Journal of Sport Science Technology and Physical Activities
ISSN : 1112-4032
eISSN 2543-3776

Vol: 19 / N°: 2 (december 2022),
p :69-82

The common mistake in choosing the appropriate statistical method for calculating reliability by test and re-test

Baazi Redhouan[1] ; Baazi Ahmed[2]

[1] M'hamed Bougherra Boumerdes Univesity, Algeria, r.baazi@univ-boumerdes.dz
[2]Abdelhamid Ben Badis University Mostaganem, Algeria, ahmedbaazi18@gmail.com

ARTICLE INFORMATION
ORIGINAL RESEARCH PAPER
RECEIVED : 14/07/2022
ACCEPTED : 03/10/2022
PUBLISHED: 01/12/2022

KEYWORDS :
Test Re-test reliability, correlation coefficient, interclass correlation coefficient (ICC)

Auteur correspondant : Baazi Redhouan, r.baazi@univ-boumerdes.dz

ABSTRACT

This study aimed to identify the statistical methods used in calculating reliability by test and re-test by presenting three models with a comparison between them by calculating each of the Pearson correlation coefficient (R) and the intraclass correlation coefficient (ICC). The results revealed that the Pearson correlation coefficient is not suitable as a test statistic for calculating stability as it gives the same value if the results differ in contrast to the (ICC) coefficient

doi.org/10.5281/zenodo.15270678

1. Introduction

Research tools are the main source for obtaining information and data about the phenomenon studied, and are divided according to the nature of the data into two types: quantitative tools, such as (closed questionnaire, tests, and measurements...), and qualitative tools such as (observation and interview...)
In order for the researcher to be objective, the research group must understand the purpose of the test and its instructions, and there should be no more than one interpretation of the required questions and answers.
and thus affect its stability; Since the objectivity of measurement contributes to defining concepts and crystallizing thinking for an enlightened understanding of the nature of various phenomena, and since measurement is of great importance in any science, all of them; That is, sciences seek to develop accurate objective methods for measuring the phenomena related to them in order to understand and explain these phenomena and to predict the relationships that exist between their variables and try to control and control them.
In order for the tool to be reliable and reliable in research, it must be stable; That is, it gives the same results if it is re-applied more than once to the same sample, and in the same conditions, and to verify the stability condition, there are several methods, including (testing and re-testing, equivalent images, internal consistency..
By reviewing the research literature and previous studies, the researcher noted that most of the research - especially Arabic - used the Pearson correlation coefficient as a statistical measure to calculate stability when using the test and re-test method, and this prompted the researcher to ask the research question:
Does the Pearson correlation coefficient reflect stability (the stability of results) when using the test-retest method?

1.1. Literature Review

Stability: stability is defined as a reliable and reliable test, and it also means stability Consistency (mohamed Hocine Bahi Mohamed)

Retest stability: The stability coefficient resulting from this method is called the stability coefficient, i.e. the stability of the test results during the period between the first and second application. (Madjid, 2013)
It is defined as the most widely used indicator of the reliability of survey tools, and it is measured by the presence of the same group of respondents at

two different points in time to determine the stability of the response. (LITWIN, 1995, p. 8)

In Quantitative Research:

There are three types of persistence that social researchers adopt:

- Reliability: It is related to stability over time, which means the extent to which the measures give the same results if they are applied in different periods of time.
- Representational stability: This type is related to stability across subject groups, and the question is: Will the scale remain reliable if it is applied to new groups.
- Equivalence stability: It means stability across indicators and with respect to multiple indicators in practical procedures (Shehda, 2017)

The question is: Does the scale produce consistent results according to different indicators?

Factors affecting the stability of the test:

- Number of items: The reliability coefficient value increases with the increase in the number of questions (test items).
- Drafting of items: Questions placed increase the reliability coefficient, and vague and long questions such as essay questions reduce the reliability coefficient.
- Variance questions (easy or hard) lead to a decrease in the reliability coefficient, while high-variance questions (average ease) lead to an increase in the reliability coefficient.
- Test performance time: Increasing the time leads the individual to obtain the highest degree consistent with his ability, but increasing the time to a greater degree than necessary may lead to confusion in the answer and thus reduce the stability coefficient.
- Guessing: The stability of the test decreases with the increase in the guessing rate. Therefore, true and false questions reduce the reliability coefficient for multiple choice questions.
- The health and psychological state of the individual: affects the stability coefficient. If he is tired, sick or tense, the stability decreases. (Delmi Isam hocin, 2014)

