NEEDU
NATIONAL EDUCATION EVALUATION & DEVELOPMENT UNIT

ARE YOUR ASSESSMENT ITEMS OF GOOD QUALITY?

How the use of item analysis can benefit learners and improve test quality

February 2018
Prepared by NEEDU

SUMMARY: Is the mark of 47% a true reflection of what a learner knows after an instruction or was this mark affected by the quality of assessment items, e.g. the items were flawed, bias or ambiguous?

The fairest tests for all learners are those that are valid and reliable—those that are free from bias and ambiguity. To improve the quality of tests, an *item analysis* not only examines how learners have performed in each assessment item, but it also critiques the credibility of each item in providing a valid and reliable evidence of learner academic performance.

This policy brief addresses three important questions that teachers in schools that work consider when conducting an *item analysis* to improve the quality of tests:

- **Is the item difficulty level appropriate?** Item analysis can identify items which are too difficult or too easy
- **Does the item discriminate adequately?** Item analysis can identify items that are not able to differentiate between those who have learned the content and those who have not
- **Are the distractors performing adequately?** Item analysys can indicate how effective each destructor is in a multiple choice question test item

INTRODUCTION

Teachers consistently assess learners to establish if *the instruction*, "the dose," has been successful; that is, if all learners have mastered the concepts and skills taught in class. However, unlike in the medical field, where medical practitioners use ready-made and more accurate tests to establish the effect of a medical treatment or a procedure, teachers are often left on their own to prepare and administer formal teacher-made tests to assess the impact of the instruction. But to what extent are teacher-made tests reliable and valid?

QUALITY OF TEACHER-MADE TESTS

Because tests play an important role in giving feedback to teachers on their educational actions, the quality of a test is a critical issue. Tests are indispensable tools in the educational enterprise. Unfortunately, some best practices in item and test analyses are too infrequently used in teacher-made classroom tests. As a result, this leads to misleading results.

Teachers in schools that work[1] follow an assessment loop. The third phase in the assessment loop is the *analysis of assessment results* (see Policy Brief N-04 entitled *"Do assessment practices in your school measure up? Some lessons from schools that work"*).

The purpose of this policy brief is to discuss some critical points about how these schools

[1] In April 2017, the Minister of Basic Education commissioned the National Education Evaluation and Development Unit (NEEDU) to conduct the *Schools that Work II* study. This study sought to examine the characteristics of top-performing schools in South Africa. The best practices discussed in this advocacy brief are based on the findings of that study. The full report is available on the Department of Basic Education website: www.education.gov.za.
NEEDU can be reached at (012) 357 4231

conduct an *item* analysis to improve the quality of the items in a test or an exam.

TEST ITEM ANALYSIS

Item analysis is described as a process of examining learners' responses on a class-wide performance and the examination of individual items on a test, rather than the test as a whole. Following are three common types of item analysis:

Described next are procedures that teachers in top-performing schools use to calculate the Difficulty Index, the Discrimination Index and to analyse response options.

ITEM DIFFICULTY INDEX

What is an Item Difficulty Index?

- It is the proportion of learners taking a test who answered a test item correctly
- It addresses the question: **How difficult is the item?**

How is the Difficulty Index calculated?

- **ITEM DIFFICULTY FORMULA:**
- ① **Count** the number of learners who got the correct answer
- ② **Divide** by the total number of learners who took the test
- **PLEASE NOTE:**
- Values can range from 0.0 to 1.0 and are calculated using the Pearson Correlation Coefficient
- When multiplied by 100, a p-value converts to a percentage, which is the percentage of learners who got the item correct
- A p-value is the item difficulty index of an assessment item in a test (p = proportion)

Table 1 below displays results of a ten-question test. The symbol "+" indicates a correct answer in the question; an "x" indicates an incorrect answer.

Table 1: Computing the difficulty of an item

LEARNER NAME	QUESTIONS									
	1	2	3	4	5	6	7	8	9	10
ZAMI	+	+	+	+	+	+	+	+	+	+
MAPHORI	+	x	+	x	x	+	x	+	x	x
ROSE	+	+	x	+	+	+	+	+	x	x
BHEKI	+	+	+	+	+	+	+	x	+	+
SAHEDA	+	+	x	+	x	+	x	+	x	x
BONGI	+	+	x	+	x	+	+	x	+	x
GUGU	+	x	x	+	x	+	x	x	+	x
BRUCE	+	x	x	x	+	x	x	+	x	x
RONALD	+	+	+	+	+	x	+	x	+	+
BARBARA	x	x	x	x	x	+	x	+	x	x

Applying the Item Difficulty Index formula described above, the difficulty indices for Questions 1 and 10 can be calculated as follows:

EXAMPLE 1: HOW AN ITEM DIFFICULTY INDEX IS CALCULATED:

- The difficulty of **Question No. 1** (i.e. the number of learners who got this question correctly is 9 out of 10 learners or 9/10 or 90% or 0.90). The p-value for Question 1 is therefore **0.90**
- The difficulty of **Question No. 10** (i.e. number of learners who got this question correctly is 3 out of 10 learners or 3/10 or 90% or 0.30). The p-value for Question 10 is therefore **0.30**

How are the results of the Difficulty Index analysis interpreted?

