Sapporo Medical Journal
Volume 55, Issue 05, May 2021

Selecting and Reporting Reliability of Continuous Measurements: A Practical Guideline for Clinical Researchers

Muhamamd Amirul Mat Lazin[1], Mohamad Arif Awang Nawi[1]

[1]School of Dental Sciences, Health Campus, Universiti Sains Malaysia, Kubang Kerian, 16150 Kota Bharu, Kelantan, Malaysia

Abstract— The purpose of this article is to guide researchers to understand the basic concept of ICC and example ICC analysis will provide for easier to understanding the ways the process analysis in SPSS software. Selection the correct ICC reliability should include software information, "model" and "type," selections. For model selection focusing on one-way random, two-way random and two-way mixed model. Besides that, for type selection only focusing on "absolute agreement" and "consistency". Interpretation of the result from ICC must follow the guideline because currently a lack of standard for reporting ICC in the clinical research community. It is important for researchers to report detailed information about their ICC estimates by inserting software information, "Model", "Type," selections and 95% confidence intervals.

Keywords— ICC, one-way random, two-way random, two-way mixed model, absolute agreement, consistency

1. Introduction

What is reliability? These questions are addressed through the understanding of reliability. There are many definitions of reliability among researchers and the easy-to-understand meaning is defined a fundamental element in the evaluation of a measurement instrument. Before any measurement instruments or assessment tools can be used for research or clinical applications, their reliability must be established [1]. There are several reliability analyzes used in research, such as Internal Consistency Reliability (Cronbach's Alpha), Intraclass Correlation Coefficient (ICC) Analysis and Cohen's Kappa Analysis. In this research paper, we are focusing on Intraclass Correlation Coefficient (ICC). R. A. Fisher first introduced the concept of ICC in 1954 as a modification of Pearson correlation coefficient [2]. With changing times and circumstances, modern ICC is calculated by mean squares (i.e. estimates of the population variances based on the variability among a given set of measures) obtained through analysis of variance.

Nowadays, ICC is frequently used in medical and health field to evaluate three aspects of reliability are usually of interest such as intrarater reliability, interrater reliability and test–retest reliability [3][4][5]. These evaluations are fundamental to clinical assessment because without three aspects of reliability, the researcher will have no confidence in their measurements and researcher cannot draw any rational conclusions from measurements. There are different aspects of ICC that can give different results when applied to the same set of data, and the ways for reporting ICC may vary between researchers. It is important that researchers are aware of the correct application of each aspect of ICC, use the appropriate aspect in their analyses, and accurately report the which ICC they used [6]. What kind of situation would allow researchers to use interrater, intrarater and test-retest reliability? Description of all three analyses as shown in the Figure 1.

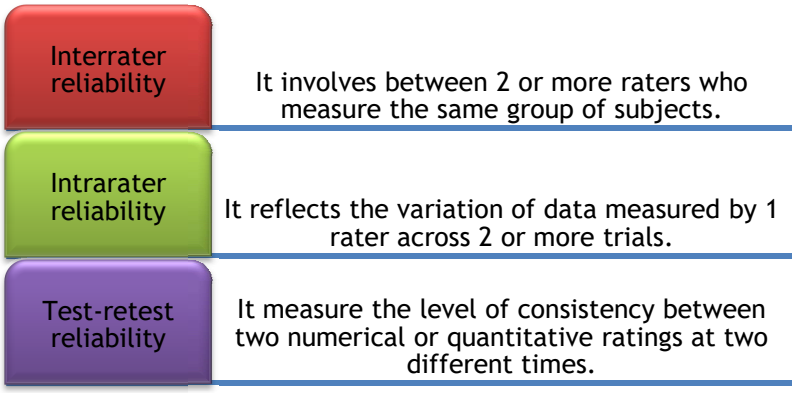

Figure 1. Description of ICC Reliability

The purpose of this article is to guide researchers to understand the basic concept of ICC and example ICC analysis will provide for easier to understanding the ways the process analysis in SPSS software. As conclusion, researcher can apply it to better interpret the reliability data.

