ISSN 2234-7658 (print) / ISSN 2234-7666 (online)
http://dx.doi.org/10.5395/rde.2013.38.2.98

Statistical notes for clinical researchers: Evaluation of measurement error 1: using intraclass correlation coefficients

Hae-Young Kim

Department of Dental Laboratory Science & Engineering, Korea University College of Health Science, Seoul, Korea

Evaluation of measurement error is a fundamental procedure in clinical studies, field surveys, or experimental researches to confirm the reliability of measurements. If we examine oral health status of a patient, the number of caries teeth or degree of periodontal pocket depth need to be similar when an examiner repeated the measuring procedure (intra-examiner reliability) or when two independent examiners repeated the measuring procedure, to guarantee the reliability of measurement. In experimental researches, confirming small measurement error of measuring machines should be a prerequisite to start the main measuring procedure for the study.

1. Measurement error in our daily lives

We meet many situations which might be subject to measurement error in our daily lives. For example, when we measure body weight using a scale displaying kilograms (kg) to one decimal point, we disregard body weight differences less than 0.1 kg. Similarly, when we check time using a hand-watch with two hands indicating hours and minutes, we implicitly recognize there may be errors ranging up to a few minutes. However, generally we don’t worry about these possible errors because we know such a small amount of error comprises a relatively small fraction of the quantity measured. In other words, the measurements may still be reliable even under consideration of the small amount of error. The degree of measurement error could be visualized as a ratio of error variability to total variability. Similarly, degree of reliability could be expressed as a ratio of subject variability to total variability.

2. Reliability: Consistency or absolute agreement?

Reliability is defined as the degree to which a measurement technique can secure consistent results upon repeated measuring on the same objects either by multiple raters or test-retest trials by one observer at different time points. It is necessary to differentiate two different kinds of reliability; consistency or absolute agreement. For example, three raters independently evaluate twenty students’ applications for a scholarship on a scale of zero to 100. The first rater is especially harsh and the third one is particularly lenient, but each rater scores consistently. There must be differences among the actual scores which the three raters give. If the purpose is ranking applicants and choosing five students, the difference among raters may not make significantly different results if the ‘consistency’ was maintained during the entire scoring procedure. However if the purpose is to select students who are rated above or below a preset standard absolute score, the scores from the three raters need to be absolutely similar on a mathematical level. Therefore while we want consistency of the evaluation in the former case, we want to achieve ‘absolute agreement’ in the later case. Difference of purpose is reflected in the procedure used for reliability calculation.

*Correspondence to
Hae-Young Kim, DDS, PhD.
Associate Professor,
Department of Dental Laboratory Science & Engineering, Korea University College of Health Science, San 1 Jeongneung 3-dong, Seongbuk-gu, Seoul, Korea 136-703
TEL, +82-2-940-2845; FAX, +82-2-909-3502, E-mail, kimhaey@korea.ac.kr

3. Intraclass correlation coefficient (ICC)

Though there are some important reliability measures, such as Dahlberg's error or Kappa statistics, ICC seems to be the most useful. The ICC is a reliability measure we may use to assess either degree of consistency or absolute agreement. ICC is defined as the ratio of variability between subjects to the total variability including subject variability and error variability.

If we evaluate consistency of an outcome measure which was repeatedly measured, the repetition is regarded as a fixed factor which doesn't involve any errors and the following equation may be applied:

ICC (consistency) = subject variability / (subject variability + measurement error)

If we evaluate absolute agreement of an outcome measure which was repeatedly measured, the repetition variability needs to be counted because the factor is regarded as a random factor as in the following equation:

ICC (absolute agreement) = subject variability / (subject variability + variability in repetition + measurement error)

Reliability based on absolute agreement is always lower than for consistency because a more stringent criterion is applied.

4. ICC for a single observer and multiple observers

If multiple observers assessed subjects, the average of repetition variability and error variability are applied in calculating ICC. Use of average variability results in higher reliability compared to use of any single rater, because the measurement error is averaged out. When k observers were involved, the ICC equations need to be changed as following:

ICC (consistency, k raters) = subject variability / (subject variability + measurement error/k)

ICC (absolute agreement, k raters) = subject variability / [subject variability + (variability in repetition + measurement error) / k]

5. An Example: evaluation of measurement errors

1) Repeated scores of oral health-related quality of life (OHRQoL)

Table 1 displays repeatedly measured scores of the oral health impact profile for children (COHIP), one of the measures for OHRQoL which was obtained among ten 5th-grade school children. The COHIP inventory is a measure ranging from 0 (lowest OHRQoL) to 112 (highest OHRQoL), which assesses level of subjective oral health status by asking questions mainly about oral impacts on daily lives for children. Let's assume that the repeated measurements were obtained by a rater with an appropriate interval to assess test-retest reliability.

