International Journal of Research and Review
Vol.7; Issue: 5; May 2020
Website: www.ijrrjournal.com
Short Communication *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Classical and Modern Methods in Item Analysis of Test Tools

Z. A. Ashraf[1], Jaseem K[2]

[1]Associate Professor of Statistics, Govt. Arts & Science College, Calicut, Kerala
[2]Assistant Professor of Clinical Psychology, IMHANS, Kozhikode

Corresponding Author: Z. A. Ashraf

ABSTRACT

Item analysis is a process which examines the examinee's responses to individual test item (questions) in order to assess the characteristics of those items and of the test as a whole. Item analysis is especially valuable in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration. There are different approaches in item analysis. In all approaches general goal is to arrive at tests having minimum items that will yield necessary degree of reliability.

Classical Test Theory (CTT) and Item Response Theory (IRT) are the two broad methodology of test theory. In CTT framework, using the selected sample, some indices like item difficulty, item discriminations are calculated for each item. The quality of item will be decided on the basis of these values. In IRT, which is also known as modern test theory, the item characteristics are decided based on values taken by the parameters of the model chosen for the item response. The parameters are estimated from the samples chosen for the item analysis. Based on the values taken by the parameters for each item, the quality of the item will be decided.

This paper tries to explain the item analysis procedure in both classical and Item Response Theory frameworks

Key words: - Item analysis, Classical test theory, Item Response theory, item difficulty, item discrimination

1. INTRODUCTION

Many different facets are involved in the process of test construction. One must go through a series of steps in order to create a test that suits best for assessing the trait to be measured. These steps include test conceptualization, test construction, test try out, analysis and revision. All these come under the process of item analysis.

Item analysis is a process which examines the examinee's responses to individual test item (questions) in order to assess the characteristics of those items and of the test as a whole. Item analysis is especially valuable in improving items which will be used again in later tests, but it can also be used to eliminate ambiguous or misleading items in a single test administration.

French (2001) considers item analysis as statistical procedure to analyze test items that combines methods used to evaluate the important characteristic of test items.

Within the item analysis all the possible test items are subjected to stringent series of evaluation procedures, individually and within the context of the whole test. Then sufficient samples of subjects are to be collected from the targeted population (for whom the test is made) for the process of item analysis.

In CTT framework, using the selected sample, some indices like item difficulty, item discriminations are calculated for each item. The quality of item will be decided on the basis of these values. The quality of the test as a whole will be determined on the basis of some coefficients for reliability and validity

In IRT framework the item characteristics are decided based on values taken by the parameters of the model chosen for the item response. The parameters are estimated from the samples chosen for the item analysis. Based on the values taken by the parameters for each item, the quality of the item will be decided. This paper tries to explain the item analysis procedure in both classical and Item Response Theory frameworks

2. Classical Item analysis

The conventional method of test construction and its interpretation is Classical Test Theory (CTT) methods. Classical test theory methods are widely used in almost all areas of social, behavioral, medical and many other fields of study. The conceptual foundations, assumptions, and extensions of the basic premises of CTT have allowed for the development of psychometrically sounds scales over several decades.

In both classical and modern test theory, general goal is to arrive at tests having minimum items that will yield necessary degree of reliability. The process applied to get a set of good items to measure a particular trait is known as item analysis.

The most common statistics reported in an item analysis are the item difficulty, which is a measure of the proportion of examinees who responded to an item correctly, and the item discrimination, which is the measure of how well the item discriminates between examinees who are having different level of trait (inherent capacity) and those who are not.

Both item difficulty and item discrimination are item statistics; for each item one can find indices for item difficulty and item discrimination. There are some statistics such as reliability coefficient which are test statistics rather than item statistics. It means they will give some information on the tests as a whole rather than item.

