EDUCATION & SKILLS

Considerations in Classroom Assessment Design

How to choose and design tools that gather data about where students are in their learning

Dr Nathanael Reinertsen
Directorate of Education and Skills

23/10/2024

PISA for Schools

OECD
BETTER POLICIES FOR BETTER LIVES

EDUCATION & SKILLS

1

Validity, Reliability, Objectivity, Inclusiveness and Fairness

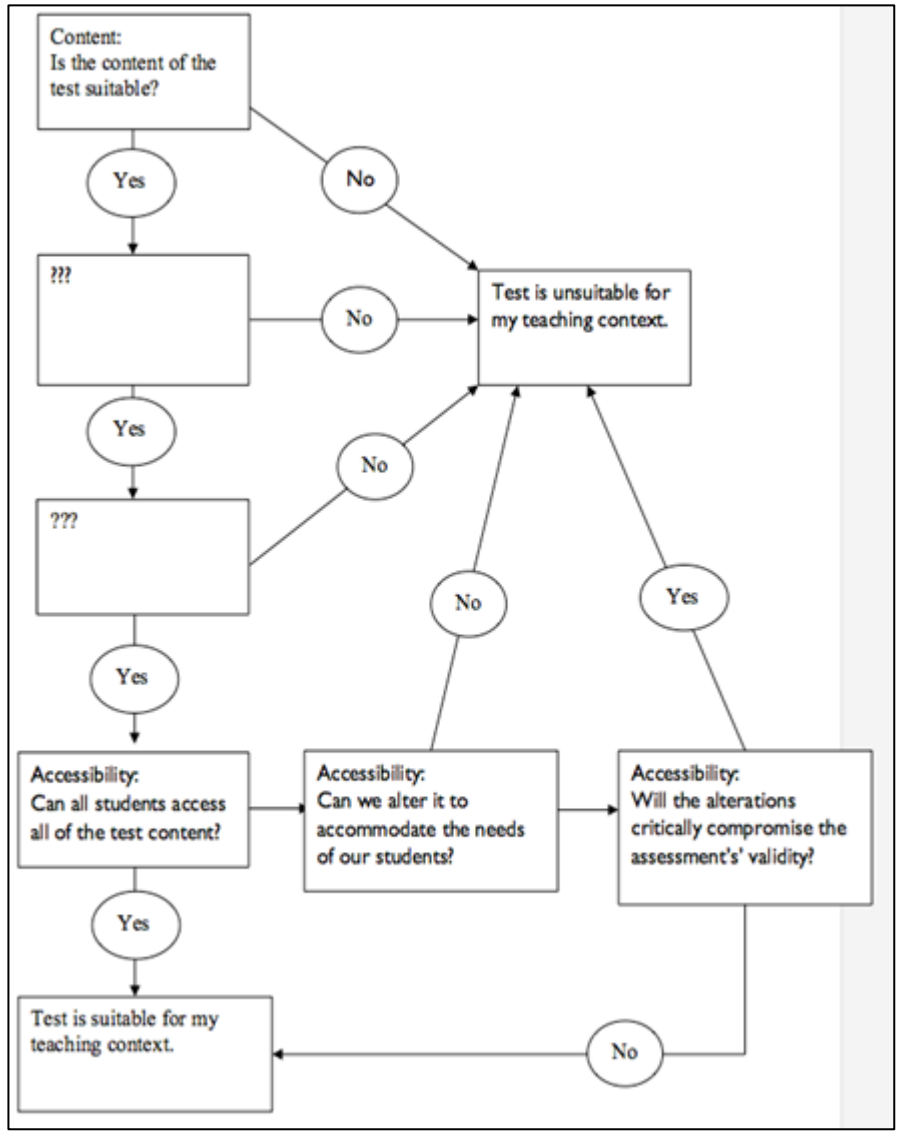
PISA for Schools

Activity/Discussion: What makes a ‘good’ assessment?

Sketch a decision tree or flowchart that represents *your* thought process when you decide whether an assessment is ‘good’ or not, for your students.

Compare and discuss in your small group.

After, we will ask each group to report back what their top 3 criteria are for deciding if a test is “good”.



Assessment Design Concepts

In planning (or evaluating) an assessment task there are several important aspects. These include:

- Validity
- Reliability
- Objectivity
- Fairness
- Inclusivity
- Feasibility

Validity

Are we testing what we intend to test?

Validity

Aspects of validity:

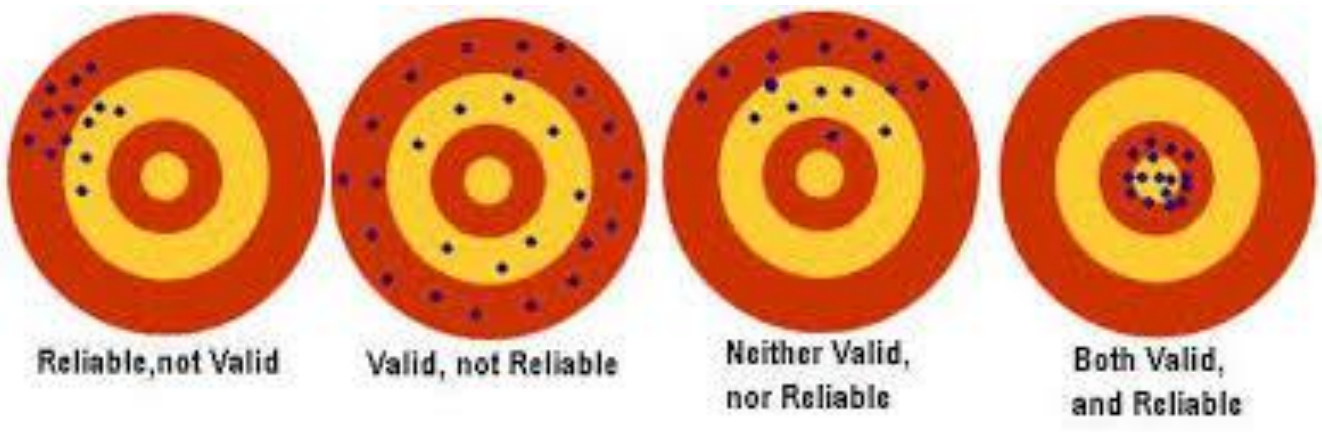
- Face validity
- Criterion validity
- Construct validity
- Content validity

Reliability

Is the assessment consistent when

- given to different people,
- marked by different markers,
- sat at different times, and
- administered in different ways or places?