2. Methodology

2.1 Correct ICC Selection Guide for Interrater Reliability Studies

Selecting the correct ICC for Interrater reliability based on the model and types as shown in Figure 2. The selection of model and type should be done with prudence in accordance with the data, the circumstances of the rater or judgment, difference trial and time required in the study.

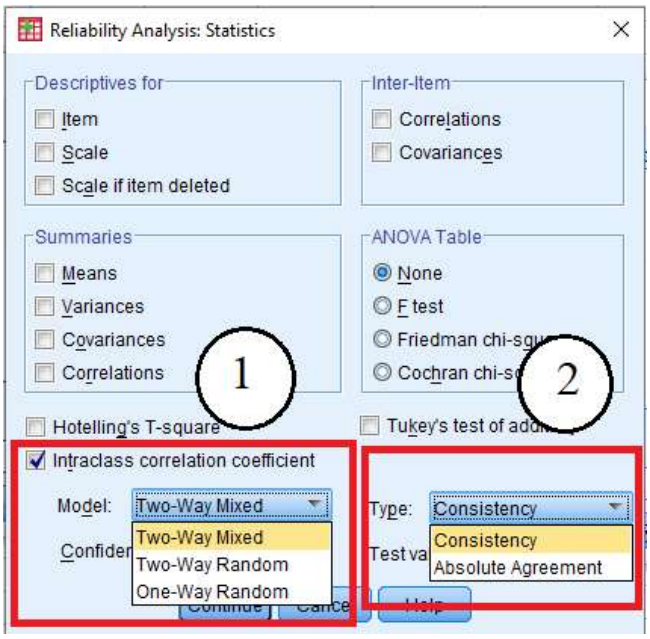

Figure 2: Model and Type of Reliability

"Model" Selection,

i. **One-way Random-Effects Model** can use when each subject is rated by a different raters and randomly chosen from a larger population of possible raters. Example scenario for this model is a rater may assess a subgroup of subjects in center A and another rater may assess a subgroup of

subjects in center B.

ii. **Two-way Random-Effects Model** used when the researcher randomly selects raters from a larger population and consider same/similar characteristics when selecting the raters (for example; the same years of working experience). This model is suitable for evaluating rater-based clinical assessment methods. This model is commonly used in **interrater reliability** analysis.

iii. **Two-way Mixed-Effects Model** used when researcher select the specific rater (when raters was fit) in the reliability experiment. With this model, the results only represent the reliability of the specific raters involved in the reliability experiment. This model is **less commonly** used in interrater reliability analysis.

"Type" Selection,

There are two definitions of ICC when researcher using 2-way random- and 2-way mixed-effects models namely "absolute agreement" and "consistency". Which selection should be made? This selection depends on how the measurement protocol will be conducted in actual application. Absolute agreement concerns if different raters assign the same score to the same subject. Instead, consistency defining concerns if raters' scores of the same group of subjects are correlated in an additive manner [7].Consider an interrater reliability study of 2 raters as an example. In this case, consistency defining concerns the degree to which one rater's score (y) can be equated to another rater's score (x) plus a systematic error (c) (ie, y = x + c). Whereas an absolute agreement concerns about the extent to which y equals x [6].

2.2 Correct ICC Selection Guide for for Test-Retest and Intrarater Reliability Studies

The ICC selection process of the test-retest and intrarater reliability is easier and not complicated. 2-way mixed-effects model is appropriate for testing intrarater reliability with multiple scores from the same rater, as it is not reasonable to generalize one raters' scores to a larger population of raters [8]. Similarly, 2-way mixed-effects model should also be used in test-retest reliability study because repeated measurements cannot be regarded as randomized samples [9]. In addition, Koo & Li [6] stated that the absolute agreement definition should always be chosen for both test-retest and intrarater reliability studies because measurements would be meaningless if there is no agreement between repeated measurements [6].

3. Result

The low result from ICC such as low degree or measurement agreement relate to the lack of variability among the sampled subjects, the small number of subjects, and the small number of raters being tested [10] [9]. Under such conditions, researcher suggest that ICC values based on the Table 1.