2.1 Item Difficulty in CTT

Item difficulty index is the proportion of number of examinees who get an item correct to the total number of examinees (Ansthasi and Urbina, 2004). It means item difficulty is a measure of the proportion of examinees that answered the item correctly. The item difficulty index, symbolized as *pi* for an item *i,* can be computed simply by dividing the number of test takers who answered the item correctly by the total number of students who answered the item. Therefore easier the item, larger the proportion will be. Hence a numerical problem correctly answered by 30 percentage of subjects (p=0.30) will be considered harder than an item answered by 75 percentage of subjects (p=0.75) .the value of *pi* ranges from 0 to 1. It takes 0 when no examinees answered the item correctly and 1 when all examinees answered the item correctly. However if pi approaches either end of the spectrum less information about the group is revealed. The item difficulty index is also known as item endorsement index. In some literature 1-*pi* is called item facility. For item with one correct alternative worth a single point, the item difficulty is simply the percentage of students who answer an item correctly. In this case, it is also equal to the item mean. When an alternative is worth other than a single point, or when there is more than one correct alternative per question, the item difficulty is the average score on that item divided by the highest number of points for any one alternative. In our discussion we will consider only items having one correct alternative.

In an achievement test, item difficulty is relevant for determining whether students have learned the concept being tested. It also plays an important role in the ability of an item to discriminate between the students who know the tested material and those who do not know. The item will have low discrimination. If it is so difficult that almost everyone gets it wrong or guesses, or so easy that everyone gets it right. Hence all standardized test have generally been designed to elicit maximum differentiation among individuals at all levels. For this the difficulty index of each

item will be computed and items falls outside a desirable value of *pi* will be rejected from the test tool.

When items are dichotomous, i.e. when items are scored either 0 or 1, simplest index of its difficulty is its mean item score. There are different criteria for fixing this desirable value. Different authors have suggested different criteria based on their arguments for determining the ideal value of *pi* . Generally we can say that the optimum value of item difficulty is decided by the test developer based on the objectives of measurements.

Usually an item with difficulty index nearer to 0.5 is treated as a good item, as 0.5 is the value for item difficulty where 50% of subjects responded correctly. But there are no strict rules in deciding the admissible variation from 0.5. Kaplan and Succuzzo (2001) states that, for most tests, items in the difficulty range of 0.30 to 0.70 tend to maximize information about the difference among individuals. Chung (1985) state that a good item usually has a difficulty that lies between 40% and 70%. All these are only some thump rules.

Kaplan and Succuzzo (2001) put forward a set of ideal values for item difficulty by considering guessing factor. They suggest that optimum difficulty level for items is usually about half way between 100% of respondent getting the item correct and the level of success expected by chance alone. Thus the optimum level of a four-choice item is 0.625. The optimum item difficulty index for n item with three-choice will be 0.666 and that of an item with only two-choice will be 0.75. Guilford (1982) suggests some correction to proportion and presented a table to facilitate the correction for ideal item difficulty.

There is no strict criterion for deciding the allowed variation of item difficulty from the value. It is decided by the test developer with his personal judgment based on the need and situation. Fareed and Ashraf (2008) used proportion test using the statistic.

$$Z = \frac{p_i - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Where p_i is the item difficulty of the i^{th} item, p_o is the optimum value of the item difficulty and *n* is the sample size, to decide the significance of item difficulty index.

Besides the item difficulty of each item one can define the item difficulty of the whole test as the average test item difficulty of entire items (French, 2001). According to Cohen et al. (1996) the optimal average item difficulty is approximately 0.50.

2.2 Item Discrimination in CTT

Item discrimination refers to the degree to which an item differentiates correctly among test takers in the behavior that the test is designed to measure (Anasthasi and Urbina, 2004). It is an index that measures how well an item is able to distinguish between examinees who are knowledgeable and those who are not, or between masters and non-masters. Cohen and Swerdlik (2005) define it as a statistic designed to indicate how adequately a test item separates or discriminates between high and low scorers.

