IOSR Journal of Research & Method in Education (IOSR-JRME)
e-ISSN: 2320–7388,p-ISSN: 2320–737X Volume 3, Issue 1 (Sep. –Oct. 2013), PP 01-08
www.iosrjournals.org

Best Split-Half and Maximum Reliability

Satyendra Nath Chakrabartty[1]
(Prof, Galgotias Business School, India)

Abstract: *The paper addresses an iterative method by which a test can be dichotomized in parallel halves and ensures maximum split-half reliability. The method assumes availability of data on scores of binary items. Since, it was aiming at splitting a test in parallel halves, no assumption was made regarding form or availability of reference test. Empirical verification is also provided. Other properties of the iterative methods discussed. New measures of degree of parallelism given. Simultaneous testing of single multidimensional hypothesis of equality of mean, variance and correlation of parallel tests can also be carried out by testing equality of regression lines of test scores on scores of each of the parallel halves, ANOVA or by MahalanobisD^2. The iterative method can be extended to find split-half reliability of a battery of tests. The method thus provides answer to much needed question of splitting a test uniquely in parallel halves ensuring maximum value of the split-half reliability. The method may be adopted while reporting a test.*

Key words:*Split-half reliability, Parallel tests, Dichotomization, Battery of tests*

I. Introduction:

Reliabilityofatesthasbeendefinedintheliteratureinvarious ways aiming at to measure consistency, precision, repeatability, trustworthiness, etc. of a test. The terms are operationalized differently resulting in confusions since each approach leads to a different value of reliability of the same test. Further confusion stems from the fact that reliability, conceptualized as consistency, consists of both absolute consistency and relative consistency (Safrit, 1976). Absolute consistency concerns with consistency of scores of individuals, whereas relative consistency concerns with the consistency of the positions or ranks of individuals in the group (Weir, 2005). Strictly speaking, reliability is a property of the scores of a measure rather than the measure itself and thus is sample dependent. However, reliability coefficient of a test can be viewed as information about the dependability of the measurement by a number lying between zero and one. Dawis (1987) viewed that reliability is influenced by the instrument used, the sample and other related features. Estimate of reliability of a test may vary from one sample to other.Vacha – Hasse (1998) opined that if a test is administered 100 times, it may yield 100 different reliability coefficients. Webb et. al.(2006) explained that even if a test is administered to the same sample for more than once, individual scores and rankings may vary. Such situation may arise depending on value of reliability, type of reliability, heterogeneity of sample, item characteristics of the test, etc.

Split half method is a popular method of assessing reliability of a test primarily for the advantage of single administration of the test and use of one sample. Stages involved are

* Single administration of a test to a sample
* Splitting the test i.e. dividing the items of the test in halves so the two sub-tests are parallel
* Correlate scores on one half of the test with scores on the other half of the test.
* Find reliability of the test as the correlation between two parallel tests as proved by Lord and Novick (1968).

However, researches like (Kaplan and Saccuzzo, 2001) recommend finding split-half reliability of the entire test using Spearman-Brown formula

$r_{tt} = \frac{2r_{gh}}{1+r_{gh}}$ where r_{tt} denote estimated reliability of the entire test and r_{gh} is the correlation between two halves which are assumed to be strictly parallel.

Large number of split-half reliability coefficients have been developed historically that relax the assumption of parallel tests (e.g., Flanagan's formula, which requires only the essential tau equivalence assumption). In fact, Cronbach's alpha is based on a weaker assumption, that of essential tau equivalence. Webb, et.al.(2006) expressed that the assumption of strictly parallel items is too restrictive.

There could be several ways of splitting a test. Each method of split-half gives a different value of reliability. It is acknowledge that a test consisting of 2n-number of items can be split in half in $2n_{c_n}$ number of ways and that each method of splitting the test in halfs will result in a distinct value of split-half reliability even for the same sample. In this context, it merits mention that it is of crucial importance to identify the way of splitting a test in two parallel halves.

