Consequential validity, transparency and accountability: Value and use of large-scale assessment studies

RESEP conference on quantitative education research
5 & 6 September 2023

Anil Kanjee
anil.kanjee@gmail.com
KanjeeA@tut.ac.za

Tshwane University of Technology
We empower people

Tshwane University of Technology
We empower people

Purpose

- Share some ideas/thoughts with a view of getting comments, critique, ideas/suggestions regarding
  - how to develop/promote more effective use of assessment evidence for improving **learning for ALL learners**

- Unique opportunity GIVEN current audience

Approach

- Raise 4 issues for consideration
- Am aware – this is a Quantitative Conference
- Not going to present findings, or results or sample sizes
- But am going to present my OWN RCT

- RANDOMISED CONTROL THOUGHTS

Context

for this presentation

Background/Context

1. Measurement/performance driven system – limited focus on learning
2. To date – 106 LSAS in post-apartheid era –Greater use
3. More critical of ILAS results
4. Enhance use of our data
   - Heading a TAG for DBE
   - Capacity Development Programme for key National, provincial and district level decision makers on Assessment and Data literacy & decision making

What lies behind South Africa's improvements in PIRLS?

An Oaxaca-Blinder analysis of the 2011 and 2016 data

MARTIN GUSTAFSSON AND STEPHEN TAYLOR

FEBRUARY 2022

Tshwane University of Technology

AfL approach

- Emphasis – use of assessment evidence to
  - Improve learning (and teaching) – **QUALITY**
  - For ALL - **EQUITY**

Consequential Validity

- Messick (1995) defined consequential validity to be "evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term."

Trigger - Release of PIRLS results in April 2022

PIRLS 2021

- Controversy / Delays in release of results
- DBE set up an International TAG to:
  - Advise on implications of SA PIRLS results
  - Provide support for developing capacity of officials to enhance the use LSA results – Systemic, ELNA, PIRLS, TIMSS

ALL information reported here is publicly available

Key issues to be consider - 2021

1. Differential impact of COVID
2. Possible floor effects
3. Information to replicate results at national levels
4. Information on how DIF addressed

Other issues - not addressed

1. Translation
2. Content and face validity - e.g. Octopus reading passage
3. Common items across years

Differential Impact of COVID

Summary of data collection period for participating countries

As scheduled (5 year trend)		Delayed - (6 months)	Delayed - One year	
Oct-Nov 2021	Feb - July 2021	Sep - Dec 2022	Aug-Dec 2021	April-July 2022
2 countries	38	17 countries	4	3
		4th Grade Cohort - Beginning of 5th Grade	Southern Hemisphere	Northern Hemisphere

- Exhibit 5: PIRLS 2021 Countries by Chronological Order of Data Collection
- Included benchmarking data

