Archives of Orofacial Sciences
The Journal of the School of Dental Sciences, USM

Arch Orofac Sci (2017), **12**(1): 1-11.

Review Article

A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review

Mohamad Adam Bujang[a,b]*, Nurakmal Baharum [a]

[a] Biostatistics Unit, National Clinical Research Centre, Ministry of Health Malaysia, 1st Floor, MMA Building, 124 Jalan Pahang, 53000 Kuala Lumpur, Malaysia.
[b] Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara, 40450 Shah Alam, Selangor, Malaysia.

* Corresponding author: adam@crc.gov.my

Submitted: 06/02/2017. Accepted: 25/04/2017. Published online: 25/04/2017.

Abstract Intraclass correlation coefficient (ICC) measures the extent of agreement and consistency among raters for two or more numerical or quantitative variables. This review paper aimed to present several tables that could illustrate the minimum sample sizes required for estimating the desired effect size of ICC, which is a measurement of the magnitude of an agreement. Determination of the minimum sample size under such circumstances is based on the two fundamentally important parameters, namely the actual value of the ICC and the number of observations made by each subject. The sample size calculations are derived from Power Analysis and Sample Size (PASS) software where the alpha and minimum required power is fixed at 0.05 and higher than 0.80 respectively. A discussion on how to use these tables for determining sample sizes required for each of the various scenarios and the limitations associated with their use in each of these scenarios is provided.

Keywords: coefficient, correlation, intraclass, sample size.

Introduction

Intraclass correlation coefficient (ICC) is a statistical estimate that measures the extent of agreement between at least two quantitative measurements. While kappa statistic measures the extent of agreement for categorical variables, ICC measures the extent of agreement for numerical or quantitative variables. Apart from measuring the extent of agreement, ICC is also designed to measure the degree of reliability, consistency and stability. The concept, theory and the application of ICC have been well described previously (Bartko, 1966; Bartko, 1976; Shrout and Fleiss, 1979; Hunt, 1986; Taylor, 2010).

Sample size estimation is an important initial step when researchers are planning the design and conduct of their study. However, it can be difficult for researchers to estimate empirically the minimum sample size requirement if they are not statisticians. To the best of our knowledge, there is a lack of research conducted on how to estimate a minimum sample size required in determining the value of ICC. Although a sample size formula is available for this purpose, researchers who are not mathematicians and/or statisticians would prefer to use a table to determine the minimum sample sizes required for their studies. The purpose of the present review paper is to provide a simple guide in the form of a table to estimate a minimum sample size required to obtain the desired value of intraclass correlation coefficient, which is also the effect size of ICC.

Several tables are presented as a guide to assist researchers in determining the minimum sample size required for estimating the desired effect size of ICC. This review paper will cover both the methodology on which the sample size determination for obtaining a desired effect size of ICC is based, and discussion on how to use the tables for sample size determination in various circumstances.

Sample size calculation using PASS software

In this review paper, the calculation of the minimum sample size to estimate the value of ICC was performed by using Power Analysis and Sample Size (PASS) software (version 11.0.7; PASS, NCSS, LLC). The formula for minimum sample size (*n*) estimation using the PASS software is derived from other previous studies (Walter *et al.,* 1998; Winer *et al.,* 1991).