Table 1. Determine Level of ICC Reliability

ICC value	Level of reliability
Less than 0.5	Poor reliability
0.5 – 0.75	Moderate reliability
0.76 – 0.9	Good reliability
Greater than 0.9	Excellent reliability

In this paper demonstrates how to determine inter-rater reliability with the intraclass correlation coefficient (ICC) in SPSS. As part of the reliability analysis, SPSS statistical package version 18 computes not only an ICC value, but also its 95% confidence interval.

3.1 The Steps for Conducting an Inter-rater Reliability in SPSS

A comparison of the reliability of measurements from three physiotherapists was performed. Data from real-time ultrasound imaging for rehabilitation that involves observation of muscle contraction to provide feedback in 30 participants, one reading per one physiotherapist and physiotherapist selection are specifically based on the same years of working experience. Do the three physiotherapists produce 'reliable' readings? Steps to perform the Inter-rater reliability Analysis using SPSS software shown as Figure 3.

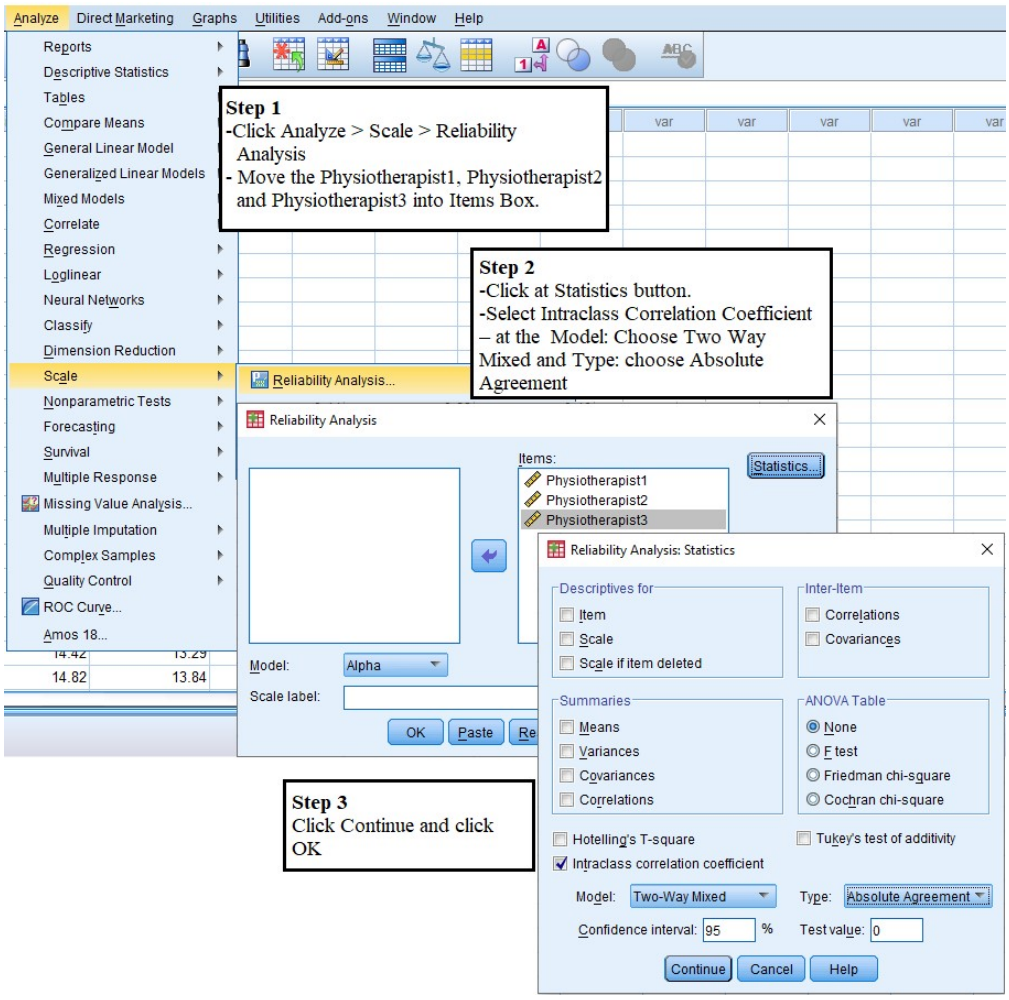

Figure 3: The Steps for Inter-rater reliability Analysis in SPSS

3.2 Interpreting the SPSS output for the Inter-rater reliability

Interpretation of the result from ICC must follow the guideline because currently a lack of standard for reporting ICC in the clinical research community. By giving different types of ICC involve distinct assumptions in theircalculation can cause different interpretations. It is important for researchers to report detailed information about their ICC estimates. The researcher suggests that the best practice of reporting ICC should include software information, “Model”, and “Type,” selections. In addition, both ICC estimates and their 95% confidence intervals should be reported.

Table 2. Hypothetical Example Showing Results of ICC Calculation in SPSS Using Absolute Agreement, 2-Way Mixed-Effects Model with 3 raters across 30 subjects

		95% Confidence Interval		F Test with True Value 0			
	Intraclass Correlation[a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	0.916[b]	0.853	0.956	33.417	29	58	0.000
Average Measures	0.970[c]	0.946	0.985	33.417	29	58	0.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Table 2 shows a sample output of an inter-rater reliability analysis from SPSS. In this hypothetical example, the obtained ICC was computed by an absolute-agreement, 2-way mixed-effects model with 3 raters across 30 subjects. Our estimated reliability between physiotherapists is 0.970 with 95% CI (0.946, 0.985), which is excellent reliability. As conclusion, we have evidence to support the reliability of this measurement between the three physiotherapists.

