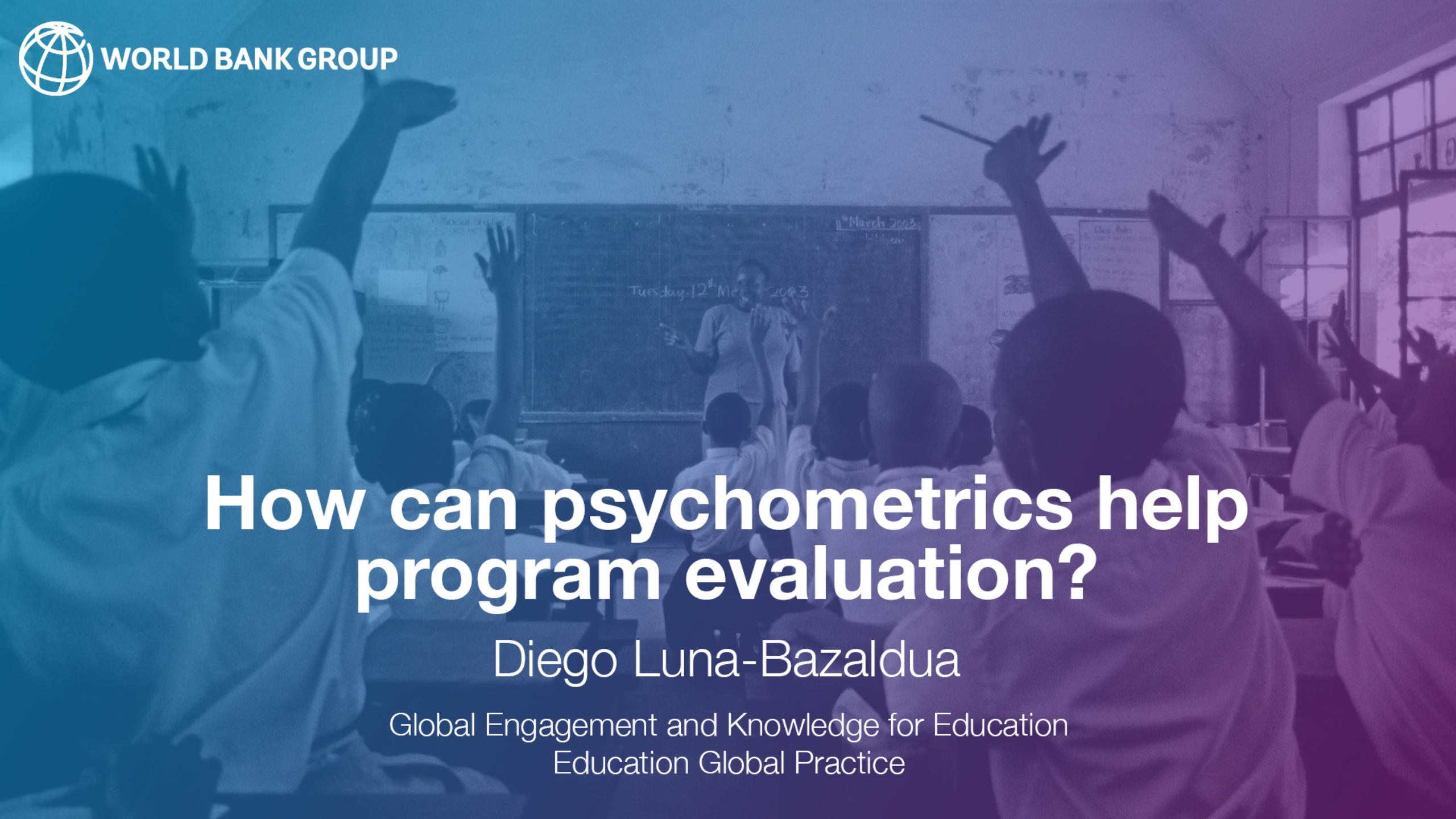


WORLD BANK GROUP

How can psychometrics help program evaluation?

Diego Luna-Bazaldua

Global Engagement and Knowledge for Education
Education Global Practice

Roadmap

- Motivation
- Measurement Error
- Reliability
- Validity
- Conclusions and Q&A

WORLD BANK GROUP
Education

Psychometrics – a brief definition

- Also known as *psychological measurement.*
  - Development and use of **quantitative models** to explain **psychological processes**.
  - Analysis of psychological data.
  - Identification of properties of measurement tools.

- Dual role of Quantitative Psychologists:
  - Develop measurement theory and models.
  - Translate into concrete applications for psychology, health, and education.

WORLD BANK GROUP
Education

Why psychometrics?

- Economics and psychology, distinct scientific disciplines.
  - Both focus on human behavior at small and large scales
  - Common ground: Behavioral Economics, Game Theory.

- Lack of interdisciplinary work has resulted in the development of methodological approaches for program evaluation that overlook the principles of psychometrics.
- Applied economists: What psychometric properties should be used when developing or selecting measurement tools?
  - E.g.: questionnaires, checklists, rubrics, or standardized tests.