Possible floor effects

PIRLS Report - 2021

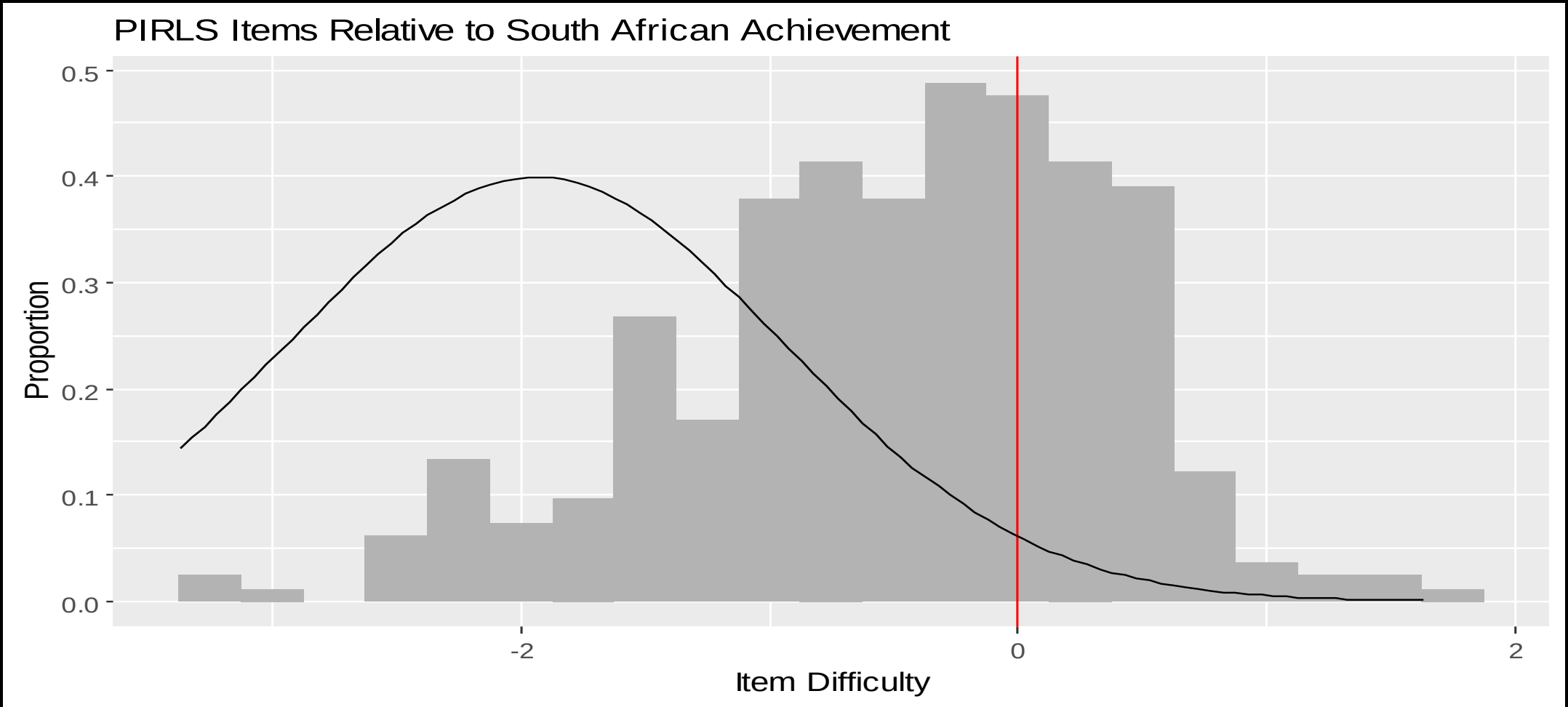
1. South Africa (+ 6 other countries) – data collected a year later than originally planned (i.e. in Sep 2022) which impact of COVID was "*greater????*"
2. *FOOTNOTE*
    1. Ж Reservations about reliability because the percentage of students with achievement too low for estimation exceeds 25%.
    2. South Africa continued investigating its PIRLS 2021 results at the time of publication and will deal with the findings through its national report.

Tshwane University of Technology
We empower people

Some questions to address

- What is the impact if GREATER 25% of learner response could NOT be estimated?
  - Results
  - Comparisons
  - trends
- How differential impact of COVID accounted for?
- How others issues noted addressed?

PIRLS 2021 item difficulty and achievement distribution – SA



- Ruthowski, May 2023, DBE symposium on

Tshwane University of Technology

Implications

- Rutkowski concludes
- A typical South African learner has a very low chance of correctly answering a typical PIRLS item (less than 10% probability).
- In other words, most South African learners are measured by few or no items.

Transparency -

Provide information to allow for replication of results

Calculation of Plausible Values

Overview – Matrix Sampling

Learner	Item number														
	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15
A	✗	✗	✗	✗	✗						✗	✗	✗	✗	
B	✗	✗	✗	✗	✗						✗	✗	✗	✗	✗
C	✗	✗	✗	✗	✗						✗	✗	✗	✗	✗
D	✗	✗	✗	✗	✗						✗	✗	✗	✗	✗
E	✗	✗	✗	✗	✗						✗	✗	✗	✗	✗
F	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗					
G	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗					
H	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗					
I	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗					
J	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗					
K						✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
L						✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
M						✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
N						✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
O						✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Plausible Values Generated

ASRREA01	ASRREA02	ASRREA03	ASRREA04	ASRREA05
377.189630	433.283420	444.866680	457.192330	407.451010
410.288410	487.721470	433.048080	453.352990	447.625680
483.695440	471.590330	485.254930	508.804950	454.784040
624.711100	612.569210	621.475620	590.886220	609.630980
573.275140	484.765690	541.731670	573.371500	512.403880
589.648820	571.448650	560.612300	571.499970	583.239580
588.871690	561.428600	572.827360	584.552540	522.992650
556.704940	527.502320	608.608220	573.528430	513.620550
528.897620	510.696600	543.798710	555.683210	442.280730
550.907060	533.822580	509.571960	516.800060	511.456870
464.071960	470.611580	480.103850	482.172750	443.339440
518.679150	510.910920	504.704050	435.685750	474.219860
574.335510	526.036180	571.177250	565.236800	589.064140
462.450310	542.152120	519.522010	536.603030	532.244540
415.562670	461.968520	483.779220	472.518680	441.432900
554.775130	541.054140	603.847620	536.281150	572.233760
618.538600	578.171780	671.375390	617.880230	559.660810

When achievement scores are used, the analyses are performed five times (once for each plausible value) and the results are aggregated to produce accurate estimates of achievement and standard errors that incorporate both sampling and imputation errors.