$$n = 1 + \frac{2\,(Z_{\alpha} + Z_{\beta})^2\,k}{(\ell n\ C_0)^2\,(k-1)}$$

where,

$$C_0 = \frac{1 + k\theta_0}{1 + k\theta_1}$$

$$\theta_0 = \frac{R_0}{1 - R_0}\ ;\quad \theta_1 = \frac{R_1}{1 - R_1}$$

Power is pre-specified to be at least 0.80 and 0.90. The value of alpha is pre-specified to be 0.05 (which represents the probability of a type I error). As mentioned earlier, the concept of ICC arises from a need to quantify the extent of agreement among raters when the ratings are in the form of at least two quantitative measurements. These measurements can be made by a person (either rater or observant) or by an instrument. Thus, calculations were made to obtain the minimum sample size required for determining the value of ICC when the ratings are made by raters or instruments. In this paper, the number of raters is denoted as (*k*), and it can range between 2 to 10. However, the number of raters can be as high as 20, 30, 40, 50, 60, 70, 80, 90 and 100, etc. especially for a larger scale study. Two other parameters that also require to be taken into account when determining the minimum sample size for ICC are the values of R_0 and R_1. R_0 is the value of ICC that is pre-specified in the null hypothesis if it is true, while the value of R_1 is the value of ICC that is pre-specified in the alternative hypothesis. Sometimes the values of R_0 and R_1 are also denoted as acceptable and expected reliability, respectively.

The values of R_0 and R_1 are pre-specified in the two opposite conditions such as:

(i) When the agreement in the null hypothesis (R_0) is pre-specified to be equal to 0.0 while R_1=0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.
(This is meant to test whether or not there is a statistically significant extent of agreement when it is initially assumed there is no agreement exists between the ratings).

(ii) When the agreement in the null hypothesis (R_0) is pre-specified to be not equal to 0.0 such as; R_0=0.3 vs. R_1=0.5, R_0=0.4 vs. R_1=0.6, R_0=0.5 vs. R_1=0.7, R_0=0.7 vs. R_1=0.9, R_0=0.9 vs. R_1=0.95 and R_0=0.9 vs. R_1=0.97.
(This is meant to test whether or not there is a statistically significant extent of agreement when it is initially assumed there is already a certain extent of agreement exists between the ratings).

The two different settings above are meant to illustrate the two opposite scenarios for sample size planning that are necessary for conducting both reliability and agreement studies. The sample size calculations based on these two different settings are presented as a guide for researchers to determine the desired sample size required for conducting both reliability and agreement studies. To illustrate how the above formula could be used, let’s take for an example there are three raters (k=3) who measure the reliability of their measurements by pre-specifying an acceptable reliability and an expected reliability of 0.0 and 0.2 respectively (and where power is set to be at least 80% while the value of alpha is set to be 0.05). Thus, the minimum sample size required for this case is calculated as follows:

$$k = 3$$
$$R_0 = 0.0$$
$$R_1 = 0.2$$
$$\alpha = 0.05$$

Power sets at 80%, thus $\beta = 1 - 0.8 = 0.2$

$$\theta_0 = \frac{0.0}{1-0.0} = 0.0\,;\ \theta_1 = \frac{0.2}{1-0.2} = 0.25$$

$$C_0 = \frac{1+3(0.0)}{1+3(0.25)} = 0.5714$$

$$n = 1 + \frac{2(Z_{0.05} = -1.65 + Z_{0.2} = -0.84)^2 3}{(\ell n\ 0.5714)^2 (3-1)} = 60.383$$

Hence, the minimum sample size required for the above is calculated to be approximately 60 or 61 patients.

Interpretation of the results and review of their significance

Determination of the value of intraclass correlation coefficient (ICC) does not usually require a large sample, especially if the aim is to determine a high level of agreement with a large value of ICC, when it is initially assumed to be no agreement exists between the ratings (i.e. when the agreement in the null hypothesis (R_0) is pre-specified to be equal to 0.0) (Table 1a and Table 1b). For example, with a pre-specified value of alpha with 0.05 and a pre-specified power of at least 0.8, a minimum sample size of 152 is required to detect the smallest possible value of 0.2 for ICC when it is initially assumed there is no agreement exists between the ratings [i.e. when the agreement in the null hypothesis (R_0) is pre-specified to be equal to 0.0] and there are at least two observations made by each subject. On the other hand, in order to detect the smallest possible value of 0.7 for ICC, a minimum sample size of only 10 is required, as shown in Table 1a.

