Metadata

Title: Measuring of Holistic Skills

Purpose: To disseminate findings from the 2023 report "Measurement of Holistic Skills in RCTs: Review and Guidelines"

Date created: 3/28/24

Created by: Jessica Williams

Last edited on: 4/23/24

Last edited by: Jess Williams

Notes/guidelines:

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Measurement of Holistic Skills in RCTs: Review and Guidelines

Jessica Williams, J-PAL

with Karen Macours (Paris School of Economics) and Samuel Wolf (MIT)

CEGA Psychology and Economics of Poverty Convening

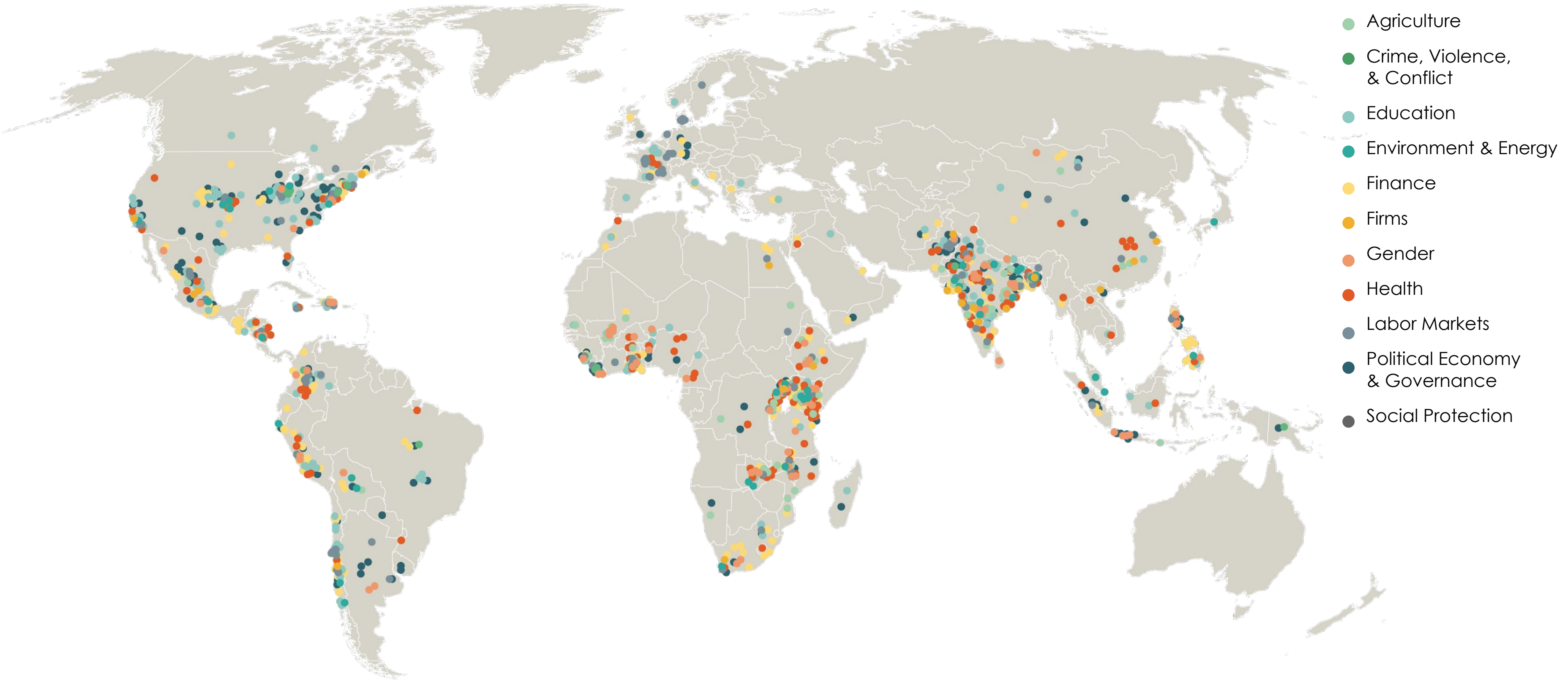
April 26, 2024

Roadmap

1. Motivation and Data
2. Results
3. Discussion & Action

J-PAL's mission is to **reduce poverty** by ensuring that **policy** is informed by **evidence**, and **research** is translated into **action**.

