Reliability Measurements: Methods and Estimation in Healthcare Research

Pritha Sarkar[1], Dr. Sunita Srivastava[2], Dr. Hariprasath Pandurangan[3*], Dr. Anil Kumar [4]

[1] General Manager- Clinical, Arjo Huntleigh Healthcare India Private Limited, Mumbai, India

[2]Principal, Amity College of Nursing, Amity University Haryana, Gurugram, India

[3]Asst. Professor, Amity College of Nursing, Amity University Haryana, Gurugram, India

[4]Assistant Professor, School of Legal Studies, Central University of Kashmir, India

Email: [3]hpandurangan@ggn.amity.edu

Abstract

Reliability refers to how much the test, process, or instrument produces similar results under different conditions under similar conditions. Reliability is crucial for tests intended to be stable over time. Although it's impossible to determine Reliability precisely, various techniques exist to assess it. This article focuses on methods for computing reliability in quantitative data, including ratio and interval data. The primary purpose of this paper is to discuss the idea of Reliability and present the calculation of Reliability for commonly used research instruments in simple language with examples. The article presents methods and measures of statistical Reliability. It includes Stability, internal consistency, and equivalence measurement. The authors estimated the Stability of the instrument using Karl Pearson's Coefficient of Correlation by adopting the test and retest method. Internal consistency of the instrument was estimated by Spearman-Brown Prophecy, Kuder-Richardson 20, Kuder-Richardson 21, and Cronbach's alpha formulas. The Cohen kappa correlation coefficient and Fleiss kappa correlation coefficient estimated the equivalence of the instrument. It is concluded that young and inexperienced researchers should know the significance of Reliability, its measures and how to ascertain it correctly. A greater understanding of score reliability will help them to avoid misunderstandings and write and discuss cautiously about Reliability estimates.

Keywords: Reliability estimation, Measurement methods, Internal consistency, Test-retest reliability, Psychometric assessment, Quantitative data.

Introduction

In the realm of research, ensuring the precision of an apparatus constitutes a crucial preliminary step before its utilization for data collection. The researcher is responsible for minimizing errors to the utmost degree to attain the highest degree of accuracy for the given tool or device. The significance of Reliability underscores the selection of a measuring instrument, as it guarantees the constancy and coherence of research outcomes over time, even in the presence of variations within the population or Sample.[1,2] This multifaceted issue necessitates engagement through diverse methodologies. An array of instruments encompassing mechanical and electrical equipment, questionnaires, schedules, opinionnaires, and rating scales is employed in research for data collection. Confirming the Reliability of these instruments establishes their credibility and trustworthiness. Researchers must exercise prudence in using instruments known for their Reliability, as this augments the robustness of the results. In empirical research, diverse techniques are deployed to gauge Reliability, encompassing commonplace approaches such as test-retest Reliability, parallel/alternative

forms, and internal consistency. Internal consistency assessments manifest in three variants: split-half, items-total correlations, and the alpha reliability coefficient.[3-5]

Researchers employ the previously mentioned methods in studies focused on developing scales to assess their trustworthiness. However, in the case of established tools, solely conducting an internal consistency test suffices. The prevalent method for evaluating internal consistency is utilizing the Alpha reliability coefficient. From a statistical perspective, the reliability computation employs the correlation coefficient formula, yielding a numerical range between 0 and 1. Elevated values within this range indicate heightened consistency among the measurements.[6-8]

Reliability plays an undeniable, pivotal role in healthcare research. Nevertheless, the existing body of literature exposes significant gaps that underscore the necessity for an all-encompassing review. Notably, the comprehension of Reliability within current resources often needs to be more cohesive, leaving researchers with incomplete and disjointed insights that impede a holistic understanding of the topic. A lack of standardized approaches for applying methods to estimate reliability compounds this issue, leading to confusion and hindering the establishment of universally recognized best practices. The inadequate attention to addressing reliability concerns during the study design phase is equally noteworthy. This oversight can result in potential weaknesses during data collection and subsequent analysis, ultimately diminishing the strength of research findings.

Efforts to bridge these gaps in healthcare research can elevate the overall quality of research, bolster its credibility, and furnish decision-makers with dependable data, culminating in more trustworthy outcomes.

