# National Examinations in Secondary Schools in Ethiopia: Validity and Achievement Disparity

By

Tamirie Andualem Adal (Presenter)

**41ST African Educational Assessment Association (AEAA): Transforming Educational Assessment: Towards Quality learning and Informed decision making**

August 26, 2025

This paper is part of 10th Round Addis Ababa University Thematic Research entitled:

**Classroom Assessment, National Examinations and Academic Cheating in Secondary Schools in Ethiopia: Practices, Challenges and Interventions**

**By**

Research Theme: Kassahun H. (PI), Tamirie A., Arega M., Abera T., Abebaw M., Daniel T., Mulat A., Seleshi Z., & Yekoyealem D.

# The Main Issue

**National exam result as national shock: Pass rate, 3 – 5% in the past three years**

**Standardized testing as cost effective, while assessment as more important, but expensive to educational outcome**

**The test/exam being valid, but with the result being disparate across sex, region,…**

# 1. Background

- National Examination in Ethiopia is a standardized educational tool carried out throughout the country for the purpose of certifying student for their completion of secondary school education and admission to higher education.

- The National Examination dated back to **1946,** given at the completion of grade six; for grade 12, in **1954**, named as School Leaving Certificate Examination (ESLCE)

# Background (Cont’d)

- የኢትዮጵያ የትምህርትና ሥልጠና ፍልስፍና ከአገራዊ ፍላጎትና ተጨባጭ ሁኔታ የሚመነጭ እንደሁኔታውና እንደአግባቡ ባለብዙ ዘርፍ የትምህርትና ሥልጠና ፍልስፍና (Eclecic) የሚከተል ሲሆን ዋና ግብ ያደረገውም በመልካም ሥነምግባር የታነፁ፣ በራስ መተማመናቸው የዳበረ፣ ቴክኖሎጂን የሚጠቀሙ፣ ለራሳቸውና ለአገራቸው ብልፅግና የሚተጉ በሁለንተናዊ መልኩ የዘመኑና ችግር ፈቺ የሆኑ ኢትዮጵያዊ ዜጎችን መፍጠር ነው። (MEd, 2023)

= The New Education and Training Policy based on eclectic educational philosophy and national need and practical reality aims at producing Ethiopian citizens with productive and ethical values, self-reliance, use of technology & problem solving skills (my translation).

# Background (Cont’d)

- Though the above national goal is not directly assessed by standardized test at different levels, learning outcomes (e.g., understanding and application of concepts) assessed, partly, through standardized test are necessary for the attainment of such grand goals.

# Validity

Unlike reliability, validity is complex and evolving in time.

Reliability is a statistical component that supports validity, which is sometimes called a “reliability is a necessary, but not sufficient condition for validity”

Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. (AERA, APA, NCME, 2014, p. 11).

# Validity …

Five sources of evidence for validity:

- Content,
- Response process,
- Structure (internal relationship, including reliability),
- Relationship (external) with other variables,
- Consequences and fairness.

Content related evidence is easily and widely used educational testing and assessment through table of specification. Identification of experts in assessment of the quality of the items should be carefully made

# Reliability

In classical test theory, the consistency of test scores is evaluated mainly in terms of reliability coefficients, defined in terms of the correlation between scores derived from replications of the testing procedure on a sample of test takers (Standards, 2014, p37).

Practically reliability means the degree to which individuals' deviation scores, or z-scores, remain relatively consistent over repeated administration of the saine test or alternate test forms.

# Reliability ….

- There are four forms of estimating reliability coefficient:

1) test-retest method,

2) alternative forms method,

3) internal consistency methods including Cronbach α and Split-Half, and

4) interrater reliability

Cronbach α is widely used, and is affected by the number of items and the similarity (redundancy) of items

Low reliability coefficient (e.g., Cronbach alpha) doesn't necessarily mean poor test or assessment

# Objectives of this study

- Validity: To determine psychometric qualities the national examinations held in the AY of 2014 and 2015.

- Disparity: To explore difference in the score of students based on their gender, school type, area of residence, and subjects

# Method

- Design: This study employed a cross-sectional descriptive design which is based primarily on the two-year score of students taken from the office of  Educational Assessment and Examinations Service (EAES), Addis Ababa University record office, and secondary school teachers.
- Participants: The score of a total of 1,741,619 students in the national examination (i.e., 896,520 students in 20014 AY & 845,099 students in 2015 AY) was obtained from EAES.