Global knowledge: 2,200+ completed randomized evaluations in more than 96 countries



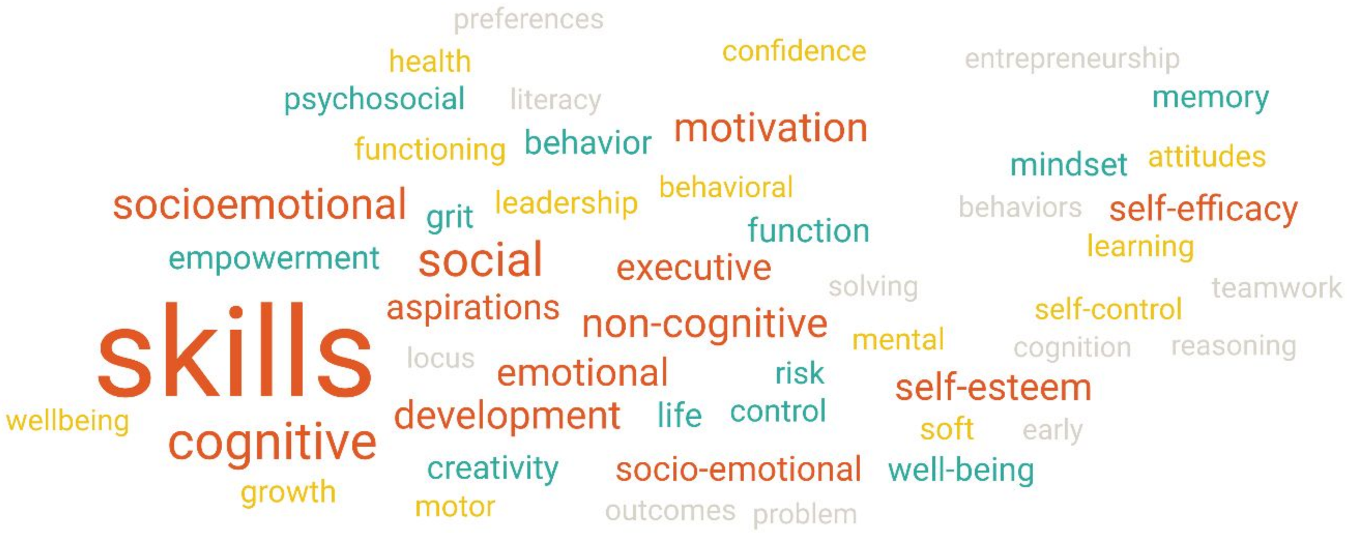
Motivation

- Which *interventions* are most effective at *improving* holistic skills in children?

Motivation

- Which *interventions* are most effective at *improving* **holistic skills** in children?

Holistic Skills



To accurately assess whether an **intervention** can **improve** a certain skill, that evaluation must be able to measure the skill **validly** and **reliably**.

To accurately assess whether an **intervention** can **improve** a certain skill, that evaluation must be able to measure the skill **validly** and **reliably**.

Without establishing validity, one can **claim effects** on an outcome, when in reality, this is not the actual outcome that is being affected.

Holistic Skill Challenges

- More “fuzzy” than literacy/numeracy
- More closely dependent on local social/cultural contexts
- Smaller evidence base
- Concentrated in high-income countries

Motivation

- Which *interventions* are most effective at *improving* holistic skills in children?

Motivation

- ~~Which *interventions* are most effective at *improving* holistic skills in children?~~
- How are researchers *measuring* holistic skills in children?

Motivation

- ~~Which *interventions* are most effective at *improving* holistic skills in children?~~
- How are researchers *measuring* holistic skills in children?
  - What kinds of skills are researchers measuring?

Motivation

- ~~Which *interventions* are most effective at *improving* holistic skills in children?~~
- How are researchers *measuring* holistic skills in children?
  - What kinds of skills are researchers measuring?
  - **What kinds of measures are researchers using?**

Motivation

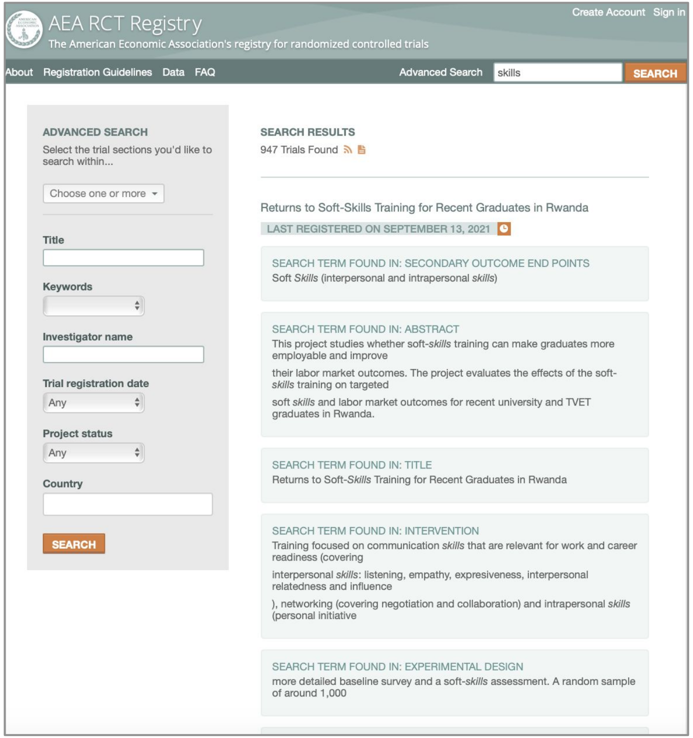
- ~~Which *interventions* are most effective at *improving* holistic skills in children?~~
- How are researchers *measuring* holistic skills in children?
  - What kinds of skills are researchers measuring?
  - **What kinds of measures are researchers using?**
  - **How are they establishing the validity and reliability of those measures in the contexts they are working in?**

Motivation

- ~~Which *interventions* are most effective at *improving* holistic skills in children?~~
- How are researchers *measuring* holistic skills in children?
  - What kinds of skills are researchers measuring?
  - **What kinds of measures are researchers using?**
  - **How are they establishing the validity and reliability of those measures in the contexts they are working in?**
  - What should future researchers continue doing or do differently?

