Azim Premji University

An Introduction to Classical Test Theory

An Introduction to Classical Test Theory[i]

Context

Assessment is conducted to gather information about students' learning. This information can be used for identifying student's misconceptions, trace learning gaps, report learning levels, certify students and select them for future academic programs or scholarships. When a high-quality assessment is designed, it provides maximum information about students learning. Quality in assessment is characterized by validity and reliability.

A valid test should measure the attainment of learning outcomes. Validity is the property of a test that enables us to make valid interpretation and decision based on assessment data. For example, if a test is designed to measure a student's ability to solve problems, then the test paper should cover a wide range of content that can assess problem solving. A class 5 student should be able to solve problems using all four arithmetic operations when given in numerical format or word problems. If that happens, we can draw a valid interpretation about a student's ability to solve problems. However, if the vocabulary used in word problems is complex or needs a lot of comprehension abilities, then the reading demands associated with the test paper lowers the validity. Similarly, there is a threat to validity when the test paper has solved items from textbook and student is able to recall/ reproduce the answers instead of using the skill of problem solving.

Reliability deals with the consistency and accuracy in measuring the attainment of learning outcomes. For example, if a class 5 student is asked to design a poster on importance of trees to assess her skill of creative writing, it is important to clearly define the indicators of success in the form of rubrics. Creativity here can be assessed through indicators like colours/ patterns, layout, graphics/ photos, titles/ sub-titles, quality of information, writing and grammar/ spelling. Different levels can be defined for each indicator and students can be assessed against each on these indicators. When the same poster is assessed in the absence of these indicators, different teachers may assess it differently. It is also possible to miss assessing the actual cognitive processes which a student needs to attain to master a specific skill.

Quality of Question Paper

The usual approach taken to measure students' ability level in certification and large-scale assessments is to develop a question paper consisting of a number of items. Each of these items measure some facet of the particular ability (e.g., problem solving, creativity, etc.) of interest.

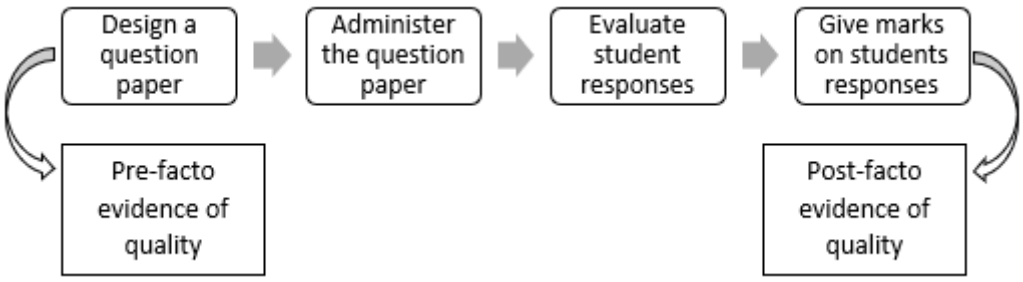


Figure 1. Gathering evidence of quality in a question paper

Pre-facto evidence of quality

It is ensured by having a detailed assessment framework, a balanced blueprint and following some quality protocols for developing items and scoring guides.

- Assessment framework defines measurable and observable learning outcomes for the test construct.
- A balanced blueprint will help to ensure that all the content domains are uniformly addressed and there is an equal distribution across all the domains.
- Some quality checks are considered while developing items- they should be aligned well with the learning outcomes, factually and conceptually correct and unambiguous.
- Scoring guides are designed to clearly communicate the expected response to teachers as well as students.

Post-facto evidence of quality

After we design the question paper, we administer it to our students, evaluate their responses and then give them some marks. Post facto evidence of quality is gathered using student marks. The important question here is why is it important to gather such evidence? In spite of following all the protocols for developing a good quality question paper, often there is a possibility of presence of unintended reasons or irrelevant clues that affects the quality of items. Analysing the items after students have attempted them helps to identify such irrelevant cues and ensure that the question paper is balanced and assesses what it attempts to assess.

In the field of educational measurement and evaluation, two theories dominate namely, Classical Test Theory (CTT) and Item Response Theory (IRT) for gathering post facto quality evidence. These theories link observable variables, such as test scores and item scores, to unobservable variables, such as students' ability level. Students' ability level is defined in relation to the construct that is measured through a test paper, e.g., a student may be highly capable in reading comprehension

while may have low aptitude level in mathematical ability. Therefore, this student's performance will be poor in Mathematics test as compared to a Language test.