As the total number of observations made by each subject increases, the minimum sample size required will decrease. The minimum sample size required will not differ too greatly if the total number of observations made by each subject is large (especially 20 or more), no matter what the desired effect size for the ICC can be. For example, the minimum sample size required could range from two to five when the total number of observations made by each subject is at least 20 (Table 1b).

When power is set to be at least 80.0% and *p*-value is set to be equal to 0.05, the number of subjects required would be affected by the total number of observations made by each subject (as mentioned earlier) and also by the actual values of effect size for ICC (i.e. R_0 and R_1) (Tables 2a, 2b and 2c). For example, to detect a sizeable strong level of agreement of 0.7 when there is only a low existing level of agreement within the null hypothesis (that is assumed to be 0.5), a minimum sample of 63 is required (Table 2b). On the other hand, a minimum sample of only 50 is required if the aim is to detect a very strong agreement of 0.95 when there is already a high existing level of agreement within the null hypothesis that is assumed at 0.9 (Table 2c). This illustrates that in order to detect a higher level of agreement, a smaller sample size will be required if there is already a high existing level of agreement.

Discussion

Sample size of ICC for test-retest reliability

Test-retest reliability studies usually measure the level of consistency between two numerical or quantitative ratings at two different times. Some studies have used Pearson’s correlation coefficients to measure the level of test-retest reliability (Feldman *et al*., 1982; Lemasney *et al*., 1984; Mann *et al*., 1985). However, the use of Pearson’s correlation coefficients to assess the level of consistency can be misleading because Pearson's product-moment correlation coefficients are only measuring the correlation between two different ratings and do not take into account the presence of any systematic biases in both ratings (Bartko, 1976). Therefore, a more accurate method to measure the level of statistical consistency is ICC when two ratings are made from numerical or quantitative variables.

Test-retest reliability is usually applied to determine the level of consistency for the purpose of validating a questionnaire design, especially during the initial pilot test. In a validation study of Children

Depression Inventory, ICC was used to measure the test-retest reliability of the total score for evaluating depression in children (Tan *et al*., 2013). This is in order to determine to what extent the total score for evaluating depression in children are found to be consistent, despite obtaining the total scores at two different times. Researchers usually aim to achieve a high level of consistency between the two total scores, in order to ensure that the questionnaire has a high degree of reliability. Since this test-retest reliability will only involve two observations, therefore the minimum number of sample required will be 22, 15 and 10 for detecting the values of ICC of 0.5, 0.6 and 0.7 respectively (Table 1a).

If a researcher plans to determine the level of agreement for a particular score in a questionnaire between two responses in time 1 and time 2; the proposed statement for deriving its sample size would be as follows: "The objective of this study is to determine the level of agreement for the score that assesses the level of satisfaction of the same respondents at two different periods (time 1 and time 2) by determining its test-retest reliability." Sample size calculation will be derived from formula of ICC test using the PASS software. When alpha and power are fixed at 0.05 and lower than 80% respectively, a minimum sample size of 22 is sufficient to detect the value of 0.50 for the ICC (Table 1a). An additional twenty percent of drop-out rate is usually included to make up for those respondent(s) who would fail to attend the follow-up session (i.e. re-test). Hence the number of sample size required would be inflated to 28 (i.e. 22/0.8 = 27.5).

A small sample size is usually required for estimation of ICC and this is preferable because test-retest reliability usually is conducted during an initial pilot study involving only a small sample (Tan *et al*., 2013). In addition, it can be costly to perform a reliability estimation study by assessing the test-retest reliability as it may necessitate rewarding each respondent an incentive to encourage them for participating again during the follow-up.

Sample size requirement for estimating ICCs when the value of ICC in the null hypothesis can be assumed equal to zero

This scenario usually occurs when researchers aim to demonstrate that scores obtained from certain observations or performances have found to be consistent when it is reasonable to initially assume no consistency in the first place. In other words, the researchers aim to demonstrate that a certain level of agreement exists between two consecutive scores because the level of agreement between them is found not to be zero.

