Towards less abstract, more direct phrasing
  - Vignettes?
  - "observational" measures?

Lessons Learned so Far (1)

- Cognitive skills can be measured reliably
  - High correlation between measures
    - a subset or shorter tests can provide a good proxy
    - Value added compared to education?
- Technical skills very noisy but predictive and coherent.
  - Addressing measurement error helps
  - But obtaining good and stable measure remains difficult, possibly due to idiosyncratic nature of agricultural knowledge
- Non-cognitive skills are more challenging to measure
  - Standardized scales from developed country settings may not be valid due to non-random measurement error.
  - Factor structure hard to identify
  - Predictive power depends on corrections

Lessons Learned so Far (2)

- Economics has some catch up to do with other disciplines regarding measurement and testing validity
  - Testing internal consistency can be done with most data

- Improvement can be reached with:
  - Balancing Enumerators (and homogenizing)
  - Plan ahead for possible need to correct for answering patterns
    - E.g. including reverse coded questions key
  - Letting the data tell you how to aggregate Method
    - IRT
    - Factor analysis

Thank you!

- r.laajaj@uniandes.edu.co
- karen.macours@psemail.eu