Validity typology in the Cook & Campbell framework

- Likely familiar with Cook & Campbell's work on methodology and causal inference.
- The Cook & Campbell tradition vs. the Rubin tradition vs. the Pearl tradition in CI.
- **Validity** as an approximate truth of an inference.
  - When we say something is valid, we make a judgment about the extent to which relevant **evidence supports that inference** as being true or correct

Validity typology in the Cook & Campbell framework

Type of validity	Research/Methodological question
Internal: Causal relations. • ***Main focus in Program Evaluation***	Is the covariation causal, or would the same covariation have been obtained without the treatment?
External: Generalizability of results. • ***Focus in Program Evaluation***	Which general constructs are involved in the persons, settings, treatments, and observations used in the experiment?
Statistical conclusion: Proper use of statistics to explain covariation. • ***Focus in both Program Evaluation and in Psychometrics***	How large and reliable is the covariation between the presumed cause and effect?
Construct: understanding constructs and assessing them • ***Focus in in Psychometrics***	How generalizable is the locally embedded causal relationship in a study to the higher-order constructs **measured** in the study?

Let’s start with a familiar equation!

$$Y = \beta_0 + \beta_1 X + E$$

Dependent Variable	Independent Variable
Learning outcomes • Test scores in national or international assessments.	Socioemotional skills. • Scores on an observation rubric.
Preference for STEM-related fields in preparation for higher education. • Answers to a checklist of aptitudes and interests.	Attitude towards traditional vs. non-traditional gender roles. • Scores to items in a Likert scale questionnaire.

Linear Regression

- How is this related to psychometrics?

$$Y = \beta_0 + \beta_1 X$$

- The **more measurement error** in an observed total score (i.e., in either X or Y), the **less psychometrically reliable** that score is.

- Moreover, the presence of measurement error reduces (i.e., attenuates) the correlation between X and Y.

Correction for Attenuation

- The regression coefficient β_1 is not affected by measurement errors in the dependent variable Y, but it is attenuated by measurement errors in the independent variable X.

$$Y = \beta_0 + \frac{\beta_1}{\rho_x} X$$

ρ_x is the reliability coefficient for the score X, which can take values between 0 and 1. A value of 1 would indicate that X has no measurement error.

- Thus, if we are not careful about the psychometric properties of the tools we use in program evaluation, economics and psychology, we could end up drawing incorrect conclusions about our target variable Y given X.

Example

- A Bank project on gender.
- Supported by theory and former research in economics and psychology.

Training on entrepreneurship-related skills → $\beta = 0$ → **Sense of personal empowerment**

- Why is the TTL getting these odd results?
- Could it be due to **measurement error**?

WORLD BANK GROUP
Education

Introduction

- Previous example:
  - Linkage of program evaluation in Economics and topics from psychometrics in Psychology.
  - Concepts: Reliability, Measurement Error (and Validity).

- Other threats associated with low evidence of validity and/or reliability:
  - lack of consistency of within-subject scores over time,
  - construct confounding,
  - mono-operation and mono-method bias,
  - reactivity to stimulus/tool,
  - ...

Reliability

Test reliability

- Mathematical definition in the Classical Test Theory:
  - Ratio of true score variance and observed score variance.

- Reliability coefficients in psychometrics:
  - Ratios that quantify the mathematical definition.
  - The closest the value to **1**, the higher reliability of the tool.
  - The closest the value to **0**, the lower reliability of the tool.

Test reliability

- Day 1: 70.8 kg (156.08 pounds).
  Dinner: 10 Beef tacos and beer.
- Day 2: 66.3 kg (146.16 pounds).
  Dinner: Greek salad and non-fat yogurt.
- Day 3: 82.1 kg (180.99 pounds).
  Dinner: Leftover Greek salad and non-fat yogurt.
- Day 4: 75.6 kg (166.67 pounds).
  Dinner: Fish and chips.

Would you say that this scale **reliable**? Why?

Image taken from fitbit.com

Image taken from fitbit.com

Test reliability

- Scale 1: 70.8 kg.

- Scale 2: 70.6 kg.

- Scale 3: 86.8 kg.

- Scale 4: 70.7 kg.

- Scale 5: 70.8 kg.

- Scale 6: 70.7 kg.

WORLD BANK GROUP
Education

Would you say that these scales are **reliable**? Why?

Test reliability

- Three coefficients to quantify the reliability for questionnaires, checklists, multiple-choice standardized tests:
  - Temporal stability coefficient.
  - Equivalence coefficient.
  - Internal consistency coefficient

- There are two coefficients to quantify reliability for rubrics or observation protocols that require experts ratings:
  - Inter-rater coefficients (Cohen's Kappa and ICC).

Test reliability

Source of error	Reliability coefficient	Data collection procedure	Statistical analysis
1. Examinee changes over time.	Temporal stability coefficient.	One test form administration, wait, retest with the same test form.	Pearson correlation coefficient. $\rho_{test\text{-}retest}$
2. Content sampling between two test forms measuring the same construct.	Equivalence coefficient.	Administration of forms A and B to the same examinees.	Pearson correlation coefficient. $\rho_{A\text{-}B}$
3. Content sampling in a single test form with potentially biased items.	Internal consistency coefficient	Administration of one test form in a single occasion.	a. Cronbach's Alpha coefficient, Guttman's Lambda coefficient, Split-half methods with Spearman correction…

Example:

- Alaka is interested in using a questionnaire to measure attitudes towards women empowerment in a country project. Since this will be a four-year project with multiple measures over time, she is concerned that participants may not express consistent attitudes towards women. What aspects of reliability should Alaka take into consideration when selecting a tool that measures attitudes toward women empowerment?
  - Internal consistency.
  - Test forms equivalence.
  - Temporal stability.

