IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Reliability Measurements: Methods And Estimation In Healthcare Research

N Balasubramanian[1], S Sudha Rani[2],

[1]Principal cum Professor, Facultyof Nursing, Constituent College of Desh Baghat University, Mandi Gobindgarh, Punjab

[2] Jai Ambey Institute of Nursing, Mukheri, Punjab

Abstract

The primary purpose of this paper is to discuss the idea of reliability for research. The authors present methods for calculating statistical reliability. Reliability is crucial for tests intended to be stable across time. That means that if a user conducted a test multiple times, its result should be identical every time. Although it's impossible to determine reliability precisely, there are ways to assess it. Each instrument has an element of reliability. This article focuses on methods for computing reliability in quantitative information, including ratio and interval data. Interval data has equally spaced intervals between the numbers; however, they aren't connected to the actual zero point and do not accurately represent the quantity. Ratio data comprises numbers that represent units that have equal intervals. They are measured from a true zero. Reliability refers to how much the test, process, or instrument (such as an instrument or questionnaire) produces similar results under different conditions as long as nothing else has changed. Stability, internal consistency, and equivalence are all reliability measures employed in research. Stability is the term used to describe the instrument's capacity to give the same results using repeated measurements. Internal consistency refers to all components of the apparatus having an identical concept, feature, or characteristic. Equivalence signifies that the device yields exact results when similar or parallel procedures or instruments are employed.

Keywords: Reliability, Stability, Internal Consistency, Equivalence, Quantitative Data.

1. Introduction

In research, the accuracy of an instrument is vital before its implementation help to the researchers/study participants. It's the researcher's responsibility to minimize the error to the maximum extent to achieve the accuracy of the tool or instrument. Reliability is essential when choosing a measuring device because it ensures its results are stable and consistent. It means that the agent can be relied on to give similar results when applied at different times, even though there may be some variation due to factors such as changes in the population or sample. [1,2]

Reliability is multifaceted and complex and addressed in different approaches. Apart from the mechanical and electrical equipment, the tools in the research are opinionnaire, schedules, questionnaires, rating scales, etc. reliability proves its trustworthiness and dependability. The reliability of the measuring instrument significantly impacts the study's results. Researchers should, therefore, take care to use a reliable measuring instrument. Empirical research uses different methods to determine reliability. The most commonly used techniques are tests-retest reliability as well as parallel/alternative forms, and internal consistency. Internal consistency tests are of three types (split-half, items-total correlations, and the alpha reliability coefficient).[3-5]

In studies on scale development, researchers evaluate the credibility of the scales through reliability tests such as test-retest reliability as well as alternative forms and internal consistency. Meanwhile, researchers who use scales already developed and assessed for reliability only need to do one internal consistency test. Alpha reliability is one of the famous test internal consistencies. Statistically, users compute reliability using the correlation coefficient formula. Its range is from 0 to 1. The higher the coefficient value, the more consistency between the measures. [6-8]

This study aims to introduce researchers the main methods used in estimation of reliability of tools such as opinionnaire, schedules, questionnaires, rating scales. By providing a general overview, we hope that researchers will develop a better understanding and awareness of studies on reliability.

The study consists of five parts. The first part of the research is the introduction part, where general information is presented to the readers. In the second part, information about three attributes Reliability and its calculation, in the third part, factors influencing Reliability and in the fourth part, ways to improve reliability are included. The last part is the discussion and conclusion part.

Measurement of reliability

Reliability measurement concerns three attributes: Stability, Homogeneity, and equivalence.

1.1. Stability

If the data analyzer obtains the same results on repeated administrations of the same instrument. That is, through the test-retest method and parallel form, the results are the same; hence, the user has established stability. The 'Test-retest' is a method through which the same instrument is administered twice to the same participants in similar circumstances to check the consistency. To avoid the error rate, the researcher follows the same steps, same comfort zone, lighting, day, etc. After the procedure, the user computes the coefficient r value by comparing two scores. The higher the coefficient, the tool is stable. [9-12]

Steps:

1. Administration of the test to a large group (ideally, above 30)
2. Administration to the same group over some periodic time.
3. Generally, the second administration should happen about two weeks after the first, although this time might differ based on the setting.
4. It is crucial to ensure that no activities have occurred between the first and second administrations that could change the measured characteristic.
5. Finding the correlation coefficient for the scores.
6. Computation of the Karl Pearson's Correlation Coefficient.

2.1.1. Formulas

A. Karl Pearson's Co-efficient of Correlation (r)

Karl Pearson's coefficient of correlation is an extensively used in reliability equation in which the numerical representation is applied to measure the level of relation between linearly related variables. The coefficient of correlation is expressed by "r". [13]

$$r = \frac{\sum (X-\underline{X})(Y-\underline{Y})}{\sqrt{\sum (X-\underline{X})^2 \sum (Y-\underline{Y})^2}} \qquad (1)$$

here $\underline{X}$ is the mean of the X values, and $\underline{Y}$ is the mean of the Y values.

