Computing Intraclass Correlations (ICC) as Estimates of Interrater Reliability in SPSS

Richard Landers[1]

[1]Old Dominion University

April 17, 2023

Abstract

Intraclass correlation (ICC) is one of the most commonly misused indicators of interrater reliability, but a simple step-by-step process will get it right. In this article, I provide a brief review of reliability theory and interrater reliability, followed by a set of practical guidelines for the calculation of ICC in SPSS.

the WINNOWER

MATHEMATICS

Computing Intraclass Correlations (ICC) as Estimates of Interrater Reliability in SPSS

RICHARD LANDERS [1]

1. Old Dominion University

ABSTRACT

Intraclass correlation (ICC) is one of the most commonly misused indicators of interrater reliability, but a simple step-by-step process will get it right. In this article, I provide a brief review of reliability theory and interrater reliability, followed by a set of practical guidelines for the calculation of ICC in SPSS.

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:
rnlanders@odu.edu

DATE RECEIVED:
June 16, 2015

DOI:
10.15200/winn.143518.81744

ARCHIVED:
June 24, 2015

KEYWORDS:
intraclass correlation, spss, icc, reliability

CITATION:
Richard Landers, Computing Intraclass Correlations (ICC) as Estimates of Interrater Reliability in SPSS, *The Winnower* 2:e143518.81744 , 2015 , DOI: 10.15200/winn.143518.81744

Recently, a colleague of mine asked for some advice on how to compute interrater reliability for a coding task, and I discovered that there aren't many resources online written in an easy-to-understand format – most either 1) go in depth about formulas and computation or 2) go in depth about SPSS without giving many specific reasons for why you'd make several important decisions. The primary resource available is a paper by Shrout and Fleiss (1979), which is quite dense. So I am taking a stab at providing a comprehensive but easier-to-understand resource.

Reliability, generally, is the proportion of "real" information about a construct of interest captured by your measurement of it. For example, if someone reported the reliability of their measure was .8, you could conclude that 80% of the variability in the scores captured by that measure represented the construct, and 20% represented random variation. The more uniform your measurement, the higher reliability will be.

In the social sciences, we often have research participants complete surveys, in which case you don't need ICCs – you would more typically use coefficient alpha. But when you have research participants provide something about themselves from which you need to extract data, your measurement becomes what you get from that extraction. For example, in one of my lab's current studies, we are collecting copies of Facebook profiles from research participants, after which a team of lab assistants looks them over and makes ratings based upon their content. This process is called coding. Because the research assistants are creating the data, their ratings are my scale – not the original data. Which means they 1) make mistakes and 2) vary in their ability to make those ratings. An estimate of interrater reliability will tell me what proportion of their ratings is "real", i.e. represents an underlying construct (or potentially a combination of constructs – there is no way to know from reliability alone – all you can conclude is that you are measuring *something* consistently).

An intraclass correlation (ICC) can be a useful estimate of inter-rater reliability on quantitative data because it is highly flexible. A Pearson correlation can be a valid estimator of interrater reliability, but only when you have meaningful pairings between two and only two raters. What if you have more? What if your raters differ by ratee? This is where ICC comes in (note that if you have qualitative data, e.g. categorical data or ranks, you would not use ICC).

Unfortunately, this flexibility makes ICC a little more complicated than many estimators of reliability. While you can often just throw items into SPSS to compute a coefficient alpha on a scale measure, there are several additional questions one must ask when computing an ICC, and one restriction. The restriction is straightforward: you must have the same number of ratings for every case rated. The questions are more complicated, and their answers are based upon how you identified your raters, and what you ultimately want to do with your reliability estimate. Here are the first two questions:

1. Do you have consistent raters for all ratees? For example, do the exact same 8 raters make ratings on every ratee?

2. Do you have a sample or population of raters?

If your answer to Question 1 is no, you need ICC(1). In SPSS, this is called "One-Way Random." In coding tasks, this is uncommon, since you can typically control the number of raters fairly carefully. It is most useful with massively large coding tasks. For example, if you had 2000 ratings to make, you might assign your 10 research assistants to make 400 ratings each – each research assistant makes ratings on 2 ratees (you always have 2 ratings per case), but you counterbalance them so that a random two raters make ratings on each subject. It's called "One-Way Random" because 1) it makes no effort to disentangle the effects of the rater and ratee (i.e. one effect) and 2) it assumes these ratings are randomly drawn from a larger populations (i.e. a random effects model). ICC(1) will always be the smallest of the ICCs.

If your answer to Question 1 is yes and your answer to Question 2 is "sample", you need ICC(2). In SPSS, this is called "Two-Way Random." Unlike ICC(1), this ICC assumes that the variance of the raters is only adding noise to the estimate of the ratees, and that mean rater error = 0. Or in other words, while a particular rater might rate Ratee 1 high and Ratee 2 low, it should all even out across many raters. Like ICC(1), it assumes a random effects model for raters, but it explicitly models this effect – you can sort of think of it like "controlling for rater effects" when producing an estimate of reliability. If you have the same raters for each case, this is generally the model to go with. This will always be larger than ICC(1) and is represented in SPSS as "Two-Way Random" because 1) it models both an effect of rater and of ratee (i.e. two effects) and 2) assumes both are drawn randomly from larger populations (i.e. a random effects model).