- 330 teachers (95 females and 235 males) have participated in the study by providing self-report data on administration and related issues of Grade-12 examination.

# Method --- Subject Scores

- The score of a total of 1,741,619 students in the national examination (i.e., 896,520 students in 20014 AY & 845,099 students in 2015 AY) of all natural science and social science subjects was obtained from EAES

# Method … Questionnaire as Tool

This tool was developed to assess teachers' perceptions of how the national examination is scored and administered. To achieve this, the researchers developed eight close ended items to be rated from "not a problem", "minor problem", "moderate problem", and "serious problem."

Following an exploratory factor analysis, two dimensions that could be named as "testee-related" problem and "test-administration-and-scoring problem" were extracted. Cronbach alpha reliability for each of the dimensions were carried out and found to be 0.93 (for the three items of testee-related problems), and 0.90 (for five items of the "test-administration-and-scoring problems).

# Method …. Data analysis

- Cronbach alpha reliability
- Validity analysis: Analysis of convergent and discriminant validity, predictive validity
- Comparison of groups using t-test ,and ANOVA

# Result

- Reliability as necessary for Validity Cronbach

Reliability ranges from 0.49 (Math SS) to 0.92 (English for NS)

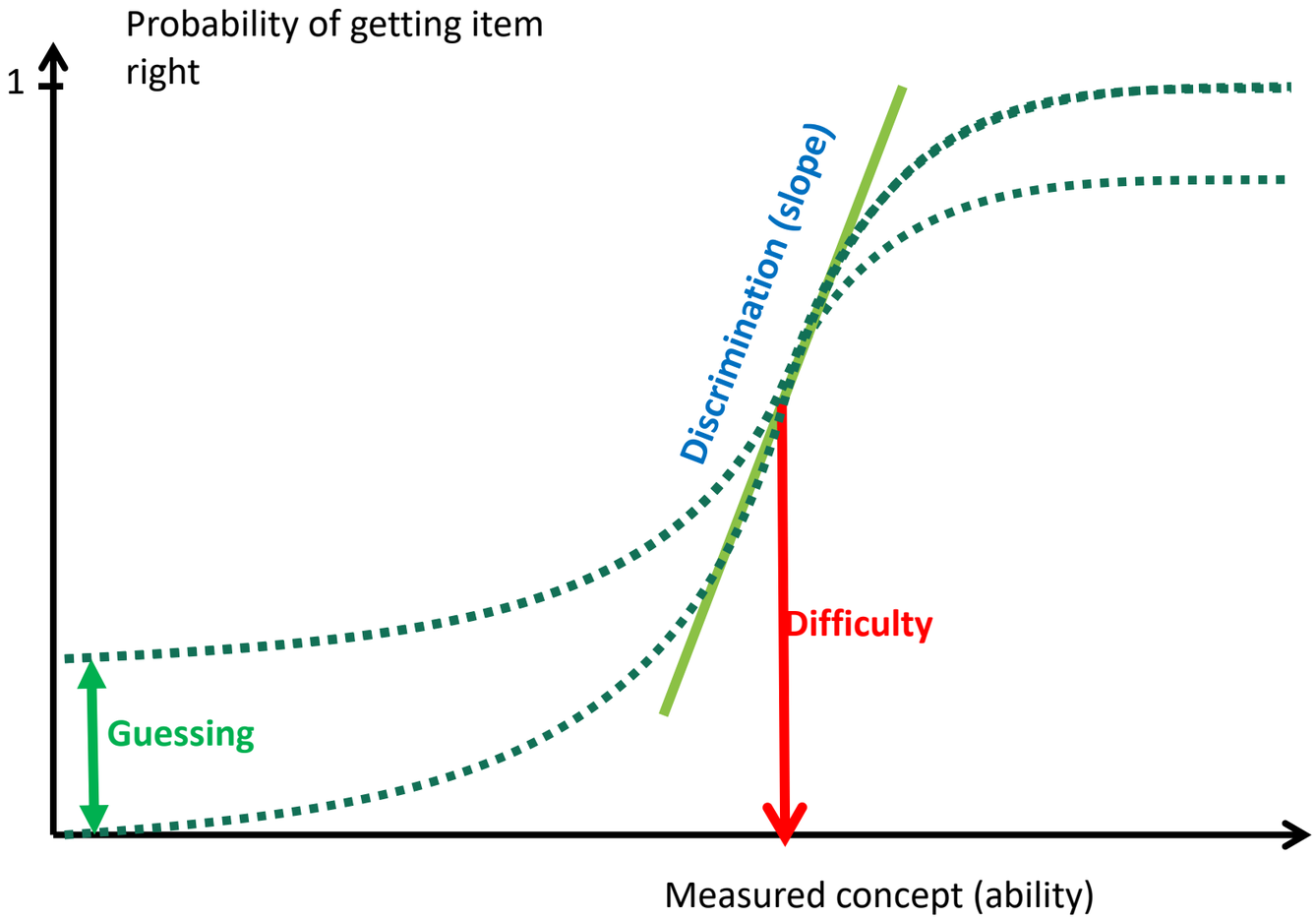
# Reliability of the EUEEs administered (2023) as indicated by Cronbach's alpha: 0.78 as median reliability