1.2. Advantages of the stability method:

The method is appropriate to measure attributes that don't change over time and are the same no matter who counts them. It also ensures that the results are consistent and produce similar results every time. The sample of items or stimulus situations is held constant to ensure that we are only measuring the traits we want to measure [14].

2.3. Disadvantages of the stability method:

Subjects can learn from taking a test, which can impact how they perform the second time. If there's less time between the first and second test, though, maturation can occur. Development is when subject factors or respondents change over time, impacting the measurements taken at different points. The test-retest method can be affected by reactivity when measuring something that causes a change in the phenomenon [14].

Table 1: Computation of Karl Pearson coefficient using fictitious data.

Sample No	Test X	Retest Y	$X - \underline{\square}$	Y-$\underline{\square}$	$(X - \underline{\square})$ (Y-$\underline{\square}$)	$(X-\underline{\square})^2$	$(Y-\underline{\square})^2$
1	55	57	0.2	2.2	0.44	0.04	4.84
2	49	46	-5.8	-8.8	51.04	33.64	77.44
3	78	74	23.2	19.2	445.44	538.24	368.64
4	37	35	-17.8	-19.8	352.44	316.84	392.04
5	44	46	-10.8	-8.8	95.04	116.64	77.44
6	50	56	-4.8	1.2	-5.76	23.04	1.44
7	58	55	3.2	0.2	0.64	10.24	0.04
8	62	66	7.2	11.2	80.64	51.84	125.44
9	48	50	-6.8	-4.8	32.64	46.24	23.04
10	67	63	12.2	8.2	100.04	148.84	67.24

Here:

$\underline{\square}$ = 54.8, $\underline{\square}$ = 54.8 $\sum (X - \underline{\square})$ (Y-$\underline{\square}$) = 1152.6, $(X-\underline{\square})^2$ =1285.6, $(Y-\underline{\square})^2$ = 1137.6, $\sqrt{\sum (X - \underline{X})^2}$ = 35.85, $\sqrt{\sum (Y - \underline{Y})^2}$ = 33.72

By substituting all the values in the given formula, we get the reliability r= 0.953. This value indicates a very high correlation.

Table 1 shows the fictitious data of test and retest of score of participants. Karl Pearson correlation coefficient formula was used to calculate the reliability of the test score. The obtained result (r=0.953) shows that there was strong positive correlation between the test and retest scores. This formula helps in knowing how strong the relationship between the two variables. Depending on the direction of the relationship between variables, correlation can be of three types, namely – positive Correlation (0 to +1), negative Correlation (0 to -1), zero Correlation (0).[15]

2.4. Internal consistency or Homogeneity

It measures consistency within the instrument. Commonly, the Split-half method is used for determining internal consistency. This test can be possible using any tool with more than two response choices. Odd-even method is the most acceptable method. The scores of the two sets, i.e., odd and even, are used to compute a correlation coefficient. The Spearman-Brown Prophecy formula is applied in this method to adjust the correlation coefficient of the entire test. Another split-half technique is the first and second half of the tool, which is rare in use. The coefficient alpha, such as Kuder- Richardson and Cronbach's alpha, is another method to estimate internal consistency. The analyzer uses Kuder-Richardson on questions with two answers, e.g., true or false / yes or no/ dichotomous measurements with a score of 0 or 1. Here, all correct responses scored as +1 and incorrect responses as zero.

In most cases, the analyzer utilizes Cronbach's alpha to estimate internal consistency between items in a scale, e.g., a Numerical rating scale with a 1 to 5 score. Each item in this test is expected to have an exact correlation with a few scores. Thus, coefficient alpha proves item-specific variance in uni-dimensional tests. [16-19]

Steps:

1. Application of the test to a large group that is, ideally, over
2. Division of the test question randomly into two parts. For example, select items into equal halves or odd-even.
3. Calculation of the correlation coefficient for the two halves.
4. Compute Spearman-Brown/Kuder-Richardson/ Coefficient of alpha depends on the tool.

2.5. Formulas

2.5.1. Spearman- Brown Prophecy formula

This formula is related to psychometric reliability to test length and used by psychometricians to predict the reliability of a test after changing the test length. The method was published independently by Spearman (1910) and Brown (1910).[20]

$$r_{tt} = \frac{2r_h}{1+r_h} \qquad (2)$$

where r_{tt}= reliability of the entire test, r_h= reliability calculated through Karl Pearson's formula.

Table 2: Computation of Spearman-Brown Prophecy coefficient using fictitious data.