Guliksen (1987) observed that for a test with 40 items, the first part with 20 items and the second part with last 20 items may not be parallel since responses from the first half may be systematically different from responses

in the second half due to increasing level of item difficulty and fatigue. In addition, speed of work may be different in the two halves thus formed. Other strategy could be splitting a test on the basis of odd and even items. However, it is well known that during the stage of development of test, there is no accepted rule regarding numbering of items. Hence, odd-even splitting may not ensure that the two parts are parallel. Moreover, if a different version of a test is prepared by keeping the same set of items but by assigning different numbers to the items haphazardly, the odd-even reliability may be changed despite the fact that the new version is parallel to the original test. One can also choose random halves of a test and have different values of reliability for different method of splitting of the test. Rudner et.al. (2002) observed that split-half reliability is a function of how the test was split. A solution to the problem is provided by Cronbach's alpha which is interpreted by many researchers as the average of all possible split-half correlations (Cortina, 1993). Cronbach's alpha also assumes that average covariance among non- parallel items is equal to the average covariance among all parallel items. Limitations of Cronbach's alpha have also been reported by Hattie (1985), Eisinga et.al.(2012), Ritter, (2010) .

Thus, it appears there are confusions over the inconsistencies amongst different available ways of splitting a test and it is of crucial importance to identify the way of splitting a test that ensures the two sub-tests are really parallel. Thisstateofaffairsmotivatesaneedto split a test in two parallel halves and estimate split-half reliability fromasingleadministrationofthetest.

Gullicksen defines two tests "g" and "h" are parallel, if it does not matter which test is used
i. e.

(i) True score of an individual in the g-th test and h-th test are same
(ii) Error SD's are same i.e. $S_{e_g} = S_{e_h}$

It is well known that parallel tests have equal mean, variance and correlation. Thus, to know whether two or more tests are parallel, equal means, equal variances and equal covariance are required to be tested empirically.

In a slightly different context, there are needs to construct parallel tests from a pool of items especially in large testing programmes. Considerable literature exists on methods of construction of parallel tests or equivalent forms using mathematical programming. Van der Linden and Luecht (1998) proposed an IRT-based method for constructing strongly parallel tests by matching items on item response functions. They used 0-1 mathematical programming to generate parallel forms from a given pool of items. Construction of parallel tests using 0-1 linear programming were also undertaken by Boekkooi-Timminga(1990), Rao (1985), Salkin(1975), Taha(1975), Wagner(1972), Theunissen (1985), etc. Armstrong and Jones (1992) considered Polynomial algorithms for item matching.

Regarding extent of parallelism, Van der Linden and Luecht (1998) defined that if two tests are exactly parallel, they must have identical moments of $P(\theta)$ for all values of θ and thus they have identical distributions of $P(\theta)$ for all values of θ. However, the strongly parallel index is rather a strict criterion for statistical parallelism. Based on a less stringent criterion, Van der Linden &Luecht(1998) worked with an alternative index called Smirnov statistic *T*. McDonald (1999) defined two test forms as item-parallelif they consist of paired items with identical item parameters andtarget test characteristic curves (TCC) or functions. The definition of item parallelismproposed by Van der Linden and Luecht (1998) and McDonald (1999) is a more stringent constraint for construction of equivalent test forms than are those of classical-related parallelism (e.g., equivalent means and standard deviations of observed scale distributions). Tests are defined to be weakly parallel if their information functions are identical (Samejima, 1977). Tests are strongly parallel if they have the same test length and if they have exactly the same test characteristic function (Lord, 1980). An exact definition of the concept of information is given by Birnbaum (1968, Chapter 17). Here it is assumed that maximum-likelihood estimation is used for subjects' abilities so that the test information function is the sum of the item information function

Similarities of the methods include among others

i) Assumption of availability of target test information function or target test characteristic function for the test(s) to be constructed at some pre-chosen ability level.
ii) Large amount of CPU-time and large number of decision variables in the model

Much research has been carried out to develop approximations which require less of CPU time with reduced number of constraints for construction of parallel tests sequentially or simultaneously. O'hEigeartaigh, et.al. (1985) gives a comprehensive review in this context.

However, following points merit consideration:

- Test information functions are only related to the asymptotic error variance of proficiency estimates on the θ -scale rather than the true score distribution.

- Even though the definition of parallelism involves all m- moments, in a practical sense, the important moments to examine are the first and the second central moments.

- Number of the IRT-based constraints is much greater than that of the CTT-based constraints for automated test assembly

- The IRT-based methods may produce less optimal tests, and thereby less parallel ones as observed by Lin (2008). He concluded that the Classical Test Theory approach performed at least as well as the IRT approaches.