In test construction theory there are many indices to determine the property of item discrimination. Some of these assume normal distribution of the underline trait. Despite of different procedures, most of the item discrimination indices provide closely similar result (Anasthasi and Urbin, 2004). A common practice in computing item discrimination is to compare the proportion of cases that pass an item in contrasting criterion groups. This method compares people who have done very well with those who have done very poorly on a test (Kaplan and Saccuzzo, 2001). In this method there are three simple steps in calculating item discrimination index D*i*. First, those who have the highest and lowest overall test scores are grouped into upper and lower groups. The upper group is made up of the 25% to 33% who are the poorest performers' (have the lowest overall test scores). Cureton (1957) suggested to use the top and bottom 27% of the distribution in creating these extreme groups, as this is the critical ratio that separates the tail from the mean of the standard normal distribution of

response error. Step two is to examine each item and determine the *p* levels for the upper and lower groups, respectively. Step three is to subtract the *p* levels of the two groups; this provides the D*i*.

Another way to find the discrimination index of items is to find the correlation between performance on an item and the performance on the total test (Kaplan and Saccuzzo, 2001). One situation , which occurs frequently in item analysis , is when the test developer is interested in how closely performance on a test item scored 0 to 1 is related to performance on the total test scores (Crocker and Algina, 1986). A simplified formula used in this situation is point biserial correlation, which are defined correlations between item score and total score.

This statistic looks at the relationship between examinees performance on the given item (correct or incorrect) and the examinees score on the overall test. For an item that is highly discriminating, in general the examinees who responded to the item incorrectly also tend to do poorly on the overall test. Item discrimination indices must always be interpreted in the context of the type of test, which is being analyzed. Item with low discrimination indices are often ambiguously worded and should be examined. Items with negative indices should be examined to determine why a negative value was obtained. Test with high intensity consistency consists of items with mostly positive relationship with total test score.

Chung (1985) states that a good item usually has a discriminating index that exceeds 0.40.There are various other methods also for computation of discrimination index of test items. Obviously in some situations, because of the scoring of the variables one techniques may be more appropriate than others (see Croker and Algina, 1986).

3. Item Response Theory

Item Response Theory (IRT) is an area of test theory which provides probabilistic approach to overcome some of the limitations of classical methods. IRT is a statistical technique involving models expressing the probability of a particular response to a scale item as a function of the ability (more precisely trait) of the subject. IRT models are widely used in the preparation and standardization of test items. For more basic discussion on IRT see Chang and Reeve (2005) and Baker (2001).

In IRT the term trait means the characteristic of the subject to be measured, which is latent or unobservable. This variable is often something intuitively understood like intelligence. When one says somebody is highly intelligent or very poor in intelligence the listener has some idea as to what the speaker is conveying. Although this type of variables are easily understood and one can list its characteristics, they cannot be measured directly as one can measure height or weight.

Kapaln and Saccuzzo (2001) defines trait as relatively enduring disposions (tendencies to act, thin k or feel in a certain manner in any given circumstances) that distinguishes one individual from another.

Although Item Response Theory (IRT) methods have been in existence for over three decades, only recently have they begun to achieve widespread popularity in psychological assessment. One very practical reason for this belated popularity is the fact that IRT technique tends to far more computationally demanding than methods of test construction and scoring that are based on classical test theory.

In the fields of education and psychology, now IRT methods are increasingly being applied to personality, attitude, aptitude and similar inventories containing items that are scored in a dichotomous fashion, such as checklists and inventory type items. Recently increased attention has also been devoted to IRT models that are capable for analyzing items that are rated using either ordered-category scales such as Likert- type or unordered, nominal scales. Nowadays in medical research also IRT technique is widely used

With item response theory the test developer assumes that the response to the item on attest can be accounted for by latent traits. Indeed most applications of the theory assume that a single latent trait account for the response to items on a test (Crocker and Algina, 1986). Generally trait is a single entity or multiple entity. But in practical situations it is considered as a single trait and is measured through test.