Statistical methods used to calculate stability:

- Stability factor: (retest method)
- Equivalence factor: (equivalence images)
- Internal consistency coefficient: (halved split) (Madjid, 2013)

Pearson's correlation coefficient (r):

It is one of the correlation coefficients to measure the relationship between two variables with different measuresIt (Wiley & Sons, 2005) is calculated according to the following equation: $$r = \frac{\sum xy - \sum(x)\sum(y)}{\sqrt{[n\sum x^2 - n\sum(x)^2][n\sum y^2 - n\sum(y)^2]}}$$

Statistical properties of correlation coefficient (r)

- The value of the numerical correlation coefficient does not exceed one and all values are within ±1
- The correlation coefficient is not affected by changing the unit of measurement or adding or subtracting a fixed amount that is not equal to zero to or from each degree of one or both of the two distributions. The point of origin and the unit of measure.
- The value of the correlation coefficient depends on the characteristics of the sample, as the sample size does not affect the significance of the correlation coefficient.
- The value of the correlation coefficient is affected by the extent of the degree of variation in both distributions. The value of the correlation coefficient calculated from a set of varying degrees to a large extent is greater than its value if the set of scores are close in one or both variables.
- The nature of the relationship between two variables is assumed to be linear. (Hasan el jendi)

Interclass Correlation Coefficient (ICC):

It is a statistical method that refers to measuring the difference between subjects, and it is expressed by a numerical value that is between (0,1), and it is one of the measures recommended for measuring stability. (David, Elfving, & oaldsen, 2019, p. 1)

Table 1 represents the three models of the inter-category correlation coefficient ICC (David, Elfving, & oaldsen, 2019, p. 9)

Sample ICC	model
(1) $ICC = \frac{MSBS - MSWS}{MSBS + (K-1)MSWS}$	**Model 1** $x_{ij} = \mu + r_i + v_{ij}$ random
Agreement (A1) $ICC = \frac{MSBS - MSWS}{MSBS + (K-1) + \frac{K}{N}(MSBM - MSE)}$ **Consistency (C1)** $ICC = \frac{MSBS - MSWS}{MSBS + (K-1)MSE}$	**Model 2** $x_{ij} = \mu + r_i + c + v_{ij}$ random
Agreement (A1) $ICC = \frac{MSBS - MSWS}{MSBS + (K-1)MSE + \frac{K}{N}(MSBM - MSE)}$ **Consistency (C1)** $ICC = \frac{MSBS - MSWS}{MSBS + (K-1)MSE}$	**Model 3** Fixed $x_{ij} = \mu + r_i + c_j + v_{ij}$ random

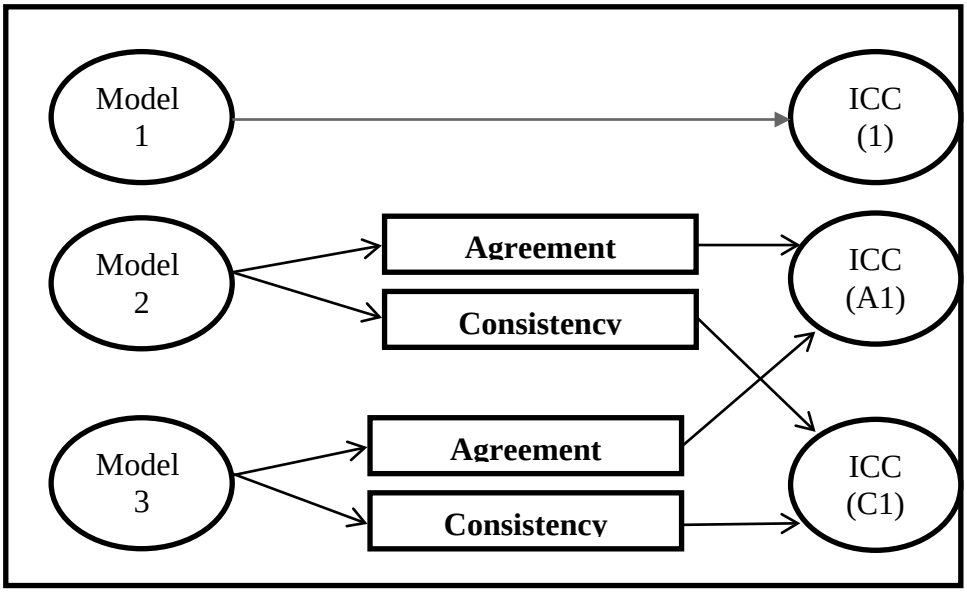
Through the previous table, it is clear that the inter-category correlation coefficient (ICC) is divided into three models:

- Model (1): Unilateral, random, without bias
- Model (2): Binary Random Bias
- Model (3) Binary Mixed

whereas:

- μ: average vocabulary score
- K: number of measures
- n: number of vocabulary
- r: reaction of measurements
- V: measurement errors
- c: measurement bias
- MSBS: mean of squares between groups
- MSWS: mean of squares within groups
- MSE: Mean Squared Errors
- Consistency
- Agreement

Figure 1 represents modulus models and their relationship to agreement and consistency equations



(David, Elfving, & oaldsen, 2019)

2. Method

The researcher used the survey method to suit the objectives of the research through comparison between three groups

2.1. Participants

Table No. 02 represents the proposed numerical models

SAMPLE	(01) MODEL		(02) MODEL		(03) MODEL	
	Test	Retest	Test	Retest	Test	Retest
01	6	8	6	5	6	6
02	8	10	8	10	8	8
03	10	12	10	15	10	10
04	12	14	12	20	12	12
05	14	16	14	25	14	14

3. Results

Table No. 03 represents the arithmetic averages and standard deviations of the application and re-application (model 1)

Item Statistics			
	Mean	Std. Deviation	N
Test	10,00	3,162	5
Retest	12,00	3,162	5

We note from Table No. 03 of (Form No. 01) that the arithmetic mean for the first application was 10.00 with a standard deviation of 3,162. As for the re-application, the arithmetic average reached 12.00 with a standard deviation of 3,162

Table No. 04 represents the Pearson correlation coefficient between application and re-application (model 1)

Inter-Item Correlation Matrix		
	Test	Retest
Test	1,000	1,000
Retest	1,000	1,000

We note from Table No. 04 that the Pearson correlation coefficient amounted to 1.00, which indicates the existence of a complete direct correlation between the first application and the second application, that is, the higher the degrees of the first application, the higher the degrees of the second application

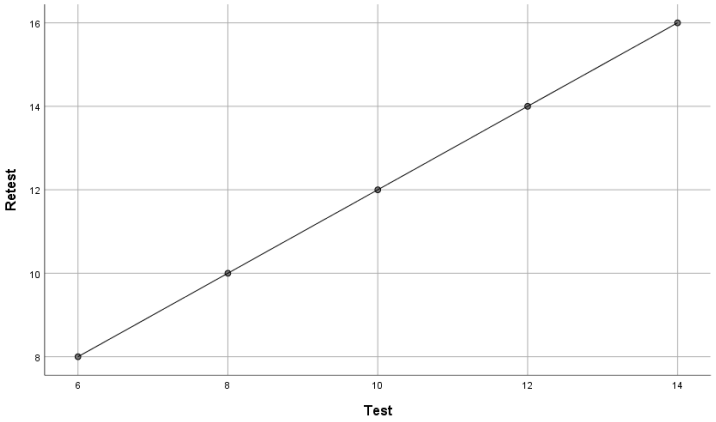


Figure No. 02 shows the spread form for degrees of application and re-application (Model No. 01)

We note through the diffusion form of (Model No. 01) that there is a complete linear correlation between the test and retest, that is, the higher the test value by one degree, the higher the test value by one degree, and vice versa.

Table No. 05 represents the inter-category correlation coefficient (ICC) between application and re-application (Model No. 01)

Intraclass Correlation Coefficient							
	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,833[a]	,006	,984	.	4	.	.
Average Measures	,909	,011	,992	.	4	.	.
Two-way random effects model where both people effects and measures effects are random.							
a. The estimator is the same, whether the interaction effect is present or not.							
b. Type A intraclass correlation coefficients using an absolute agreement definition.							

We note from Table No. 05 of (Model No. 01) that the correlation coefficient (ICC) between the categories is 0.833 between cases, while the correlation coefficient (ICC) between the averages was 0.909, which

indicates a high stability between application and re-application when compared with the relationship indicator.