Paradoxically, despite the abundance of printed and online literature dedicated to Reliability and its computation, novice researchers often need help to grasp its intricacies. This paper fulfills the purpose of acquainting researchers with the primary methodologies for assessing the Reliability of tools such as opinionnaires, schedules, questionnaires, and rating scales. Presented in plain and accessible language, augmented by illustrative examples, this paper aims to demystify these techniques for novice researchers. By doing so, it endeavors to foster a clearer comprehension of Reliability and its computation, ultimately aiding researchers in developing a more profound understanding of this essential concept.

Measurement of Reliability

Reliability measurement encompasses three key attributes: Stability, Homogeneity, and Equivalence.

Stability refers to the consistency of results obtained when the same instrument is administered repeatedly. It dictates that the instrument should yield consistent outcomes across its multiple applications. Stability can be established through the test-retest method and parallel or alternate instrument forms. To assess the consistency, the test-retest method involves administering the same instrument twice to participants under similar conditions over a defined period. To minimize potential errors, researchers must ensure uniformity in procedures, environment, lighting, and time of day. Subsequently, the coefficient "r" value is computed by comparing two scores. A higher coefficient signifies greater Stability of the tool.[8-12]

The procedure involves the following **Steps:**

1. Administer the test to a sufficiently large group, ideally comprising more than 30 participants.
2. Re-administer the test to the same group after a defined interval.
3. Aim for a second administration around two weeks after the initial one, although this time frame may vary depending on the context.
4. Ensure that no intervening activities between the two administrations could influence the measured characteristic.

5. Calculate the correlation coefficient for the obtained scores.
6. Compute the Karl Pearson's Correlation Coefficient.

By adhering to these steps, researchers can ascertain the Stability of an instrument and its consistency over time.

Karl Pearson (r) Formula

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

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

The advantage of the Stability Method

The approach is suitable for gauging characteristics that exhibit Stability over time and remain unaffected by the person conducting the measurement. It also guarantees uniform outcomes. A consistent set of items or stimulus scenarios ensures that only the intended traits are being assessed.[13]

The disadvantage of the Stability Method

Subjects can gain insights from participating in a test, influencing their subsequent performance. This phenomenon can affect the results during the second testing session. Maturation can transpire if the interval between the initial and follow-up tests is too extended. Maturation refers to alterations in subject-related factors or respondents over time, leading to measurement variations captured in different instances. The test-retest method is susceptible to reactivity, wherein the measurement changes the observed phenomenon.[13]

Table 1: Computation of Karl Pearson coefficient using fictitious data

Sample No	**Test X**	**Retest Y**	**X-$\overline{X}$**	**Y-$\overline{Y}$**	**(X-$\overline{X}$) (Y-$\overline{Y}$)**	**(X−$\overline{X}$)2**	**(Y-$\overline{Y}$)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

$\overline{X}$ = 54.8, $\overline{Y}$ = 54.8 $\sum$ **(X** − $\overline{X}$**) (Y-**$\overline{Y}$**)** = 1152.6, **(X**−$\overline{X}$**)2** =1285.6**, (Y-**$\overline{Y}$**)2** = 1137.6, √(**X** − **X**$\overline{)}$2 = 35.85**,** √(**Y** − **Y**$\overline{)}$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.

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 taken using any instrument 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. Kuder-Richardson is used on questions with two answers, e.g. true or false / yes or no/ dichotomous measurements with a score of 0 or 1. All correct responses are scored as +1, and incorrect responses as zero.

In most cases, Cronbach's alpha is utilized 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.[14-19]

Steps:

1. Administer the test to a substantial group, preferably exceeding a certain threshold.
2. Randomly divide the test questions into two segments. It could involve separating items into halves of equal size or using an odd-even approach.
3. Calculate the correlation coefficient for the two divided sections.
4. Calculate the appropriate statistical measure based on the tool employed, which could involve the Spearman-Brown formula, Kuder-Richardson coefficient, or Cronbach's alpha coefficient. The selection depends on the specific measurement tool being used.