- All testing programs cannot use IRT methods to calibrate item response data because 0 - 1 item response data may be unavailable and the test assembler may have to rely on classical item statistics alone in creating parallel forms

- There are situations under which Classical Test Theory (CTT) item statistics are the only data available to test developers e.g. availability of only item scores or item difficulty values or item discrimination values. Under these situations, it is necessary to develop parallel tests based on CTT item statistics.

Chakrabartty (2011) gave an iterative method based on Classical Test Theory to split a test in two parallel sub-tests with almost equality of means and variances. The method assumed availability of data on scores of binary items. Since, it was aiming at splitting a test in parallel halves, no assumption was made regarding form or availability of reference test

Objectives of the paper was to examine nature of correlation and benefits of such dichotomization given by Chakrabartty (2011)

II. Method:

As per classical theory, two tests "g" and "h" are parallel if

$T_{ig} = T_{ih}$ …… ……. (1)

and $S_{eg} = S_{eh}$ …… ……. (2)

If tests "g" and "h" are parallel, then $\bar{X}_g = \bar{X}_h$ and $S_g^2 = S_h^2$ i.e. parallel tests have equal means and equal variances in terms of observed scores. Chakrabartty (2011) gave the following algorithm to split a real test into parallel halves, 'g' and 'h', that have equal means and equal variances in terms of the observed scores.

Step I. Find item wise total score for each item.

Step II. Sort the item-wise total scores in an ascending order: $S_1, S_2, \ldots\ldots, S_n$

Step III. Choose the item with highest total score and allocate it to the g-th test. The item with second highest total score to be allocated in the h-th test. Put the item with the third highest scores to h-th test and the fourth highest in the g-th test and so on. In other words, allocation of items to be such that the following structure is realised.

test g	test h	difference in elements of 2 sub- tests
S_1	S_2	$S_1 - S_2 \geq 0$
S_4	S_3	$S_4 - S_3 \leq 0$
..	till all the items are accommodated either in the g-th or h-th test.	

Step IV. Find the sum of the entries in the g-th and h-th tests, difference of which will depend on values of third column in the above table. If the difference is close to zero, stop the process, otherwise go to next step.

Step V. Find the row of the above table that contains the highest difference between the elements of the two sub- tests.

Let this row number be ρ^*. Swap the two entries in the g-th and h-th tests in row ρ^* i.e. replace the entry of the g-th test in row $\rho^*$with that of the h-th test. Calculate sum of all the entries of the revised g-th test and h-th test. If the difference of sum is close to zero, stop the process otherwise proceed to the next step.

Step VI. Repeat Step V.

III. Empirical verification

3.1 **Data**: A Selection Test was administered to 911 candidates. The test had 50 items and maximum time given was 90 minutes. Scores of those 911 candidates were considered for empirical verification of the foregoing method. Here, for the test N = 911 and number of items n = 50

3,2 Splitting of the test in parallel halves:

Usual convention of splitting a test by considering odd and even items was carried out. The results are given in Table - 1

TABLE - 1

Splitting half with respect to Odd – Even items.

g-th test		**h-th test**		
Item	Score	Item	Score	Difference (g – h)
1	670	2	283	387
3	411	4	519	-108
5	158	6	325	-167
7	171	8	654	-483
9	570	10	294	276
11	256	12	493	-237

13	393	14	417	-24
15	222	16	143	79
17	243	18	285	-42
19	239	20	194	45
21	273	22	410	-137
23	348	24	534	-186
25	520	26	248	272
27	273	28	191	82
29	386	30	221	165
31	630	32	310	320
33	645	34	470	175
35	595	36	507	88
37	601	38	452	149
39	491	40	30	461
41	113	42	375	-262
43	187	44	672	-485
45	558	46	551	7
47	197	48	385	-188
49	230	50	328	-98
Sum	9380		9291	
Average	10.29638		10.19868	0.0977
Sum of Square	4291966		4056429	
SD	67.86202		65.94473	1.91729
r_{gh}				0.882243

Results of splitting a test as per the proposed iterative method are given in Table – 2