For example, a researcher aims to determine the consistencies of the Glasgow Coma Scale (GCS) scores given by the medical officers who assess patients with traumatic injury. The scores given by them could range between three and 15, where a higher score would indicate a more severe form of traumatic injury. The initial assumption is that there is no consistency or agreement found between the scores given by the medical officers; which mean that the level of agreement between them is set to be 0. However, the researchers aim to determine whether the level of consistency or agreement between the scores could be as high as 0.5 or even higher than 0.5, hence the level of agreement between them is set to be 0.5. In a hypothetical case, there are a total of five junior medical officers in a department who would be assessing traumatic injury and each of them would give his or her Glasgow Coma Scale (GCS) score. Therefore, the statement for sample size determination would be as follows: "The aim of this study is to determine the level of an inter-rater agreement of Glasgow Coma Scale (GCS) score for patients with traumatic injury rated by five junior medical officers". When each medical officer is allowed five chances of rating (for assessing patients with traumatic injury), a minimum sample size of 6 patients with traumatic injury would be required to be assessed by each medical officer to achieve the statistical significance for an alpha-value set at 0.05 and with the minimum power of at least 80.0% (Table 1a).

Sample size requirement for estimating ICCs when the value of ICC in the null hypothesis can be assumed not equal to zero

Researchers usually pre-specify that the $R_0 \neq 0$ within the null hypothesis when they aim to establish the fact that a certain level of agreement already exists by inter-rater or intra-rater assessment. This means that the researchers have assumed that the ratings are already known to be consistent in the first place and therefore, the R_0 is pre-specified to be more than zero. In this scenario, researchers will usually aim for detecting a higher level of agreement between the ratings and thus R_1 is always been set to be higher than R_0. For example, a researcher would like to determine the level of agreement found in the scores given by the medical officers who assess an X-ray image of a head injury. The range of scores is between 0 and 10 where a higher score indicates a more severe form of a head injury. The initial assumption is that the scores given by the medical officers are found to be consistent; hence level of agreement between them is set to be 0.5. However, the researcher claims that this level of agreement could possibly be as high as 0.7. In a hypothetical case, there are a total of nine medical officers in a department who would be assessing head injury and each of them would give his or her score for the severity of the head injury. The statement for sample size determination would be as follows: “The aim of this study is to determine the level of an inter-rater agreement of a score that assesses the severity of head injury by nine medical officers based on an X-ray image”. When each medical officer is allowed nine chances of rating (for assessing patients with head injury), a minimum sample size of 23 patients with head injury would be required to be assessed by each medical officer to achieve the statistical significance for an alpha-value set at 0.05 and with the minimum power of at least 80.0% (Table 2b).

In this particular situation when the value of ICC in the null hypothesis can be assumed not equal to zero, for the sake of brevity, only a few possible values for both the R_0 and the R_1 were tabulated (Tables 2a, 2b and 2c). This is because there are so many possible different values for both the R_0 and the R_1 and therefore re-calculation will be necessary if the researcher aims to determine the estimated sample size required for detecting the various effect sizes of the ICC apart from those already presented in the tables.

Sample size requirement for estimating ICCs which are assessed from ratings obtained from two different rating methods or instruments

Previous studies had already demonstrated both the utility and applicability of using the ICC to compare the consistency of ratings obtained from two different rating methods or instruments (Bland and Altman, 1986; Bland and Altman, 1990). Consider a scenario where a new weighing machine “A” had been developed and a researcher is interested to find out to what extent the measurements obtained from machine “A” would agree with those obtained from the existing weighing machine “B” which is currently regarded as the gold standard. In this situation, it is recommended that the researchers to pre-specify a high value (for ICC) of R_0 of at least 0.90 in the null hypothesis and then aim for an even higher value (for ICC) of R_1 of at least 0.95 or 0.97 in the alternative hypothesis. The minimum sample size required for this purpose would then range between 18 and 50. Therefore, the statement for sample size determination would be as follows: “The aim of this study is to determine a high level of agreement between readings obtained from weighing machine A and weighing machine B”, which means that two observations would be made for each subject). It is recommended to pre-specify a high value (for ICC) of R_0 at 0.90 in the null hypothesis and a higher value (for ICC) of R_1 at 0.97 in the alternative hypothesis. This to ensure that the study has indicated that a minimum level of agreement as shown by ICC = 0.90 is expected in the first place, but the aim is to establish that the targeted level of agreement should in fact be much higher, as shown by the value of ICC which exceeds 0.97. Therefore, based on only two observations made on each subject, a sample size of at least 18 is required to achieve statistical significance for an alpha-value set to be 0.05 and with a power of more than 80.0% (Table 2c).