Relationship between Reliability & Validity



Reliability is a necessary but not sufficient condition for validity.

Image from https://webcourses.ucf.edu/courses/1140056/pages/chapter-5-measurement-concepts

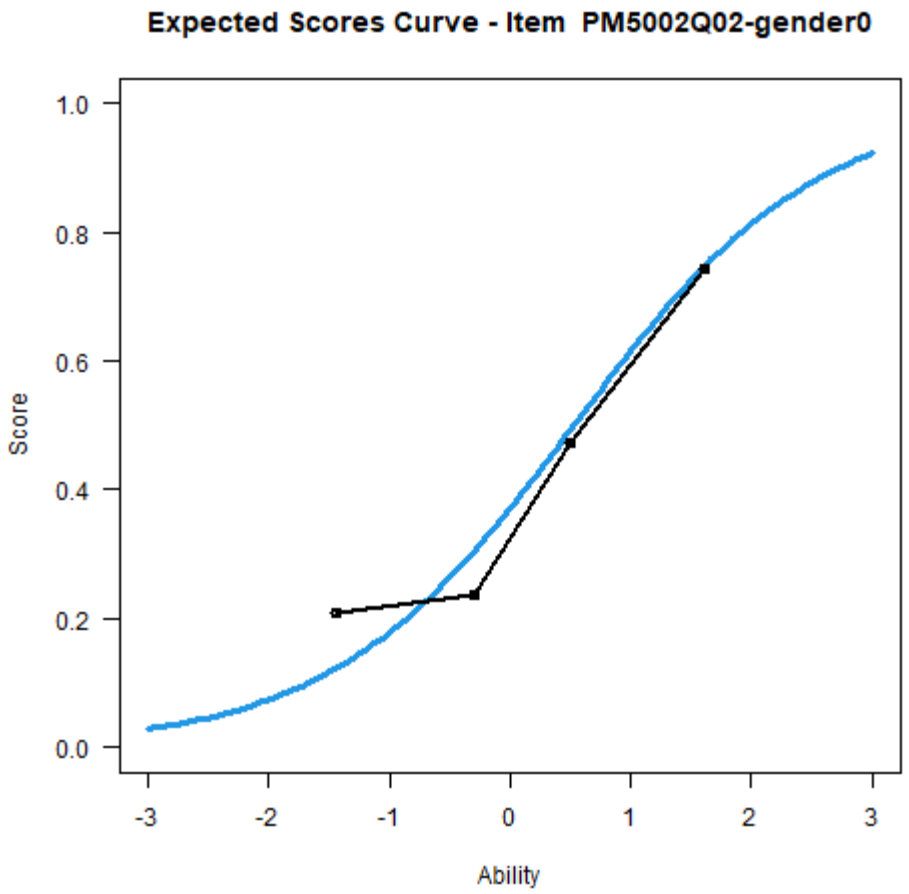
Objectivity

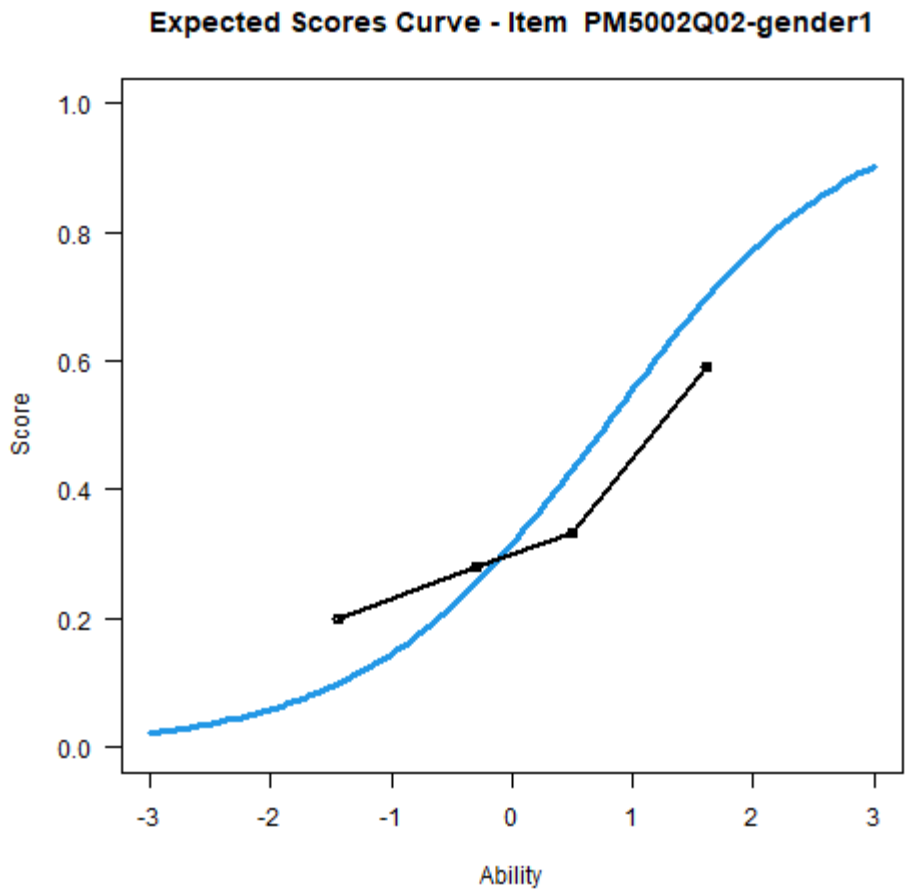
- A generally-held idea about objectivity in assessment is that it is the same as reliability: the results should not be influenced by ‘subjective’ judgments or choices.
- This is true: the results of measuring a pencil with a ruler shouldn’t be different because of who is using the ruler.
- But it is not the whole story, when it comes to educational assessment: sometimes, sub-groups of students have different response probabilities to an item

Fairness

Are there sub-groups of students who are advantaged or disadvantaged by the content or the nature of the task?