Test reliability

- In some instances the measurement tool is focused on the agreement among third party ratings (i.e., experts) based on the examinee's performance.
  - Overall score of a student's essay.
  - Manager's score on a rubric of employee's performance.
  - Physician's judgement of a patient's mental health.
  - Panelists' scores on a rubric of an applicant during a job interview.

Image taken from clifdiving.com

Coefficients of inter-rater reliability:

- Cohen's Kappa coefficient. Used for nominal or ordinal data.

- ICC (Intra-class correlation) coefficient. Used for interval/numerical data.

Checklist: Reliability

- ❑When selecting a tool:
  - ❑Select tools that have reported reliability from previous studies.
  - ❑The higher value for the reliability coefficient, the better.
  - ❑Reliability coefficients above 0.90 are recommended if the assessment will have an impact on the examinees or others.
- ❑When developing a new tool:
  - ❑Do a pilot study and calculate the reliability coefficient of interest.
- ❑Reliability coefficients and methods for raters ≠ reliability coefficients for checklists, standardized tests, or questionnaires.

Validity

Validity and the validity evidence framework

- The current definition of validity defines it as the degree to which **evidence** and **theory** support **interpretations of test scores for proposed uses of the tests**.

- Validity requires the **continuous accumulation of evidence** to support or dispute interpretations of tests scores, uses of tests, as well as the intended (and unintended consequences) of such uses.

- A test is not itself valid.

Validity is not…

- The extent to which the test measures what it intends to measure. This definition is no longer accepted in psychometrics.

- A static test property. (e.g., "*the test is valid*").

- A purely quantitative coefficient to rank measures "from more to less valid" (which happens to be the case of CTT reliability). There is also qualitative evidence of validity.

Sources of validity evidence

Source of Validity evidence	Method	Examples
Based on test content.	Mostly qualitative	• Item review by experts in the domain of the assessment.
Based on cognitive processes.	Qualitative and quantitative	• Think-aloud interviews to examinees. • Examinee feedback. • Multicomponent latent trait analysis.
Based on internal structure.	Quantitative	• Exploratory and confirmatory factor analyses. • Multidimensional scaling.
Based on relationship between test scores and external variables.	Quantitative	• Correlation and regression models. • Structural equation models.
Based on consequences of the use of the test.	Qualitative and quantitative	• Alignment between test objectives and use. • Assessment and policy analysis.

Example:

- Adelle has developed a measurement tool to assess early childhood development (ECD) in domains of literacy and numeracy.
- Adelle submits her work for review. A reviewer replies indicating that the tool is not measuring ECD but rather exposure to school contents.
- Embedded in the feedback, there is a validity claim for the tool: "the tool is measuring a different construct."
- If Adelle wants to reply with a validity claim highlighting that the tool is measuring ECD, what would you suggest to her as evidence to document in a psychometric research agenda?

Example:

- Adelle's psychometric research agenda:
  - Document the alignment of every item in the tool to early childhood literacy and numeracy developmental milestones described in psychological theory.
  - Document cognitive interviews to children in which the researcher describes their cognitive processes and behavior when answering each one of the items in the tool. In psychology we know that children at different stages of development provide qualitatively different answers to cognitive and socioemotional tests.
  - Document the correlation patterns of the items. Psychologists expect higher correlations among literacy items and among numeracy items, and lower correlations between items that measure different domains.
  - Document the association between each item and children's age. If the items are measuring development, regression coefficients for age should be positive when controlling for school enrollment factors.
  - Document standard setting studies in which other experts in early childhood provide meaning to the literacy and numeracy total scores highlighting the intended uses for the tool.

Example:

- Adelle's psychometric research agenda:
  - Document the alignment of every item in the tool to early childhood literacy and numeracy developmental milestones described in psychological theory.
  - Document cognitive interviews to children in which the researcher describes their cognitive processes and behavior when answering each one of the items in the tool. In psychology we know that children at different stages of development provide qualitatively different answers to cognitive and socioemotional tests.
  - Document the correlation patterns of the items. Psychologists expect higher correlations among literacy items and among numeracy items, and lower correlations between items that measure different domains.
  - Document the association between each item and children's age. If the items are measuring development, regression coefficients for age should be positive when controlling for school enrollment factors.
  - Document standard setting studies in which other experts in early childhood provide meaning to the literacy and numeracy total scores highlighting the intended uses for the tool.

Example:

- Adelle's psychometric research agenda:
  - Document the alignment of every item in the tool to early childhood literacy and numeracy developmental milestones described in psychological theory.
  - Document cognitive interviews to children in which the researcher describes their cognitive processes and behavior when answering each one of the items in the tool. In psychology we know that children at different stages of development provide qualitatively different answers to cognitive and socioemotional tests.
  - Document the correlation patterns of the items. Psychologists expect higher correlations among literacy items and among numeracy items, and lower correlations between items that measure different domains.
  - Document the association between each item and children's age. If the items are measuring development, regression coefficients for age should be positive when controlling for school enrollment factors.
  - Document standard setting studies in which other experts in early childhood provide meaning to the literacy and numeracy total scores highlighting the intended uses for the tool.