TABLE – 2

Splitting half as per the proposed iterative process

Iteration				
g-th test		**h-th test**		**Difference(g-h)**
Item	Score	Item	Score	
41	113	40	30	83
5	158	16	143	15
7	171	43	187	-16
20	194	28	191	3
47	197	30	221	-24
49	230	15	222	8
19	239	17	243	-4
11	256	26	248	8
21	273	27	273	0
2	283	18	285	-2
10	294	32	310	-16
50	328	6	325	3
23	348	42	375	-27
29	386	48	385	1
13	393	22	410	-17
14	417	3	411	6
38	452	34	470	-18
39	491	12	493	-2
36	507	4	519	-12
24	534	25	520	14
46	551	45	558	-7
35	595	9	570	25
37	601	31	630	-29
8	654	33	645	9
1	670	44	672	-2
Sum	9335		9336	-1
Mean	10.24698		10.24808	0.0011
Sum of square	4149085		4199310	
SD	66.70404		67.11585	-0.4181
r_{gh}				0.99858

Thus, splitting half of the test by the iterative process is better since it gives

- Means for the g-th and h-th tests are almost equal (much less in comparison to odd-even split)
- Marginal difference (0.4181) between the SD's of the g-th and h-th tests (much less than the same obtained from odd – even split half).
- Correlation between the scores of two halves obtained from the odd-even split was 0.88224 whereas the same obtained from the iterative method was 0.99858, which is little less than unity because of marginal

difference of means and SDs. Reliability of the entire test as per Spearman- Brown formula works out to be 0.93744 for odd-even split and 0.99929 for the iterative method.

Accordingly, splitting half as per the iterative process was considered better for almost equality of means and SDs and higher split-half reliability.

IV. Properties of the iterative method:

The following two theorems are relevant to the said iterative method:

Theorem 1: For binomially distributed item scores, if $\bar{X}_g = \bar{X}_h$ then $S_g^2 = S_h^2$

Proof: Here, distribution of the i-th item is Binomial with parameters N and p_i where N denotes the sample size and p_i denotes probability of the answering the i-th item correctly. For the i-th and j-th items, means are Np_i and Np_j. If $Np_i = Np_j$

then $(1 - p_i) = (1 - p_j)$ which implies $Np_i(1 - p_i) = Np_j(1 - p_j)$

Thus, variance of the i-th item and j-th item are equal.

In other words, dichotomisation with respect to item scores or difficulty values or item means is equivalent to dichotomisation with respect to Item variances

Theorem 2: If $\bar{X}_g = \bar{X}_h$ and $S_g^2 = S_h^2$ then r_{gh} is maximum

Proof: Let the regression line of g *on* h is given by $g = \alpha_1 + \beta_1 h$

where, $\beta_1 = \frac{r_{gh} S_g}{S_h}$

Now $S_g^2 = S_h^2$ implies $\beta_1 = r_{gh}$

Similarly, the regression line of h *on* g is given by $h = \alpha_2 + \beta_2 g$

where $\beta_2 = r_{gh}$

Now, $\alpha_1 = \bar{g} - r_{gh}\bar{h}$ and $\alpha_2 = \bar{h} - r_{gh}\bar{g}$.

So $\alpha_1 - \alpha_2 = 0$. since $\bar{g} = \bar{h}$

or $\alpha_1 = \alpha_2$

Thus, regression coefficient of g on h and h on g are same and is equal to r_{gh} and intercepts of the two regression lines are also same. Therefore, the two regression lines with equal β coefficients coincide, which is possible only if $r_{gh} = 1$.

Thus, departure from $r_{gh} = 1$ implies departure from $S_g^2 = S_h^2$

The following observations may merit consideration:

(i) If $r_{gh} = 1$ then $r_{gy} = r_{hy}$ i.e. parallel tests are equi-correlated with a third variable. In other words, if g-th test and h-th are parallel, then they are equi-correlated with the criterion score i.e. parallel tests have equal validity.

(ii) Let X_i denotes score of the i-th individual in the entire test. Here, $X_i = X_{gi} + X_{hi}$

If "g" and "h" are parallel, the regression line of X on X_g is same as regression line of X on X_h.

Proof: Let the two regression lines are $X = \alpha_1 + \beta_1 X_g$ *and* $X = \alpha_2 + \beta_2 X_h$

where $\beta_1 = \frac{r_{Xg} S_X}{S_g}$ and $\beta_2 = \frac{r_{Xh} S_X}{S_h}$.