Formula: Spearman-Brown Prophecy

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

Where rtt = Reliability of entire test, rh = reliability calculated through Karl Pearson 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 − $\bar{X}$**	**Y-$\bar{Y}$**	**(X − $\bar{X}$) (Y-$\bar{Y}$)**	**(X−$\bar{X}$)2**	**(Y-$\bar{Y}$)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 rh, we have to use the Karl Pearson formula,

$\bar{X}$ = 27.5, $\bar{Y}$ = 26.3, $\sum$ **(X − $\bar{X}$) (Y-$\bar{Y}$)** = 242.5, **(X−$\bar{X}$)2** =248.5**, (Y-$\bar{Y}$)2** = 398.1, $\sqrt{}$(**X − $\bar{X}$**)2 = 15.76**,** $\sqrt{}$(**Y − $\bar{Y}$**)2 = 19.95

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

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

Kuder-Richardson 20 Formula

The KR20 is a statistical measure that allows to compute Reliability for items with varying difficulty. For example, in Multiple choice questions, some items might be straightforward, and others may be more difficult. The limitation of the KR-20 formula is that it cannot be applied to scales such as *Likert* and *Visual Analogue*.

$$KR_{20} = \frac{K}{K-1}\left[1 - \frac{\sum pq}{\sigma^2 X}\right]$$

Here, K = Number of items; p = Proportion of right answer; q = Proportion of wrong answer; σ2X = 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 responses	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), ∑pq= 2.14, σ2X=5.04

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

Kuder-Richardson 21 Formula

It is used for a test where the items are all about the same difficulty level. For example, True or False, Yes or No type of questions.

$$K-R21=\frac{k}{k-1}\left(1-\frac{M(k-M)}{kS^2}\right)$$

K = Number of items; M= Mean of Total Score; S^2 = Variance or SD^2

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.

Cronbach's Alpha Formula

This formula is used to find the scale's Reliability, for example, the Likert Scale.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\Sigma Vi}{Vtest}\right)$$

- n = number of questions
- Vi = variance of scores on each question
- 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.

The advantage of the 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.

The disadvantage of the 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 form / 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 administered over a period of time. It avoids carryover biases. The first form and second form of the test are similar but not identical.[20-21]

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 of Form B) over time.
4. The finding of the correlation coefficient for the two halves.
5. Compute the Karl Correlation Coefficient (Refer to Test and Retest method) Form 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 are asked 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.[22-25]

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. 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).

Cohen Kappa Correlation Coefficient

$$K = \frac{\text{Number of agreements}}{\text{Number of agreements + Number of disagreements}}$$

Table 6: Computation of Cohen Kappa Coefficient using fictitious data

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

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

Fleiss Kappa Correlation Coefficient

$$K = \frac{Po - Pe}{1 - Pe}$$

Where Po = Observed Agreement, Pe = Expected Agreement.

Pe = (Proportion of Agreement)^2 + (Proportion of Disagreement)^2

$$\mathbf{Po} = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\right)$$ Where 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	X2	Y2	X2+ Y2
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**			

pe = 0.780+0.013 = 0.793, N= 20, n= 3, n2 = 152

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

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

Advantages of the Parallel form method:

1. The same test needs to be avoided.
2. Minimizing Memory, practice, carryover effects and recall factors without affecting the scores.
3. The method combines two types of Reliability as the reliability coefficient obtained is a measure of temporal Stability and consistency of response in the case of different item samples or test forms.
4. It proves useful in the Reliability of achievement tests.
5. The method is appropriate for determining the Reliability of educational and psychological tests.

Disadvantages of the Parallel form method:

1. There is difficulty in having two parallel forms of a test. In certain situations (like in the case of Rorschach), it proves to be impossible.
2. The comparison of two sets of scores that may be obtained from these tests may lead to flawed decisions, mainly when the tests are different in content difficulty and length.
3. Practice and carryover factors cannot be fully controlled.
4. The simultaneous administration of the two forms may create boredom. A single administration of the test is the 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.

Generally, the second form of test scores is high.

Table 8: Summary of Reliability Statistics

	Karl Pearson	**Spearman Brown Prophecy**	**KR-20**	**KR-21**	**Cronbach Alpha**	**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	