Latent trait refers to a statistical construct; there is no implication that it is a psychological or physiological entity with an independent existence. In cognitive tests, the latent trait is generally called the ability measured θ denote the latent to be measured based on a test which consists of a finite number of items. People at higher levels of θ have a higher probability of responding correctly to n item. Obviously, as θ is a latent construct, it cannot be directly observed or measured, and thus tests do not measure it in an absolute sense, like a ruler measures length. Instead what can be determined is relative positions of individual test takers on the θ continuum

3.1 Item Characteristic Curve

In an item response theory approach, for each item on test there will be a curve which characterizes the nature of responding to an item, which is known as Item Characteristic Curve (ICC). It describes the probability of getting each particular item right given the ability level of each test taker. Baker (2001) treats the ICC as the basic building blocks if item response theory; all the other constructs of the theory depend upon this curve (Baker, 2001). Croker and Algina (1986) consider ICC as the central concept of IRT.

Let θ be the latent trait and $pi(\theta)$ be the probability that an examinee with trait θ will give correct answer to the items I, then $pi(\theta)$ can be plotted as the function of θ and the resulting S –shaped curve will give Item Characteristic Curve. i.e. the *ICC* of *i*th item is the graph of $pi(\theta)$ verses θ. A typical ICC is given in the figure. Here θ is represented on X-axis and $pi(\theta)$ is represented on Y-axis.

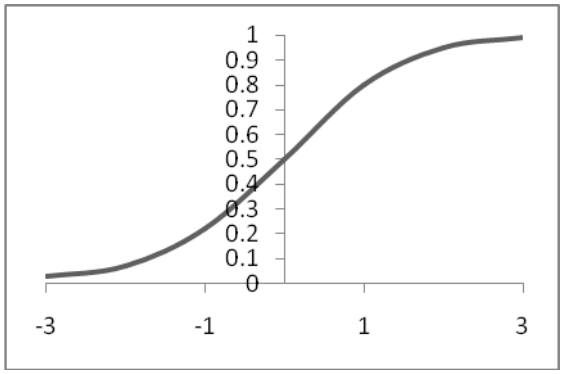

Figure 1: Item Characteristics Curve (ICC)

Since $p_i(\theta)$ increases with θ and has values ranges from 0 to 1, $p_i(\theta)$ can be assumed to have the nature of cumulative distribution function (cdf) with asymptotic, in the sense that $p_i(\theta)$ never touches its lower and upper ends; i.e.., no person has either no ability or complete ability to bring to bear on a given item (Henson, 1999). Baker (2001) point out the two technical properties of an ICC that are used to describe it as item difficulty and item discrimination.

3.2 Item Difficulty Parameter in IRT

In all IRT models it involves certain number of parameters. These parameters have its own psychical importance for making decision on items. In IRT the difficulty of an item describes where the item functions along the ability scale. For example an easy item functions among the low ability examinees and a hard item functions among the high ability examinees. This means that difficulty can be considered as the location index. It analogs the item difficulty index defined in classical approach, that indicates the proportion of numbers of examinees who get an item correct to the total number of examinees. Usually the item difficulty parameter is denoted as b_j for j^{th} item.

In an ICC, parameter b_j defines the location of the curve's inflection point along the X-axis. If two parameter logistic model is considered for $p_i(\theta)$ as in equation

$p_{ij} = P\{ Y_{ij} = 1/ \ \Theta = \theta \} = [\frac{1}{1+e^{-a_j(\theta - b_j)}}]$

the parameter b_j stands for item difficulty index of an item j. the figure gives ICC of a

2PL model for different values of *bj*. Lower the value of *bj* will shifts the curve left and higher the value of b_j will shifts the curve right. The b_j does not affect the shape of the curve.

When b_j =0, the probability of correct response to an item is 0.5 for those individual have their trait as 0. If bj is greater than 0 it indicates the item is harder. One has to choose items with a desirable level of item difficulty. Items with difficulty index near to 0 will give more information on latent trait. Generally one can choose an item with difficulty index lies between-0.5 and 0.5.