Sample No	Total Score	Odd items Score X	Even Items Score Y	$X - \underline{\square}$	Y-□	$(X - \underline{\square})$ (Y-□)	$(X-\underline{\square})^2$	$(Y-\underline{\square})^2$
1	55	28	27	0.5	0.7	0.35	0.25	0.49
2	49	26	23	-1.5	-3.3	4.95	2.25	10.89
3	78	36	42	8.5	15.7	133.45	72.25	246.49
4	37	18	19	-9.5	-7.3	69.35	90.25	53.29
5	44	23	21	-4.5	-5.3	23.85	20.25	28.09
6	50	30	20	2.5	-6.3	-15.75	6.25	39.69
7	58	30	28	2.5	1.7	4.25	6.25	2.89
8	62	33	29	5.5	2.7	14.85	30.25	7.29
9	48	23	25	-4.5	-1.3	5.85	20.25	1.69
10	57	28	29	0.5	2.7	1.35	0.25	7.29

For estimating of $r_{h,}$ we have to use Karl Pearson's formula,

$\underline{\square}$ = 27.5, $\underline{\square}$ = 26.3, $\sum (X - \underline{\square})$ (Y-□) =242.5, $(X-\underline{\square})^2$ =248.5, $(Y-\underline{\square})^2$ =398.1, $\sqrt{\sum (X - \underline{X})^2}$ = 15.76, $\sqrt{\sum (Y - \underline{Y})^2}$ = 19.95

By substituting all the values in the given formula, we get the reliability r_h= 0.770

Then replacing the r_h = 0.770 in the Spearman-Brown Prophecy formula, we get r= 0.870. This value indicates a good correlation.

Table 2 shows the fictitious data of test scores of participants. Spearman-Brown Prophecy coefficient formula was used to calculate the reliability of the test score. The obtained result (r=0.870) shows that there was strong positive correlation between the test and retest scores. The Spearman-Brown prophecy formula may help the researchers accurately predict the effects of adding or removing items on score reliability, and to make immediate improvements to the psychometric quality and functioning of the tool.[21]

2.5.2. Kuder-Richardson 20 Formula

The KR20 is a statistical measure that allows you to compute reliability for items with varying difficulty. For example, in Multiple choice questions, some things might be elementary, and others may be more difficult. The limitation of the KR-20 formula is that the user cannot apply it to scales such as the Likert scale or visual analog scale.[22]

$$KR_{20} = \frac{K}{K-1}\left(1 - \frac{\sum pq}{\sigma^2 X}\right) \qquad (3)$$

Here

K = Number of items.

p = Proportion of right answer

q = Proportion of the wrong answer

σ^2X = Variance.

Table 3: Computation of KR 20 coefficient using fictitious data.

Samples	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Total Score
1	1	1	1	1	1	1	1	1	1	1	1	11
2	1	1	1	1	1	1	1	1	0	1	0	9
3	1	0	1	1	1	1	1	1	1	0	0	8
4	1	1	1	0	1	1	0	1	1	0	0	7
5	1	1	1	1	1	0	0	0	1	0	0	6
6	0	1	1	0	1	1	1	1	0	0	0	6
7	1	1	1	1	0	0	1	0	0	0	0	5
8	1	1	1	1	1	0	0	0	0	0	0	5
9	0	1	0	1	1	0	0	0	0	1	0	4
10	1	0	0	1	0	1	0	0	0	0	0	3
11	1	1	1	0	0	0	0	0	0	0	0	3
12	1	0	0	1	0	0	0	0	0	0	0	2
13	0	1	0	1	1	0	0	0	0	1	1	5
14	0	1	1	0	1	1	1	0	1	0	0	6
15	1	1	0	1	1	0	0	1	0	1	1	7
16	0	1	1	1	0	0	1	0	0	0	0	4
17	1	1	1	0	0	0	1	0	0	0	1	5
18	0	0	1	0	1	1	1	1	1	0	1	7
19	0	1	0	0	0	0	0	0	1	0	0	2
20	1	1	1	1	1	0	0	0	0	1	1	7
No. of correct response	13	16	14	13	13	8	9	7	7	6	6	**5.04**
p	0.65	0.80	0.70	0.65	0.65	0.40	0.45	0.35	0.35	0.30	0.30	
q	0.35	0.20	0.30	0.35	0.35	0.60	0.55	0.65	0.65	0.70	0.70	
pq	0.23	0.16	0.21	0.23	0.23	0.24	0.25	0.23	0.23	0.21	0.21	**2.14**

K= 11, p= Number of right correct answers for each item/Number of samples, e.g., Number correct responses for item 1= 13, sample size= 20, So, 13/20 = 0.65, q= 0.35 (1-p), $\sum pq$= 2.14, σ^2X=5.04

By substituting all the values in the given formula, we get **r= 0.633.** This value indicates questionable correlation.