4. Conclusion

As a conclusion, ICC is a useful measure for describing reliability within a set of data and reflects both degrees of correlation and agreement between measurements. It has been often used in medicine, dental, and education field to evaluate interrater, test-retest, and intrarater reliability of continuous measurements/variable. Given that the types of ICC and every each involves distinct assumptions in their calculations and will lead to different interpretations, it is important for researchers and readers understand the principles of selecting an appropriate ICC. Because the ICC estimate obtained from a reliability study is only an expected value of the true ICC, it is more appropriate to evaluate the level of reliability based on the 95% confident interval of the ICC estimate, not the ICC estimate itself.

5. Acknowledgments

The authors would like to express their gratitude to Universiti Sains Malaysia (USM) for providing the research funding (Short Term Grant No.304/PPSG/6315410, School of Dental Sciences, Health Campus, Kelantan, Malaysia).

6. References

[1]. Daly, L. E. & Bourke, G. J. 2000. Interpretation and use of medical statistics. Oxford: Blackwell Science Ltd. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1954. 10.

[2]. Fisher, R. A. 1954. Oliver and Boyd; Edinburgh. Statistical methods for research workers.

[3]. Houweling, T., Bolton, J. & Newell, D. 2014. Comparison of two methods of collecting healthcare usage data in chiropractic clinics: patient-report versus documentation in patient files. Chiropr Man Ther.22:32.

[4]. Battaglia, P.J., Maeda, Y., Welk, A., Hough, B. & Kettner, N. 2014. Reliability of the Goutallier classification in quantifying muscle fatty degeneration in the lumbar multifidus using magnetic resonance imaging. J Manipulative Physiol Ther;37:190–7.

[5]. Russell, B.S., Muhlenkamp, K.A., Hoiriis, K.T. & Desimone, C. M. 2012. Measurement of lumbar lordosis in static standing posture with and without high-heeled shoes. J Chiropr Med;11:145–53.

[6]. Koo, T.K.& Li, M.Y. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 15, 155–163.

[7]. McGraw, K. O. & Wong, S. P. 1996. Forming inferences about some intraclass correlation coefficients. Psychol Methods; 1:30–46.

[8]. Shrout, P. E. & Fleiss, J. L. 1979. Intraclass correlations: uses in assessing rater reliability. Psychol Bull; 86:420–8.

[9]. Portney, L. G. & Watkins, M. P. 2000. Foundations of clinical research: applications to practice. New Jersey: Prentice Hall.

[10]. Lee, K. M. Lee, J., Chung, C.Y., et al. 2012. Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. Clin Orthop Surg;4:149–55.