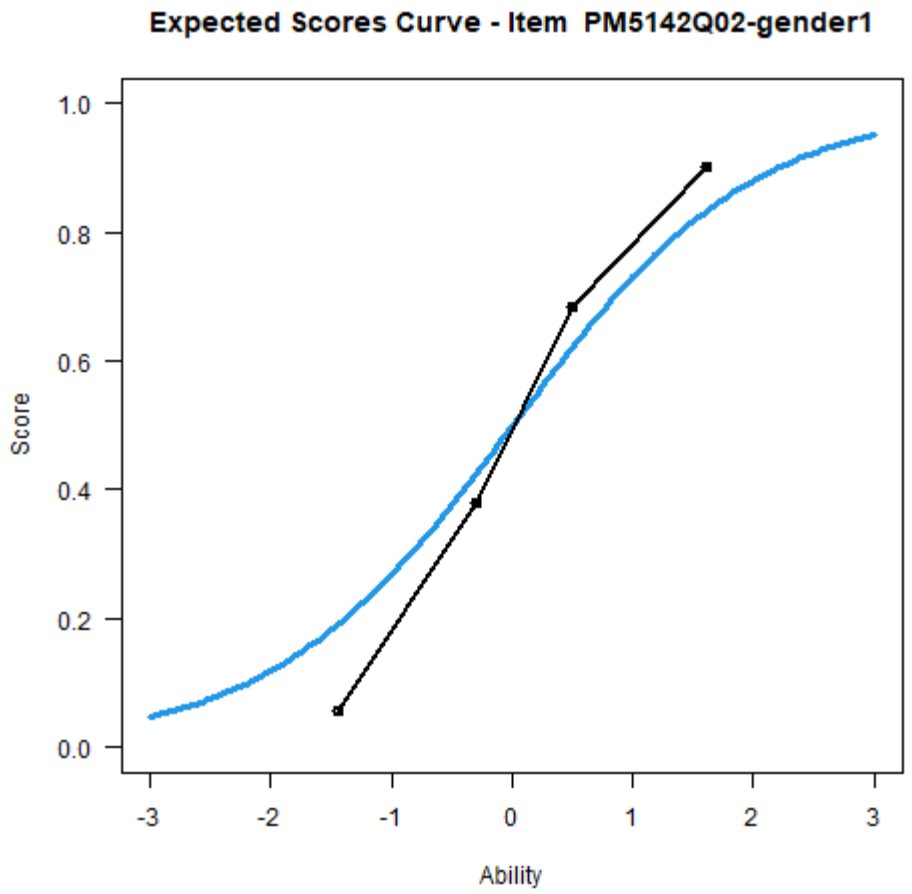
Examining DIF graphically





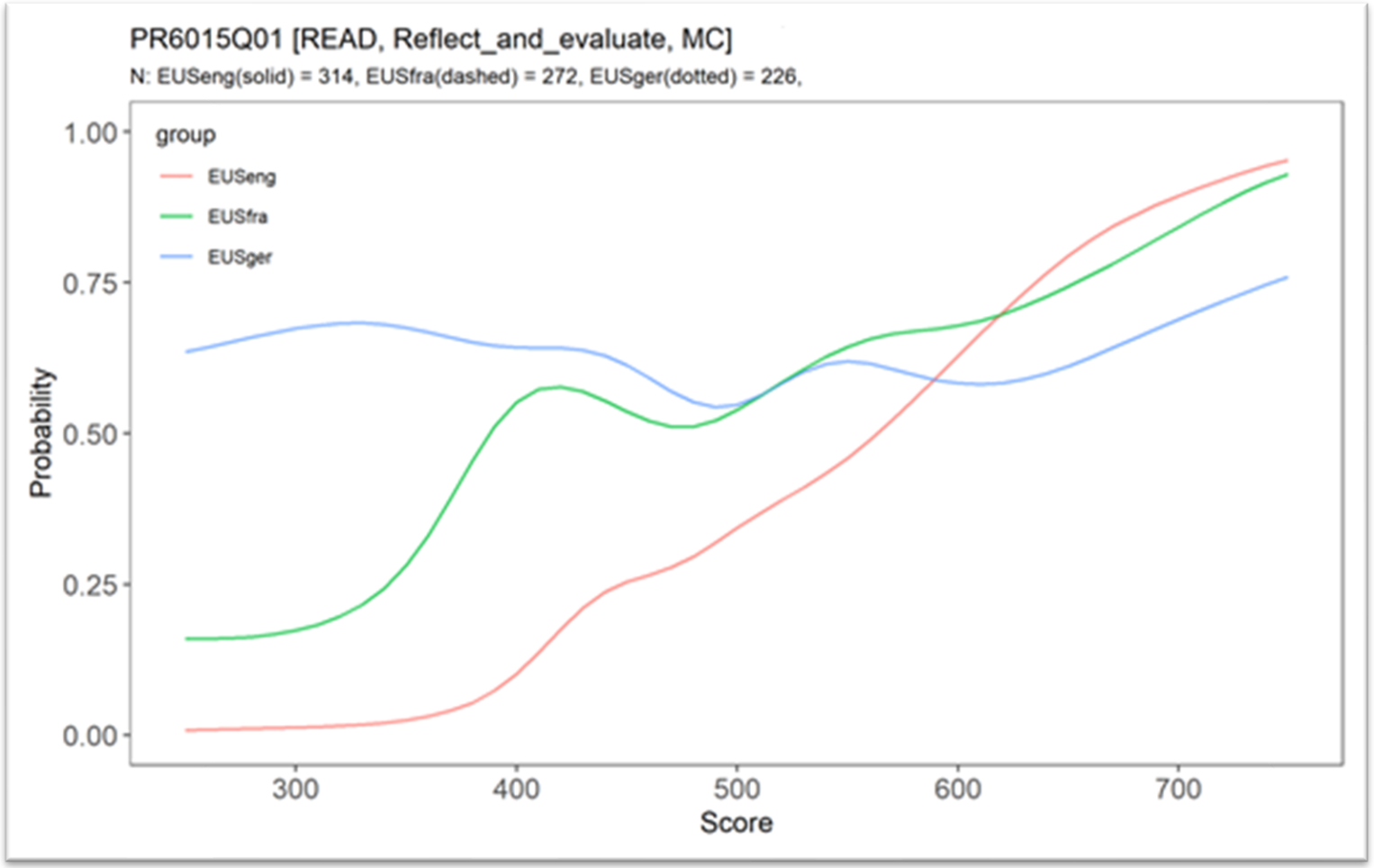
n.b. the magnitude of this DIF isn't enough to 'flag' the item