(Fishbein, B., Yin, L., & Foy, P. (2023). *PIRLS 2021 User Guide for the International Database*. Boston College, TIMSS & PIRLS International Study Center. https://pirls2021.org/data)

Tshwane University of Technology

Estimating learner scores - Plausible Values

1. Model Estimation
2. Compute the Proficiency Distribution
3. Consider Background Variables
4. Principal Component Analysis (PCA)
5. Create Conditional Proficiency Distributions:
6. Draw Plausible Values: Analysis

This information must be reported to allow for replication of results – Currently NOT AVAILABLE

Calculation of DIF

Overview of DIF

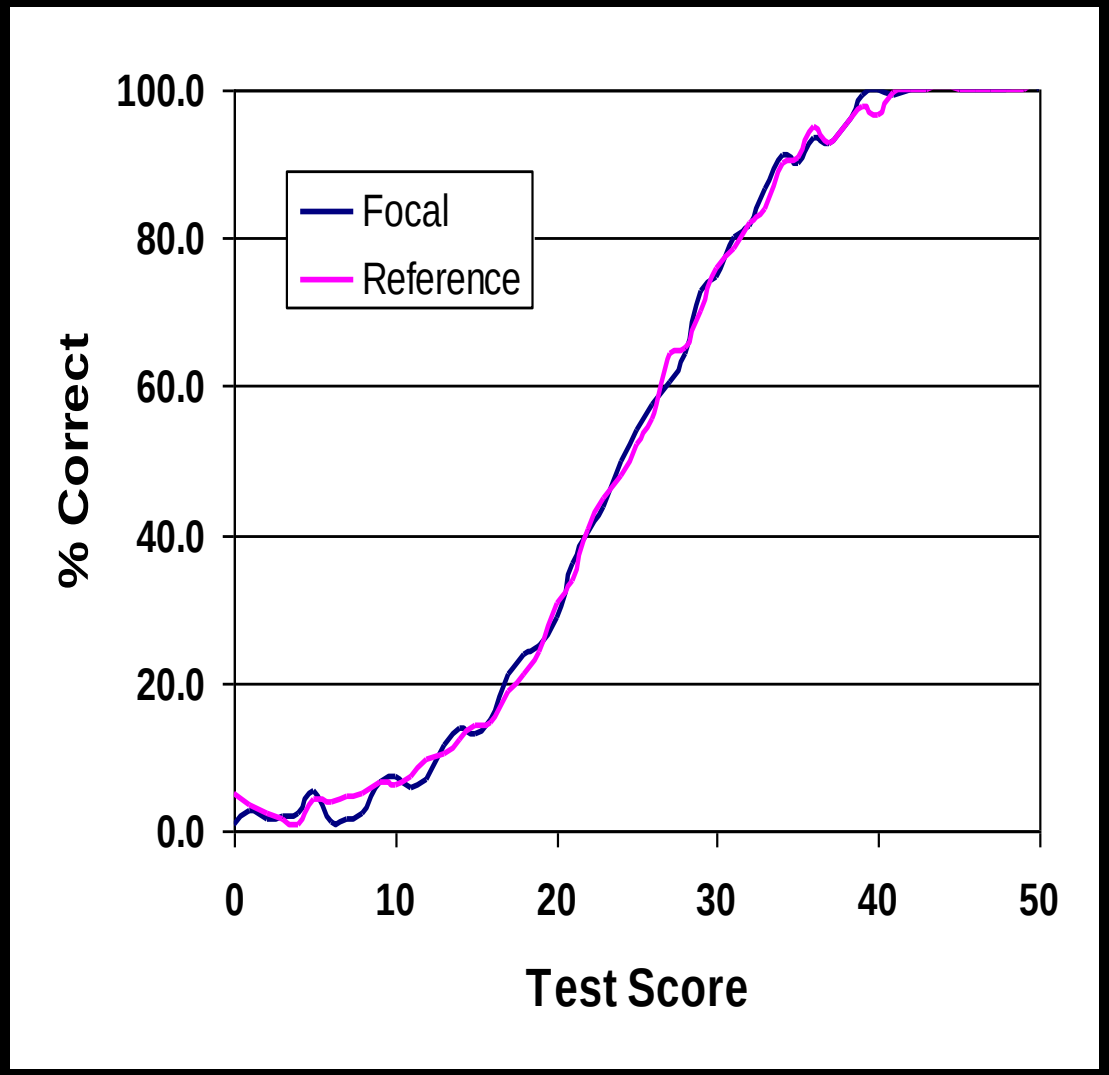
- Estimation models used - assumption items are equivalent across the measured populations

- In contrast, an item is said to suffer from differential item functioning (DIF), if for two examinees of identical proficiency, the probability of a correct answer is <u>NOT</u> the same.