The Classical Test Theory is the most prevalent and widely used in our country and uses simple mathematical analysis which is easy to interpret. However, this theory has a few limitations. Item response theory was developed in order to overcome such limitations of Classical Test Theory.

Classical Test Theory (CTT)

In CTT, items are analysed to improve their quality so that they provide maximum information about students' ability levels. Such analysis helps us to estimate the difficulty level of items, if the item is discriminating well between low and high ability level students and the options in multiple choice questions are plausible. Difficulty level, discrimination index and distractor analysis are conducted in CTT. Item responses can only be dichotomous, meaning only two possible answers are acceptable – right or wrong. Such responses are recorded for all types of multiple-choice questions.

Difficulty level

Difficulty level is a measure of individual test item difficulty. It is calculated by finding proportion of students who answered correctly out of the total number of students.

$$Diffciulty\ level\ = \frac{Number\ of\ correct\ responses}{Total\ number\ of\ students}$$

Table 1 shows data for students score for a test paper with 10 items of 1 mark each. Yuzvendra could attempt all items correctly, while Ishant made a mistake in two items. Difficulty level of the items can be calculated by counting the number of students those who have answered an item correctly divided by the total number of students. Item 1 is answered correctly by most of the students except Jasprit, Hardik, Rohit and Virat. Therefore, the difficulty level of item 1,

$$Diffciulty\ level\ = \frac{Number\ of\ correct\ responses}{Total\ number\ of\ students}$$

Greater the number of students attempting an item correctly, lesser is the difficulty level of items. In this case, item 2 is the easiest as all students have attempted it correctly while item 9 is most difficult since just 4 students could attempt it correctly.

Also, ability level is defined in reference to the number of items attempted correctly - greater the number of items attempted correctly, higher is the ability level of the student. Here, Yuzvendra has the highest ability level.

Student Name	Total Score (%)	Items									
		1	2	3	4	5	6	7	8	9	10
Yuzvendra	100	1	1	1	1	1	1	1	1	1	1
Pujara	90	1	1	1	1	1	1	1	1	0	1
Ishant	80	1	1	0	1	1	1	1	1	0	0
Jasprit	70	0	1	1	1	1	1	0	1	0	1
Rahul	70	1	1	1	0	1	1	1	0	0	1
Jadeja	60	1	1	1	0	1	1	0	1	0	0
Hardik	60	0	1	1	0	1	1	0	1	0	1
Rohit	50	0	1	1	1	0	0	1	0	1	0
Dhoni	40	1	1	1	0	0	0	0	0	1	1
Virat	30	0	1	0	0	0	1	0	0	1	0
Difficulty		**0.6**	**1**	**0.8**	**0.5**	**0.7**	**0.8**	**0.5**	**0.6**	**0.4**	**0.6**

Table 1. Difficulty level of items

Threats to difficulty level

Ideally, the item difficulty should relate to the skill being assessed, and not due to unfamiliar terminology, complex item format, unclear instructions, ambiguous items, confusing diagrams or text and wrong key. For example, if the following item is given to Indian students, some of them will not be able to answer this correctly. Not because they do not know the answer, but because they may not be familiar with Halloween as a festival. The item in Example 2 will do a better job to assess the same learning outcome.

Example 1

In which season do we celebrate Halloween?
A.Winter B.Rainy C.Summer D.Autumn

Example 2

In which season do we celebrate Independence Day?
A.Winter B.Rainy C.Summer D.Autumn

Often students find an item easy due to irrelevant clues provided in the item which in turn helps them guess the answer. Such clues can be in the form of a key that stands out in an MCQ,

implausible distractors and inclusion of a keyword in the distractors. In Example 3, it is very easy to guess the right answer. Option A is too small while option C and D are too big to be the right answer. If the same item is asked in the manner shown in Example 4 where every distractor is included to identify students' misconceptions, then the inherent difficulty of an item is retained.