Table No. 06: Arithmetic averages and standard deviations represent application and re-application (Model No. 02)

Item Statistics			
	Mean	Std. Deviation	N
Test	10,00	3,162	5
Retest	15,00	7,906	5

We note through Table No. 06 of (Form No. 02) that the arithmetic mean for the first application was 10.00 with a standard deviation of 3,162. As for the re-application, the arithmetic mean reached 15.00 with a standard deviation of 7,906

Table No. 07: Represents the Pearson Correlation Coefficient (R) between application and re-application (Model No. 02)

Inter-Item Correlation Matrix		
	Test	Retest
Test	1,000	1,000
Retest	1,000	1,000

We note through Table No. 07, which shows the relationship between the application and re-application of (Form No. 02) that the Barson correlation coefficient (R) between the test and re-test was 1.00, which indicates the existence of a full direct correlation relationship, that is, the higher the degrees of the first application, the higher the scores The second application.

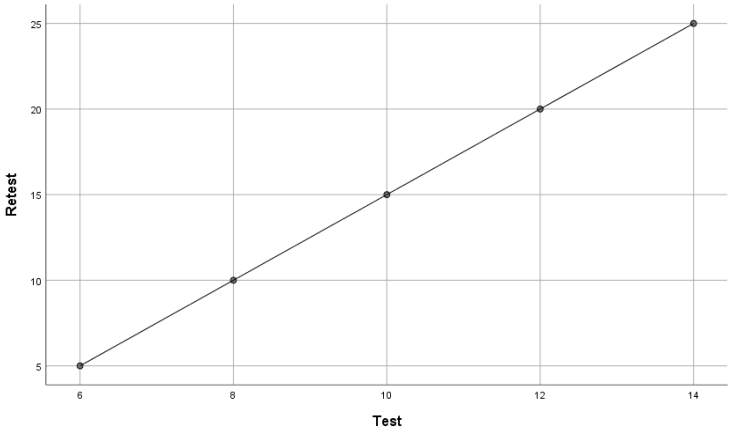


Figure No. 03: shows the spread form of test scores and retesting (Form No. 01)

We note through the diffusion form of (Model No. 02) that there is a complete linear correlation between the application and re-application, that is, the higher the test value, the higher the test re-test value, but not to a fixed degree and vice versa.

Table No. 08: Represents the Inter-Category Correlation Coefficient (ICC) between application and re-application (Model No. 02)

Intraclass Correlation Coefficient							
	Intraclass Correlation [b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,538[a]	-,169	,932	5,444	4	4	,065
Average Measures	,699	-,407	,965	5,444	4	4	,065
Two-way random effects model where both people effects and measures effects are random.							
a. The estimator is the same, whether the interaction effect is present or not.							
b. Type A intraclass correlation coefficients using an absolute agreement definition.							

We note through Table No. 08 of (Model No. 02) that the correlation coefficient (ICC) between the categories is 0.538 between the cases, and the correlation coefficient (ICC) between the averages was 0.699, which indicates a low stability between application and re-application when compared with the relationship indicator.

Table No. 09: Arithmetic averages and standard deviations represent application and re-application (Model No. 03)

Item Statistics			
	Mean	Std. Deviation	N
Test	10,00	3,162	5
Retest	10,00	3,162	5

We note through Table No. 09 of (Form No. 03) that the arithmetic mean for the first application was 10.00 with a standard deviation of 3,162. As for the re-application, the arithmetic mean reached 10.00 with a standard deviation of 3,162

Table No. 10 represents the Pearson correlation coefficient (R) between application and re-application (model No. 3)

Inter-Item Correlation Matrix		
	Test	Retest
Test	1,000	1,000
Retest	1,000	1,000

We note from Table No. 10 that the Pearson correlation coefficient amounted to 1.00, which indicates the existence of a complete direct correlation between the first application and the second application, that is, the higher the degrees of the first application, the higher the degrees of the second application.