Data and Inclusion Criteria

- American **Economic** Assn. (AEA) **RCT** registry
  - Almost all projects had at least 1 economist PI
  - 98 projects (41%) had interdisciplinary PI teams



AEA RCT Registry

The American Economic Association's registry for randomized controlled trials

ADVANCED SEARCH

Select the trial sections you'd like to search within...

Choose one or more

Title

Keywords

Investigator name

Trial registration date: Any

Project status: Any

Country

SEARCH

SEARCH RESULTS

947 Trials Found

Returns to Soft-Skills Training for Recent Graduates in Rwanda

LAST REGISTERED ON SEPTEMBER 13, 2021

SEARCH TERM FOUND IN: SECONDARY OUTCOME END POINTS

Soft *Skills* (interpersonal and intrapersonal *skills*)

SEARCH TERM FOUND IN: ABSTRACT

This project studies whether soft-*skills* training can make graduates more employable and improve

their labor market outcomes. The project evaluates the effects of the soft-*skills* training on targeted

soft *skills* and labor market outcomes for recent university and TVET graduates in Rwanda.

SEARCH TERM FOUND IN: TITLE

Returns to Soft-*Skills* Training for Recent Graduates in Rwanda

SEARCH TERM FOUND IN: INTERVENTION

Training focused on communication *skills* that are relevant for work and career readiness (covering

interpersonal *skills*: listening, empathy, expresiveness, interpersonal relatedness and influence

), networking (covering negotiation and collaboration) and intrapersonal *skills* (personal initiative

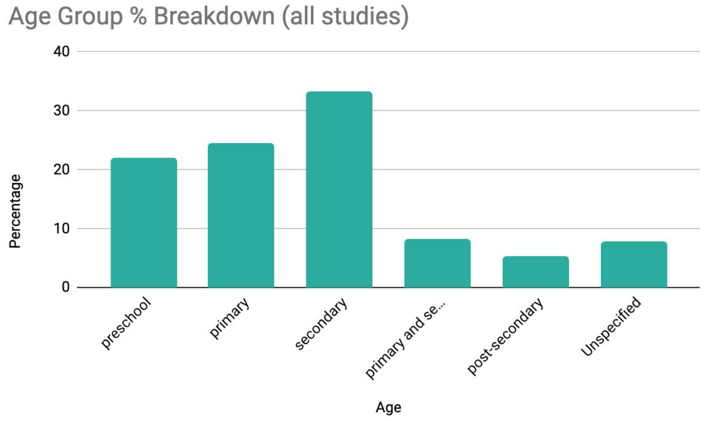
SEARCH TERM FOUND IN: EXPERIMENTAL DESIGN

more detailed baseline survey and a soft-*skills* assessment. A random sample of around 1,000

www.socialscienceregistry.org

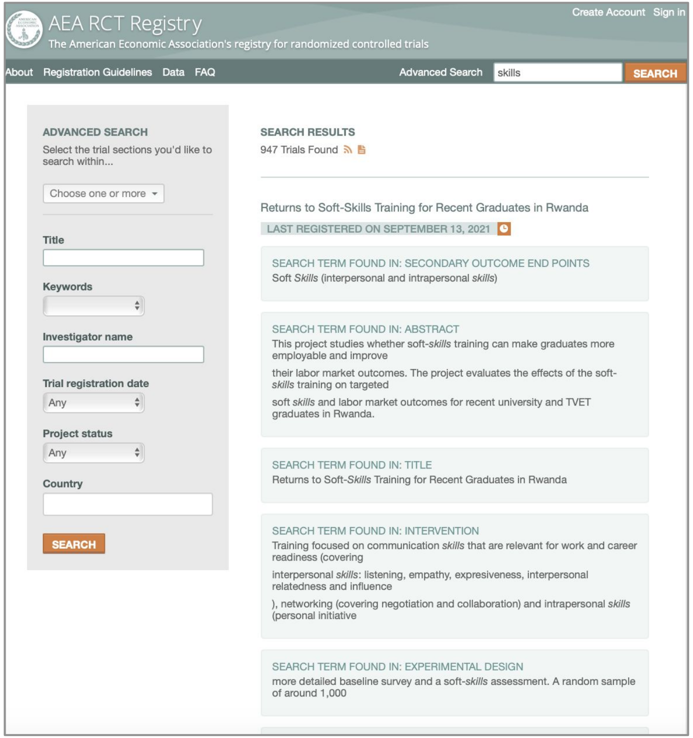
Data and Inclusion Criteria

- American **Economic** Assn. (AEA) **RCT** registry
  - Almost all projects had at least 1 economist PI
  - 98 projects (41%) had interdisciplinary PI teams
- Age range (3-18 years old)
- Outcomes measured in children