Example:

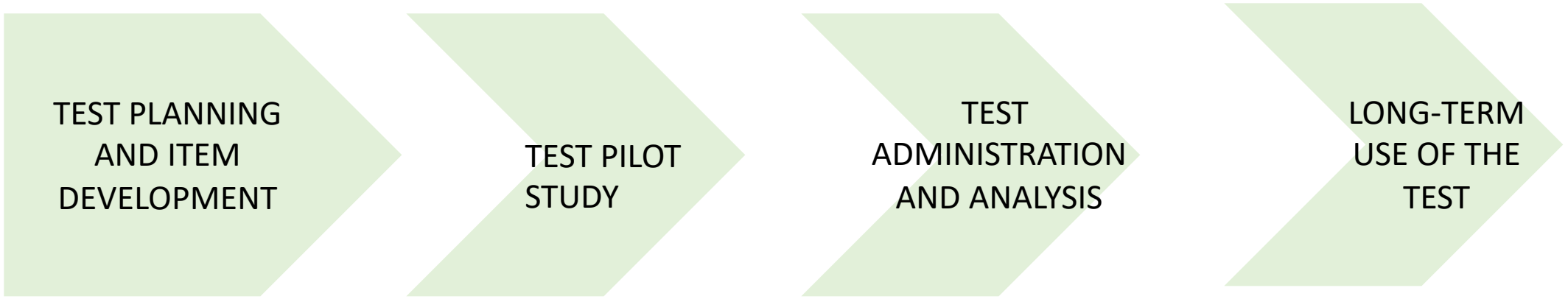
- Adelle's psychometric research agenda:
  - Document the alignment of every item in the tool to early childhood literacy and numeracy developmental milestones described in psychological theory.
  - Document cognitive interviews to children in which the researcher describes their cognitive processes and behavior when answering each one of the items in the tool. In psychology we know that children at different stages of development provide qualitatively different answers to cognitive and socioemotional tests.
  - Document the correlation patterns of the items. Psychologists expect higher correlations among literacy items and among numeracy items, and lower correlations between items that measure different domains.
  - Document the association between each item and children's age. If the items are measuring development, regression coefficients for age should be positive when controlling for school enrollment factors.
  - Document standard setting studies in which other experts in early childhood provide meaning to the literacy and numeracy total scores highlighting the intended uses for the tool.

Example:

- Adelle's psychometric research agenda:
  - Document the alignment of every item in the tool to early childhood literacy and numeracy developmental milestones described in psychological theory.
  - Document cognitive interviews to children in which the researcher describes their cognitive processes and behavior when answering each one of the items in the tool. In psychology we know that children at different stages of development provide qualitatively different answers to cognitive and socioemotional tests.
  - Document the correlation patterns of the items. Psychologists expect higher correlations among literacy items and among numeracy items, and lower correlations between items that measure different domains.
  - Document the association between each item and children's age. If the items are measuring development, regression coefficients for age should be positive when controlling for school enrollment factors.
  - Document standard setting studies in which other experts in early childhood provide meaning to the literacy and numeracy total scores highlighting the intended uses for the tool.

How to prevent or reduce measurement problems in program evaluation?

- If standards and best practices are followed in the development of an assessment, the test developers and test users can prevent several validity problems.



- Validity evidence: Items are developed aligned to the construct intended to be measured

- Item analysis to identify item psychometric properties.
- Validity evidence: cognitive processes.

- Analysis to calculate reliability coefficients.
- Validity evidence: internal structure, cognitive processes, and relationship with external variables.

- Validity evidence: external variables and consequences of test use.

How to prevent or reduce measurement problems in program evaluation?

- Validity evidence is gathered at **the different stages of the test development process**.

- Validity evidence has to be systematically analyzed and documented to support test uses and interpretations.

The standards for educational and psychological testing

- Approximately every 10 years, psychometricians gather to update the standards in psychometrics.

- Essential document on best practices in educational and psychological measurement:
  - **Reliability.**
  - **Validity.**
  - Fairness in testing.
  - Operations in assessment.

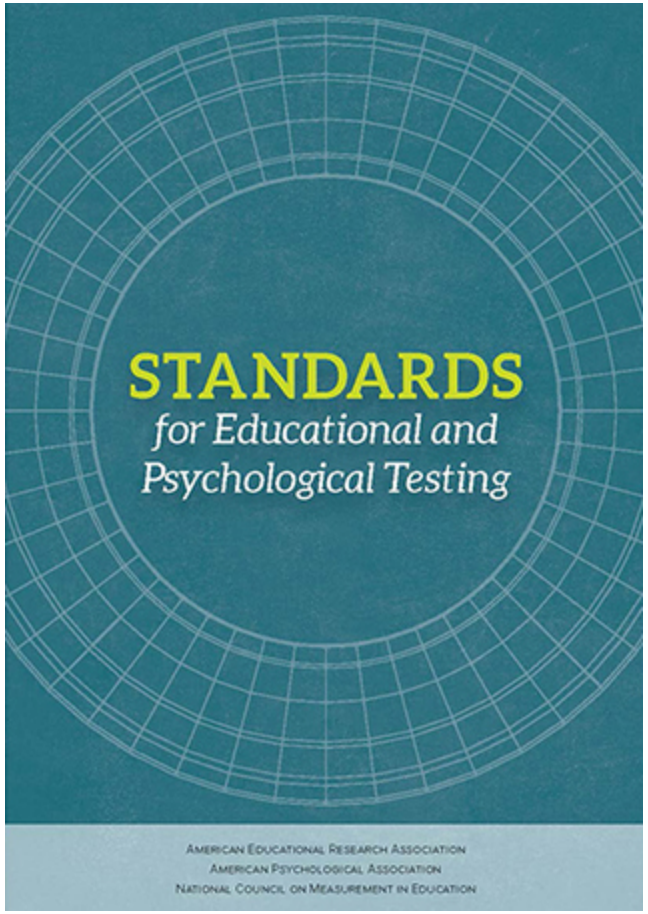


Image taken from apa.com

The standards from ITC for test translation and adaptation

Image taken from intestcom.org

- Essential process as part of the work of many teams in the Bank.
- The AERA, APA and NCME standards also discuss issues related to test translation and adaptation.
- Objective: produce 18 guidelines for adapting psychological and educational tests for use in various different linguistic and cultural contexts (Van de Vijver & Hambleton, 1996).