Table 3 shows the fictitious data of item wise test scores of participants. KR 20 coefficient formula was used to calculate the reliability of the test score. The obtained result (r=0.633) shows that there was weak correlation between the items of test score. KR-20 is derivative of the Cronbach formula, with the advantage to Cronbach that it can handle both dichotomous and continuous variables. [23]

2.5.3 Kuder-Richardson 21 Formula

The analyzer uses for a test where the items are all about the same difficulty level. For example: True or False, Yes or No type of questions.[22]

$$KR_{21} = \frac{k}{k-1}\left(1 - \frac{M(k-M)}{S^2 k}\right) \qquad (4)$$

K= Number of items

M= Mean of Total Score

S^2 = Variance.

Table 4: Computation of KR 21 coefficient using fictitious data.

Samples	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Total Score
1	1	1	1	1	1	1	1	1	1	1	1	11
2	1	1	1	1	1	1	1	1	0	1	0	9
3	1	0	1	1	1	1	1	1	1	0	0	8
4	1	1	1	0	1	1	0	1	1	0	0	7
5	1	1	1	1	1	0	0	0	1	0	0	6
6	0	1	1	0	1	1	1	1	0	0	0	6
7	1	1	1	1	0	0	1	0	0	0	0	5
8	1	1	1	1	1	0	0	0	0	0	0	5
9	0	1	0	1	1	0	0	0	0	1	0	4
10	1	0	0	1	0	1	0	0	0	0	0	3
11	1	1	1	0	0	0	0	0	0	0	0	3
12	1	0	0	1	0	0	0	0	0	0	0	2
13	0	1	0	1	1	0	0	0	0	1	1	5
14	0	1	1	0	1	1	1	0	1	0	0	6
15	1	1	0	1	1	0	0	1	0	1	1	7
16	0	1	1	1	0	0	1	0	0	0	0	4
17	1	1	1	0	0	0	1	0	0	0	1	5
18	0	0	1	0	1	1	1	1	1	0	1	7
19	0	1	0	0	0	0	0	0	1	0	0	2
20	1	1	1	1	1	0	0	0	0	1	1	7
											Mean	10.21
											Variance	5.04

k=11, *M*= 10.21, S^2 = 5.04

By substituting all the values in the given formula, we get **r= 0.939.** This value indicates an excellent correlation.

Table 4 shows the fictitious data of item wise test scores of participants. KR 21 coefficient formula was used to calculate the reliability of the test score. The obtained result (r=0.939) shows that there was strong correlation between the items of test score. The difference in KR20 and KR21 reliability coefficients was significantly greater when the range of item difficulty values was .30 or more. Nevertheless, KR21 was a good estimate of

KR20 when the range of item difficulty was relatively narrow. Implications for test selection are suggested. When KR21 has been used to estimate a test's reliability, the user should note that the test has a lower bound of internal consistency reliability, particularly when the item difficulty range is great. [24]

2.5.4. Cronbach's Alpha Formula

Cronbach's alpha is calculated by taking the score from each scale item and correlating them with the total score for each observation and then comparing that with the variance for all individual item scores. Cronbach's alpha is best understood as a function of the number of questions or items in a measure, the between pairs of items average covariance, and the overall variance of the total measured score.[25]

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\Sigma \quad V_i}{V_{test}}\right) \qquad (5)$$

n = Number of questions

V_i = variance of scores on each question Mean of Total Score

Vtest = total variance of overall scores (not %'s) on the entire test.

Table 5: Computation of Cronbach's Alpha coefficient using fictitious data.

Samples	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Total Score
1	2	2	2	2	2	2	2	2	2	1	2	21
2	3	3	3	3	2	4	4	4	4	4	4	38
3	3	3	3	3	3	3	3	3	4	4	1	33
4	4	4	4	4	3	4	4	4	4	4	4	43
5	3	3	4	4	3	4	4	4	4	3	3	39
6	2	2	2	2	2	2	2	2	4	4	1	25
7	3	3	3	3	3	4	4	4	4	4	4	39
8	3	3	3	3	3	4	4	4	4	4	4	39
9	2	2	2	2	2	2	2	2	4	4	4	28
10	2	2	2	2	2	2	2	2	4	4	1	25
11	4	4	4	4	3	4	4	4	4	4	4	43
12	2	3	3	3	3	2	3	3	4	4	1	31
13	4	4	4	4	5	4	4	4	4	5	5	47
14	2	2	2	2	2	2	2	2	2	1	2	21
15	2	2	2	2	2	2	2	2	4	4	1	25
16	4	4	4	4	3	4	4	4	4	4	4	43
17	4	4	4	4	5	4	4	4	4	5	5	47
18	4	4	4	4	5	4	4	4	4	5	5	47
19	2	2	2	2	2	2	2	2	4	4	1	25
20	4	4	4	4	5	4	4	4	4	5	5	47
Variance	**0.75**	**0.70**	**0.75**	**0.75**	**1.20**	**0.93**	**0.86**	**0.86**	**0.36**	**1.13**	**2.45**	
Sum of Variance	**(Summation variance of Item 1 to 11)**											**10.73**
Variance of Total Score												**84.21**

n=11, $\sum Vi$=10.73, V*test* = 84.21

By substituting all the values in the given formula, we get r= 0.959. This value indicates an excellent correlation.