Nevertheless, ICC is not recommended to be the only statistical measure for use as an indicator of agreement between two ratings. “Technical Error of Measurement” (TEM), has been adopted by the International Society Standardization Advancement in Kinanthropometry (ISAK) for the accreditation of anthropometrics practice in Australia (Perini *et al*., 2005). Thus, for this example, researchers will need to calculate TEM to further estimate the level of statistical precision in the data analysis. TEM was also used in many other fields of studies (Duthie *et al*., 2002; Sheppard *et al*., 2006; Jamaiyah *et al*., 2010).

TEM is considered as one of the more reliable indicators for measuring the level of agreement than ICC because a higher value of ICC does not necessarily mean there is less variability among the ratings (Lee *et al*., 2012). TEM also provides an indication of the presence of variability among the ratings, which is not provided by ICC. A commonly acceptable range for the value of relative TEM is less than 2.0% (Perini *et al*., 2005), which means that the level of variability among the ratings is still within acceptable limits. Therefore, it is strongly advisable for researchers to measure the relative TEM after the required sample size has been obtained, and the relevant statistics have been calculated for the reliability estimation study.

Other considerations

Although the present review paper offers a simplified guide to estimate the minimum sample size required for determination of the value of ICC, it is often recommended for researchers to obtain much bigger data than the minimum sample size had suggested. For example, if a minimum sample size requirement is 10, therefore researchers would be recommended to collect an additional of 20% to 30% to make up for any possible loss of data due to drop-outs or missing data.

Usually, in the conduct of a pilot study, only a small sample size is required; therefore it is likely for a high level of variability to be found in the responses. In a test-retest reliability study, a researcher who wants to achieve the ICC value of at least 0.7 would obtain the minimum sample size of 10 subjects (Table 1a). However, due to the presence of high level of variability in the way the subjects would response to the questions; the researcher might have to obtain a larger sample, of at least 15 to 20 subjects, in order to offset the high level of variability found in the responses. The specific advantage of recruiting a much larger sample for a reliability estimation study is to enable the researchers to detect with statistical significance a much smaller value of ICC, such as 0.6. However, if it is possible to minimize the level of variability in the ratings obtained by ensuring them to be generated by an instrument or machine, then researchers can then depend on the simplified guide to obtain an estimate of the required minimum sample size.

If a researcher would like to conduct a reliability estimation study which aims to estimate a value of ICC that has not been provided by the guide (Tables 1a, 1b, 2a, 2b and 2c); it is always valid to recommend that the researcher to first identify the value of ICC from the guide which is closest to what the researcher has aimed for. However, a larger sample size than what this guide specifies would be required in this instance. For example, if a researcher would like to estimate a value of ICC to be 0.75, and the minimum sample size for estimating this particular value of ICC has not been provided by this guide; therefore it is recommended for researchers to determine the minimum sample size required for estimating a value of ICC to be 0.7, since it will invariably yield a larger sample. By obtaining a larger sample size than necessary, this would ensure it will have sufficient power to estimate this particular value of ICC for a particular pre-specified alpha-value.

Determining the minimum sample size required for estimating the value of ICC is usually based on the research objectives as shown by the various examples described previously. However, the use of the ICC could be confused with correlation tests in measuring the strength of association. This is because correlation test aims to address different research objectives and therefore, a different formula is required to estimate the minimum sample size (Bujang and Baharum, 2016). In general, the minimum sample size required for estimating the desired value of ICC is small, especially when a researcher aims to estimate a very high value of ICC.