Table 1. Repeatedly measured scores of the oral health impact profile for children (COHIP)

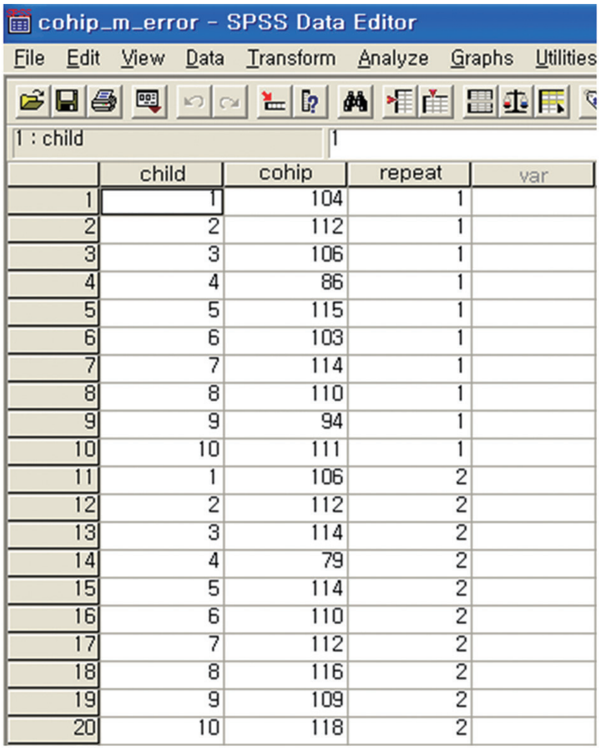
Children	COHIP score 1	COHIP score 2	Mean
1	104	106	105
2	112	112	112
3	106	114	110
4	86	79	82.5
5	115	114	114.5
6	103	110	106.5
7	114	112	113
8	110	116	113
9	94	109	101.5
10	111	118	114.5
Mean	105.5	109	107.25

http://dx.doi.org/10.5395/rde.2013.38.2.98

2) Intraclass correlation coefficient (ICC) measuring consistency and absolute agreement

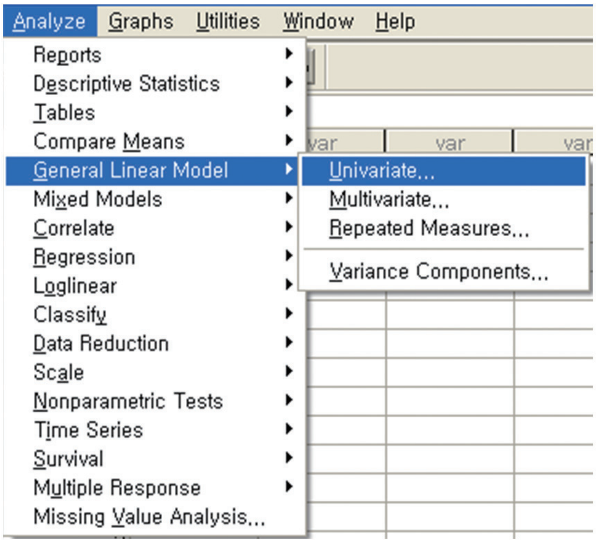
To calculate ICC, we need to obtain subject variability and error variability using a statistical package, such as SPSS, as shown in the following procedures.

(a) Data

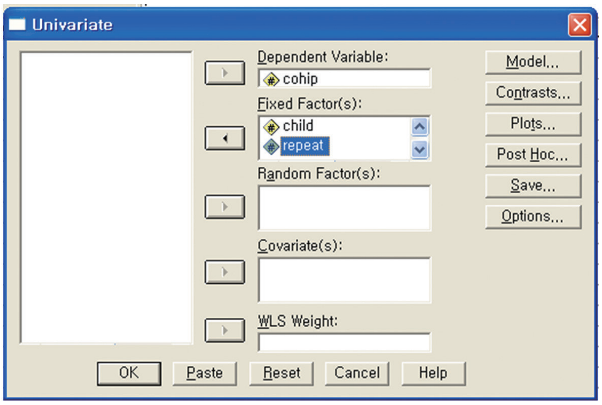


	child	cohip	repeat	var
1	1	104	1	
2	2	112	1	
3	3	106	1	
4	4	86	1	
5	5	115	1	
6	6	103	1	
7	7	114	1	
8	8	110	1	
9	9	94	1	
10	10	111	1	
11	1	106	2	
12	2	112	2	
13	3	114	2	
14	4	79	2	
15	5	114	2	
16	6	110	2	
17	7	112	2	
18	8	116	2	
19	9	109	2	
20	10	118	2	