Table 4 shows the fictitious data of item wise test scores of participants. of Cronbach's Alpha coefficient formula was used to calculate the reliability of the test score. The obtained result (r=0.959) shows that there was strong correlation between the items of test score. Instruments with greater Cronbach alpha should be used for any type of research since they have smaller measurement error and thus have greater statistical power for any research settings, cross-sectional or longitudinal.[26]

Advantages of Internal Consistency Method: Administration of the same test can only be applied once, and there is no place for two different versions of the same test.

Disadvantages of Internal Consistency Method: Insurance of homogeneity and Different subsections of the same test

Equivalence: It is estimated in two methods (i) Parallel form / alternate form (ii) inter-rater/inter-rater observer reliability.

(i) **Parallel/alternate form**: The same test is administered randomly to the same individual. It is similar to the test-retest method; here, randomly selected two sets of items from an item pool are distributed over a while. It avoids carryover biases. The first form and second form of the test are similar but different [27].

Steps:

1) Administration of the test to a large group (ideally, over about 30).
2) Division of the test question randomly into two parts. For example, select items into equal halves.
3) Administration of the first half (Set 1 of Form A) initially and the second half (Set 2 or Form B) over a time
4) The finding of the correlation coefficient between the two given sets
5) Compute the Karl Correlation Coefficient (Refer to Test and Retest method) from Set 1 = X, Set 2 = Y.

(ii) **Inter-rater/inter-observer**: This includes determining the level of agreement among two or more observers. Here, the observers ask to give a score for every item on an instrument, and the consistency in their scores would relate to the instrument's inter-rater reliability level. [28-30]

Steps:

1. Administration of the test to a large group (ideally, over about 30).
2. Observation of the participants at a time by two or more observers.
3. Scoring the task observed.
4. The finding of the correlation coefficient.
5. Computation of the Kappa Correlation Coefficient (Two observers) and Fleiss Kappa Correlation Coefficient or Intra-class Correlation Coefficient (More than two observers)

2.5.5. Cohen Kappa Correlation Coefficient

Cohen Kappa is used to measure inter-rater reliability (and also intra-rater reliability) for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance.[31]

$$K = \frac{Number\ of\ agreements}{Number\ of\ agreements + Number\ of\ disagreements} \quad (6)$$

Below is a checklist containing 20 items given by two observers. The observer will rate each item on a scale of 1 to 3.

Table 6: Computation of Cohen Kappa Coefficient using fictitious data.

Items	Observer 1	Observer 2	Difference
1.	1	1	0
2.	2	2	0
3.	3	2	1
4.	2	2	0
5.	2	3	1
6.	2	1	1
7.	3	3	0
8.	3	2	1
9.	1	1	0
10.	1	1	0
11.	1	1	0
12.	3	3	0
13.	3	2	1

14.	2	1	1
15.	2	2	0
16.	3	3	0
17.	1	2	1
18.	1	1	0
19.	2	2	0
20.	3	3	0

Note: 0= Agreement, 1= Disagreement

So, Number of Agreements= 13, Number of Disagreement= 7

By substituting all the values in the given formula, i.e., 13/ 13+ 7 = 13/20= 0.65,

Therefore, Cohen Kappa Coefficient is = **0.65.** This value indicates a Substantial Agreement.

Table 6 shows the fictitious data of 20 item check list rated by two observer on a scale of 1 to 3 for handing procedure of health care students. Cohen Kappa coefficient formula was used to calculate the reliability of the test score. The obtained result (r=0.65) shows that there was substantial agreement. Kappa is an index that considers observed agreement with respect to a baseline agreement. However, investigators must consider carefully whether Kappa's baseline agreement is relevant for the particular research question. Kappa's baseline is frequently described as the agreement due to chance, which is only partially correct. [32]

2.5.6. Fleiss Kappa Correlation Coefficient

Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. This contrasts with other kappas such as Cohen's kappa, which only work when assessing the agreement between not more than two raters or the intra-rater reliability (for one appraiser versus themself).[33]

$$K = \frac{P_O - P_e}{1 - P_e} \qquad (7)$$

here $P_o = Observed\ Agreement, P_e = Expected\ Agreement$

P_e = (Proportion of Agreement)^2 + (Proportion of Disagreement) ^{2} and

$P_O = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N} \quad \sum_{j=1}^{k} \quad n_{ij}^2 - Nn\right).$

here N= Number of items, n= Number of observer, n^2 = Summation of (Agreement^2 + Disagreement^2).