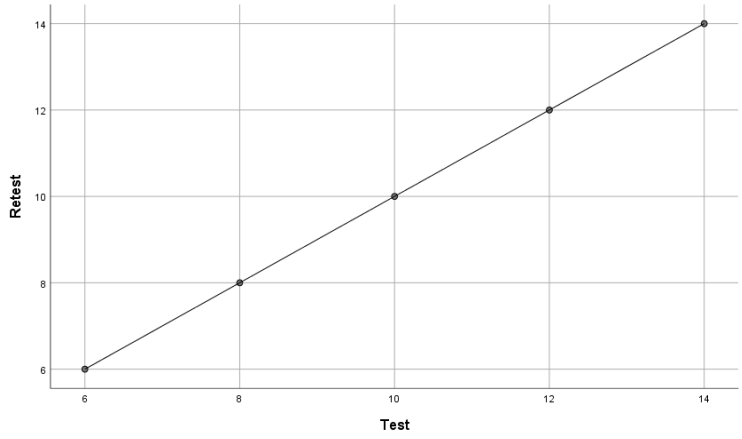


Figure No. 04: shows the spread form of test scores and retesting (Form No. 03)

We note through the diffusion form of (Model No. 02) that there is a complete linear correlation between the test and retest, that is, the higher the test value, the higher the test retest value, but not to a fixed degree and vice versa.

Table No. 11 represents the inter-category correlation coefficient (ICC) between application and re-application (Model No. 03)

Intraclass Correlation Coefficient							
	Intraclass Correlation [b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	1,000[a]	.	.	.	4	.	.
Average Measures	1,000	.	.	.	4	.	.
Two-way random effects model where both people effects and measures effects are random.							
a. The estimator is the same, whether the interaction effect is present or not.							
b. Type A intraclass correlation coefficients using an absolute agreement definition.							

We note from Table No. 11 of (Model No. 03) that the correlation coefficient (ICC) between the categories is 1.00 between the cases, and the correlation coefficient (ICC) between the averages was 1.00, which indicates that there is strong stability between the application and re-application when compared relationship index.

4. Discussion

The results of (Form No. 01) revealed that the arithmetic mean for the first application was 10.00 with a standard deviation of 3.162, as for the re-application, it reached 12.00 with a standard deviation of 3.162. With regard to the Pearson correlation coefficient, it reached 1.00 The correlation coefficient ICC) was 0.83 among the cases and 0.909 between the means.

As for (Form No. 02), the arithmetic mean for the first application was 10.00 with a standard deviation of 3,162, and for re-application, the mean was 15.00 with a standard deviation of 7,906.

The value of the Pearson correlation coefficient did not change, while the coefficient of (ICC) decreased; Among the cases, there was 0.538 and the averages were 0.699.

And after the arithmetic mean for (Model No. 03) for the first application was 10.00 with a standard deviation of 3.162, the arithmetic mean after re-application was 10.00 with a standard deviation of 3.162.

As for the Berson correlation coefficient, it kept the same value, while the (ICC) coefficient became equal to 1.00.

5. Conclusion

In light of the results that were reached, we note that the Pearson correlation coefficient did not change across the three models despite their differences, while the coefficient (icc) differed from one model to another.

We conclude that the Pearson correlation coefficient is not affected by changing the unit of measure or adding or subtracting a fixed amount that is not equal to zero to or from each degree of one or both of the two distributions of the two variables. Its value does not change with the change of the point of origin and the unit of measurement scale.

Accordingly, it is recommended to use the (icc) coefficient when calculating the stability of the application and re-application, which shows the agreement between the degrees, in contrast to the Berson correlation coefficient, which reflects the consistency between them.

References

David, L., Elfving, B., & oaldsen, K. S. (2019). Intraclass corrolation-Adiscution and demonstration of baisic features. (U. o. Ferdinando Chaicchio, Ed.) *PLOS ONE | https://doi.org/10.1371/journal.pone.*, 9.

Delmi Isam hocin, A. i. (2014). *Scientific research foundations and methods.* Aman: Dar Al Radwan.

fareg, s. (2017). *social search* (1 ed.). The Arab Center for Research and Policy Studies.

Hasan el jendi, H. D. (s.d.). *IBM Statistics.* Anglo-Saxon library.

LITWIN, M. (1995). *how to measure survey reliability and validity.* CALIFORNIA.

Madjid, S. C. (2013). *The foundations of constructing psychological and educational tests and measures.* Bagdad.

mohamed hocine bahimohamed, a. a. (n.d.). *Applied Statistics.* (A.-S. library, Ed.) Egypt.

Salem, A. a. (s.d.). *Applied Statistics.* Egypt: Anglo-Saxon library.

Shehda, F. (2017). *Social research.* The Arab Center for Research and Policy Studies.

Wiley, J., & Sons, L. (2005). *Encyclopedia of statistics in Behavioral Science.*