| Exam Types | Number of Items | Reliability |
|---|---|---|
| English for NS | 120 | 0.92 |
| English for SS | 120 | 0.66 |
| Math for NS | 65 | 0.86 |
| Math for SS | 65 | 0.49 |
| SAT for NS | 60 | 0.84 |
| SAT for SS | 60 | 0.59 |
| Physics | 50 | 0.75 |
| Geography | 100 | 0.81 |

# Item Response Theory (IRT) analysis

| *Differential Item Functioning (DIF) Analysis* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Type of Exam** | **Number of ILRs** | **Total Number of Items** | **Average DIF** | **A Flag (-0.5 to + 0.5 negligible DIF %** | **B Flag (-1 to +1) moderate DIF%** | **C Flag (< -1.0 or > 1.0) considerable DIF%** | **Critical C Flag (< -2.0 or > 2.0) %** | **Percentages of Items < -2 & > 2.0** |
| English NS | 3,418 | 120 | -0.461 | 20 | 25.8 | 54.2 | 31.7 | < -2=22.5% & > 2.0= 9.2% |
| English SS | 4,964 | 120 | -0.153 | 21.7 | 18.3 | 60 | 27.5 | < -2.0 =15 & > 2.0= 12.5% |
| Math NS | 3,972 | 65 | -0.499 | 26.2 | 21.5 | 52.3 | 29.2 | < -2.0 = 20% & > 2.0=9.2% |
| Math SS | 3,880 | 65 | 0.078 | 12.3 | 13.9 | 73.8 | 40 | < -2.0 = 21.5% & > 2.0= 16.9% |
| SAT NS | 4,194 | 60 | -0.494 | 21.7 | 16.7 | 61.7 | 36.7 | < -2.0 = 21.7% & > 2.0= 15% |
| SAT SS | 4,738 | 60 | 0.131 | 18.3 | 8.3 | 73.3 | 50 | < -2.0 = 23.3% & > 2.0= 26.7% |
| Physics | 3,918 | 50 | -0.151 | 12 | 20 | 68 | 40 | < -2.0 = |

# Item Response Function




**Parameters:**

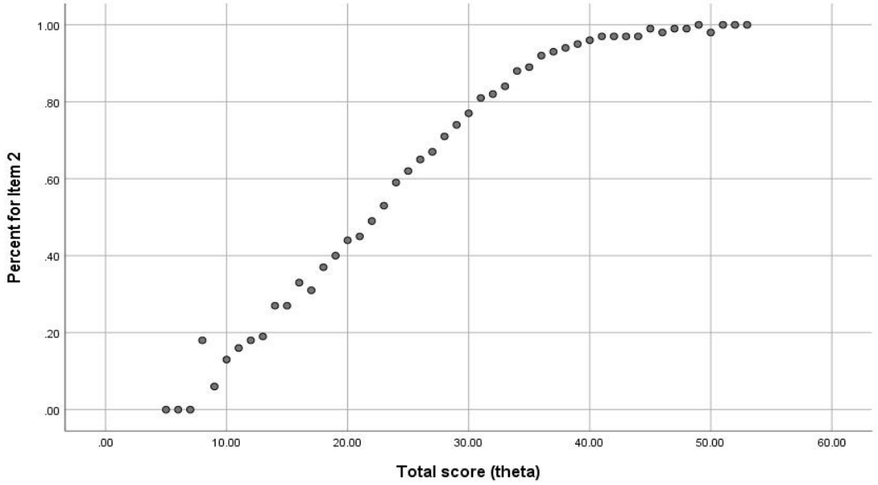
- **Difficulty**
- **Discrimination**
- **Guessing**