(b) General Linear Model - Univariate



(c) Dependent vs Fixed factors



Dependent Variable: cohip

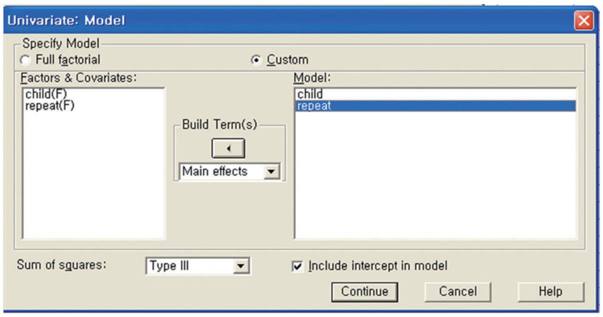
Fixed Factor(s): child, repeat

Random Factor(s):

Covariate(s):

WLS Weight:

(d) Custom – Main effects



Specify Model

○ Full factorial ● Custom

Factors & Covariates: child(F), repeat(F)

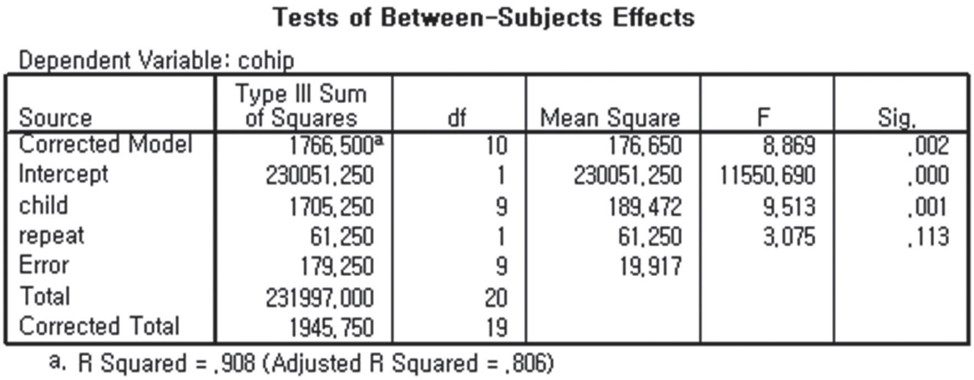
Build Term(s): Main effects

Model: child, repeat

Sum of squares: Type III

☑ Include intercept in model

(e) ANOVA table



Tests of Between-Subjects Effects

Dependent Variable: cohip

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1766.500a	10	176.650	8.869	.002
Intercept	230051.250	1	230051.250	11550.690	.000
child	1705.250	9	189.472	9.513	.001
repeat	61.250	1	61.250	3.075	.113
Error	179.250	9	19.917		
Total	231997.000	20			
Corrected Total	1945.750	19			

a. R Squared = .908 (Adjusted R Squared = .806)

From the ANOVA table, we use Mean Square (MS) to calculate variances of subject (σ^2_{child}), repetition (σ^2_{repet}), and error (σ^2_{error}) as following:

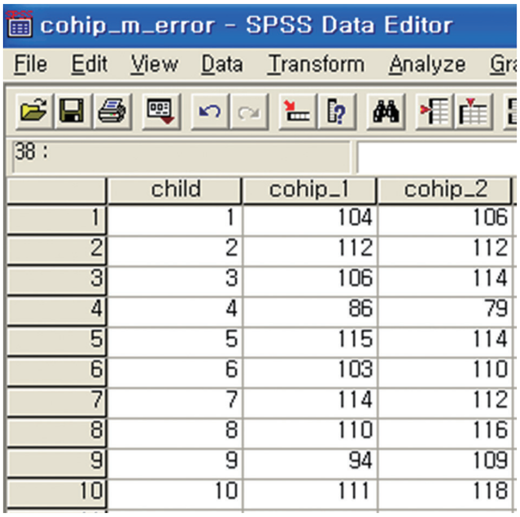
MS (child) = 2 (number of repetition) $* \sigma^2_{child} + \sigma^2_{error} = 189.47$
MS (repetition) = 10 (number of children) $* \sigma^2_{repet} + \sigma^2_{error} = 61.25$
MS (error) = $\sigma^2_{error} = 19.92$

From the equations above we obtain the variance among children (σ^2_{child}) as 84.78, and variance of repetition (σ^2_{repet}) as 4.13. Then the ICC measuring consistency may be calculated as the proportion of subject (children) variability among total variability excluding variability of repetition which is regarded as a fixed factor.