3.3 Item Discrimination Parameter in IRT

The item discrimination indicates the extent to which success on an item corresponds to success on the whole test. It describes how well an item can differentiate between examinees having the trait below the item location and those having the trait above the item location. In ICC the item discrimination property essentially reflects the steepness of the curve in its middle section. The steeper the curve, the better the item can discrimination.

In the case of two parameter logistic model is considered for $pi(\theta)$

$$p_{ij} = P\{ Y_{ij} = 1/ \ \Theta = \theta \} = [\frac{1}{1+e^{-a_j(\theta - b_j)}}]$$

The parameter a_j stands for item discrimination index for an item j. From the curve one can say that the change in the values of aj changes the shape of the item response function and does not change its location. Also it is noted that higher values of aj will give more information on item j that item.

Normally the value of a_j will be positive. If a_j is negative, it results in monotonically decreasing item response function (Rudner.1998). it means that people having higher θ will have lower probability of correctly responding to the item and people having lower θ will have higher probability to answer the item correctly.

Theoretically items with higher values of a is thought to be better items. But if the value of very high as Masters (1988) pointed out, it can be a symptom of a special kind of measurement disturbance introduced by that item that gives persons of high ability a special advantage over and above their higher abilities. Generally an item with a value of $0.75 \leq aj \leq 1.75$ will be accepted to the final test tool.

4. Advantage of IRT over CTT

IRT methods have many advantages over CTT based method of test development and scoring. Item analysis techniques within the classical test theory approach are generally crude in nature. The common method is to determine the values of some pre-defined statistics and based on these values decision is taken to reject or accepts an item, without considering the nature, form or characteristics of the population.

Consistent with its origins in tests of educational achievement and aptitude, IRT methods are already well known among educational researchers. Item response theories have gained popularity due to their promise to provide greater precision and control in measurement involving both achievement and aptitude instruments (Henson, 1999). IRT has also achieved wide use among industrial and organizational psychologist, in part due to its ability to quantify the degree to which test exhibit consistent bias with respect to race, sex, age or other demographic factors.

5. REFERENCES

1. Anastasi, A., & Urbina, S. (2004). Psychological Testing, 7th Edition, Pearson Education, New Delhi.
2. Baker, Frank B (2001), The Basics of Item Response Theory. Second Edition.,ERIC
3. Chang, C. H., & Reeve, B. B. (2005). Item Response theory and its applications to Patient reported outcome measurement. Evaluation & the Health professions Vol. 28 No. 3 264-282.
4. Cohen, R. J., & Swerdlik, M. E. (2005). Psychological Testing and Assessment: An introduction to tests and measurement. McGraw Hill- New York.
5. Cohen, R. J., Swerdlick, M. E., & Philips, S. M. (1996). Test development in Psychological testing and Assessment: an

Introduction to tests and Measurement, 3rd Edition, Mountain view, CA, Mayfield.
6. Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Toronto: Holt, Rine Hart, and Winston, Inc.
7. Cureton, E. E., (1957). The upper and lower twenty seven percent rule. Psychometrika 22, 293-296
8. French, C.L. (2001). A review of classical methods of Item Analysis. Paper presented in annual meeting of the Southwest Educational Research Association, New Orleans, February 1-3, 2001. ERIC.
9. Guilford, J. P. (1982). Psychometric Methods. Mc Graw Hill , New York
10. Henson, R. K., (1999). Understanding the one parameter Rasch Model of Item Response Theory. ERIC.
11. Kaplan, R. M., & Saccuzoo, D.P. (2001). Psychological Testing - Principles, Applications and Issues -5th Edition, Wadsworth, Stanford.
12. Masters, G., N. (1988). Item Discrimination: When more is Worse, Journal of Educational Measurement, 25, 1,15,1988.

How to cite this article: Ashraf ZA, Jaseem K. Classical and modern methods in item analysis of test tools. International Journal of Research and Review. 2020; 7(5): 397-403.

******