Clearly, $\beta_1 = \beta_2$ since for parallel tests, $S_g = S_h$ and $r_{Xg} = r_{Xh}$ (as per observation (i))

Now, $\alpha_1 = \bar{X} - \beta_1 \overline{X_g}$ and $\alpha_2 = \bar{X} - \beta_2 \overline{X_h}$. Clearly $\alpha_1 = \alpha_2$

Thus, the two regression lines are same.

4.1 Summary of properties of the iterative method:

(i) Splitting a test by the above iterative process based on item scores is equivalent to those obtained from item difficulty values or item means or item variances.

(ii) The iterative method gives near equality of means and SDs of the g-th test and h-th test and ensures the resulting sub-tests are parallel. Difference of item means can be reduced to as low as one wants by the said iterative process depending on availability of adequate number of items. Since parallel tests have equal means and variances, the item scores of the two sub-tests can be taken as coming from same population with same density function having two parameters namely mean and variance.

(iii) The method ensures maximum correlation between the two sub-tests and thus gives maximum split-half reliability. Thus, the algorithm provides a unique way to split a test in parallel halves ensuring maximum split-half reliability. Maximum split-half reliability(r_{tt}) implies minimum error variance of the entire test (S_E^2) since $S_E^2 = S_X^2(1 - r_{tt})$

(iv) Almost equal mean and variance of the g-th test and the h-th test implies that distribution of the two sub-tests will be almost same under the assumption of Normal distribution. Even for small number of items, total score of the g-th test will follow Binomial distribution because of convolution property of Binomial distribution. The same is true for the distribution of total score of the h-th test. Parameters of two such Binomial distributions will be equal because of equality of means of the g-th and h-th test

(v) The iterative method also ensures that the two sub-tests are equi-correlated or almost equi-correlated with a third variable. Thus, parallel tests have almost equal validity

(vi) If the items of the g-th test are arranged in increasing order of item scores, corresponding items of the h-th test are also arranged in increasing order. After such arrangement, a score of$g_0$of the g-th test is equivalent to a score of h_0 on the h-th test if $\sum_1^k X_{gi} - \sum_1^k X_{hi} \cong 0$ i.e. difference of cumulative sum of scores of first k-items of the g-th tests and h-th test $\cong 0$

(vii) An item of a sub-test corresponds uniquely to another item of the other sub-test and vice versa. A test with 2n-number of items results in n-pairs of items of such that$|s_{gi} - s_{hi}| \approx 0 \; \forall \; i$ where s_{gi} denotes total score of the i-th item of the g-th test and S_{hi} denotes total score of the i-th item of the h-th test
In other words, the algorithm divides the items of the test in two sub-tests so that absolute deviation of scores of an item in one sub-test with its corresponding item in the other sub-test is close to zero

(ix) If distribution of g-th and h-th test are same, Mahalanobis distance between the g-th and h-th test will be closed to zero since $d_i = \bar{X}_{gi} - \bar{X}_{hi}$is closed to zero

V. Test of parallelism

The hypotheses of equality of mean, variance and correlation of parallel tests can be tested separately and/or simultaneously as a single multidimensional hypothesis in the framework of simultaneous equation models via AMOS (Arbuckle, 1997), EQS (Bentler, 1995), LISREL 8 (Jöreskog&Sörbom, 1998), MPLUS (Muthén&Muthén, 1998), MX (Neale, 1997), RAMONA (Browne &Mels, 1998), SEPATH (Steiger, 1995), and others. In addition, toknow whether two sub-tests "g' and "h" are parallel, one may use either of the following two well-known statistical tests

* Test equality of regression lines of X on X_g and X on X_h by ANOVA (Rao, 1952). Significance of the ratio of mean sum of square due to deviation from hypothesis to residual due to separate regression along with corresponding degrees of freedom may help to accept or reject the hypothesis.

* Test the hypothesis of no difference in mean values of n-items for the g-th and h-th sub-tests using Mahalanobis$D^2 = d^T S^{-1} d$ which is distributed as a variance ratio, where $d_i = \bar{X}_{gi} - \bar{X}_{hi}$and S^{-1} is the inverse of the item variance-covariance matrix

VI.Degree of parallelism

Degree of parallelism can be indexed by various methods including the following:

a) Coefficient of variation: Parallel tests should have equal coefficient of variation (CV) because of equality of mean and SD. In practice, the CVs may differ marginally. Without loss of generality, if $CV_1 > CV_2$. Index of parallelism could be defined as$\frac{CV_1 - CV_2}{CV_1}$

b) Split-half reliability: High positive correlation between two parallel tests implies either the tests have same distribution, or scores of g-th test and h-th test are linearly dependent by equation of the form $X_g = \beta X_h$ or $X_g = \alpha + X_h$or$X_g = \alpha + \beta X_h$
Under the Classical theory set up, the forms of last two equations may contradict Equation 1 and 2 above. Thus, split-half reliability itself reflects degree of parallelism.

c) Euclidian distance between g-th and h-th test i.e.$\sqrt{(X_{gi} - X_{hi})^2}$.