If your answer to Question 1 is yes and your answer to Question 2 is "population", you need ICC(3). In SPSS, this is called "Two-Way Mixed." This ICC makes the same assumptions as ICC(2), but instead of treating rater effects as random, it treats them as fixed. This means that the raters in your task are the only raters anyone would be interested in. This is uncommon in coding, because theoretically your research assistants are only a few of an unlimited number of people that could make these ratings. This means ICC(3) will also always be larger than ICC(1) and typically larger than ICC(2), and is represented in SPSS as "Two-Way Mixed" because 1) it models both an effect of rater and of ratee (i.e. two effects) and 2) assumes a random effect of ratee but a fixed effect of rater (i.e. a mixed effects model).

After you've determined which kind of ICC you need, there is a second decision to be made: are you interested in the reliability of a single rater, or of their mean? If you're coding for research, you're probably going to use the mean rating. If you're coding to determine how accurate a single person would be if they made the ratings on their own, you're interested in the reliability of a single rater. For example, in our Facebook study, we want to know both. First, we might ask "what is the reliability of our ratings?" Second, we might ask "if one person were to make these judgments from a Facebook profile, how accurate would that person be?" We add ",k" to the ICC rating when looking at means, or ",1" when looking at the reliability of single raters. For example, if you computed an ICC(2) with 8 raters, you'd be computing ICC(2,8). If you computed an ICC(1) with the same 16 raters for every case but were interested in a single rater, you'd still be computing ICC(2,1). For ICC(#,1), a large number of raters will produce a narrower confidence interval around your reliability estimate than a small number of raters, which is why you'd still want a large number of raters, if possible, when

estimating ICC(#,1).

After you've determined which specificity you need, the third decision is to figure out whether you need a measure of absolute agreement or consistency. If you've studied correlation, you're probably already familiar with this concept: if two variables are perfectly consistent, they don't necessarily agree. For example, consider Variable 1 with values 1, 2, 3 and Variable 2 with values 7, 8, 9. Even though these scores are very different, the correlation between them is 1 – so they are highly consistent but don't agree. If using a mean [ICC(#, k)], consistency is typically fine, especially for coding tasks, as mean differences between raters won't affect subsequent analyses on that data. But if you are interested in determining the reliability for a single individual, you probably want to know how well that score will assess the real value.

Once you know what kind of ICC you want, it's pretty easy in SPSS. First, create a dataset with columns representing raters (e.g. if you had 8 raters, you'd have 8 columns) and rows representing cases. You'll need a complete dataset for each variable you are interested in. So if you wanted to assess the reliability for 8 raters on 50 cases across 10 variables being rated, you'd have 10 datasets containing 8 columns and 50 rows (400 cases per dataset, 4000 total points of data).

A special note for those of you using surveys: if you're interested in the inter-rater reliability of a scale mean, compute ICC on that scale mean – not the individual items. For example, if you have a 10-item unidimensional scale, calculate the scale mean for each of your rater/target combinations *first* (i.e. one mean score per rater per ratee), and then use that scale mean as the target of your computation of ICC. Don't worry about the inter-rater reliability of the individual items unless you are doing so as part of a scale development process, i.e. you are assessing scale reliability in a pilot sample in order to cut some items from your final scale, which you will later cross-validate in a second sample.

In each dataset, you then need to open the Analyze menu, select Scale, and click on Reliability Analysis. Move all of your rater variables to the right for analysis. Click Statistics and checkIntraclass correlation coefficient at the bottom. Specify your model (One-Way Random, Two-Way Random, or Two-Way Mixed) and type (Consistency or Absolute Agreement). Click Continue and OK. You should end up with something like Figure 1.

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	.169[b]	-.021	.439	1.814	19	57	.043
Average Measures	.449	-.090	.758	1.814	19	57	.043

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

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.

Figure 1. Results of a Two-Way Random Consistency ICC Calculation in SPSS.

In Figure 1, I computed an ICC(2) with 4 raters across 20 ratees. You can find the ICC(2,1) in the first line – ICC(2,1) = .169. That means ICC(2, k), which in this case is ICC(2, 4) = .449. Therefore, 44.9% of the variance in the mean of these raters is "real".

So here's the summary of this whole process:

1. DECIDE WHICH CATEGORY OF ICC YOU NEED.

· Determine if you have consistent raters across all ratees (e.g. always 3 raters, and always the same 3 raters). If not, use ICC(1), which is "One-way Random" in SPSS.

· Determine if you have a population of raters. If yes, use ICC(3), which is "Two-Way Mixed" in SPSS.

· If you didn't use ICC(1) or ICC(3), you need ICC(2), which assumes a sample of raters, and is "Two-Way Random" in SPSS.

2. DETERMINE WHICH VALUE YOU WILL ULTIMATELY USE.

· If a single individual, you want ICC(#,1), which is "Single Measure" in SPSS.

· If the mean, you want ICC(#,k), which is "Average Measures" in SPSS.

3. DETERMINE WHICH SET OF VALUES YOU ULTIMATELY WANT THE RELIABILITY FOR.

· If you want to use the subsequent values for other analyses, you probably want to assess consistency.

· If you want to know the reliability of individual scores, you probably want to assess absolute agreement.

4. RUN THE ANALYSIS IN SPSS.

· Analyze>Scale>Reliability Analysis.

· Select Statistics.

· Check "Intraclass correlation coefficient".

· Make choices as you decided above.

· Click Continue.

· Click OK.

· Interpret output.

REFERENCES

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428. DOI: 10.1037/0033-2909.86.2.420