However, some studies do require large sample size so that the sample statistics will have closer approximation to the actual population parameters. This is often true when conducting a survey where there are many research objectives and statistical analyses involved (Bujang *et al*., 2012; Bujang *et al*., 2015).

Conclusion

This review article has demonstrated the sample size guidelines for ICC. These guidelines are useful for a quick sample size planning with regards to research question that require the use of ICC to answer the particular research question. For studies that aim to measure a very high agreement, TEM has also to be incorporated in the result besides ICC.

Acknowledgment

The authors would like to acknowledge the Director General of the Ministry of Health for his support and permission to publish this study. A special thanks to Mr John Hon Yoon Khee for his proofread of this paper.

Table 1a Sample size requirement for intraclass correlation with power = 80% and 90%; alpha = 0.05, observation per subject from 2 to 10 and R_0 is set at 0

Observation per Subject	ICC	Number of subjects (power=80%)	Number of subjects (power=90%)
2	0.2	152	210
	0.3	66	91
	0.4	36	50
	0.5	22	30
	0.6	15	20
	0.7	10	13
	0.8	7	9
	0.9	5	6
3	0.2	60	83
	0.3	28	39
	0.4	17	23
	0.5	11	15
	0.6	8	10
	0.7	6	8
	0.8	4	6
	0.9	3	4
4	0.2	35	49
	0.3	18	24
	0.4	11	15
	0.5	8	10
	0.6	6	8
	0.7	5	6
	0.8	4	5
	0.9	3	4
5	0.2	24	34
	0.3	13	18
	0.4	8	12
	0.5	6	8
	0.6	5	6
	0.7	4	5
	0.8	3	4
	0.9	3	3
6	0.2	18	26
	0.3	10	14
	0.4	7	10
	0.5	5	7
6	0.6	4	6
	0.7	4	5
	0.8	3	4
	0.9	3	3
7	0.2	15	21
	0.3	9	12
	0.4	6	8
	0.5	5	6
	0.6	4	5
	0.7	3	4
	0.8	3	4
	0.9	3	3
8	0.2	13	18
	0.3	8	11
	0.4	6	8
	0.5	4	6
	0.6	4	5
	0.7	3	4
	0.8	3	3
	0.9	2	3
9	0.2	11	15
	0.3	7	9
	0.4	5	7
	0.5	4	5
	0.6	4	5
	0.7	3	4
	0.8	3	3
	0.9	2	3
10	0.2	10	14
	0.3	6	9
	0.4	5	6
	0.5	4	5
	0.6	3	4
	0.7	3	4
	0.8	3	3
	0.9	2	3

Table 1b Sample size requirement for intraclass correlation with power = 80% and 90%; alpha = 0.05, observation per subject from 20 to 100 (gap of every 10) and R_0 is set at 0

Observation per Subject	ICC	Number of subjects (power=80%)	Number of subjects (power=90%)
20	0.2	5	7
	0.3	4	5
	0.4	3	4
	0.5	3	4
	0.6	3	3
	0.7	3	3
	0.8	2	3
	0.9	2	3
30	0.2	4	6
	0.3	4	4
	0.4	3	4
	0.5	3	3
	0.6	3	3
	0.7	2	3
	0.8	2	3
	0.9	2	2
40	0.2	4	5
	0.3	3	4
	0.4	3	4
	0.5	3	3
	0.6	3	3
	0.7	2	3
	0.8	2	3
	0.9	2	2
50	0.2	4	5
	0.3	3	4
	0.4	3	3
	0.5	3	3
	0.6	2	3
	0.7	2	3
	0.8	2	3
	0.9	2	2
60	0.2	3	4
	0.3	3	4
	0.4	3	3
	0.5	2	3
60	0.6	2	3
	0.7	2	3
	0.8	2	3
	0.9	2	2
70	0.2	3	4
	0.3	3	3
	0.4	3	3
	0.5	2	3
	0.6	2	3
	0.7	2	3
	0.8	2	2
	0.9	2	2
80	0.2	3	4
	0.3	3	3
	0.4	3	3
	0.5	2	3
	0.6	2	3
	0.7	2	3
	0.8	2	2
	0.9	2	2
90	0.2	3	4
	0.3	3	3
	0.4	2	3
	0.5	2	3
	0.6	2	3
	0.7	2	3
	0.8	2	2
	0.9	2	2
100	0.2	3	4
	0.3	3	3
	0.4	2	3
	0.5	2	3
	0.6	2	3
	0.7	2	3
	0.8	2	2
	0.9	2	2