Data and Inclusion Criteria

- American **Economic** Assn. (AEA) **RCT** registry
  - Almost all projects had at least 1 economist PI
  - 98 projects (41%) had interdisciplinary PI teams
- Age range (3-18 years old)
- Outcomes measured in children
- Search term: "skills" in the abstract, intervention, outcomes sections
  - Added search term "preschool" to better capture studies in ECE age group
- = 237 RCT registry entries, 122 papers
- 20 peer reviews → qualitative insights



AEA RCT Registry

The American Economic Association's registry for randomized controlled trials

skills

ADVANCED SEARCH

Select the trial sections you'd like to search within...

Title

Keywords

Investigator name

Trial registration date: Any

Project status: Any

Country

SEARCH RESULTS

947 Trials Found

Returns to Soft-Skills Training for Recent Graduates in Rwanda

LAST REGISTERED ON SEPTEMBER 13, 2021

SEARCH TERM FOUND IN: SECONDARY OUTCOME END POINTS

Soft *Skills* (interpersonal and intrapersonal *skills*)

SEARCH TERM FOUND IN: ABSTRACT

This project studies whether soft-*skills* training can make graduates more employable and improve

their labor market outcomes. The project evaluates the effects of the soft-*skills* training on targeted

soft *skills* and labor market outcomes for recent university and TVET graduates in Rwanda.

SEARCH TERM FOUND IN: TITLE

Returns to Soft-*Skills* Training for Recent Graduates in Rwanda

SEARCH TERM FOUND IN: INTERVENTION

Training focused on communication *skills* that are relevant for work and career readiness (covering

interpersonal *skills*: listening, empathy, expresiveness, interpersonal relatedness and influence

), networking (covering negotiation and collaboration) and intrapersonal *skills* (personal initiative

SEARCH TERM FOUND IN: EXPERIMENTAL DESIGN

more detailed baseline survey and a soft-*skills* assessment. A random sample of around 1,000

www.socialscienceregistry.org

Results

Results: What kinds of measures are researchers using?

Results: What kinds of measures are researchers using?

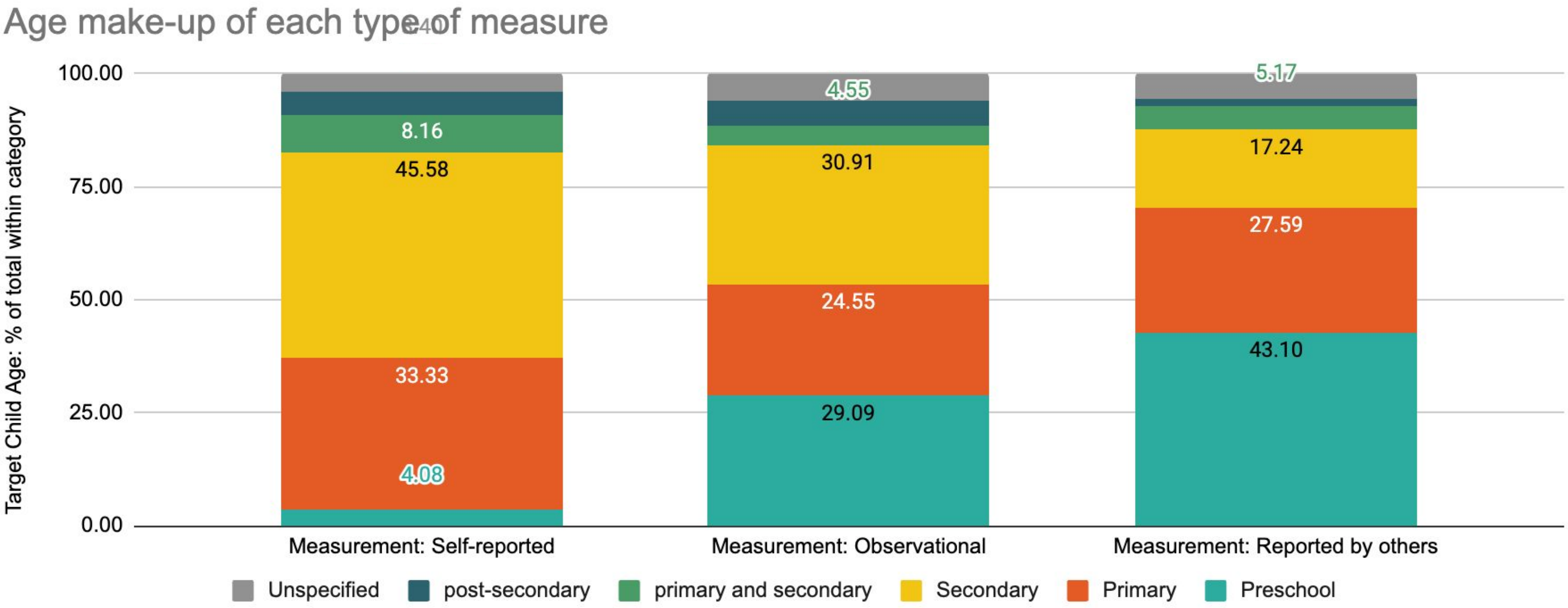
Self-report measures	**Observed / direct assessment**	**Reported by others**
(49%)	(38%)	(11%)