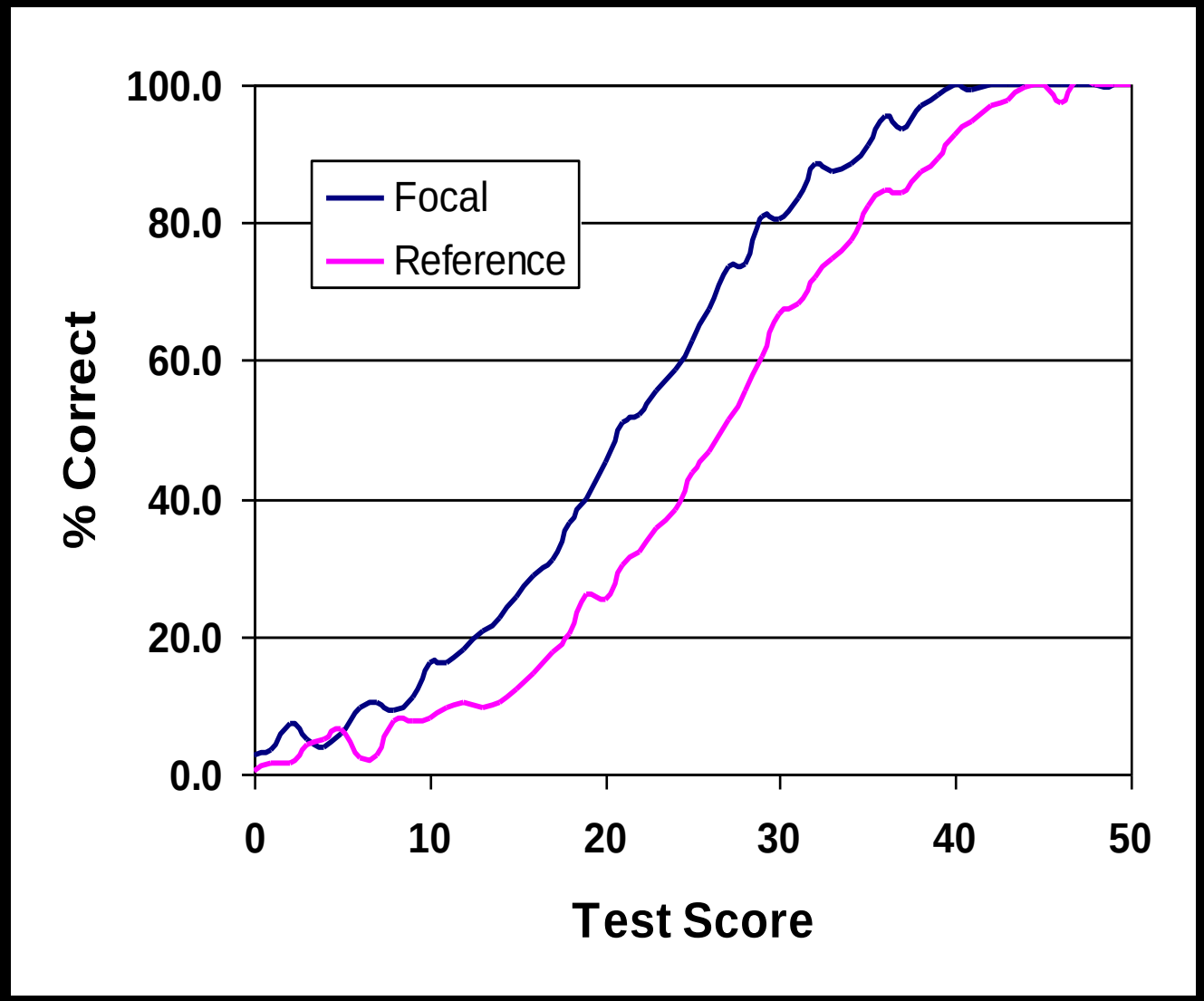
Overview of DIF

- if an item seems harder (or easier) for a group of examinees, we would wrongly infer that those examinees do not (or do) know the content associated with that item.
- A consequence is that their score on that item would be lower (higher) than it should be.
- If DIF is limited to a single item, its impact is limited.
- When DIF exists for many items, it can have a substantial biasing effect on achievement estimates