- The higher the p-value, the easier the items. This means, the higher the difficulty index, the easier the item is understood to be
- A rough "rule-of-thumb" is that:
- ① **Easy Item** = 70% and above or a p-value of **0.70** and above
- ② **Moderate Item** = 30-69% or a p-value of **0.30-0.69**
- ③ **Difficult Item** = 29% and below or a p-value of **0.29** and below

These interpretation guidelines can be applied to the analysis of results in Table 1 as follows:

EXAMPLE 2: HOW THE RESULTS OF ITEM DIFFICULTY INDEX RESULTS FOR QUESTIONS 1 AND 10 IN TABLE 1 CAN BE INTERPRETED

- The p-value for Question 1 is **0.90**—indicating that this test item was very easy and/or that most learners grasped the concept/skill that was tested
- The p-value for Question 10 is **0.30**—suggesting that this test item was much more difficult or confusing or ambiguous

ITEM DISCRIMINATION INDEX (DI)

Why calculate a Discrimination Index ?

- To enable teachers to identify items that are not able to differentiate between learners who have learned the content and those who have not
- To improve the quality of the items in a test, thereby improving both reliability and validity of the test

What is an Item Discrimination Index?

- It is a measure of an item's ability to discriminate or differentiates between those who scored high on the total test and those who scored low
- It is the point biserial correlation between getting the item right and the total score on all other items
- It seeks to establish whether getting an item correct or not is due to learners' level of knowledge or ability and not due to something else such as chance or a test bias
- It addresses the question:
  Does the item discriminate between high and low achievers?

How is Discrimination Index calculated?

- **ITEM DISCRIMINATION FORMULA:**
- After marking the test:
- ① **Arrange** or sort the test by total marks from the highest marks to the lowest marks so that the highest total marks appear at the top (see Table 2)
- ② **Create** two groupings of tests: (a) the *upper group* with high marks and (b) the *lower group* with low marks (see Table 2)
- ③ **Count** the number of learners who got each item correct in each group, i.e. upper and lower
- ④ **Calculate** the Difficulty Index for each item in the test and get a p-values for each group
- ⑤ **Subtract** the Difficulty Index for the lower group from the Difficulty Index for the upper group to get the DI (i.e. Discrimination Index).
- **PLEASE NOTE:**
- Values are calculated using the Pearson Correlation Coefficient.
- Values can range as follows:
  Range = (+1) --- 0 --- (-1)
  Maximum size Zero Minimum size
- A *positive discrimination index* (between 0 and 1) indicates that learners who received a high total mark chose the correct answer for a specific item more often than the learners who had a lower overall mark.
- If, on the other hand, more of the low-performing learners got a specific item correct, then the item has a *negative discrimination index* (between -1 and 0).

Table 2 displays the same results as in Table 1 but here results are arranged with the top overall performers at the top of the table (shaded in yellow) and the lower group at the bottom of the table (shaded in grey).

Table 2: Computing the DI of an item

LEARNER NAME	QUESTIONS										TOTAL MARK (%)
	1	2	3	4	5	6	7	8	9	10	
ZAMI	+	+	+	+	+	+	+	+	+	+	100
BHEKI	+	+	+	+	+	+	+	x	+	+	90
RONALD	+	+	+	+	+	x	+	x	+	+	80
ROSE	+	+	x	+	+	+	+	+	x	x	70
BONGI	+	+	x	+	x	+	+	x	+	x	60
SAHEDA	+	+	x	+	x	+	x	+	x	x	50
GUGU	+	x	x	+	x	+	x	x	+	x	40
MAPHORI	+	x	+	x	x	+	x	+	x	x	40
BRUCE	+	x	x	x	+	x	x	+	x	x	30
BARBARA	x	x	x	x	x	+	x	+	x	x	20

"+" indicates the answer was correct; "x" indicates it was incorrect.

Applying the DI formula described above, the discrimination indices for Questions 1 and 10 in Table 2 can be calculated as follows:

EXAMPLE 3: HOW DISCRIMINATION INDEX (DI) IS CALCULATED:

DI CALCULATION STEPS	QUESTION 1		QUESTION 10	
① Sort by total marks from highest to lowest	As done in Table 2		As done in Table 2	
② Create upper and lower groups	As done in Table 2		As done in Table 2	
	Upper Group	Lower Group	Upper Group	Lower Group
③ Calculate difficulty indices for each group	5/5 or 100% or p-value =1	4/5 or 80% or p-value =0.80	3/5 or 60% or p-value =0.60	0/5 or 0% or p-value =0
④ Subtract the Difficulty Index for the lower group from the Difficulty Index for the upper group	p-value **1** minus p-value **0.80 = 0.20** DI/discrimination value for Question 1 = **0.20**		p-value **0.60** minus p-value **0 = 0.60** DI/discrimination value for Question 10 = **0.60**	

How are the results of the Discrimination Index analysis interpreted?