ICC (consistency, single rater) = $\sigma^2_{child} / (\sigma^2_{child} + \sigma^2_{error})= 84.78 / (84.78 + 19.92) = 0.810$
ICC (consistency, two raters) = $\sigma^2_{child} / (\sigma^2_{child} + \sigma^2_{error}/2)= 84.78 / (84.78 + 19.92 / 2) = 0.895$
ICC (absolute agreement, single rater) = $\sigma^2_{children} / (\sigma^2_{child} + \sigma^2_{repet} + \sigma^2_{error}) = 84.78 / (84.78 + 4.13 + 19.92) = 0.779$
ICC (absolute agreement, two raters) = $\sigma^2_{children} / [\sigma^2_{child} + (\sigma^2_{repet} + \sigma^2_{error}) / 2]= 84.78 / [84.78 + (4.13 + 19.9) / 2] = 0.876$

The same ICC for consistency may be obtained using SPSS, following procedure:

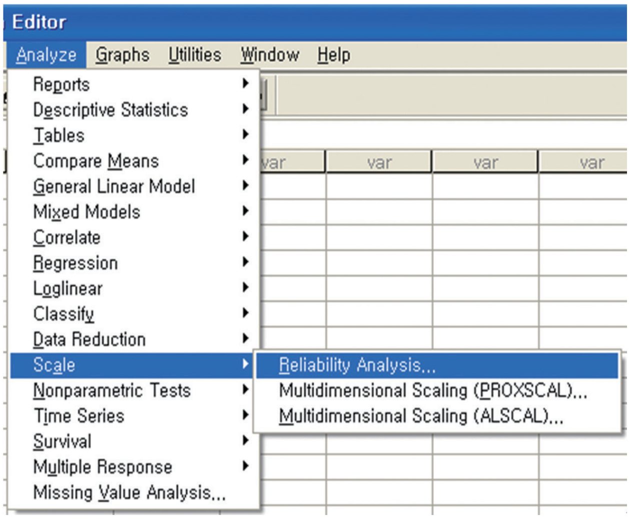
(f) Data arrangement



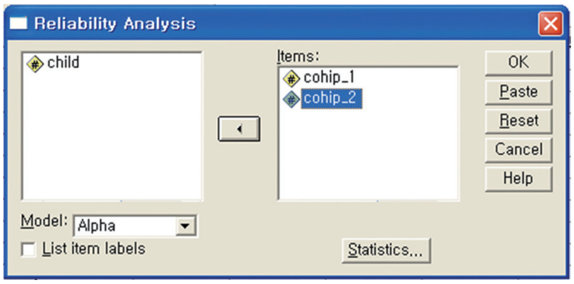
cohip_m_error - SPSS Data Editor
File Edit View Data Transform Analyze Gr
38 :

	child	cohip_1	cohip_2
1	1	104	106
2	2	112	112
3	3	106	114
4	4	86	79
5	5	115	114
6	6	103	110
7	7	114	112
8	8	110	116
9	9	94	109
10	10	111	118

(g) Scale – Reliability Analysis



(h) Items



Reliability Analysis

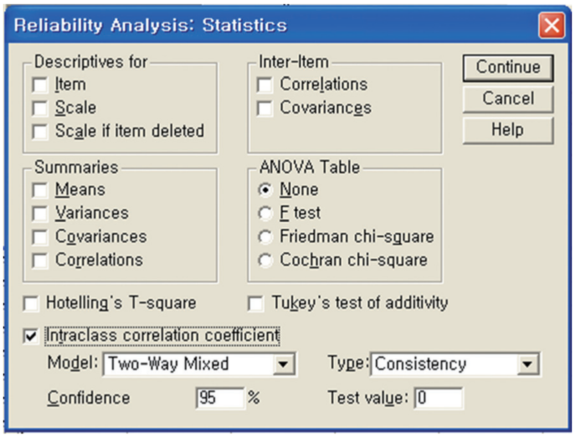
child

Items:
cohip_1
cohip_2

Model: Alpha

☐ List item labels

(i) ICC - Type (Consistency)