Table 2a Sample size requirement for intraclass correlation for $R_0 \neq 0$ vs R_1 (R_0=0.3 vs R_1=0.5 and R_0=0.4 vs R_1=0.6) and alpha=0.05

R_0=0.3 vs R_1=0.5			R_0=0.4 vs R_1=0.6		
Observation per subject	Number of subjects (power=80%)	Number of subjects (power=90%)	Observation per subject	Number of subjects (power=80%)	Number of subjects (power=90%)
2	109	151	2	87	120
3	60	83	3	51	71
4	45	62	4	40	56
5	37	52	5	35	48
6	33	47	6	31	44
7	30	43	7	29	41
8	29	40	8	28	39
9	27	38	9	27	38
10	26	37	10	26	36
20	22	31	20	22	32
30	20	29	30	21	30
40	20	28	40	21	29
50	19	27	50	21	29
60	19	27	60	20	29
70	19	27	70	20	29
80	19	27	80	20	28
90	19	26	90	20	28
100	19	26	100	20	28

Table 2b Sample size requirement for intraclass correlation for $R_0 \neq 0$ vs R_1 (R_0=0.5 vs R_1=0.7 and R_0=0.7 vs R_1=0.9) and alpha=0.05

R_0=0.5 vs R_1=0.7			R_0=0.7 vs R_1=0.9		
Observation per subject	Number of subjects (power=80%)	Number of subjects (power=90%)	Observation per subject	Number of subjects (power=80%)	Number of subjects (power=90%)
2	63	87	2	19	25
3	39	55	3	13	18
4	32	45	4	11	16
5	28	40	5	10	14
6	26	37	6	10	14
7	25	35	7	10	13
8	24	33	8	9	13
9	23	32	9	9	13
10	22	32	10	9	13
20	20	28	20	9	12
30	19	27	30	8	12
40	19	27	40	8	11
50	19	26	50	8	11
60	19	26	60	8	11
70	18	26	70	8	11
80	18	26	80	8	11
90	18	26	90	8	11
100	18	26	100	8	11

Table 2c Sample size requirement for intraclass correlation for $R_0 \neq 0$ vs R_1 (R_0=0.9 vs R_1=0.95 and R_0=0.9 vs R_1=0.97) and alpha=0.05

R_0=0.9 vs R_1=0.95			R_0=0.9 vs R_1=0.97		
Observation per subject	Number of subjects (power=80%)	Number of subjects (power=90%)	Observation per subject	Number of subjects (power=80%)	Number of subjects (power=90%)
2	50	68	2	18	24
3	36	50	3	13	18
4	31	44	4	12	16
5	29	41	5	11	15
6	28	39	6	10	14
7	27	38	7	10	14
8	26	37	8	10	14
9	26	36	9	10	14
10	25	36	10	10	13
20	24	34	20	9	13
30	23	33	30	9	13
40	23	33	40	9	12
50	23	33	50	9	12
60	23	33	60	9	12
70	23	33	70	9	12
80	23	32	80	9	12
90	23	32	90	9	12
100	23	32	100	9	12

References

Bartko JJ (1966). The intraclass correlation coefficient as a measure of reliability. *Psychol Rep*, **19**(1): 3-11.