Examining DIF graphically



n.b. the magnitude of this DIF isn't enough to 'flag' the item

Low discrimination for one language



Feasibility

Does the assessment represent value for

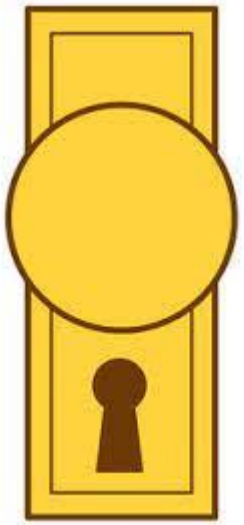
- time,
- money, and
- effort?

Inclusivity

To what extent is the assessment accessible to all test takers?

In the physical environment

Doorknobs are less accessible than **door handles**, which are less accessible than **automatic doors.**



This is true for **everyone**, not just people who have impaired mobility.

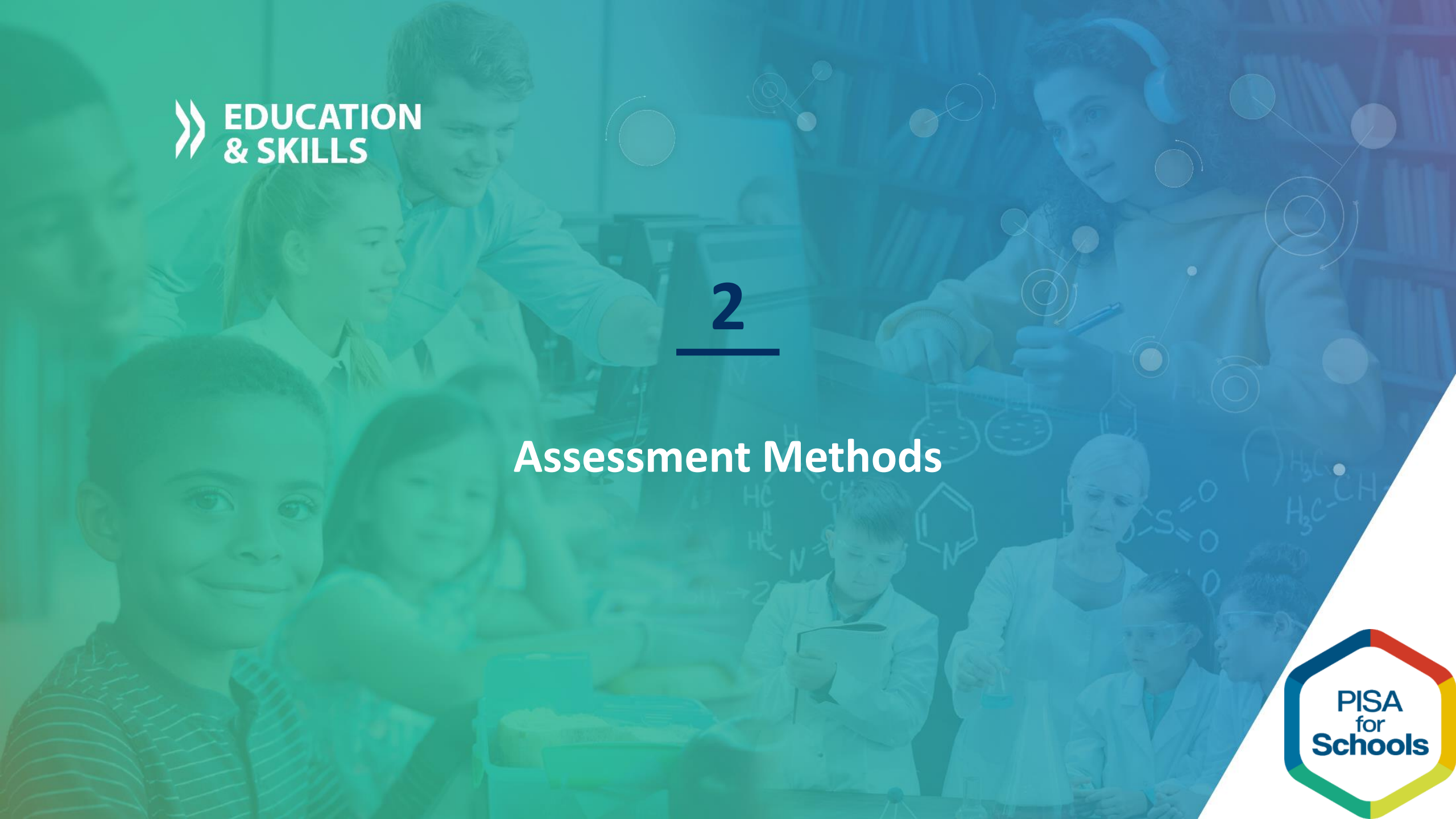
UD is not about ‘special’ treatment for some; it is about best design for all.

In assessment design

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

These are not special changes for a minority of students.

They are good design for *all* students.