* These are broad categories and not exclusive (one study can use more than one type of measure)

Photo credits: (self report) Photo: shutterstock; (observed) Young1ove, Botswana; (reported by others) Kyle Murphy, J-PAL, Haryana, India

Results: What kinds of measures are researchers using?

Self-report measures	Observed / direct assessment	Reported by others



Results: What checks are researchers using to assess the validity and reliability of their measures?

Definitions

- **Reliability** - the degree to which an instrument produces the same results under unchanged conditions

Photo: Lundberg (2006)

Definitions

- **Reliability** - the degree to which an instrument produces the same results under unchanged conditions
- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)

Photo: Lundberg (2006)

Definitions

- **Reliability** - the degree to which an instrument produces the same results under unchanged conditions
- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)

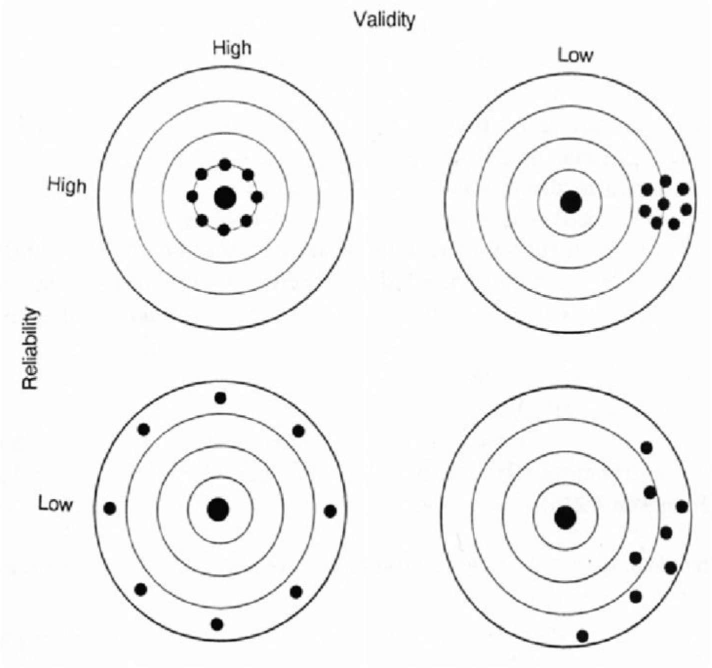


Photo: Lundberg (2006)

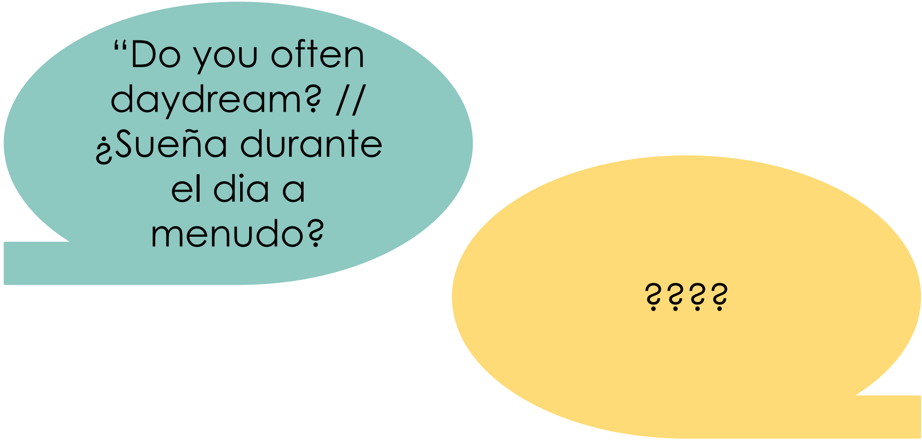
Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
  - **Face validity** - does the participant understand the assessment?

Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
  - **Face validity** - does the participant understand the assessment?

Example: Measuring **Openness** in a New Context



"Do you often daydream? // ¿Sueña durante el dia a menudo?

????

Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
  - **Face validity** - does the participant understand the assessment?
  - **Content validity** - does the assessment measure the concept in full?

Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
  - **Face validity** - does the participant understand the assessment?
  - **Content validity** - does the assessment measure the concept in full?