Table 7: Computation of Fleiss Kappa Coefficient using fictitious data.

Items	Observer 1	Observer 2	Observer 3	Number of Agreement X	Number of Disagreement Y	X^2	Y^2	$X^{2+} Y^2$
1.	1	1	1	3	0	9	0	9
2.	2	2	2	3	0	9	0	9
3.	3	2	3	2	1	4	1	5
4.	2	2	2	3	0	9	0	9
5.	2	3	2	2	1	4	1	5
6.	2	1	2	2	1	4	1	5
7.	3	3	3	3	0	9	0	9
8.	3	2	3	2	1	4	1	5
9.	1	1	1	3	0	9	0	9
10.	1	1	1	3	0	9	0	9
11.	1	1	1	3	0	9	0	9
12.	3	3	3	3	0	9	0	9
13.	3	2	3	2	1	4	1	5
14.	2	1	2	2	1	4	1	5
15.	2	2	2	3	0	9	0	9
16.	3	3	3	3	0	9	0	9
17.	1	2	1	2	1	4	1	5
18.	1	1	1	3	0	9	0	9
19.	2	2	2	3	0	9	0	9
20.	3	3	3	3	0	9	0	9
	Sum			53	7	**Sum**		152
	Proportion			53/60= **0.883**	7/60= **0.116**			
	Proportion^2			**0.780**	**0.013**			

P_e = 0.780+0.013 = 0.793, N= 20, n= 3, n^2 = 152

P_O = 1/ 20*3 (3-1) [152- 20*3] = 0.008* 92= 0.736

By Subsisting the value of □□ and □□ in the above formula, we get Fleiss correlation coefficient = 0.280. This value indicates fair agreement.

Table 7 shows the fictitious data of 20 item check list rated by three observer on a scale of 1 to 3 for handing procedure of health care students. Fleiss Kappa coefficient formula was used to calculate the reliability of the test score. The obtained result (r=0.280) shows that there was fair agreement.

2.6. Advantages of the Parallel form method:

1) The analyzer need not repeat the same test.

2) The minimization of Memory, practice, carryover effects, and recall factors without affecting the scores.

3) The method combines two types of reliability as the reliability coefficient obtained measures temporal stability and response consistency in the case of different item samples or test forms.

4) It proves helpful in the reliability of achievement tests.

5) The method is appropriate for determining the reliability of educational and psychological tests.

2.7. Disadvantages of the Parallel form method:

1) There is difficulty in having two parallel forms of a test. In certain situations (like Rorschach's case), it proves impossible.

2) Comparing two scores from these tests may lead to flawed decisions, mainly when the tests differ in content difficulty and length.

3) The user needs help to control practice and carryover factors fully.

4) The simultaneous administration of the two forms may create boredom. A single administration of the test is a preferred method.

5) The testing conditions in administering Form B may not always be the same, and the test may not be in an identical physical, mental, or emotional state during both instances of administration.

6) Generally, the second form of test scores is high.

2.8. Factors Influencing Reliability

The scores of an instrument are influenced by several intrinsic and extrinsic factors while computing reliability. Singh, 1986 mentioned different intrinsic and extrinsic factors influencing reliability. Outside factors influence reliability is guessing, environment, and group variability. Intrinsic factors influence reliability is the length of the test/tool, the spread of scores, homogeneity of items, difficulty value of things, and reliability scores. [35-36]

A. **Guessing**: Guess works tend to give high total scores and falsely give high reliability. It can yield different scores for different individuals and may cause measurement errors. Using guessing, one may get a high score, and another may get a lower score, thus giving lower reliability. For example, one person may get 75% and another 50%. These differences in scores may cause error scores and provide lower levels of reliability.
B. **Environment**: The testing environment should remain the same. If any variation is there, then it can lower the reliability. Lighting, noise, temperature, etc., are the environmental factors that may affect reliability.
C. **Group variability**: The reliability may be low if the groups are homogeneous. If heterogeneous groups, then the reliability may be high.
D. **Length of test/tool**: A longer tool will provide a sufficient sample of behavior measurement, and it avoids speculation. Thus it yields a high-reliability coefficient. Therefore, short device yields lower reliability.
E. **Spread of scores**: The more extensive spread yields a high reliability.
F. **Homogeneity of items**: item correlation is higher when all items measure the same trait, and the test is high and vice versa.
G. **Difficulty value of items**: if the things are too easy or too tricky can affect reliability. The items having indices of difficulty 0.5 or close result in higher reliability.
H. **Reliability scores**: if the agreement is low, the reliability will be low.