no DIF Item

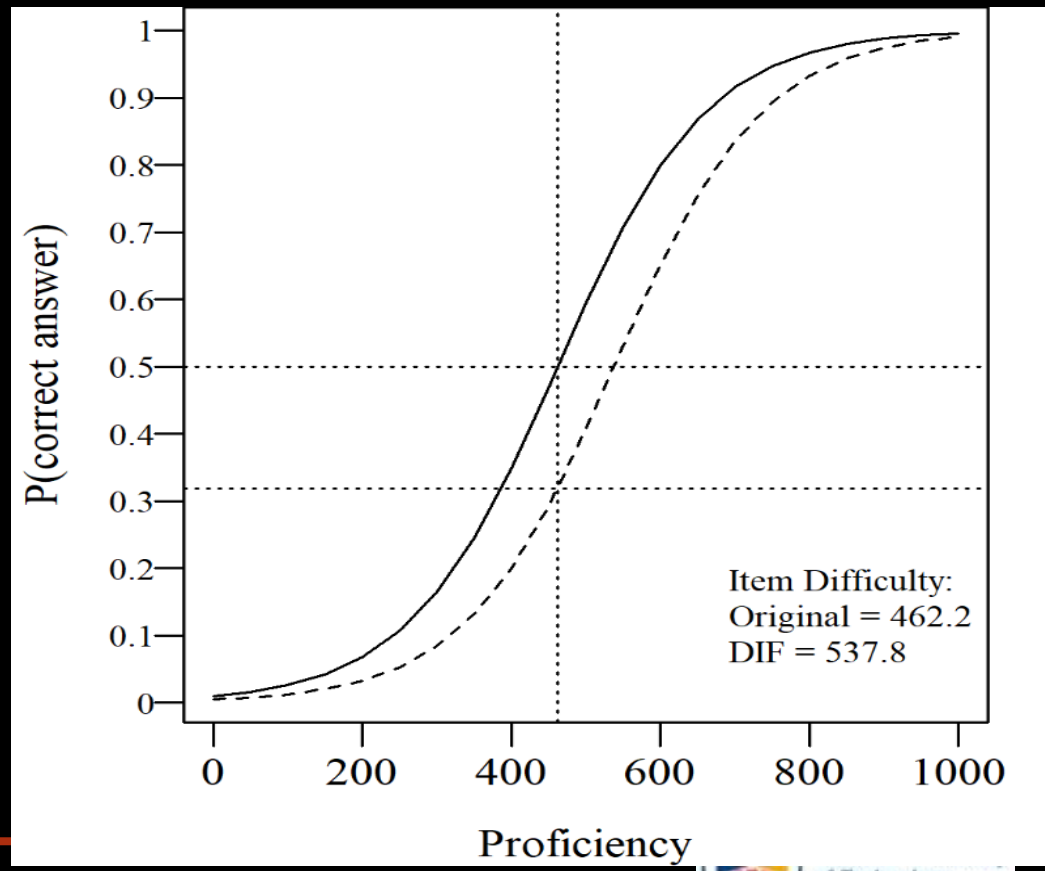


DIF Item



Measuring DIF - Differential item functioning

- DIF –Revised/remove items – Not feasible for ILAS
- Treat DIF items as fixed – not feasible that means “make impact of DIF the same of all countries
- OR Freely estimate items –i.e. do account for country specific DIF – procedure used in PIRLS (TIMSS & PISA)



Issues to consider/address

- Critical for any country to identify and address DIF as it impact results
- Large number of DIF results impact on reliability and validity of results
- How DIF items treated may also impact on fit of IRT model
- In South Africa, additional complexity of 11 languages

Way forward

So what does this mean?

-

- Messick (1995) defined consequential validity to be "evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term."

Implications???

- Additional analysis of the PIRLS results – in South Africa –meaning and implications?
- Call for more detail technical information to be reported to allow for countries to replicate results
- Extend similar analysis and interrogation to other national LSAS
- Need to enhance understanding - the value of the data and its effective use

Questions ?
Suggestions !
Comments
Ideas!

anil.kanjee@gmail.com
KanjeeA@tut.ac.za