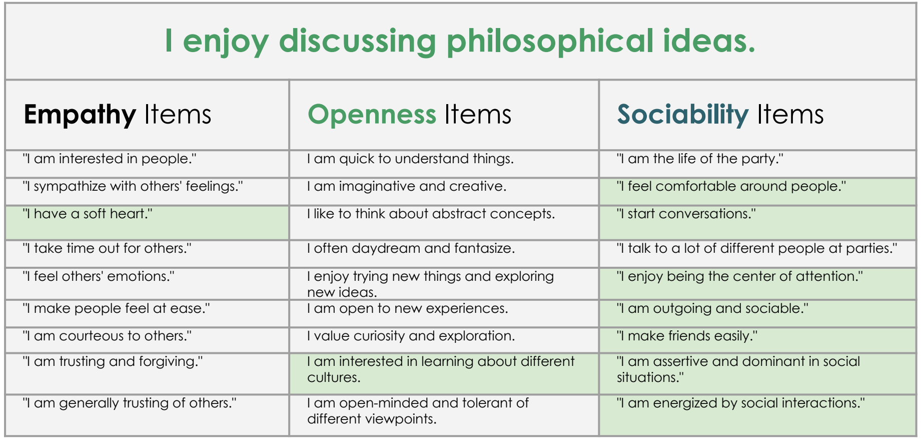
Validity of **Openness** Items		
❓	openness to ideas	I am quick to understand things.
❓	openness to ideas	I am imaginative and creative.
❓	openness to ideas	I like to think about abstract concepts.
❓	openness to ideas	I often daydream and fantasize.
✅	openness to experiences	I enjoy trying new things and exploring new ideas.
✅	openness to experiences	I am open to new experiences.
✅	openness to experiences	I value curiosity and exploration.
✅	openness to people	I enjoy discussing philosophical ideas.
✅	openness to people	I am interested in learning about different cultures.
✅	openness to people	I am open-minded and tolerant of different viewpoints.

Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
    - **Face validity** - does the participant understand the assessment?
    - **Content validity** - does the assessment measure the concept in full?
    - **Construct validity** - is the assessment correlated with other items or assessments that try to measure the same concept?

Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
  - **Face validity** - does the participant understand the assessment?
  - **Content validity** - does the assessment measure the concept in full?
  - **Construct validity** - is the assessment correlated with other assessments or items that try to measure the same concept?



I enjoy discussing philosophical ideas.

Empathy Items	**Openness** Items	**Sociability** Items
"I am interested in people."	I am quick to understand things.	"I am the life of the party."
"I sympathize with others' feelings."	I am imaginative and creative.	"I feel comfortable around people."
"I have a soft heart."	I like to think about abstract concepts.	"I start conversations."
"I take time out for others."	I often daydream and fantasize.	"I talk to a lot of different people at parties."
"I feel others' emotions."	I enjoy trying new things and exploring new ideas.	"I enjoy being the center of attention."
"I make people feel at ease."	I am open to new experiences.	"I am outgoing and sociable."
"I am courteous to others."	I value curiosity and exploration.	"I make friends easily."
"I am trusting and forgiving."	I am interested in learning about different cultures.	"I am assertive and dominant in social situations."
"I am generally trusting of others."	I am open-minded and tolerant of different viewpoints.	"I am energized by social interactions."

Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
  - **Face validity** - does the participant understand the assessment?
  - **Content validity** - does the assessment measure the concept in full?
  - **Construct validity** - is the assessment correlated with other assessments or items that try to measure the same concept?

Wolf SEL Test
Openness Score: 💯

Macours Life Skills Test
Openness Score: 💯

Williams Holistic Skills Test
Openness Score: 30/100

I enjoy discussing philosophical ideas.

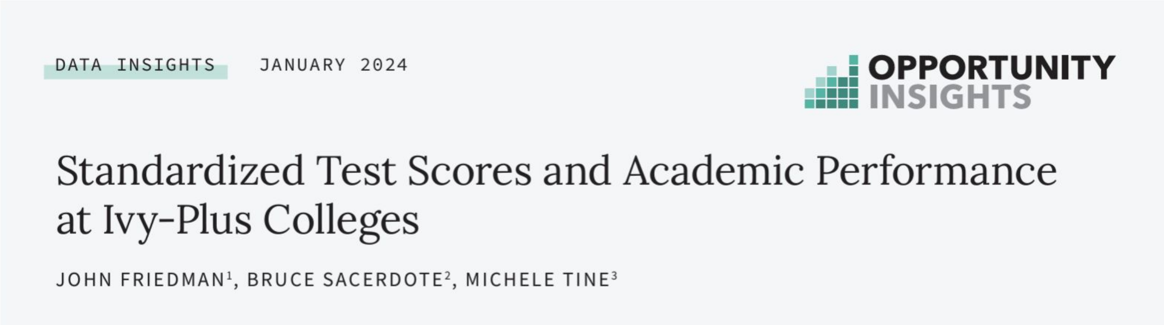
Empathy Items	**Openness** Items	**Sociability** Items
"I am interested in people."	I am quick to understand things.	"I am the life of the party."
"I sympathize with others' feelings."	I am imaginative and creative.	"I feel comfortable around people."
"I have a soft heart."	I like to think about abstract concepts.	"I start conversations."
"I take time out for others."	I often daydream and fantasize.	"I talk to a lot of different people at parties."
"I feel others' emotions."	I enjoy trying new things and exploring new ideas.	"I enjoy being the center of attention."
"I make people feel at ease."	I am open to new experiences.	"I am outgoing and sociable."
"I am courteous to others."	I value curiosity and exploration.	"I make friends easily."
"I am trusting and forgiving."	I am interested in learning about different cultures.	"I am assertive and dominant in social situations."
"I am generally trusting of others."	I am open-minded and tolerant of different viewpoints.	"I am energized by social interactions."

Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
  - **Face validity** - does the participant understand the assessment/concept?
  - **Content validity** - does the assessment measure the concept in full?
  - **Construct validity** - is the assessment correlated with other assessments or items that try to measure the same concept?
  - **Predictive validity** - are the assessment results associated with an outcome that, ex-ante, one would expect the concept to correlate with?

Definitions

- **Validity** - the degree to which evidence and theory support the interpretation of test scores for the proposed use of tests (AERA, APAP, NCME)
  - **Face validity** - does the participant understand the assessment/concept?
  - **Content validity** - does the assessment measure the concept in full?
  - **Construct validity** - is the assessment correlated with other assessments or items that try to measure the same concept?
  - **Predictive validity** - are the assessment results associated with an outcome that, ex-ante, one would expect the concept to correlate with?



DATA INSIGHTS JANUARY 2024

OPPORTUNITY INSIGHTS

Standardized Test Scores and Academic Performance at Ivy-Plus Colleges

JOHN FRIEDMAN[1], BRUCE SACERDOTE[2], MICHELE TINE[3]

Results: What checks are researchers using to assess the validity and reliability of their measures?

Results: What checks are researchers using to assess the validity and reliability of their measures?

- Validation checks were relatively scarce
  - (21% of all entries, 45% of papers w/ appendix)

Results: What checks are researchers using to assess the validity and reliability of their measures?

- Validation was relatively scarce
  - (21% of all entries, **45% of papers w/ appendix***)
- **Cronbach's alpha** (17%*, 8% of total) - assessing reliability
- **Factor analysis** (16%*, 7% of total) - assessing content and construct validity
- **Piloting** - (11%*, 6% of total) assessing face validity
- **Correlations** (9%*, 4% of total) - assessing predictive validity

* out of all studies with a paper and appendix.

Borrowing measures from other contexts

- Are researchers developing their own measures or **borrowing** existing measures?

Borrowing measures from other contexts

- Are researchers developing their own measures or **borrowing** existing measures?
  - 60% of studies borrowed measures,14% developed original measure, 26% both

Borrowing measures from other contexts

- Are researchers developing their own measures or **borrowing** existing measures?
  - 60% of studies borrowed measures,14% developed original measure, 26% both
- If they are citing others, is the cited study a **validation paper** or a regular paper signaling precedence?

Borrowing measures from other contexts

- Are researchers developing their own measures or **borrowing** existing measures?
  - 60% of studies borrowed measures,14% developed original measure, 26% both
- If they are citing others, is the cited study a **validation paper** or a regular paper signaling precedence?
  - 37% (of papers, 19% of all entries) cite a validation paper

Borrowing measures from other contexts

- Are researchers developing their own measures or **borrowing** existing measures?
  - 60% of studies borrowed measures,14% developed original measure, 26% both
- If they are citing others, is the cited study a **validation paper** or a regular paper signaling precedence?
  - 37% (of papers, 19% of all entries) cite a validation paper
- If they are citing a validation paper, was the tool validated in the **same context**?

Borrowing measures from other contexts

- Are researchers developing their own measures or **borrowing** existing measures?
  - 60% of studies borrowed measures,14% developed original measure, 26% both
- If they are citing others, is the cited study a **validation paper** or a regular paper signaling precedence?
  - 37% (of papers, 19% of all entries) cite a validation paper
- If they are citing a validation paper, was the tool validated in the **same context**?
  - 18 papers cited a validation study which matched the age range in the current intervention
  - 11 papers had all their tools match the regional contexts of their cited validation papers

How are RCT researchers (*economists) *measuring* holistic skills in children?

- **What kinds of measures are researchers using?**
  - self-report
- **How are they establishing the validity and reliability of those measures in the contexts they are working in?**
  - By referencing past studies
  - lack of public information about the different validity/reliability checks performed for holistic skills measures
  - limited context-specific validity and reliability testing being done

Why?

Why is there a lack of validity and reliability reporting?

Execution challenges

- Lack of guidance
- Lack of examples from previous research in similar contexts
- Cost

Why is there a lack of validity and reliability reporting?