- A question is a good discriminator when learners who answer the question or an assessment item correctly also do well on the test as a whole
- The higher the DI, the better the test item discriminates between the learners with higher marks and those with lower marks
- Items with low discrimination indices are often ambiguously worded and should be examined
- The more difficult or easy the item, the lower its discriminating power

How are the results of the Discrimination Index analysis interpreted?

- The greater the positive value, i.e. the closer it is to 1.0, the stronger the relationship is between overall test performance and performance on that item
- A rough "rule-of-thumb" is that:
- ① If DI is **0.40 and above**, then the item is functioning satisfactorily, i.e. it is differentiating between learners with higher and lower levels of knowledge
- ② If DI is **0.30 to 0.39**, then little or no revision is required. The item is differentiating between learners with higher and lower levels of knowledge
- ③ If DI is **0.20 to 0.29**, then the item is marginal and needs revision. The item may be of low quality or marked incorrectly
- ④ If DI is **0.19 and below**, then the item should be eliminated or completely revised

- A Good Achievement Test should have:
- 50% of Items = DI of 0.40 and above
- 40% of Items = DI of 0.39 to 0.20
- 10% of Items = DI of 0.19 to 0.00

These interpretation guidelines can be applied to the analysis of results in Table 2 as follows:

EXAMPLE 4: HOW THE RESULTS OF DI RESULTS FOR QUESTIONS 1, 8 AND 10 IN TABLES 1 AND 2 CAN BE INTERPRETED

QUESTION 1		QUESTION 8		QUESTION 10	
Difficulty Index	DI	Difficulty Index	DI	Difficulty Index	DI
0.90	0.20	0.60	-0.40	0.30	0.60
• The difficulty index of **0.90** means this item was **quite easy** • Low discriminatory power (**0.20**) suggests that this item does not discriminate well between top and low performers		• The difficulty index of **0.60** means this item was **moderately easy** • Negative DI of **-0.40** means that the low-performing learners were more likely to get this item correct		• The difficulty index of **0.30** means this item was **moderately difficult** • DI of **0.60** means that this item is a good discriminator • This is the best" overall question	
The test analysed in Tables 1 and 2 is not a good test in that while 50% of the items had a DI of **0.40** and above (good), 30% had a DI of **≤0.19**					

ANALYSIS OF RESPONSE OPTIONS

Why conduct Distractor Analysis?

- To decrease the chance that random guessing in multiple choice questions (MCQ) could result in credit for a correct answer.
- The greater the number of plausible ***distractors***, the more accurate, valid, and reliable the test typically becomes
- A ***distracter*** is an incorrect alternative option on a multiple choice item

How is Distractor Analysis conducted?

- ① **Count** the proportion of learners choosing each response option in the MCQ test
- ② **Divide** the number of learners who chose each option as an answer by the number of learners taking the test to get the Difficulty Index for each item

Table 3 shows the number of learners in a class of 30 who selected each answer choice for Questions 1 and 2 in a multiple choice test with four answer choices (A, B, C and D).

Table 3: Conducting Distractor Analysis

QUESTION	A	B	C	D
No. 1	0	8	19*	3
No. 2	10	6*	8	6

* Denotes the correct answer

EXAMPLE 5: HOW THE RESULTS OF DISTRACTOR ANALYSIS FOR QUESTIONS 1 AND 2 IN TABLE 3 CAN BE INTERPRETED

- The p-value for Question 1 (Option C) is **0.63**, i.e. 19 out of 30 learners chose the correct option. This indicates this test item was moderately easy
- The p-value for Question 2 (Option B) is **0.20**, i.e. 6 out of 30 learners chose the correct option. This suggests that this test item was much more difficult
- No learner chose Option A in Question 1. This means that Option A does not act as a good distractor
- In Question 2, more learners selected an incorrect option (A) than the correct option (B). In addition, learners chose between all four options almost evenly. This makes guessing correctly more likely, which affects the validity of this item

USING THE RESULTS OF THE ITEM ANALYSIS

Teachers in schools that work use the results of item analysis to:

- Adjust the instruction, including modifications based on learner needs, pace of instruction and coverage of subject material, e.g. when most learners get an item wrong as in Question 10 in Tables 1 and 2.
- Moderate items with zero, low positive and negative discrimination indices, e.g. Question 8 in Table 2 is problematic in that only two out of five high-performing learners got it correctly while four out five low-performing got it correctly. This is why it has a negative discrimination index of **-0.40**. It must be moderated to increase the credibility of the test.
- Identify which misconceptions are shared by the majority of learners and correct them, e.g. when many learners choose an incorrect option in a multiple choice test, e.g. Question 2 in Table 3.