Linkages between Causal Inference and Psychometrics

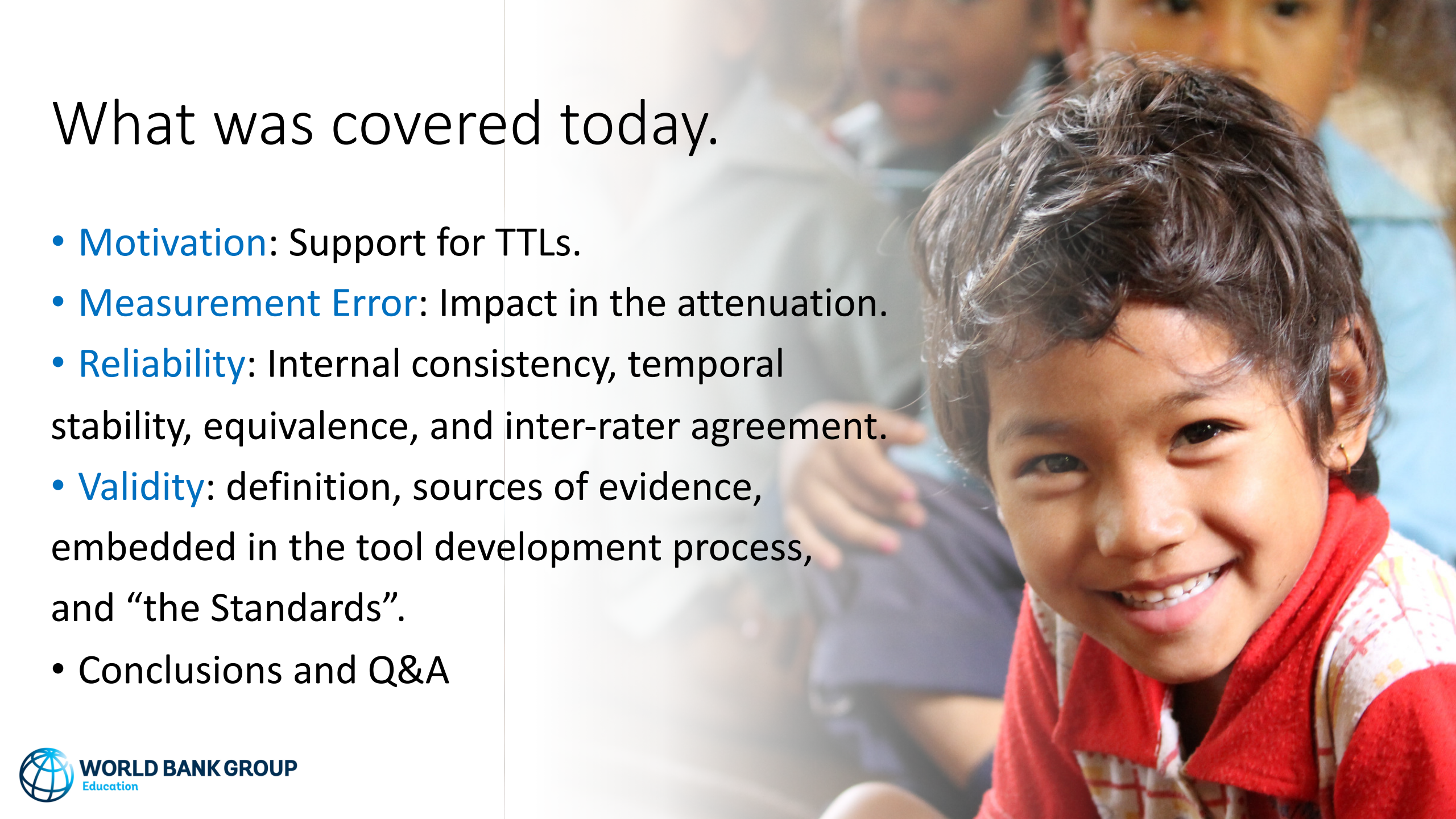
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC. CHAPTER 1

- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

- Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, *39*(1), 22-52.

What was covered today.

- Motivation: Support for TTLs.
- Measurement Error: Impact in the attenuation.
- Reliability: Internal consistency, temporal stability, equivalence, and inter-rater agreement.
- Validity: definition, sources of evidence, embedded in the tool development process, and “the Standards”.
- Conclusions and Q&A