Example 3

How much is 2^3?		
A.0.8	---------	Too small
B.8	---------	Correct answer
C.80	---------	Too big
D.800	---------	Just too big

Example 4

How much is 2^3?		
A.5	---------	(2+3)
B.6	---------	(2*3)
C.8	---------	(2*2*2)
D.9	---------	(3*3)

Discrimination index

It is a measure of the effectiveness of an item in discriminating between high and low ability students on a test. The notion is that high-ability students will tend to choose the right answer, on the other hand low-ability students will tend to choose the wrong answer. High ability or low ability is purely based on performance on the test. There is no pre-facto evidence of determining it. Discrimination index is calculated by taking the top scoring 1/3 of students and lowest scoring 1/3 of students. Therefore, we would expect that more correct answers will be given by top scoring 1/3 students as compared to lowest scoring 1/3 of students.

$$Discrimination\ index = \frac{Number\ of\ correct\ responses\ in\ higher\ group - Number\ of\ correct\ responses\ in\ lower\ group}{Total\ number\ of\ students\ in\ each\ group}$$

In Table 2, item 9 is poorly discriminating between low and high ability level students, since in high ability group just 1 student could answer this item correctly while in low ability group, 3 students answered it correctly. This indicates the presence of some irrelevant clue in the item letting students to guess the right answer. Therefore, such item should be eliminated post qualitative review as they are not giving true measure of students' ability level.

Items	# Correct (Upper group)	# Correct (Lower group)	Difficulty	Discrimination
Item 1	4	2	0.67	0.4
Item 2	5	5	1	0

Item 3	4	4	0.8	0
Item 4	4	1	0.46	0.6
Item 5	5	2	0.6	0.6
Item 6	5	3	0.73	0.4
Item 7	4	1	0.46	0.6
Item 8	4	2	0.53	0.4
Item 9	1	3	0.4	-0.4
Item 10	4	2	0.53	0.4

Table 2. Difficulty level and discrimination index of items

Distractor analysis

It is conducted to evaluate the efficiency of each distractor for a multiple-choice question. Ideally, all distractors should be equally plausible to students who do not know the right answer. It is advisable to eliminate distractors that are never chosen which means they are not working. In Example 3 that we have discussed, option A, C and D are implausible and rarely chosen by students. Such distractors should be replaced by the options shown in Example 4.

Item 1	**Group**	**A**	**B***	**C**	**D**	
	High	10	40	5	5	Difficulty = 0.6
	Low	15	20	0	5	Discrimination = 0.57
	Total	**25**	**60**	**5**	**10**	
Item 2	**Group**	**A**	**B***	**C**	**D**	
	High	5	5	0	35	Difficulty = 0.15
	Low	20	10	0	25	Discrimination = -0.14
	Total	**25**	**15**	**0**	**60**	

Table 3. Distractor analysis of items

Table 3 shows number of correct responses for each option in item 1 and item 2. For item 1, correct answer is option B. 60 students have attempted it correctly. Therefore, difficulty level is 0.6. Greater number of students in the high ability group have attempted this correctly, hence the discrimination index is positive 0.57. It will be a good idea to investigate Option C since there are very few checks on these options and include better quality distractors. For item 2, while option B is the correct answer, majority of the students have chosen option D as the right answer. Only 15 students could attempt it correctly, this is a very difficult item with difficulty level equal to 0.15. Greater number of students in the lower ability group have attempted this item correctly as compared to the higher ability group, therefore the discrimination index is negative here. All the options need to be relooked at as students seems to randomly choose the correct answer in this case.

Point biserial correlation

It is a statistic used to estimate the degree of relationship between a naturally occurring dichotomous nominal scale and an interval (or ratio) scale (the total score obtained by the examinee in the test). It is symbolized as rpbi. It thus provides an index of whether examinees who get the item correct are scoring highly, which is the hallmark of a good item.