Execution challenges

- Lack of guidance
- Lack of examples from previous research in similar contexts
- Cost

Dissemination challenges

- Perception of journal editor and referee preferences
- Field norms

Main Takeaways

1. Opportunities for **more** citing, explaining, or conducting validation & reliability checks, **relevant** to the context
2. Making thinking **public** → helps the next researcher facing similar measurement decisions in that context
3. Conducting validity checks **early** + robustness checks after data collection, with flexibility to adjust proposed measures
4. **Multidisciplinary** teams well placed to come up with better measurements that draw from the strengths of different fields or to design new measures or adapt old tools to new contexts.

What is J-PAL doing about it?

Execution challenges

- **Lack of guidance**
- **Lack of record or examples** from previous research in similar contexts
- **Cost**

Dissemination challenges

- Perception of journal editor and referee preferences
- **Field norms**

Guiding questions at the RCT Design Stage

Guiding questions at the RCT Design Stage

1. **Which skills** do I want to measure, why those skills, and **which** measurement **tools** do I plan to use?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
    a. Are there existing tools that I could consider using to measure the skill of interest, or do I need to design a new tool or measurement instrument?
    b. What are the advantages of each of the different possible measures for capturing the trait of interest? Could I use multiple measures? If so, how will I combine them?
    c. Do I have the disciplinary expertise to use or design these measures? Should I collaborate with a co-author from a different discipline?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly **capturing the latent trait** of interest in the context of the proposed study?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
   a. What other outcomes should my measure be correlated with if it truly measures the trait I would like for it to measure, and how will I test this?
   b. What can I do to assure that the proposed measure will allow separating the measurement of the latent trait of interest from other factors (e.g. other related traits or some form of response bias)?
   c. If I am planning on using a measure someone else designed, has a validation paper been published? Does the context of the validation paper match the context of my evaluation? (Tested on a similar age group, in a similar language, in a similar geography?)

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
3. How will I determine that the proposed measures will capture the latent trait with enough **precision** in the context of the proposed study?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
3. How will I determine that the proposed measures will capture the latent trait with enough precision in the context of the proposed study?
   a. What methods can I use to reduce measurement error?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
3. How will I determine that the proposed measures will capture the latent trait with enough precision in the context of the proposed study?
4. How and where will I **report** on the measurement adaptation, piloting, validity and reliability tests?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
3. How will I determine that the proposed measures will capture the latent trait with enough precision in the context of the proposed study?
4. How and where will I report on the measurement adaptation, piloting, validity and reliability tests?
   a. Should I commit to this reporting in a pre-analysis plan?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
3. How will I determine that the proposed measures will capture the latent trait with enough precision in the context of the proposed study?
4. How and where will I report on the measurement adaptation, piloting, validity and reliability tests?
5. If my measure is failing some reliability and validity checks, how will I determine if there is an **issue** with the **measure** or with the experimental **design**?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
3. How will I determine that the proposed measures will capture the latent trait with enough precision in the context of the proposed study?
4. How and where will I report on the measurement adaptation, piloting, validity and reliability tests?
5. If my measure is failing some reliability and validity checks, how will I determine if there is an issue with the measure or with the experimental design?
    a. For example, may the measure fail to capture the same trait over time or may the experimental variations themselves affect the validity of a measure (e.g. by inducing changes in response patterns/biases) ?

Guiding questions at the RCT Design Stage

1. Which skills do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly capturing the latent trait of interest in the context of the proposed study?
3. How will I determine that the proposed measures will capture the latent trait with enough precision in the context of the proposed study?
4. How and where will I report on the measurement adaptation, piloting, validity and reliability tests?
5. If my measure is failing some reliability and validity checks, how will I determine if there is an issue with the measure or with the experimental design?
6. How will I **adjust my analysis and the write-up** of the research results to reflect my findings on the validity and reliability of my measures?

Guiding questions at the RCT Design Stage

1. Which **skills** do I want to measure, why those skills, and which measurement tools do I plan to use?
2. How will I determine that the proposed measures are predominantly **capturing the latent trait** of interest in the context of the proposed study?
3. How will I determine that the proposed measures will capture the latent trait with enough **precision** in the context of the proposed study?
4. How and where will I **report** on the measurement adaptation, piloting, validity and reliability tests?
5. If my measure is failing some reliability and validity checks, how will I determine if there is an **issue** with the **measure** or with the experimental **design**?
6. How will I **adjust my analysis and the write-up** of the research results to reflect my findings on the validity and reliability of my measures?

Appendix 2 - List of Measures Observed in Review

For each tool appearing >1 time:

- # of times it appears in the review
- which RCTs use this tool
- which skill the tool measures
- larger measurement category (self-report, observed, etc)
- original paper that developed/validated the tool
- age range of the original sample
- age range the tool was used among the observed RCTs
- official website (if open-access)
- which country the original paper was in
- additional countries the tool was validated in (from observed cited validation papers from evaluations in this review)

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Thank you!

jwilliams@povertyactionlab.org