Session 12

Reliability Testing through Item Analysis of Test Results

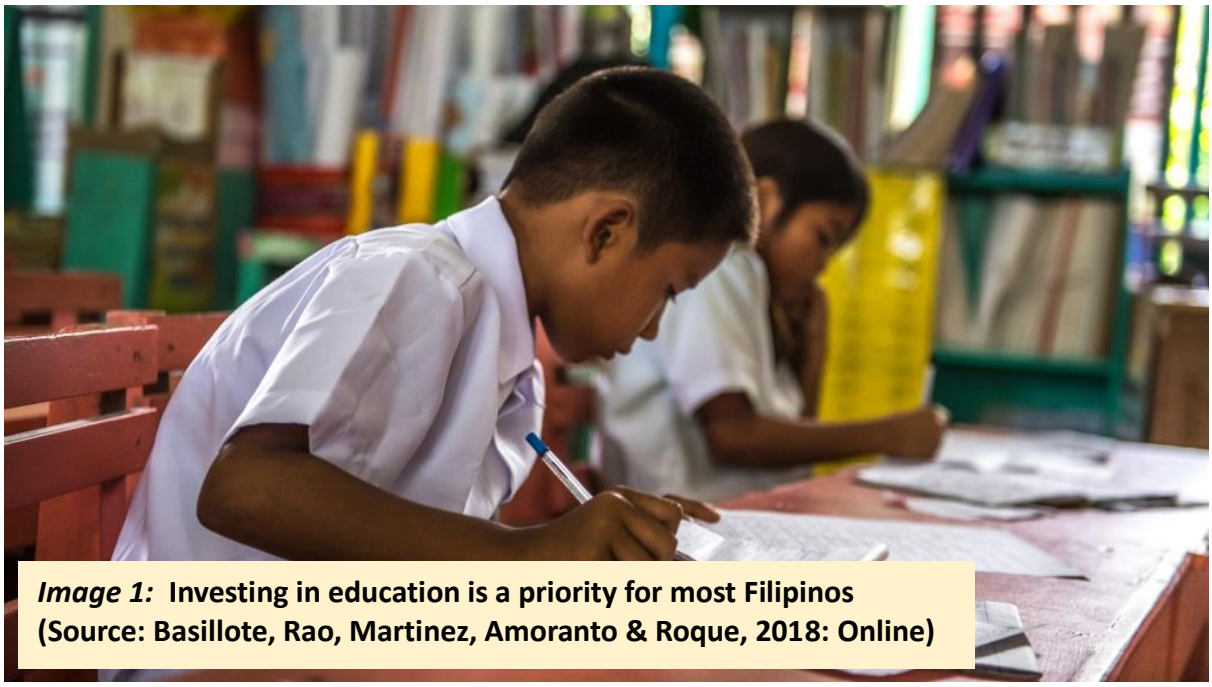

Image 1: **Investing in education is a priority for most Filipinos (Source: Basillote, Rao, Martinez, Amoranto & Roque, 2018: Online)**

A. Introduction

When teaching our pupils, the assessment will always be the final action to take before we decide what the next step should be concerning their educational pursuits. Only through assessment will we be able to determine whether or not the instructional activities in which we engaged our pupils resulted in the learning that we had hoped for them to acquire. In reality, assessment functions as the connecting link between teaching and learning.

Through a valid assessment process, we gain insights into our students' achievements as well as their shortcomings so that we can take the most appropriate and adequate measures to improve learning.

Exams offer a glimpse into students' learning discrepancies, and as a result, proper assessment provides educators with a means to strengthen student learning. In the same way that they offer questions to students about where and why they are struggling academically, exams should provide an answer for teachers.

However, once the procedure is established, item analysis becomes a scientific method that can be used to enhance testing and uphold academic honesty.

Session 11 Conclusion

The Table of Specifications (TOS) is the blueprint of the assessment instrument to be crafted. It ensures the validity and quality of the planned test, which indicates the learning content and skills to be assessed and how they are assessed.

B. Session Objectives

Right at the conclusion of this lecture, you are expected to

1. Define item analysis and its function in the assessment of learning; and,
2. Determine the three indices in analyzing test item results to describe the quality of teaching and learning.

C. Session Content

1. Analyzing Test Results

Item analysis refers to the process of examining the answers of students to specific examination questions to assess the quality of the exam. Ensuring the effectiveness and fairness of tests is a crucial aspect that necessitates the utilization of a significant tool.

Educators frequently engage in item analysis, consciously and unconsciously, as part of their regular practice. The grading process includes thoroughly examining student responses and identifying recurring errors, be it in response to a specific question or category of questions.

However, when the process is structured, item analysis transforms into a scientific methodology that enables the enhancement of examinations and the preservation of academic integrity.

In this lecture, three characteristic features determine the quality of the assessment instrument. They represent the students' learning status based on the item analysis results, which may reveal that if not the quality of learning, the test instrument could be the cause of why students could not provide actual knowledge. The three indices are (1) difficulty index (P), (2) discriminatory index (D), (3) distracter index.