EDUCATION & SKILLS

2

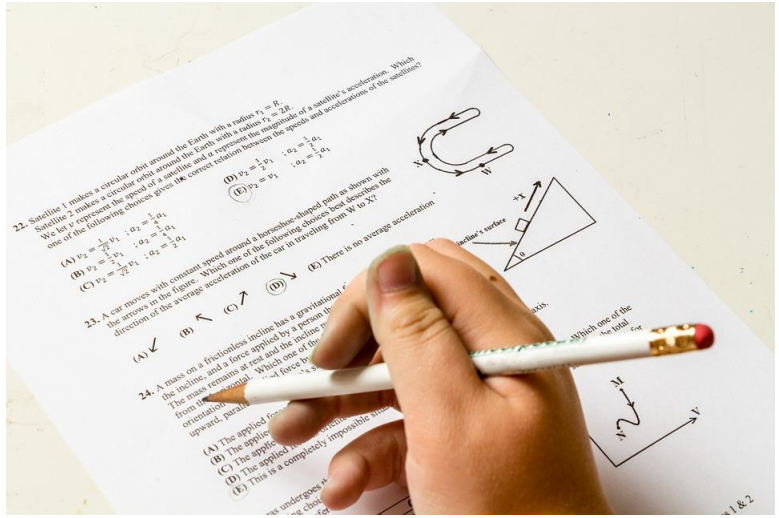
Assessment Methods

PISA for Schools

Small group discussion: Assessment methods

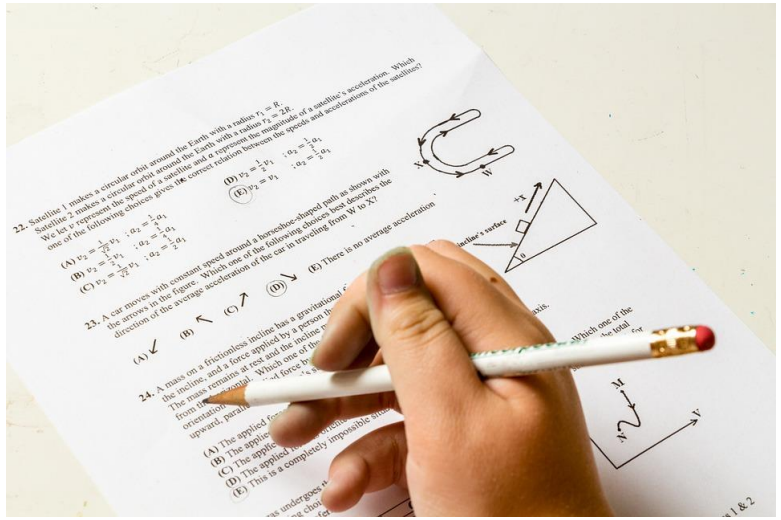
An ‘assessment method’ is what students *do* to demonstrate where they are in their learning.

Pen-and-paper



- What types are there?
- What are the advantages?
- What are the limitations?

Pen-and-paper



- Validity?
- Reliability
- Objectivity
- Inclusivity?
- Feasibility

Performance

- What types are there?
- What are the advantages?
- What are the limitations?

Performance

- **Validity**
- Reliability?
- Objectivity?
- Inclusivity?
- Feasibility?

Portfolio

- What types are there?
- What are the advantages?
- What are the limitations?

Portfolio

- Validity
- Reliability
- Objectivity
- Inclusivity
- Feasibility?