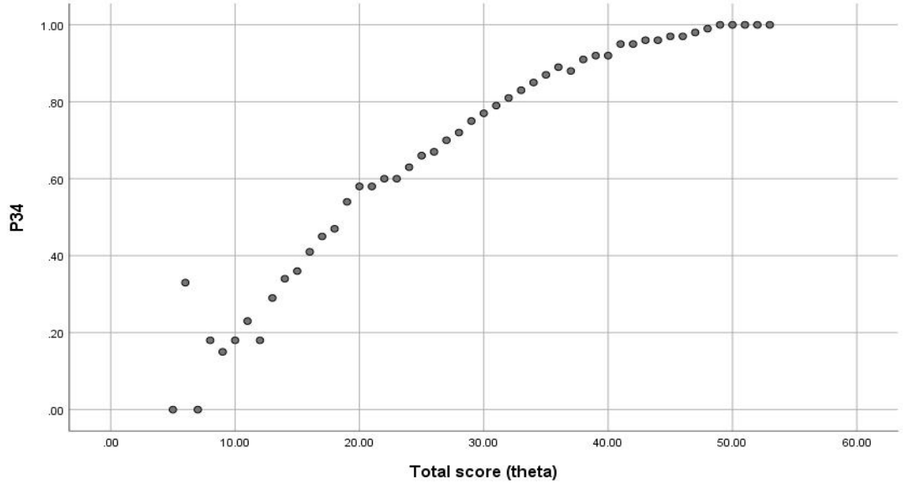
**Models:**

- **1 Parameter**
- **2 Parameter**
- **3 Parameter**

# Sample Items

## Simple Scatter plot/dot



# IRT …

Compared to CTT which is easier to use, in IRT

- Items of different difficulty and discrimination level are used
- Shorter test can be more reliable than longer tests.
- Interval scale properties are achieved by justifiable measurement models, not score distributions.

# Intercorrelation of NS subjects as evidence for validity, with the first half to 2022 & the second half to 2023

| Variable | English NS | Math NS | Physics | Chemistry | Biology | Civics NS | English NS | Math NS | Physics | Chemistry | Biology | Civics NS | English NS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Math NS | 0.85 | | | | | | | 0.85 | | | | | |
| Physics | 0.83 | 0.90 | | | | | | 0.81 | 0.89 | | | | |
| Chemistry | 0.83 | 0.94 | 0.89 | | | | | 0.81 | 0.90 | 0.87 | | | |
| Biology | 0.89 | 0.93 | 0.91 | 0.94 | | | | 0.87 | 0.90 | 0.87 | 0.91 | | |
| Civics NS | 0.91 | 0.90 | 0.86 | 0.94 | 0.93 | | | 0.89 | 0.89 | 0.84 | 0.91 | 0.92 | |
| SAT NS | 0.96 | 0.90 | 0.87 | 0.89 | 0.92 | 0.94 | 0.96 | 0.92 | 0.90 | 0.85 | 0.87 | 0.89 | 0.92 |

# Intercorrelation of SS subjects as evidence for validity, with the first half to 2022 & the second half to 2023

| Variable | Eng SS | SAT SS | Geo | Hist | Math SS | Eng SS | SAT SS | Geo | History | Math SS |
|---|---|---|---|---|---|---|---|---|---|---|
| SAT SS | 0.94 | | | | | 0.93 | | | | |
| Geo | 0.83 | 0.84 | | | | 0.85 | 0.92 | | | |
| History | 0.82 | 0.83 | 0.94 | | | 0.85 | 0.91 | 0.95 | | |
| Math SS | 0.67 | 0.71 | 0.74 | 0.72 | | 0.78 | 0.85 | 0.85 | 0.84 | |
| Civics SS | 0.88 | 0.87 | 0.94 | 0.94 | 0.68 | 0.93 | 0.95 | 0.94 | 0.94 | 0.82 |