Difficulty of the Item. When an item is uniformly answered either correctly or incorrectly by every student, it results in a drop in the dependability of the exam. When all individuals provide the exact correct response, it becomes more challenging to discern those who profoundly understand the subject matter. On the contrary, if all individuals provide an erroneous response, it becomes impossible to distinguish those who have acquired a profound understanding of the subject matter. In all sense, test items may be easy, difficult, or average, which is the ideal item in this context.

Discriminatory Power of the Item. The purpose of test questions is to assess students' diverse levels of knowledge on the subject matter, as indicated by the percentage of accurate responses on the exam. Desirable discrimination can be evaluated by examining the relationship between students' correct responses and their total test scores. Specifically, this involves comparing the performance of students who achieved high overall scores with those who obtained poor overall scores and determining if the former group demonstrates a higher proportion of correct answers on individual test items. If the top-performing individuals are distinguished from the bottom-performing individuals, which group correctly answers which questions?

Enhancing student learning can be achieved through the implementation of feedback mechanisms, as well as the careful design of examinations. Using item analysis data might significantly influence the approach to developing subsequent tests. As previously mentioned, if the evaluation of student knowledge serves as the intermediary between instruction and acquisition of knowledge, it is imperative that examinations effectively gauge the extent of the gap in student learning.

Distracters (for multiple-choice tests). Distractors are an essential part of multiple-choice tests. Can test-takers be effectively diverted away from the correct answer by exam questions? Is there a 50% probability of giving the right answer to a multiple-choice question if, for instance, there are four choices, and two are obviously wrong? Distractors lose their usefulness in gauging students' comprehension when they aren't camouflaged and instead come across as plainly false. Examinees with lower average scores will be drawn to a highly effective distractor at the expense of those with better average scores.

Importance. Not only can item analysis drive exam design, but it can also inform course content and curriculum.

Regarding item difficulty, it's important to note whether errors indicate a misunderstanding of the question or the concept the item addresses. When a large number of students answer an item incorrectly, it's notable. It may be a matter of fine-tuning a question for clarity; is the question's wording confusing? Are the answers clear?

Or it could be that the material may have to be reviewed in class, possibly with a different learning approach. Item distractor analysis is also helpful because it can help identify students' misunderstandings about the material if most students selected the same incorrect multiple-choice answer, providing insight into student learning needs and opportunities.

Whether you manually employ item analysis or via software, we think data-driven exams and curricula Distractors are essential to multiple-choice tests. Can test-takers be effectively diverted away from the correct answer by exam questions? Is there a 50/50 probability of giving the correct answer to a multiple-choice question if, for instance, there are four choices, and two are obviously wrong? Distractors lose their usefulness in gauging students' comprehension when they aren't camouflaged and instead come across as plainly false. Examinees with lower average scores will be drawn to a highly effective distractor at the expense of those with better average scores.

2. Performing the Item Analysis

Tests regarding the quality of individual items, item sets, and complete sets can be evaluated by applying statistical and expert judgment in item analysis. The focus is to analyze item performance in isolation from the rest of the exam or in comparison to an external criterion. In doing so, it enhances the quality of both items and examinations.

Although there are some similarities between the concepts used in norm-referenced and criterion-referenced item analyses, there are also important distinctions. When analyzing data from criterion-referenced tests, you should employ norm-referenced statistics for the pre-test and post-test. According to this proposal, the assumptions upon which norm-referenced statistics are based—namely, untrained individuals will know very little about pre-test material—are relevant. After students have received training, a criterion-referenced test should be administered, necessitating criterion-referenced statistics.

a) Computing the Difficulty Index (P)

As explained earlier, the difficulty index (P) is the percentage of the students who got the test item correctly. Navarro, Santos, and Corpuz (2019) suggest the following procedure for obtaining the P.

1. Arrange the test papers from highest to lowest scorers after checking and scoring them.
2. Classify the test papers by taking the 25% of the class who got the highest scores and another 25% who got the lowest scores. Some references consider 28% and 30%, especially when the total number of test takers is 30 or below. In this lesson, we will adopt the 25%.
3. Prepare a worksheet to record the number of students who got the correct answer in each item in the highest and lowest scoring groups. Record also the number of students choosing the options in each item for multiple-choice tests. Table 1 shows how to prepare the worksheet.

Table 1. Item Worksheet Template (Multiple-choice Test)

Item No.	Highest Scorers (nH=12)				Lowest Scorers (nL=12)			
	A	B	C	D	A	B	C	D
1	1	**8**	1	2	4	**4**	2	2
2	0	3	**7**	2	3	6	**1**	2
3	**5**	2	2	3	**7**	3	2	0
4	3	0	0	**12**	4	0	5	**3**
5	4	**2**	6	0	3	**5**	3	1

Cells in **RED BOLD** text are the correct options;
thus, they are also the number of students who got the items correctly.