WORLD BANK GROUP
Education

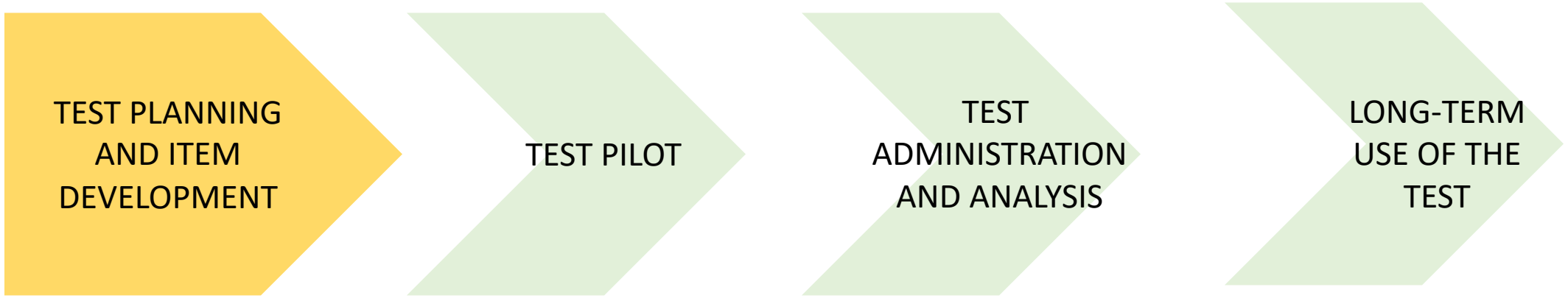
Thank you

- Takeaway of this BBL: I encourage you to engage in conversations with psychometricians and test content experts right from the beginning of any program evaluation planning.

- E-mail: dlunabazaldua@worldbank.org

WORLD BANK GROUP
Education

Annex: Additional slides on validity evidence and checklist



Test planning and item development

Evidence based on test content

- How to know whether a test is representing the construct it intends to measure?
  - Educational tests: analysis of assessment objectives and curriculum that was used to develop the assessment.
  - Psychological tests: theories on cognitive processes intended to be measured and empirical evidence that supports how to operationalize such processes in concrete tasks.

Evidence based on test content

- How well is the test representing the construct?
  - **Representativity or domain coverage**. Extent to which the test content represents all facets of the domain / construct to be measured.
  - **Relevance**. Extent to which each test item measures the defined construct in an attempt to detect the presence of irrelevant contents.

- Most validity studies focused on test content require test content experts to review and judge test items. Experts focus on the alignment between the items and the assessment objectives based on the **test blueprint**.
- Experts must inform to what extent the construct is being covered by the test items. They can rate each item either using a Likert scale or matching items with test blueprint specifications.

Checklist: Test planning and item development

For a new instrument

- ❑Develop a theoretical framework on the construct to be measured.
- ❑State general and specific objectives for the use of the instrument.
- ❑Develop items aligned with the different expressions of the construct and consistent with the instrument objectives.
- ❑Experts review the developed items.

For an instrument already developed

- ❑Gather documentation on the instrument, construct(s) it intends to measure, the alignment between items and the construct, and the use of the instrument in a contexts different than the one it was originally intended to be used for.
- ❑If there are doubts about the instrument content, request experts to review items.

Example: Depression

- Construct facets

❖Depressed mood (subjective or observed).

❖Loss of interest or pleasure.

❖Change in weight or appetite.

❖Insomnia or hypersomnia.

❖Psychomotor retardation or agitation.

❖Loss of energy or fatigue.

❖Worthlessness, hopelessness or guilt.

❖Impaired concentration.

❖Thoughts of death, suicidal ideation or attempt.

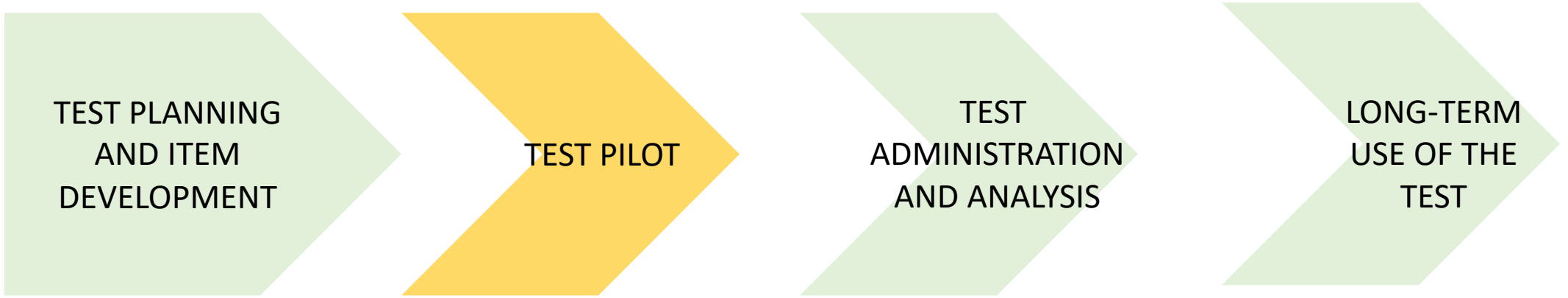
- Items

❖"I have been feeling extremely tired over the last two weeks."

❖"I have lost interest in things I used to enjoy."

❖"I have a hard time trying to fall asleep at night."

❖"I have been feeling sad despite things are going well in my personal life."



Test pilot

Purposes of test pilot

- Identify item psychometric properties before final test administration.
- Review and/or removal of items with unsatisfactory item properties.
- Feedback from examinees on clarity and lack of ambiguity in item content.
- If suitable for the use of the tool (i.e., validity evidence), analysis of cognitive processes involved in answering items.

Checklist: Test pilot

- ❑ Items present adequate psychometric properties:
  - ❑ Item difficulty.
  - ❑ Item discrimination.
  - ❑ Item guessing.
  - ❑ Item contribution to test reliability.
  - ❑ Test dimensionality and item relationship with latent construct.
  - ❑ Items are not biased against a specific group.
- ❑ Items are understood by examinees.
- ❑ Evidence of the cognitive processes involved in answering the items is gathered and documented (in consultation with a cognitive psychologist).