Euclidian distance$\cong 0 \Rightarrow$ "g" and "h" are parallel and vice versa

d) Mahalanobis$D^2 = d^T S^{-1} d$where $d_i = \bar{X}_{gi} - \bar{X}_{hi}$for the i-th item,
i= 1, 2, 3,……..,n and S^{-1} is the inverse of the item variance-covariance matrix of order n X n.

D^2indicates the generalised distance between distributions of the two sub-tests. Value of D^2 will be close to zero if "g" and "h" are obtained as per the algorithm. Thus, reciprocal of D^2 could reflect degree of parallelism of two sub-tests provided $D^2 \neq 0$

VII. Split-half reliability of battery of tests

The iterative method described in Section III can be extended to find uniquely reliability of a battery of tests. Suppose a battery consists of m-number of tests where i-th test has n_i number of items and $\sum_1^m n_i = K\ (say)$. Battery score of an individual taking the tests is either sum of his scores on the component tests (summative scores) or assigning weights to the tests and then takes weighted sum. The Step I and Step II of the iterative method may be modified in the context of Battery reliability as follows

I. Find item wise battery score for each of the K-item.

II. Arrange the item scores of the battery in increasing order and denote them by $S_1, S_2, S_3, \ldots\ldots\ldots\ldots, S_K$

Thus, split-half reliability of a test or a battery of tests can be found by splitting the test or items of the battery in parallel halves without considering homogeneity or dimensionality of items. Even if a test or a battery measures more than one factor, reliability is defined. While it may be desirable that items in a test measure something in common (i.e. exhibit uni-dimensionality), Hattie (1985) observed that a uni-dimensional scale (having an underlying latent trait), is not necessarily reliable, internally consistent or homogeneous.

VIII. Conclusions

* The iterative method provides a simple way to split a test in parallel halves ensuring almost equality of means and standard deviations

*The dichotomization of a test based on item scores is equivalent to the same based on item difficulty values or item means or item variances.

* The method ensures correlation between the halves is maximum. In other words, the iterative method gives maximum split-half reliability of a test. The method thus helps to split a test in a unique fashion ensuring maximum value of the split-half reliability and may be adopted while reporting a test.

* The iterative method also ensures that the two sub-tests are equi-correlated or almost equi-correlated with a third variable. Thus, parallel tests have almost equal validity

* For a given score of the g-th test, it is possible to compute the equivalent score of the h-th test if the sub-tests are parallel

* An item of a sub-test corresponds uniquely to another item of the other sub-test and vice versa. Thus, for a test with 2n-number of items, the algorithm results in n-pairs of ordered items so that $\left|s_{gi} - s_{hi}\right| \approx 0\ \forall\ i$ where s_{ji} denotes total score of the i-th item of the j-th test, j= g. h

* Simultaneous testing of single multidimensional hypothesis of equality of mean, variance and correlation of parallel tests can also be carried out by testing equality of regression lines of X on X_g and X on X_h by ANOVA or by Mahalanobis D^2

* New indices in terms of Coefficient of variations, Split-half reliability itself, Euclidian distance and Mahalanobis distance reflect degree parallelism.

* The iterative method can be extended to find split-half reliability and to find degree of parallelism of a battery of tests.