Discussion

In early contributions to the field, Karl Pearson[26] pioneered the linear correlation between two variables, an insight termed the product-moment correlation coefficient. This coefficient stands as a widely employed statistic in contemporary times, adept at evaluating the Stability of a test through the test-retest method. Kuder-Richardson[18] introduced the concept of internal consistency reliability for dichotomous measures, ones characterized by scores of 0 or 1. An inherent limitation of the K-R formula is its inapplicability to scales, confining researchers to employ KR-20 or KR-21 formulas solely for internal consistency estimation. In response, Cronbach[17] devised the Cronbach alpha method to gauge the internal consistency of tools featuring scale measures, such as Likert scales. This method is an extension of the KR-20 formula. Diverse strategies for determining Reliability have been embraced by researchers, with a subset discussed herein. Quinaud et al.[27] developed and validated the coach knowledge questionnaire, utilizing Karl Pearson's formula to measure the correlation coefficient via the split-half technique (odd-even). Subsequent deployment of Spearman's Brown Prophecy formula yielded a reliability value of r = 0.92. Similarly, Alam et al.[28] crafted the COVID-Vaccination Attitude Scale to evaluate attitudes towards COVID-19 vaccination. Employing the Karl Pearson correlation coefficient, they measured instrument stability and test-retest Reliability, obtaining a calculated value of 0.78. Moreover, Pandurangan & Balasubramanian[29] constructed a case vignette tool for assessing nursing students' ECG interpretation skills. Their analysis encompassed Cronbach's alpha and r values, serving to determine internal consistency and correlation. The resultant "r" value was estimated at 0.72.

Conclusion

All research measurements may involve a few errors that cannot be eliminated but can be reduced 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] Bowling A. Research methods in health. Milton Keynes: Open University Press; 2014.
[2] Drost EA. Validity and Reliability in social science research. Education Research and Perspectives. 2011;38(1):105-23.
[3] Luthans F, Avolio B, Avey J, Norman S. Positive psychological capital: measurement and relationship with performance and satisfaction. Personnel Psychology. 2007;60:541-572.
[4] Oluwatayo J. Validity and reliability issues in educational research. Journal of Educational and Social Research. 2012;2:391-395.
[5] Kaplan R, Saccuzzo D. Psychological testing: Principles, applications, and issues. USA: Nelson Education; 2017.
[6] Ruzafa-Martínez M, López-Iborra L, Madrigal-Torres M. Attitude towards Evidence-Based Nursing Questionnaire: development and psychometric testing in Spanish community nurses. Journal of Evaluation in Clinical Practice. 2011;17:664-70.
[7] Norman G, & Streiner L. PDQ statistics (Vol. 1). PMPH-USA; 2003.
[8] Henson R, Roberts J. Use of Exploratory Factor Analysis in Published Research. Educational and Psychological Measurement. 2006;66(3):393-416.
[9] Herbert C. Karl Pearson and the Human Form Divine, in Victorian Relativity: Radical Thought and Scientific Discovery. Chicago University Press; 2003.
[10] Pearson K. The grammar of science. Adam and Charles Black; 1900.
[11] Pearson K. National life from the standpoint of science. London: Adam & Charles Black. 1905;43–44.
[12] Pearson K. Introduction to The Grammar of Science. London: Water Scott. 1900;32.
[13] Cozby PC. Measurement concepts. Methods in Behavioral Research. California: Mayfield Publishing Company. 2001.
[14] Moskal B, Leydens J. Scoring rubric development: Validity and Reliability. Practical Assessment, Research, and Evaluation. 2000;7(1):10.
[15] Luiz R, Costa A, Kale P, Werneck G. Assessment of agreement of a quantitative variable: a new graphical approach. Journal of Clinical Epidemiology. 2003;56(10):963-7.
[16] Terwee C, Bot S, De Boer M, van der Windt D, Knol D, Dekker J, Bouter L. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007;60(1):34-42.
[17] Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334.
[18] Kuder G, Richardson M. The theory of the estimation of test reliability. Psychometrika. 1937;2(3):151-60.
[19] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(1):37-46.
[20] Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin. 1968;70(4):213.
[21] 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.
[22] Sim J, Wright C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy. 2005;85(3):257-68.
[23] Warrens MJ. Cohen's kappa is a weighted average. Statistical Methodology. 2011;8(6):473-84.
[24] Asaad A. Statistics as applied to education and other related fields. Manila: Rex Book Store; 2001.
[25] Calmorin L. Measurement and evaluation. Mandaluyong City: National Book Store; 2004.
[26] Pearson K. Correlation coefficient. In Royal Society Proceedings. 1895;58, 214.
[27] Quinaud RT, Backes AF, Nascimento Junior JR, Carvalho HM, Milistetd M. Development and validation of the coach knowledge questionnaire: measuring coaches' professional, interpersonal and intrapersonal knowledge. International Journal of Sport and Exercise Psychology. 2022;20(1):302-18.
[28] Alam MM, Melhim LK, Ahmad MT, Jemmali M. Public attitude towards COVID-19 vaccination: validation of COVID-Vaccination Attitude Scale (c-vas). Journal of Multidisciplinary Healthcare. 2022;15:941-54.
[29] 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.