## Correlation of EUEE with First Semester University Grade (N = 253)

| | EUEE Scores | University First Semester Grade |
|---|---|---|
| Sex | 0.08 | 0.13* |
| EUEE Scores | | 0.62** |

# Teachers Opinion *about challenges of EUEE associated with students*

| Problems associated with students | Not a problem | | Minor Problem | | Moderate Problem | | Serious Problem | |
|---|---|---|---|---|---|---|---|---|
| | Freq | % | Freq | % | Freq | % | Freq | % |
| Lack of motivation among students on the exam | 14 | 4.6 | 63 | 20.7 | 89 | 29.3 | 138 | 45.4 |
| Lack of exam readiness of students | 14 | 4.7 | 32 | 10.8 | 72 | 24.4 | 177 | 60.0 |
| Students lack of confidence in their ability to perform well on the exam | 7 | 2.3 | 31 | 10 | 91 | 29.4 | 180 | 58.3 |

# Teachers Opinion *about challenges of EUEE associated with administration and scoring of the exams*

| Problems associated with *administration and scoring of the exams* | Not a problem | | Minor Problem | | Moderate Problem | | Serious Problem | |
|---|---|---|---|---|---|---|---|---|
| | Freq | % | Freq | % | Freq | % | Freq | % |
| Invigilation related problems | 37 | 12.2 | 95 | 31.4 | 106 | 35 | 65 | 21.5 |
| Error in scoring of the exams | 32 | 10.6 | 97 | 32.2 | 87 | 28.9 | 85 | 28.2 |
| Exam malpractices such as cheating | 19 | 6.3 | 39 | 12.8 | 87 | 28.6 | 159 | 52.3 |
| Distraction or disturbance created by the examinees/students in the exam room | 33 | 10.8 | 81 | 26.6 | 104 | 34.1 | 87 | 28.5 |
| Overall exam management (leadership) problem before and during the examination process | 19 | 6.2 | 74 | 24.2 | 115 | 37.6 | 98 | 32 |

# Result: Disparity

- Subjects, Stream, Gender, Region,

## *Descriptive Statistics for the Score of Students in EUEE in 2023 EUEE by Stream and Subjects*

| Stream | Subject | N | Min | Max | Mean (M) | Standard Deviation (SD) |
|---|---|---|---|---|---|---|
| Natural Science | English | 98,100 | 1 | 93 | 32.18 | 13.16 |
| | Math | 98,395 | 3 | 97 | 30.14 | 13.19 |
| | SAT | 98,394 | 2 | 97 | 35.61 | 13.70 |
| | Physics | 98,395 | 2 | 98 | 27.75 | 10.69 |
| | Chemistry | 98,395 | 3 | 100 | 32.74 | 13.33 |
| | Biology | 98,395 | 2 | 99 | 32.19 | 13.07 |
| | Civics | 98,395 | 2 | 98 | 35.76 | 15.82 |
| Social Science | **Total** | N/A | 37 | 649 | 225.33 | 80.54 |
| | English | 183,722 | 1 | 92 | 27.41 | 7.18 |
| | SAT | 183,728 | 2 | 93 | 29.27 | 9.08 |
| | Geography | 183,730 | 2 | 93 | 27.78 | 7.71 |
| | History | 183,730 | 1 | 93 | 26.81 | 7.99 |

# *Disparity in the score of students in EUEE in Natural and Social Science Streams*

| | | N | Mean | Stan. D | t | df | p | Mean Difference | Std. Error Difference | 95% CI [Lower, Upper] | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2022 | Natural Science | 279,181 | 218.81 | 66.63 | | | | | | | |
| | Social Science | 429,965 | 167.90 | 34.04 | 373.29 | 374832.3 | < .001 | 50.91 | 0.136 | [50.64, 51.18] | 1.028 |
| 2023 | Natural Science | 98391 | 225.63 | 80.76 | | | | | | | |
| | Social Science | 183704 | 165.75 | 37.05 | 220.51 | 121003.1 | < .001 | 59.89 | 0.27 | [59.36, 60.42] | 1.064 |