References:

[1]. Armstrong, R.D., Jones, D.H., & Wang, Z. (1994)- "Automated parallel test construction using classical test theory". *Journal of Educational Statistics, 19*, 73–90

[2]. Birnbaum, A. (1968)- "Some latent trait models and their use in inferring an exam-inee's ability". *In F. M. Lord & M. R. Novick, Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

[3]. Boekkooi-Timminga, E. (1990) –"The Construction of Parallel Tests from IRT-Based Item Banks", Journal of Educational Statistics, Vol. 15, No. 2, pp. 129-145

[4]. Chakrabartty, S. N, (2011). "Measurement of reliability as per definition", in *Prooceedings of the Conference on Psychological Measurement: Strategies for the New Millenium*, School of Social Sciences, Indira Gandhi National Open University, New Delhi, 116–125

[5]. Cortina J.M. (1993) – "What is Coefficient Alpha : An Examination of Theory and Applications". *Journal of Applied Psychology*, 78 (1), 98-104

[6]. Dawis, R.V(1987) – "Scale Construction". *Journal of Counseling Psychology,* 34, 481-489

[7]. Eisinga, R.; TeGrotenhuis, M.; Pelzer, B. (2012). "The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown?" *International Journal of Public Health*. doi:10.1007/s00038-012-0416-3

[8]. Guliksen, Harold (1987) – "Theory of Mental Tests". Hilsdale, NJ : L. Erlbaum Associates, ISBN 978-0-8058-0024-1

[9] Hattie ,J.(1985)- "Methodology review: assessing uni-dimensionality of tests andItems" *Applied Psychological Measurement,*. 139–164.

[9]. Kaplan, R.M. and Saccuzzo, D.P. (2001)- "Psychological Testing: Principle, Applications and Issues" (5th Edition), Belmont, CA: Wadsworth

[10]. Lin, C.-J. (2008)- "Comparisons between Classical Test Theory and Item Response Theory in Automated Assembly of Parallel Test Forms". *Journal of Technology, Learning, and Assessment, 6*(8).http://www.jtla.org.

[11]. Lord, F. M. (1980)- "Applications of item response theory to practical testing prob-lems". Hillsdale, NJ: Erlbaum.

[12]. Lord. F. M. and Novick, M. R. (1968) – "Statistical Theory of Mental Test Scores". Reading MA : Addison-Welsley Publishing Company

[13]. McDonald R.P. (1999)- "Test theory : a unified treatment". Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

[14]. O'hEigeartaigh, M., Lenstra, J. K., &RinnooyKan, A. H. G. (Eds.). (1985) –"Combinatorial optimization: Annotated bibliographies". New York: Wiley.
[15]. Rao, C. R. (1952)- "Advanced Statistical Methods in Biometric Research". John Wiley & Sons.
[16]. Rao, S. S. (1985)- "Optimization: Theory and applications" (2nd ed.). New Delhi: Wiley Eastern Ltd.
[17]. Ritter, N. (2010).-"Understanding a widely misunderstood statistic: Cronbach's alpha". Paper presented at *Southwestern Educational Research Association (SERA) Conference 2010*, New Orleans, LA (ED526237).
[18]. Rudner, Lawrence M and Schafes, William (2002)- "Reliability : ERIC Digest". www.ericdigests.org/2002-2/reliabillity/htm.
[19]. Salkin, H. M. (1975)- "Integer programming". London: Addison-Wesley.
[20]. SAFRIT, M.J.E.(1976) - "Reliability Theory". Washington, DC: AmericanAlliance for Health, Physical Education, and Recreation,
[21]. Samejima, F. (1977) – "Weakly parallel tests in latent trait theory with some criticisms of classical test theory". Psychometrika, 42, 193-198.
[22]. Taha, H. A. (1975) – "Integer programming". New York: Academic Press.
[23]. Theunissen, T. J. J. M. (1985)- "Binary programming and test design". Psycho-metrika, 50, 411-420.
[24]. Vacha- Hasse, T. (1998) – "Reliability Generalisation : Exploring variance in Measurement error affecting Score Reliability across studies". *Educational and Psychological Measurement*, 58. 6-20
[25]. van der Linden, W.J., &Luecht, R.M. (1998). " Observed-score equating as a test assembly problem". *Psychometrika, 63*, 401–418.
[26]. Wagner,H.M. (1972)- "Principles of Operations Research :With applications to managerial decisions".London: Prentice-HallInternational.
[27]. Webb,N.M.,ShavelsonR.J.&Haertel,E.H.,(2006)-"ReliabilityCoefficientsandGeneralizability Theory",*HandbookofStatistics*,*26*,ISSN:0169-7161..
[28]. Weir, J.P. (2005) – "Quantifying Test-retest Reliability using the Intra-class Correlation Coefficient and the SEM" , Journal of Strength and Conditioning Research, 19(1), 231– 240