Reliability Analysis: Statistics

Descriptives for
☐ Item
☐ Scale
☐ Scale if item deleted

Inter-Item
☐ Correlations
☐ Covariances

Summaries
☐ Means
☐ Variances
☐ Covariances
☐ Correlations

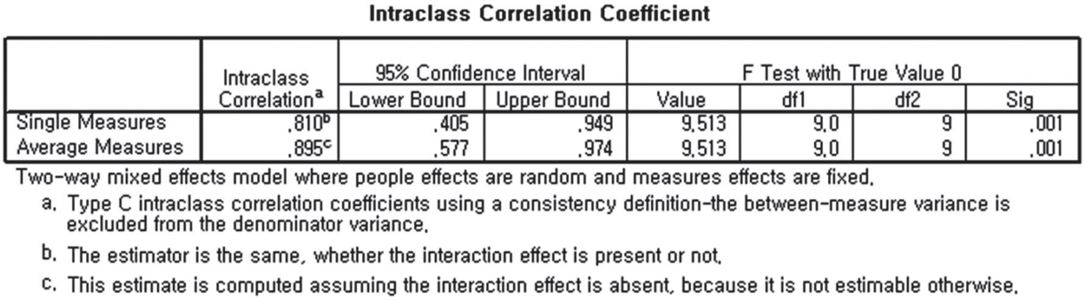
ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

☐ Hotelling's T-square
☐ Tukey's test of additivity

☑ Intraclass correlation coefficient

Model: Two-Way Mixed
Type: Consistency
Confidence 95 %
Test value: 0

(j) ICC: Consistency of single rater (0.810) and that of multiple raters (0.895)



Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.810[b]	.405	.949	9.513	9.0	9	.001
Average Measures	.895[c]	.577	.974	9.513	9.0	9	.001

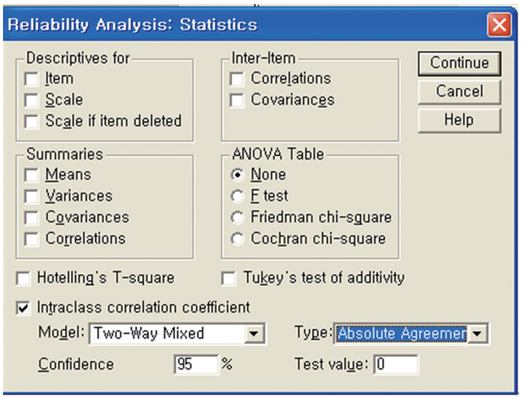
Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

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.

(k) ICC - Type (Absolute agreement)



Reliability Analysis: Statistics

Descriptives for
☐ Item
☐ Scale
☐ Scale if item deleted

Inter-Item
☐ Correlations
☐ Covariances

Summaries
☐ Means
☐ Variances
☐ Covariances
☐ Correlations

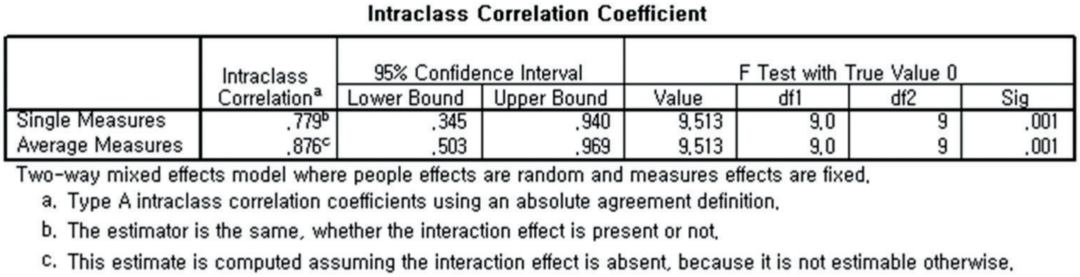
ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

☐ Hotelling's T-square
☐ Tukey's test of additivity

☑ Intraclass correlation coefficient

Model: Two-Way Mixed
Type: Absolute Agreemer
Confidence 95 %
Test value: 0

(l) ICC: Absolute agreement of single rater (0.779) and that of multiple raters (0.876)



Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.779[b]	.345	.940	9.513	9.0	9	.001
Average Measures	.876[c]	.503	.969	9.513	9.0	9	.001

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.