Example: Cognitive process behind a math item

Solve the equation 5X + 4 = 24

a) 8

b) 2

c) 4*

d) 3

5X + 4 = 24

5X + 4 - 4 = 24 - 4

5X/5 = 20/5

X = 4

- Item difficulty: 0.79
- Item-test correlation: 0.12

Example: Cognitive process behind a math item

Solve the equation $5X^2 + 4 = 24$

a) 8

b) 2*

c) 4

d) 3

$5X^2 + 4 = 24$

$5X^2 + 4 - 4 = 24 - 4$

$5X^2/5 = 20/5$

$(X^2)^{½} = (4)^{½}$

$X = 2$

- Item difficulty: 0.36
- Item-test correlation: 0.33

Example: Cognitive process behind a math item

Solve the equation $2X^2 + 4 = 12$

a) 8

b) 5

c) 4

d) 3*

$2X^2 + 4 = 12$

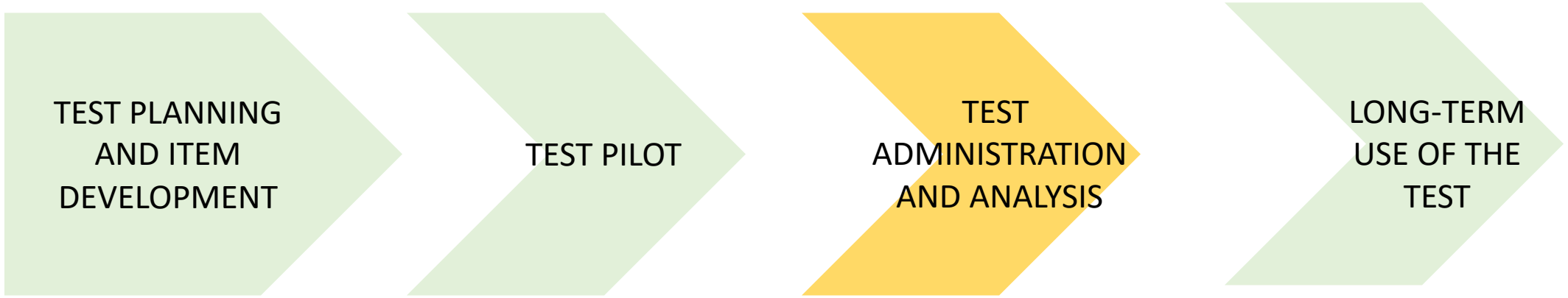
$2X^2 + 4 - 4 = 12 - 4$

$2X^2/2 = 8/2$

$(X^2)^{½} = (4)^{½}$

$X = 2$

- Item difficulty: 0.07
- Item-test correlation: -0.29



Test administration and psychometric analyses

Test administration and psychometric analyses

- In many instances, psychometricians are reached once the test administration occurred.
  - Despite recognizing the effort behind collecting data, I would strongly advice you to reach psychometricians at the earlier stages of the measurement process.
  - If the test content is not adequate to measure the construct, no post-implementation psychometric analysis will adjust for that.

- These psychometric analyses will provide validity evidence for the use of the test, determine the test reliability and items psychometric properties.

Checklist: Psychometric Analyses

- ❑ Items present adequate psychometric properties:
  - ❑ Item difficulty.
  - ❑ Item discrimination.
  - ❑ Item guessing.
- ❑ Test reliability: calculate a internal consistency coefficients (e.g., Cronbach's α).
- ❑ Test dimensionality: estimate factor analyses models to provide validity evidence of the relationship of items with the construct.