2.8.1 Measures to improve reliability

- Wide variability in the trait
- Homogenous group
- Keeping the appropriate length of the tool
- Choosing items with moderate difficulty indices.

Table 8: Summary of Reliability Statistics.

	Karl Pearson	Spearman Brown Prophecy	KR-20	KR-21	Cronbach Apha	Cohen Kappa	Fleiss Kappa
Range	0 to 1						
Interpretation	0= No Correlation 0.10-0.20= Negligible 0.21-0.40= Low 0.41-0.70= Moderate 0.71-0.90= High 0.91-0.99 = Very high	0.9 and Above = Excellent 0.80 – 0.89 = Good 0.70 – 0.79 = Average 0.60 – 0.69 = Questionable 0.50 – 0.59 = Poor Below 0.5 = Unacceptable				0 = No Agreement 0.1- 0.20= Slight Agreement 0.21- 40= Fair Agreement 0.41-0.60=Moderate Agreement 0.61-0.80=Substantial Agreement 0.81-0.90= Near perfect Agreement 1= Perfect Agreement	

Table 8 summary of reliability statistics such as karl pearson, Spearman Brown Prophecy KR-20, KR- 21 Cronbach Apha, Cohen Kappa and Fleiss Kappa's range and interpretation.

2.9. Discussion:

Karl Pearson [13] was the first to discover the linear correlation between two variables, referred to as the product-moment correlation coefficient. It is one of the most commonly used statistics today. It is used to test the stability of the test by the test and retest method. Kuder–Richardson [18] first published internal consistency reliability for dichotomous measurement measures with a score of 0 or 1. One limitation of the K-R formula is that it does not apply to scale. So the researchers were limited to estimating internal consistency by only using the KR-20 or KR -21 formula. Cronbach[17] published the Cronbach alpha method to assess the internal consistency of the tool with scale measures e.g., Likert scale. It is derived from KR- 20 formula. Different investigators adopted several ways to determine reliability. Some of the reliability determination was discussed here. Balasubramanian N et al.[36] developed a tool knowledge questionnaire for primary caregivers concerning the Home Care of Schizophrenics (KQHS). Using the Karl Pearson formula, the Split half technique (odd-even) was utilized to determine the coefficient correlation. It was followed by Spearman's Brown Prophecy formula, which established the reliability r =0.92. Balasubramanian et al.[37] also developed the caregivers' Attitude Scale on a similar subject. The authors determined the instrument's stability and test-retest reliability using the Karl Pearson correlation coefficient, and the calculated value was 0.78. Pandurangan & Balasubramanian[38] developed a case vignette tool to determine the skill of nursing students in ECG interpretations. The authors computed Cronbach's alpha and r values to assess internal consistency and correlation. The estimated value of r was 0.72.

2. Conclusion

The data analyzer cannot eliminate all the minor/significant errors from measurements, but it is possible to reduce them by employing sound measurement approaches. Reliability coefficients may bias researchers' explanations of study results. Researchers should know the importance of collecting correct data and interpreting the results. A greater understanding of score reliability might help authors avoid misunderstandings and write and speak cautiously about reliability estimates. This paper described the most commonly used reliability estimate in health care so young researchers can better understand reliability coefficients.

References

1. Dodgson JE. Why we need to critically analyze the science. Journal of Human Lactation. 2019 Aug;35(3):403-5.
2. Munn Z, Peters MD, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC medical research methodology. 2018 Dec;18:1-7.
3. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. International journal of surgery. 2021 Apr 1;88:105906.
4. Kaplan R, Saccuzzo D. Psychological testing: Principles, applications, and issues USA: Nelson Education; 2017.
5. Revelle W, Condon DM. Reliability from α to ω: A tutorial. Psychological assessment. 2019 Dec;31(12):1395.
6. Taylor JM. Reliability. Journal of Nursing Education 2021; 60: 65–66.
7. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MD, Horsley T, Weeks L, Hempel S. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine. 2018 Oct 2;169(7):467-73.
8. Urbina J, Monks SM. Validating Assessment Tools in Simulation. StatPearls Publishing, Treasure Island (FL); 2022.
9. Baumgartner R, Joshi A, Feng D, Zanderigo F, Ogden RT. Statistical evaluation of test-retest studies in PET brain imaging. EJNMMI research. 2018 Dec;8(1):1-9.
10. Lilienfeld S, Lynn SJ, Namy L, Woolf N, Jamieson G, Marks A, Slaughter V. Psychology: From inquiry to understanding. Pearson Higher Education AU; 2019 Oct 1.
11. Matheson GJ. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. PeerJ. 2019 May 24;7:e6918.
12. Peer E, Rothschild D, Gordon A, Damer E. Erratum to Peer et al. Data quality of platforms and panels for online behavioral research. Behavior Research Methods. 2022 Oct;54(5):2618-20.
13. Pearson K. Correlation coefficient. InRoyal Society Proceedings 1895 (Vol. 58, p. 214).
14. Aqil A, Saldana K, Ndu M. Reliability and validity of an innovative high performing healthcare system assessment tool. BMC Health Services Research. 2023 Dec;23(1):1-8.
15. Gullo S, Kuhlmann AS, Galavotti C, Msiska T, Nathan Marti C, Hastings P. Creating spaces for dialogue: a cluster-randomized evaluation of CARE's Community Score Card on health governance outcomes. BMC Health Services Research. 2018 Dec;18(1):1-2.