Technology enhanced assessments

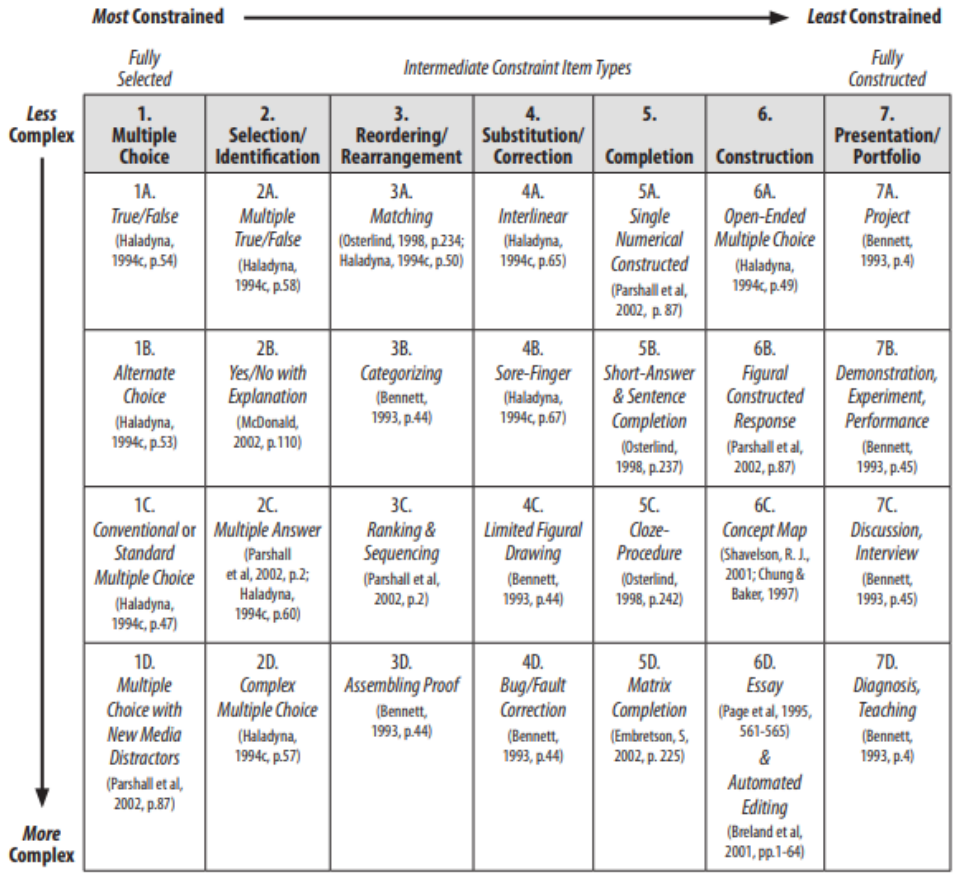
Most Constrained → *Least* Constrained

Less Complex ↓ More Complex	*Fully Selected*	*Intermediate Constraint Item Types*				*Fully Constructed*
	1. Multiple Choice	**2. Selection/ Identification**	**3. Reordering/ Rearrangement**	**4. Substitution/ Correction**	**5. Completion**	**6. Construction**	**7. Presentation/ Portfolio**
	1A. *True/False* (Haladyna, 1994c, p.54)	2A. *Multiple True/False* (Haladyna, 1994c, p.58)	3A. *Matching* (Osterlind, 1998, p.234; Haladyna, 1994c, p.50)	4A. *Interlinear* (Haladyna, 1994c, p.65)	5A. *Single Numerical Constructed* (Parshall et al, 2002, p. 87)	6A. *Open-Ended Multiple Choice* (Haladyna, 1994c, p.49)	7A. *Project* (Bennett, 1993, p.4)
	1B. *Alternate Choice* (Haladyna, 1994c, p.53)	2B. *Yes/No with Explanation* (McDonald, 2002, p.110)	3B. *Categorizing* (Bennett, 1993, p.44)	4B. *Sore-Finger* (Haladyna, 1994c, p.67)	5B. *Short-Answer & Sentence Completion* (Osterlind, 1998, p.237)	6B. *Figural Constructed Response* (Parshall et al, 2002, p.87)	7B. *Demonstration, Experiment, Performance* (Bennett, 1993, p.45)
	1C. *Conventional or Standard Multiple Choice* (Haladyna, 1994c, p.47)	2C. *Multiple Answer* (Parshall et al, 2002, p.2; Haladyna, 1994c, p.60)	3C. *Ranking & Sequencing* (Parshall et al, 2002, p.2)	4C. *Limited Figural Drawing* (Bennett, 1993, p.44)	5C. *Cloze-Procedure* (Osterlind, 1998, p.242)	6C. *Concept Map* (Shavelson, R. J., 2001; Chung & Baker, 1997)	7C. *Discussion, Interview* (Bennett, 1993, p.45)
	1D. *Multiple Choice with New Media Distractors* (Parshall et al, 2002, p.87)	2D. *Complex Multiple Choice* (Haladyna, 1994c, p.57)	3D. *Assembling Proof* (Bennett, 1993, p.44)	4D. *Bug/Fault Correction* (Bennett, 1993, p.44)	5D. *Matrix Completion* (Embretson, S, 2002, p. 225)	6D. *Essay* (Page et al, 1995, 561-565) & *Automated Editing* (Breland et al, 2001, pp.1-64)	7D. *Diagnosis, Teaching* (Bennett, 1993, p.4)

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment, 4*(6). Retrieved from http://files.eric.ed.gov/fulltext/EJ843857.pdf

PISA for Schools

Technology enhanced assessments



Most Constrained → *Least* Constrained

Less Complex	Fully Selected	Intermediate Constraint Item Types				Fully Constructed
	1. Multiple Choice	**2. Selection/ Identification**	**3. Reordering/ Rearrangement**	**4. Substitution/ Correction**	**5. Completion**	**6. Construction**	**7. Presentation/ Portfolio**
	1A. *True/False* (Haladyna, 1994c, p.54)	2A. *Multiple True/False* (Haladyna, 1994c, p.58)	3A. *Matching* (Osterlind, 1998, p.234; Haladyna, 1994c, p.50)	4A. *Interlinear* (Haladyna, 1994c, p.65)	5A. *Single Numerical Constructed* (Parshall et al, 2002, p. 87)	6A. *Open-Ended Multiple Choice* (Haladyna, 1994c, p.49)	7A. *Project* (Bennett, 1993, p.4)
	1B. *Alternate Choice* (Haladyna, 1994c, p.53)	2B. *Yes/No with Explanation* (McDonald, 2002, p.110)	3B. *Categorizing* (Bennett, 1993, p.44)	4B. *Sore-Finger* (Haladyna, 1994c, p.67)	5B. *Short-Answer & Sentence Completion* (Osterlind, 1998, p.237)	6B. *Figural Constructed Response* (Parshall et al, 2002, p.87)	7B. *Demonstration, Experiment, Performance* (Bennett, 1993, p.45)
	1C. *Conventional or Standard Multiple Choice* (Haladyna, 1994c, p.47)	2C. *Multiple Answer* (Parshall et al, 2002, p.2; Haladyna, 1994c, p.60)	3C. *Ranking & Sequencing* (Parshall et al, 2002, p.2)	4C. *Limited Figural Drawing* (Bennett, 1993, p.44)	5C. *Cloze-Procedure* (Osterlind, 1998, p.242)	6C. *Concept Map* (Shavelson, R. J., 2001; Chung & Baker, 1997)	7C. *Discussion, Interview* (Bennett, 1993, p.45)
More Complex	1D. *Multiple Choice with New Media Distractors* (Parshall et al, 2002, p.87)	2D. *Complex Multiple Choice* (Haladyna, 1994c, p.57)	3D. *Assembling Proof* (Bennett, 1993, p.44)	4D. *Bug/Fault Correction* (Bennett, 1993, p.44)	5D. *Matrix Completion* (Embretson, S, 2002, p. 225)	6D. *Essay* (Page et al, 1995, 561-565) & *Automated Editing* (Breland et al, 2001, pp.1-64)	7D. *Diagnosis, Teaching* (Bennett, 1993, p.4)