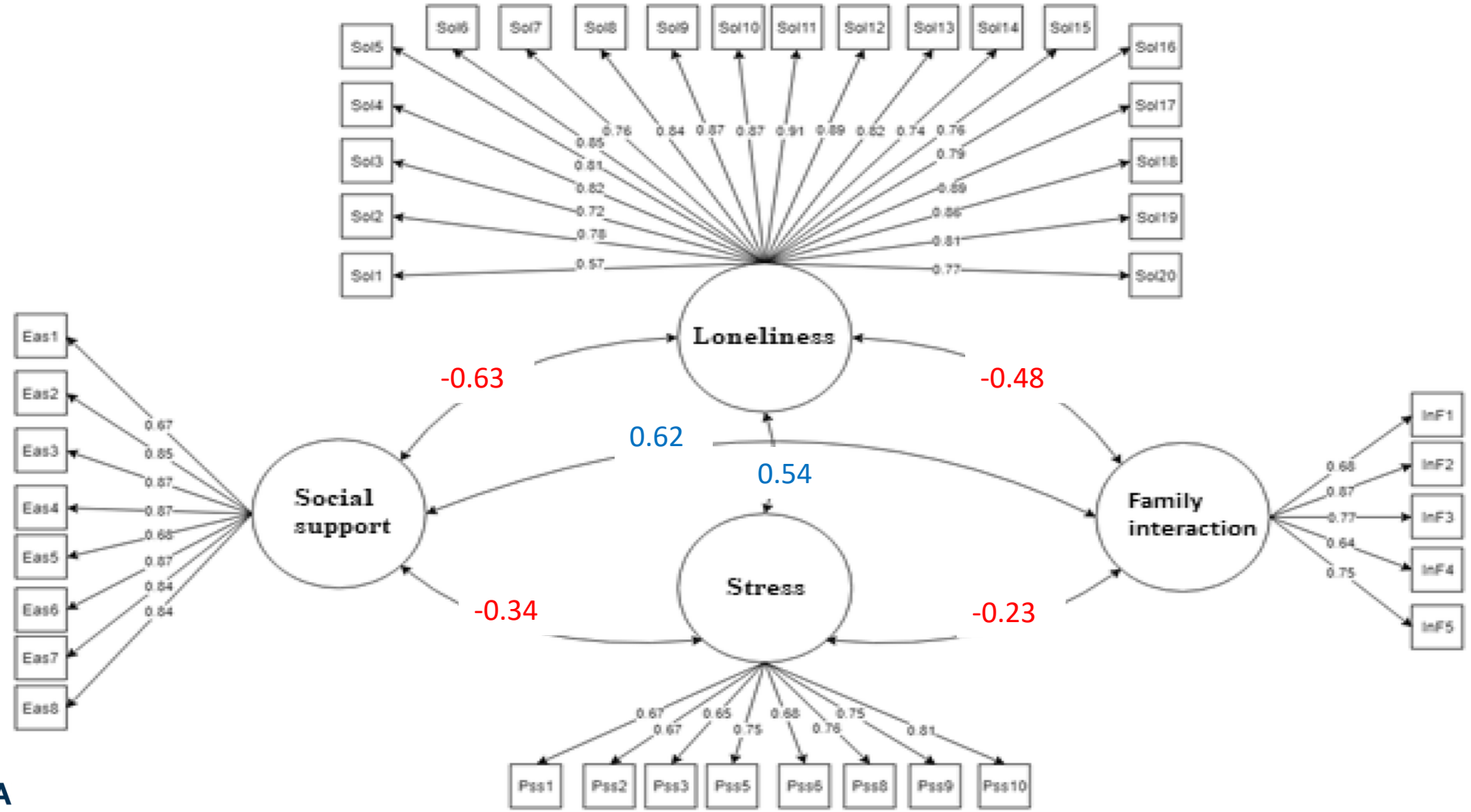
Checklist: Psychometric Analyses

- ❑ Test and item bias: estimate invariance and differential item functioning models to identify bias between groups.

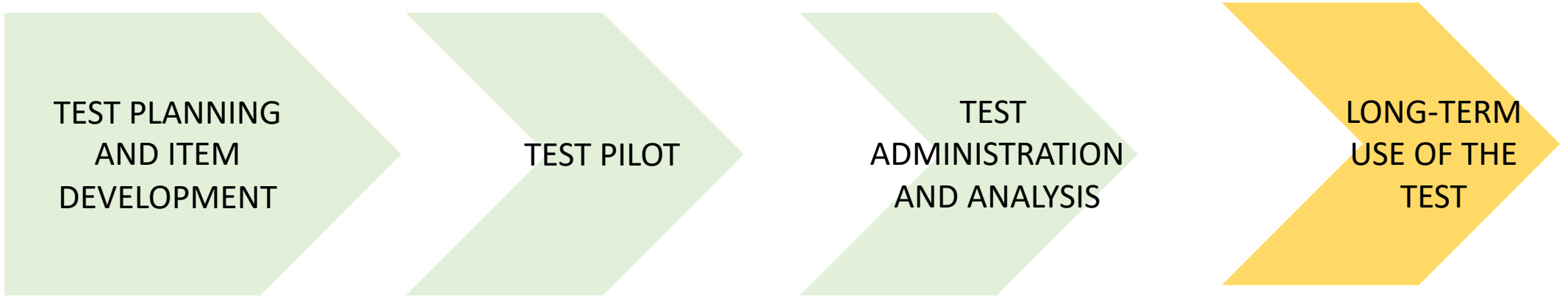
- ❑ Psychometric models to evaluate cognitive processes behind the response patterns: cognitive diagnostic models or multicomponent latent trait models.

- ❑ Relationship with external variables: correlation and regression analyses on the relationship between test scores and external variables that, according to psychological theory, should be related with the measured construct.

Example: Relationship of loneliness with external variables in a structural equation model



WORLD BA
Education



Long-term use of an instrument: intended and unintended consequences

Intended and unintended consequences

- Psychometricians, assessment specialists, and policy researchers should to analyze the intended and unintended consequences of the use a test.

- Ensuring that unintended consequences are evaluated is the responsibility of those making decisions about the use of a particular test (e.g., a national examination).

Checklist: Intended and unintended consequences

- ❑ Answer the next questions:
  - ❑Is the test **being used** for the objectives it was originally developed?
  - ❑Is the test **being used** for a purpose other than those indicated in its assessment framework?
  - ❑Are there any unintended consequences in the use of the test that were not predicted in the test development stage?

- ❑Relationship with external variables: correlation and regression analyses on the relationship between test scores and external variables over time.

- ❑Test and item bias: estimate invariance and differential item functioning models to identify bias between groups over time.

Example: Use of a teacher assessment for entry and promotion in a country's educational system

- **Country X** implemented a new teacher assessment policy four years ago. The test measures knowledge and instructional practices for elementary and middle school teachers. The test has been used to select new teachers in public schools from the pool of candidates.

- Over time, has the test been biased in favor of candidates coming from private universities over their peers from public universities?

- Since the implementation of this educational policy, has there been an unintended reduction in the net enrollment of students in teacher preparation programs?