***Differences between female and male students in their score in EUEE in 2022 and 2023***

| Years | Sex | N | Mean | SD | T | Df | p | MD | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|
| 2022 NS | F | 118669 | 208.34 | 58.32 | -74.17* | 276277.57 | < .001 | -18.20 | -0.28 |
| | M | 160540 | 226.54 | 71.18 | | | | | |
| 2023 NS | F | 42505 | 220.45 | 78.47 | -18.3* | 93019.94 | < .001 | -9.40 | -0.12 |
| | M | 55595 | 229.85 | 81.33 | | | | | |
| 2022 SS | F | 205423 | 162.55 | 28.415 | -100.89* | 414750.37 | < .001 | -10.24 | -0.30 |
| | M | 224630 | 172.79 | 37.815 | | | | | |
| 2022 SS | F | 91806 | 162.2 | 35.23 | -41.24* | 182376.30 | < .001 | -7.10 | -0.19 |
| | M | 91924 | 169.29 | 38.46 | | | | | |

***Disparities among students in their score in EUEE in 2022 and 3023 by areas of residence***

| Years | Residence | N | Mean | SD | T | df | p | MD | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|
| 2022 NS | Urban | 15043 | 283.52 | 102.34 | 81.13* | 15676.96 | < .001 | 68.40 | 1.06 |
| | Region | 264153 | 215.12 | 62.00 | | | | | |
| 2023 NS | Urban | 32252 | 264.08 | 100.02 | 95.33* | 44175.65 | < .001 | 57.66 | 0.76 |
| | Region | 66140 | 206.42 | 60.69 | | | | | |
| 2022 SS | Urban | 13454 | 198.91 | 59.86 | 61.74* | 13708.38 | < .001 | 32.01 | 0.95 |
| | Region | 416584 | 166.89 | 32.38 | | | | | |
| 2023 SS | Urban | 27906 | 196.58 | 63.59 | 94.07* | 29642.19 | < .001 | 36.35 | 1.05 |
| | Region | 155815 | 160.23 | 26.35 | | | | | |

# Achievement comparison of schools across Regions in Ethiopia (2023 EUEE)

| | Natural Science (NS) | | | Social Science (NS) | | |
|---|---|---|---|---|---|---|
| | Number (NS) | Mean (NS) | SD (NS) | N (SS) | Mean (SS) | SD (SS) |
| Tigray | 17 | 193.82 | 16.95 | 18 | 163.44 | 10.93 |
| Afar | 76 | 179.91 | 12.43 | 74 | 155.07 | 9.73 |
| Amhara | 773 | 215.96 | 42.51 | 493 | 178.86 | 26.16 |
| Oromia | 1,537 | 196.76 | 33.53 | 1,937 | 161.94 | 16.09 |
| Somali | 235 | 178.71 | 20.23 | 174 | 152.72 | 11.32 |
| BSG | 68 | 190.84 | 31.81 | 67 | 163.22 | 25.90 |
| SNNP | 664 | 208.79 | 40.82 | 728 | 162.77 | 16.55 |
| Gambela | 62 | 185.24 | 17.70 | 65 | 159.38 | 10.49 |
| Harari | 15 | 248.47 | 54.85 | 10 | 177.80 | 34.41 |
| Addis Ababa | 243 | 289.53 | 81.00 | 254 | 225.37 | 64.99 |
| Dire Dawa | 26 | 235.12 | 67.05 | 29 | 185.03 | 43.17 |
| Abroad | 4 | 310.00 | 139.57 | 2 | 344.00 | 48.08 |

# Conclusion and Suggestion

**Good standardized test outcomes can be reason and result of quality of education, with it being cost effective, compared to classroom assessment that requires huge resources including quality teachers.**

**Several evidences show that national exams can be said to be valid, which still requires further and continuous evaluation of it**

**The low test scores (which has almost become a national shock during the 2022) requires different stakeholders' active and sustained effort: Government, teachers, school management, students, and parents**

# Thank you!