Bartko JJ (1976). On various intraclass correlation reliability coefficients. *Psychol Bull*, **83**(5): 762-765.

Bland JM, Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **1**(8476): 307-310.

Bland JM, Altman DG (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med*, **20**(5): 337-340.

Bujang MA, Baharum N (2016). Sample size guideline for correlation analysis. *World J Soc Sci Res*, **3**(1): 37-46.

Bujang MA, Ghani PA, Zolkepali NA, Adnan TH, Ali MM, Selvarajah S, Haniff J (2012). A comparison between convenience sampling versus systematic sampling in getting the true parameter in a population: explore from a clinical database: The Audit Diabetes Control Management (ADCM) registry in 2009. In: *ICSSBE 2012 - Proceedings, 2012 International Conference on Statistics in Science, Business and Engineering: "Empowering Decision Making with Statistical Sciences"*, 6396615, pp. 499-503, 2012 International Conference on Statistics in Science, Business and Engineering, ICSSBE 2012, Langkawi, Kedah, Malaysia, 10-12 September 2012.

Bujang MA, Sa'at N, Joys AR, Ali MM (2015). An audit of the statistics and the comparison with the parameter in the population. *AIP Conf Proc*, **1682**: 050019.

Duthie GM, Young WB, Aitken DA (2002). The acute effects of heavy loads on jump squat performance: an evaluation of the complex and contrast methods of power development. *J Strength Cond Res*. **16**(4): 530-538.

Feldman RS, Douglass CW, Loftus ER, Kapur KK, Chauncey HH (1982). Interexaminer agreement in the measurement of periodontal disease. *J Periodont Res*, **17**(1): 80-89.

Hunt RJ (1986). Percent agreement, Pearson's correlation and kappa as measures of inter-examiner reliability. *J Dent Res*, **65**(2): 128-130.

Jamaiyah H, Geeta A, Safiza MN, Khor GL, Wong NF, Kee CC *et al*. (2010). Reliability, technical error of measurements and validity of length and weight measurements for children under two years old in Malaysia. *Med J Malaysia*, **65**(Suppl A): 131-137.

Lee KM, Lee J, Chung CY, Ahn S, Sung KH, Kim TW *et al*. (2012). Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. *Clin Orthop Surg*, **4**(2): 149-155.

Lemasney J, O'Mullane D, Coleman M (1984). Effect of fluoridation on dental health in 5- and 11-year-old Irish schoolchildren, *Community Dent Oral Epidemiol*, **12**(4): 218-222.

Mann J, Goultschin J, Call RL (1985). Assessment of inter-examiner agreement in scoring periodontal disease. *J Periodont Res*, **20**(1): 86-90.

Perini TA, de Oliveira GL, Ornellas JS, de Oliveira FP (2005). Technical error of measurement in anthropometry. *Rev Bras Med Esporte*. **11**(1): 86-90.

Sheppard JM, Young WB, Doyle TA, Sheppard TA, Newton RU (2006). An evaluation of a new test of reactive agility, and its relationship to sprint speed and change of direction speed. *J Sci Med Sport*, **9**(4): 342-349.

Shrout PE, Fleiss JL (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, **86**(2): 420-428.

Tan SMK, Loh SF, Bujang MA, Haniff J, Abd Rahman FN, Ismail F *et al.* (2013). Validation of the Malay Version of children's depression inventory. *Int Med J*, **20**(2): 188-191.

Taylor PJ (2010). *An introduction to intraclass correlation that resolves some common confusion* [unpublished manuscript]. Boston: University of Massachusetts. Available at: http://www.faculty.umb.edu/pjt/09b.pdf (Date of access: 6th March 2016).

Walter SD, Eliasziw M, Donner A (1998). Sample size and optimal designs for reliability studies. *Stat Med*, **17**(1): 101-110.

Winer BJ, Brown DR, Michels KM (1991). *Statistical Principles in Experimental Design*, 3rd edn. New York: McGraw-Hill.