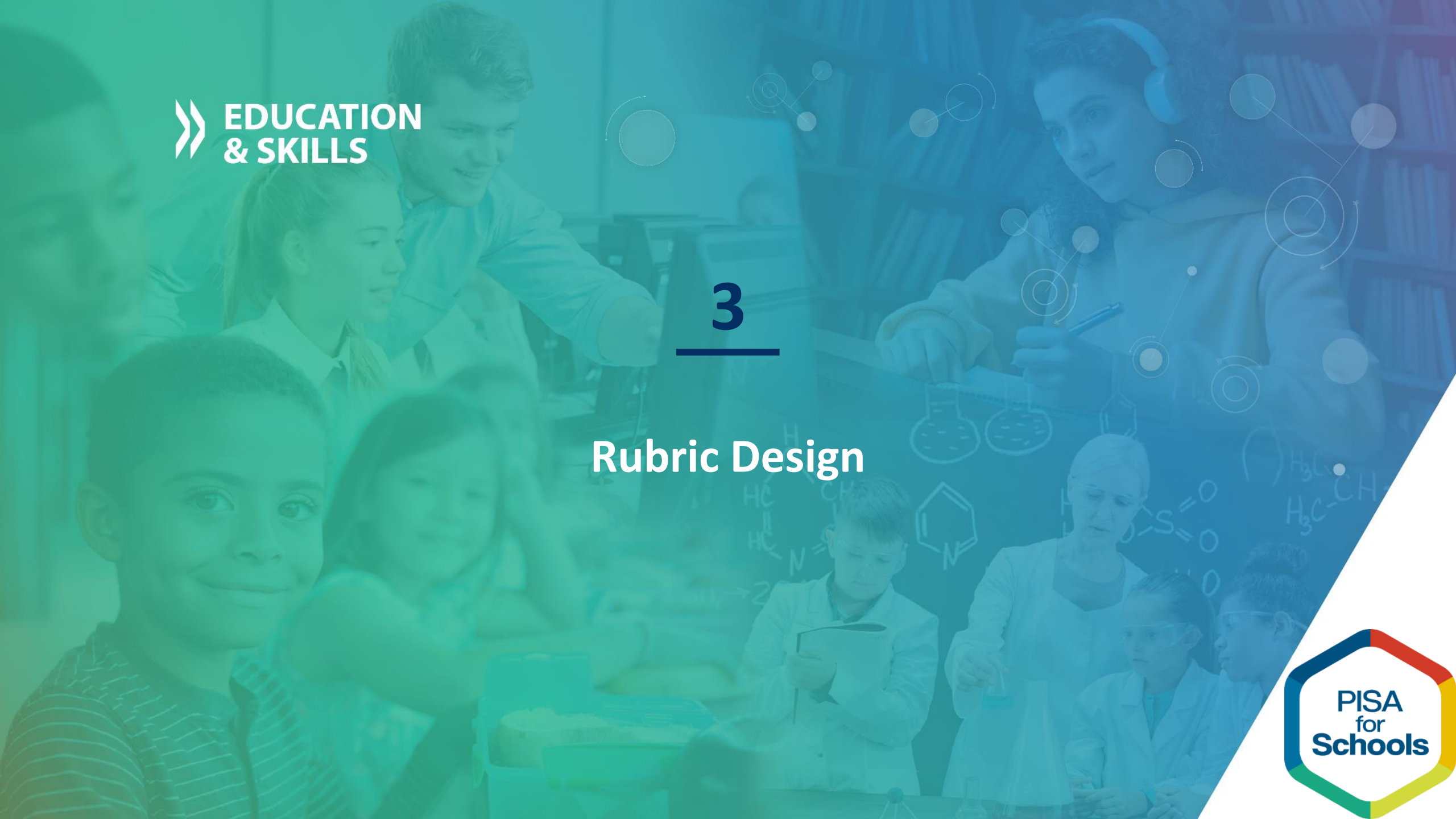
- Validity?
- Reliability?
- Objectivity?
- Inclusivity?
- Feasibility?

PISA for Schools

Methods Summary

- There are a rage of methods for collecting observations about what students know and can do.

- Selecting an assessment method should be a considered and deliberate choice, taking into account the contextual requirements for validity, reliability, objectivity, inclusivity and feasibility, in your school and classroom.



EDUCATION & SKILLS

3

Rubric Design

PISA for Schools

Defining a rubric

"…a coherent set of criteria for students' work that includes descriptions of levels of performance quality on the criteria."

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Alexandria, VA: ASCD

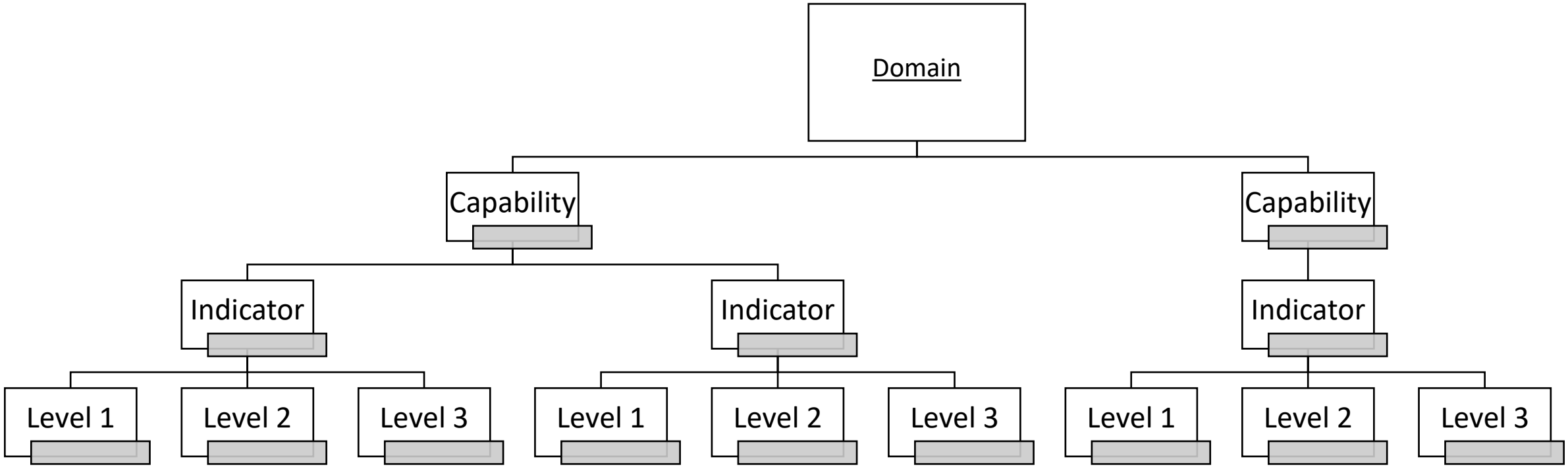
Writing a rubric

1. Articulate the domain(s) or sub-domains that are to be assessed
2. Select an appropriate developmental framework
3. Identify key capabilities
4. Break capabilities down into individual indicators
5. Separate the indicators into different levels of proficiency

Adapted from:

Griffin, P., & Robertson, P. (2014). Writing assessment rubrics. In P. Griffin (Ed.), *Assessment for teaching* (pp. 125-155). Port Melbourne, Australia: Cambridge University Press.

