EasyChair Preprint

№ 4339

Classical Test Theory - a Write Up

Layah Liz Jacob

October 10, 2020

CLASSICAL TEST THEORY – A WRITE UP

Layah Liz Jacob

Charles Spearman is known as one of the founders of classical test theory. The theory has been in use for more than a 100 years with the main understanding that errors might always be a part of any measurement (Traub, 2005). Classical Test Theory can be defined as a psychometric theory used for predicting the consistency of items on a test and quantifying measurement errors. This theory can be used to identify the relationship between the true score and observed score (Frey, 2018).

The major assumption of CTT is:

$$X = T + E$$

X= Raw Score

T= True score

E = Random error

"Raw scores are a transient description of never to be re-encountered situations" (Wright, 2005) neither are they measures. It can be understood as an appropriation score you obtain at that particular time and at that particular condition. And re-obtaining the measurement during a different time could give you a different raw score. And this is the purpose of the true score. The true score is an aggregate of all scores the person would obtain upon retaking the test multiple times. This leads us to the random error. The random error follows a bell- shaped curve (normal distribution) and hence we assume the mean to be 0.

$$E(X) = T$$

Then

$$E(E) = 0$$

To put it shortly, the assumptions of Random Errors are:

- Normally distributed
- correlated to each other (might fluctuate and may not remain consistent through all the attempts on the same test)
- Also uncorrelated to true score
- Mean of error variance is 0

Classical Test Theory covers that every item on the scale will equally contribute to the overall score. For example: The General Health Questionnaire (GHQ- 28) is a scale measuring depression, anxiety, somatic complaints and other discomforts. The scale consists of 28 items (28 statements which the test taker marks) to calculate the total score. Now, according to the classical test theory, all the 28 items on this scale will be contributing equally to the total score obtained. It is also assumed that the response options are equal interval. That is the size of the unit remains same. Furthermore, the expected error on the measurement remains the same for score obtained. For example: For the test takers of GHQ-28, whether the person has a low score of 0 or a high score of 84. The error remains the same irrespective of the score obtained.

Theories within Classical Test Theory discussed within the book (Kline, 1976):

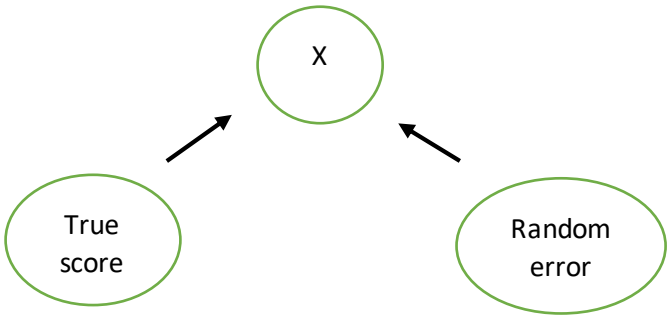
- Domain sampling theory: There are limitless amount of items that could be chosen for a test and hence, the items presented on a test are just a few out of the larger domain.

- The parallel test theory: The parallel test theory assumes that tests with parallel items but differing domains would have comparable true scores but differing error scores. For example: The GHQ-28 was discussed previously. Similar to this is the GHQ 12 questionnaire used as a screening tool for non-specific psychological morbidity. Though the items in the questionnaire are from different domains, the scores a person obtains on both the questionnaires would be similar, but the error scores will differ.

- Theory of true and error scores: It is the foundation of reliability theory and assumes that there are two components in every test. The true level (true score) the test taker is at on the overall domain of the test and the random error.

Classical test theory and reliability

Reliability is about the consistency of the test scores. That is, if an individual was to take the general health questionnaire three times, his result should be the same and not differ vastly each time he takes the test. Now, if we take the average of all the scores the individual obtained upon taking the test different times, the deviance of the scores from the mean value (average of all the scores obtained) would provide us with the variance. So therefore, a test which has good reliability will provide us with correct results, upon taking the test every time. Now connecting reliability to classical test theory



Reliability can be calculated by true score variance and total variance. Since classical test theory is not a measurement theory, we cannot use validity of a test within this concept. The true score in classical theory would not go in hand with any construct score (Lord, Novick & Birnbaum, 1968).

I – variance (error)

One major break-through is realizing that when classical test theory explains about administering the same test to a person many times for the reliability (scores obtained in each administration), this is similar to administering the same test to multiple individuals. (Allen &Yen, 2002) explains how the psychometric property of reliability can be standardized by administering the same test to multiple individuals.

Classical Test theory and assessment of test items

For descriptive statistics, when considering the mean and variance, a good mean should be near the center of distribution (showing an average value) with higher item variability. Say, we administer the personality questionnaire to the general population. For the items on the personality scale, if the items have less variability, we may not be able to use the result to make inferences about the person. Similarly, if the mean responses of individuals are normally distributed, it would be more predictive of the information we would like to obtain from the test than if the mean was positively or negatively skewed.