4. Prepare to compute the P for each item by doing the following:

   N = nH + nL

5. Compute the P (Difficulty Index) using the formula below.

$$P_i = \frac{cH+cL}{N}$$

Where:

P_i – Difficulty intext per item
cH – Number of students in the highest group who got the correct answer
cL - Number of students in the lowest group who got the correct answer
N – Total number of students considered

EXAMPLE (We will use the data in Table 1)

Item No. 1: *cH = 8; cL = 4; N = 12 + 12*

$$P_i = \frac{8+4}{24} = \frac{12}{24} = \mathbf{0.50}$$ **(Moderately Difficult)**

6. Refer to the table of interpretation (Table 2) to describe the computed number.

Table 2. Table of Interpretation for the Difficulty Index

Range	Interpretation	Action
0.0 – 0.25	Difficult	Discard
0.26 – 0.75	Moderately Difficult (Ideal result)	Retain
0.76 and above	Easy	Revise

Source: Navarro et al. (2019, p. 86)

b) Computing the Discriminatory Index (D)

A fundamental method for determining the legitimacy of an item is the discrimination index. It measures an item's capacity to discriminate between people who scored well on the entire test and those who scored poorly on individual questions.

Compute the D using the formula below with the data utilized to analyze item difficulty.

$$D_i = \frac{cH - cL}{n}$$

Where:

D_i – Discriminatory index per item
cH – Number of students in the highest group who got the correct answer
cL - Number of students in the lowest group who got the correct answer
n – 25%

EXAMPLE:

$$D_i = \frac{cH - cL}{n}$$

Item No. 1: *cH = 8; cL = 4; n = 12*

$$D_i = \frac{8-4}{12} = \frac{4}{12} = \mathbf{0.33\ (Discriminating)}$$

Table 3. Table of Interpretation for the Discriminatory Index

D value range	Interpretation
-1.00 to -0.60	Questionable
-0.59 to -0.20	Not Discriminating
-0.19 to 0.20	Moderately Discriminating
0.21 to 0.60	Discriminating (Ideal Result)
0.61 to 1.00	Very Discriminating (Most Ideal Result)

Source: Navarro et al. (2019, p. 87)

c) Making Decisions

After obtaining the P and D, refer to Table 4 for the final decision regarding the items.

Table 4. Decision Table for the Analyzed Items

Difficulty Level	Discriminating Level	Decision
Difficult / Very Difficult	Not Discriminating / Questionable	Improbable – Discard
	Moderately Discriminating	May need revision on the stem and/or choices
	Discriminating / Very Discriminating	Accept with little revision if necessary
Moderately Difficult	Not Discriminating / Questionable	Needs revision especially on the choices
	Moderately Discriminating	Accept but needs slight revision on the stem or choices
	Discriminating / Very Discriminating	Accept as it is
Easy / Very Easy	Not Discriminating / Questionable	Totally discard
	Moderately Discriminating	Needs major revision on the stem and/or choices
	Discriminating / Very Discriminating	Accepted by slightly increasing difficulty

Source: Navarro et al. (2019, p. 88)

d) Making Assumptions on the Distracter

As explained earlier, distracters in a multiple-choice test are very important because they should further engage students in analytical and critical thinking. The following assumptions based on Kubiszyn and Borich (2007) can be considered to determine whether the distracters did their job efficiently. The item analyzer will be the one to judge.

1. A distracter is inefficient when no one chooses it.
2. When the item is moderately difficult or difficult and discriminating, a distracter is very efficient when more in the lowest 25% choose it than those in the highest 25%. Otherwise, the distracter could be questionable because it might appear as essential as the correct answer.

3. When the item is rated easy or very easy regardless of the discriminating power or recommended for revision, distracters chosen by less than 10% of the total considered students are deemed inefficient.
4. On the other hand, when the test item is rated questionable or improbable, the distracter with the highest frequency may be misleading or too distracting.

D. Conclusion

With item analysis, you can see how each student in your class did on each question. The Difficulty Index, the Discriminatory Index, and the Distracter Analysis are three popular types of item analysis that give teachers distinct kinds of information for decision-making in the teaching-learning process.

E. References

Basillote, L., Rao, L.N., Martinez, A., Amoranto, G. & Roque, J.D. (2018). *Investing in education is a priority for most Filipinos* [Online Image] [Accessed on October 30, 2023] https://blogs.adb.org/blog/5-ways-make-most-philippine-education-investments

Kubiszyn, T. & Borich, G. (2007). *Educational testing and measurement: Classroom application and practice. 8th edition.* N.J., U.S.A.: John Wiley & Sons, Inc.

Navarro, R.L., Santos, R.G., & Corpuz, B.B. (2019). *Assessment in learning 1. 4th edition.* Quezon City, Philippines: LORIMAR Publishing