Rubric Hierarchy



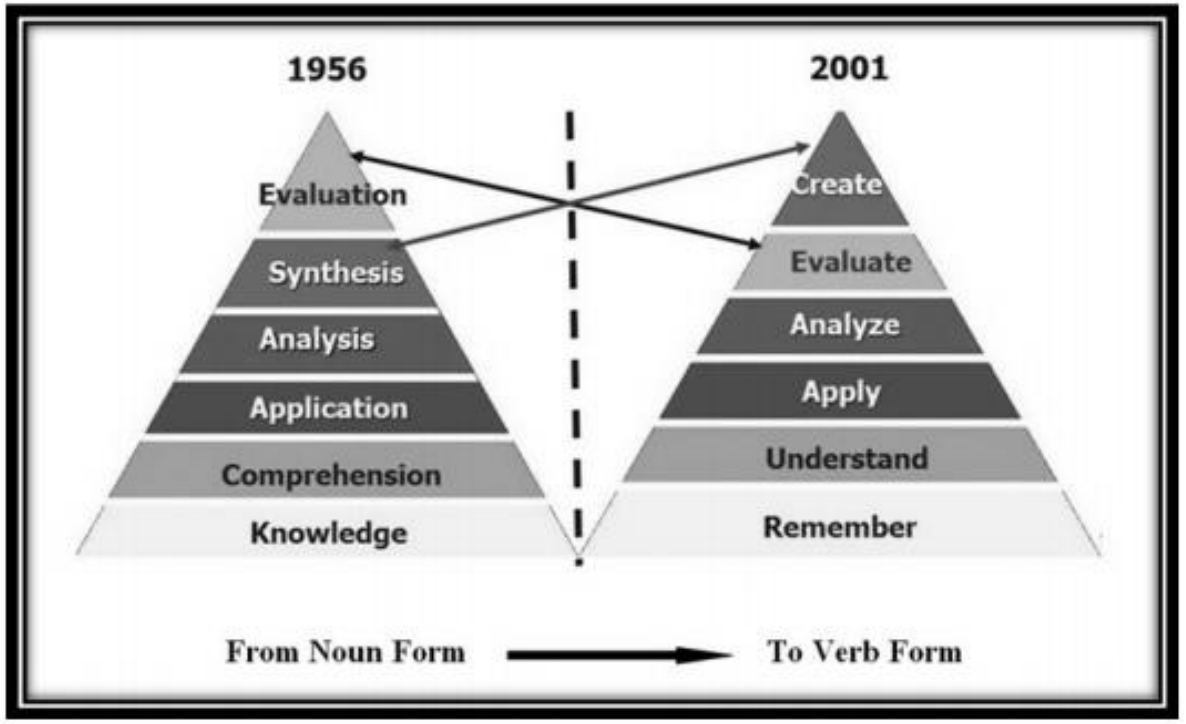
Articulating a domain

A domain can be defined by a curriculum, an assessment framework, by a department, or by a teacher.

What is essential is that the assessment designer knows *a priori* and in very specific detail what skills and knowledge they are trying to assess.

Selecting Developmental Frameworks

There are many to choose from, that you are likely to already know from your teacher training. For example, Bloom's (Revised) Taxonomy:



Anderson, L.W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* New York, NY: Longman.

Identify Key Capabilities

Based on professional experience, knowledge or judgment; or based on curriculum documents or assessment frameworks.

For example, a reading assessment in the ‘PISA-style’ would assess the following capabilities:

- Locating information
- Understanding
- Evaluating and reflecting on information and texts

Breaking Capabilities into Indicators

For a Reading assessment in the 'PISA-style':

1. Locating information
    1. Locates explicit information
    2. Links information across sentences/paragraphs
2. Understanding
    1. Interprets explicit information
    2. Identifies implied meanings
3. Evaluating and Reflecting
    1. Identifies fact/opinion
    2. Applies generalised concepts to different situations
    3. Reflects on the form of the text

Levels of Proficiency

Indicator	1	2	3
Reflects on the form of the text	Identifies authorial choices in vocabulary, syntax, layout, etc.	Identifies likely motivation/s for authorial choices in vocabulary, syntax, layout, etc.	Identifies, with clear justification, the most likely motivation for authorial choices in vocabulary, syntax, layout, etc.

For discussion

Should I or shouldn't I…

- use a different number of levels for different indicators?
- differentiate levels by using 'None', 'A few', 'Some' & 'Many'?
- use levels: 'incomplete', 'mostly complete' & 'complete' in a criteria for a project assessment?

Enhancing validity and reliability in rubrics

1. Be clear in the language used
2. Focus on a single central idea in each indicator
3. Limit yourself to four or fewer levels (not counting 0), using only as many as you can clearly define
4. Focus on progression from level to level
5. Focus on describing observable aspects of the task
6. Avoid procedural steps
7. Avoid counts or pseudo-counts
8. Resist the temptation to weight criteria

Adapted from:
Griffin, P., & Robertson, P. (2014). Writing assessment rubrics. In P. Griffin (Ed.), *Assessment for teaching* (pp. 125-155). Port Melbourne, Australia: Cambridge University Press.

PISA for Schools

Activity

In groups, consider and critique the rubric provided.

Improve it, if you can!

Bad rubrics

"Like good rubrics, bad rubrics will provide information, but they cannot be synthesised and interpreted with reference to a developmental model. They may not facilitate teaching decisions: they cannot identify the point of readiness to learn. They are often written in arbitrary language that leads to multiple interpretations, which in turn necessitates a costly process of assessment moderation, with double or triple judgments. The offer little hope of self-assessment and they provide almost no transparency of assessment, so that candidates are led into a situation in which they must play the game of 'guess what the teacher wants'. This is unethical practice, and cannot lead to valid assessment interpretation. When examining rubrics provided by other people, the rules or meta-rubrics should be used to evaluate them. Unless a rubric meets these guidelines, it should not be used.

Griffin, P., & Robertson, P. (2014). Writing assessment rubrics. In P. Griffin (Ed.), *Assessment for teaching* (pp. 125-155). Port Melbourne, Australia: Cambridge University Press.