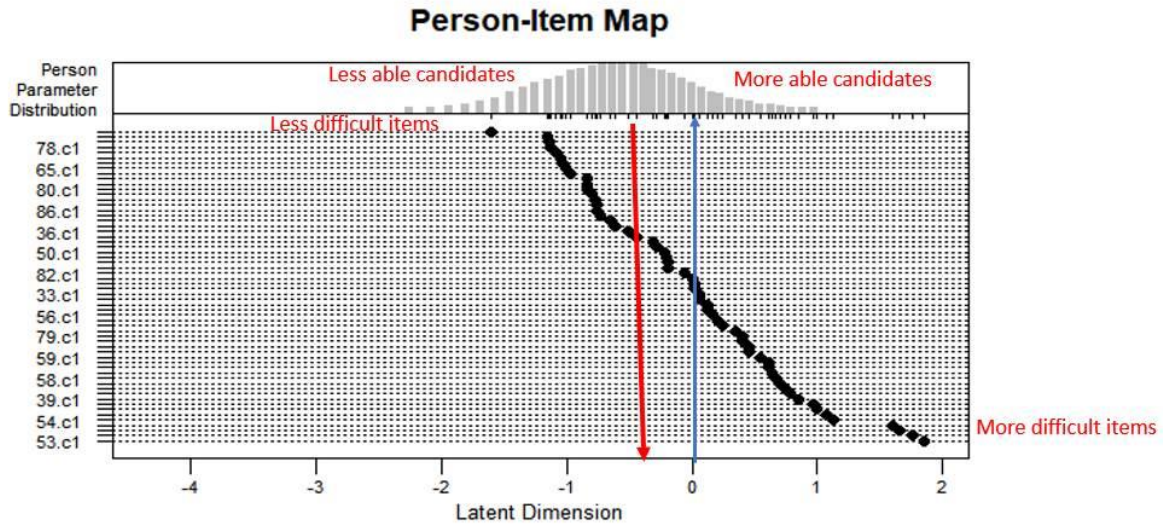
- rpbi can range from 0 to +1.00 if the two scales are related positively and from 0 to -1.00 if the two scales are related negatively.
- A rpbis of 0.0 indicates that there is no correlation, which means that there is no relationship between the score in an item and the total scores. It implies that the item is providing no information, and item responses are essentially random with respect to total scores.
- As rpbis increases, it indicates a stronger relationship between the score in an item and total score. It implies that the item discriminates well among the examines at least in terms of the way the overall test discriminates.
- A negative rpbis is undesirable. It implies that there is an inverse relationship, namely that low-scoring examinees performed better on the item than high-scoring examinees did.

Item Person Map

This map summarizes the quality of question paper by placing the difficulty of the assessment items on the same measurement scale as the ability of the examinees. Thus, it provides a comparison of ability level of examinees and difficulty level of assessment items to better understand the efficiency of the question paper. In other words, this map helps to analyse if the question paper catered well to the group of examinees it is offered to and whether the question paper could measure ability level of all the examinees with high degree of validity and reliability.

The upper section shows the distribution of the measured ability level of the examinees from most able at the right to the least able at the left. The lower section shows the difficulty level of the assessment items from the least difficult items at the left to the most difficult items at the right. Ideally, the peak of the bell curve (upper section) should map to the average difficult level as shown by the red arrow. Also, there should be good range of assessment items belonging to low and high

difficulty levels as shown by blue arrow. There should be approximately equal number of items in left and right side of the blue arrow.



Summary of CTT

Item parameters	**Interpretation**
Difficulty level	High: Less than 0.3; Moderate: 0.3 to 0.7; Low: More than 0.7
Discrimination index	Poor: Less than 0.2; Good: More than 0.2
Distractor analysis	All distractors should be plausible
Point biserial correlation	Undesirable: -1 to 0.00; Desirable: >0.00 to +1

** The range may vary for each assessment purpose.*

Limitations of CTT

One main shortcoming is that the estimates on difficulty level and discrimination index are sample dependent, and this dependency reduces their utility. CTT is most useful when the student sample is similar to the student population for whom the test is being developed. Secondly, student's ability cannot be judged just based on the number of items answered correctly, rather the item attribute, such as its difficulty level, should also be taken into account. In Table 1, both Jadeja and Hardik have attempted different set of six items correctly. However, since the difficulty level of items are not the same, having similar percentage correct does not ensure similar ability levels. Item response theory overcomes both these limitations.

Conclusion

This article provides a brief introduction to Classical Test Theory (CTT) Basic concepts and interpretations of this model is described. CTT help us to understand more about the quality of our question paper by finding difficulty level, discrimination index of items and analysing the distractors of multiple-choice questions.

One of the important applications of CTT is the selection of quality items. The final selection of items will depend on the information they contribute to the overall information provided by the question paper. Item characteristics curve helps to determine the contribution of each test item to the test information function independently of other items in the test. Good quality items are retained while the poor-quality items are reviewed and modified. A balanced question paper should have a good range of items with low, moderate, and high difficulty levels. All the items should discriminate well between low and high ability level students. Further in order to use the true power of multiple-choice questions, it is essential to use plausible distractors.

References

1. Downing, S.M., & Haladyna, T.M. (Eds.) (2006). Handbook of test development. Philadelphia: Taylor & Francis.
2. Officer, C. P. (2016). Introduction to Classical Test Theory with CITAS.

[i] This article is authored by Shilpi Banerjee. She works as Assistant Professor in the School of Continuing Education at Azim Premji University. She can be reached at shilpi.banerjee@azimpremjifoundation.org.
This article can be cited as-
An introduction to classical test theory, Assessment resources, 2021, Azim Premji University

An Introduction to Classical Test Theory