Unlike scoring items on a continuum, for a dichotomous item such as a multiple choice with two possible scores or a distractor, the mean would be calculated according to the number of individuals who correctly answered the item (denoted as p). And variance would be calculated based on how many individuals negatively answered or did not attempt the item (denoted as q)

For dichotomous items

Mean = *p* (correctly answered)

Variance = *p X q* (q is incorrectly answered or not attempted)

Standard deviation = *square root (p X q)*

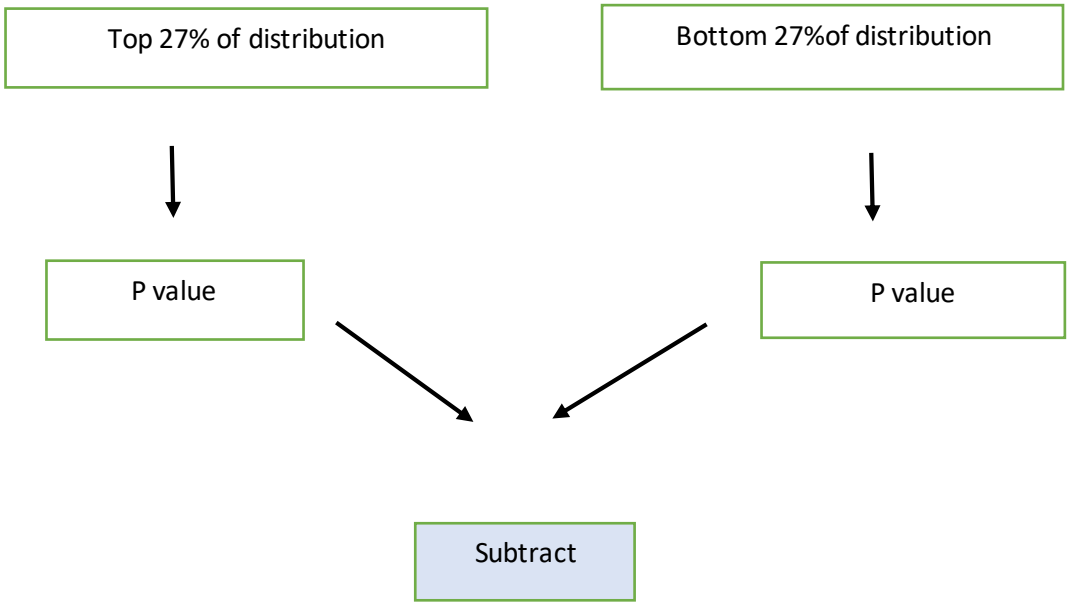
The P value is also denoted for the level of difficulty of an item. Such as for a scale measuring the intelligence quotient, we know that the items on that test have varying difficulty. Some might be easy items that can be answered by majority of the population and some has a high difficulty level, which can be passed by a small percentage of test takers. Therefore, in classical test theory, easy items have high p values and low p values are regarded as low difficulty level items. The optimum value is 0.50 which means that there is a 50% chance of fail or pass for that item, which ensures proper differentiation of the test taker's ability on that test.

The p values can also calculate the discrimination indices. The higher the discrimination value (D), the better it discriminates between the items. For the purpose of calculating the discrimination indices, the test-takers are divided into two groups: the high performers and the lowers performers. Ideally, this segregation is done by dividing the top 27% and bottom 27% (Cureton, 1957). After which the P of the high and low performers is subtracted to arrive at the discrimination index.

For calculation of discrimination index

High scorers

Low scorers



Related to the discrimination indices is the Pearson's correlation coefficient.

For **item to total correlations**, it is dependent on the number of items in the test. For tests with higher number of items, there would be little influence and the corrected score is used to tests with low items. Item to total correlations focus on how response in an individual items influence the total score.

Relatively, **the item-to-criterion correlations** are used to see if the items discriminate between the natures of participants who are test –takers. For example: a questionnaire on nutrition is administered to a group of nutritionists and to a group of mothers. Based on the responses, if it shows high correlations then the measurement is a good indication of discriminating between the population of test-takers, of those who are professionally aware of nutritional food and mothers.

In cases of **differential item weighing**, unlike in a unit weighing system, the items on the test may carry more weights to the total score. That is some items will be given more influence in determining the total score, compared to other items on the scale which may have lesser influence. The techniques used for this method include using reliability of items where items with good reliability are given more weight while the other items are assigned less weight. Or regression may be used wherein by regressing the criterion on items, the item weights are

determined. Additional two methods are by using factor analysis and item to total correlation coefficients.

Limitations of classical test theory

- Does not help with statistical estimations and hypothesis
- Exclusive focus on errors in measurement
- The parameters of classical test theory are often only applicable to the limited population. For e.g.: an IQ assessment given to the general population will not be applicable when generalizing to a student population (Franchignoni et al., 2011).

References

Allen, M., & Yen, W. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Cureton, E. E. (1957). The upper and lower twenty-seven per cent rule. *Psychometrika, 22*, 293–296. https://doi.org/10.1007/BF02289130

Frey, B. (2018). Classical Test Theory. *The SAGE Encyclopedia Of Educational Research, Measurement, And Evaluation*. doi: 10.4135/9781506326139.n112

Franchignoni, F., Ferriero, G., Giordano, A., Sartorio, F., Vercelli, S., & Brigatti, E. (2011). Psychometric properties of QuickDASH – A classical test theory and Rasch analysis study. *Manual Therapy, 16*(2), 177-182. doi: 10.1016/j.math.2010.10.004

Kline, P. (1976). *Psychological testing* (pp. 92-93). London: Malaby Press.

Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Traub, R. (2005). Classical Test Theory in Historical Perspective. *Educational Measurement: Issues And Practice, 16*(4), 8-14. doi: 10.1111/j.1745-3992.1997.tb00603.x

Wright, B. (2005). A History of Social Science Measurement. *Educational Measurement: Issues And Practice, 16*(4), 33-45. doi: 10.1111/j.1745-3992.1997.tb00606.x