16. Michaelis M, Rieger MA, Burgess S, Töws V, Abma FI, Bültmann U, Amick BC, Rothermund E. Evaluation of measurement properties of the German Work Role Functioning Questionnaire. BMC Public Health. 2022 Dec;22(1):1-9.
17. Burgess S, Junne F, Rothermund E, Zipfel S, Gündel H, Rieger MA, Michaelis M. Common mental disorders through the eyes of German employees: attributed relevance of work-related causes and prevention measures assessed by a standardised survey. International archives of occupational and environmental health. 2019 Aug 1;92:795-811.
18. Palmquist AE, Perrin MT, Cassar-Uhl D, Gribble KD, Bond AB, Cassidy T. Current trends in research on human milk exchange for infant feeding. Journal of Human Lactation. 2019 Aug;35(3):453-77.
19. Gondivkar SM, Gadbail AR, Sarode SC, Gondivkar RS, Yuwanati M, Sarode GS, Patil S. Measurement properties of oral health related patient reported outcome measures in patients with oral cancer: A systematic review using COSMIN checklist. PloS one. 2019 Jun 27;14(6):e0218833.
20. Spearman C. Correlation calculated from faulty data. British journal of psychology. 1910 Oct 1;3(3):271.
21. Thompson. What is the Spearman-Brown Formula? - Psychometric analytics. Assessment Systems, https://assess.com/spearman-brown-prediction-formula/ (2018).
22. Kuder G, Richardson M. The theory of the estimation of test reliability. Psychometrika. 1937;2(3):151-60.
23. Kuder-Richardson Formula 20. Psychology Wiki, https://psychology.fandom.com/wiki/Kuder-Richardson_Formula_20.
24. Anselmi P, Colledani D, Robusto E. A comparison of classical and modern measures of internal consistency. Frontiers in psychology. 2019 Dec 4;10:2714.
25. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.
26. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Frontiers in public health. 2018 Jun 11;6:149.
27. Shanem. Psychometrics: Validity and Reliability. Psychometrics: Validity and Reliability, https://pages.mtu.edu/~shanem/psy5220/daily/Day05/psychometrics.html (2023).
28. Reliability and Validity. Reliability and Validity, https://chfasoa.uni.edu/reliabilityandvalidity.htm.
29. Østerås N, Tveter AT, Garratt AM, Svinøy OE, Kjeken I, Natvig B, Grotle M, Hagen KB. Measurement properties for the revised patient-reported OsteoArthritis Quality Indicator questionnaire. Osteoarthritis and Cartilage. 2018 Oct 1;26(10):1300-10.Pearson K. Correlation coefficient. In Royal Society Proceedings. 1895; 58, 214.
30. Reliability and Validity | Research Methods for the Social Sciences. Chapter 7 Scale Reliability and Validity | Research Methods for the Social Sciences, https://courses.lumenlearning.com/suny-hccc-research-methods/chapter/chapter-7-scale-reliability-and-validity/.

31. Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement. 1960;20(1):37-46.
32. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin. 1968;70(4):213.
33. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement. 1973;33(3):613-9.
34. Factors Influencing the Reliability of Test Scores. Your Article Library, https://www.yourarticlelibrary.com/statistics-2/factors-influencing-the-reliability-of-test-scores/92601 (2018).
35. Validity of a Test: 5 Factors | Statistics. Your Article Library, https://www.yourarticlelibrary.com/statistics-2/validity-of-a-test-5-factorsstatistics/92589 (2016).
36. Balasubramanian N, Juliana LD, Sathyanarayana RT. Knowledge questionnaire on home care of schizophrenics (KQHS): validity and reliability. J Educ Pract. 2013;4(11):176-82.
37. Balasubramanian N. Likert technique of attitude scale construction in nursing research. Asian Journal of Nursing Education and Research. 2012 Apr 1;2(2):II.
38. Pandurangan H, Balasubramanian N, Raja A. Development of Case Vignette Tool on ECG and its Interpretation (CVECGI). International Journal of Innovative Science and Research Technology. 2021; 6(5), 527-532.