The Impact Of Using The Intraclass Correlation Coefficient On Judges' Ordinal Ratings: What Happens To The ICC If Judges Are Using The Rating Scale Differently?

Petronilla Murlita Witarsa*, Krista Breithaupt, & Bruno D. Zumbo***

***The University of British Columbia**

****AICPA**

Paper Presented at the 2003 American Educational Research Association (AERA) annual meeting in Chicago, Illinois

Introduction to the Problem

- As was stated in the first paper in this symposium (Zumbo & Ochieng, 2003) the central issue is that there is an unspoken measurement assumption when we analyse Likert / rating scale data that everyone in your population is responding using the same response process – that is, the same thresholds for a given variable.
- In our case we are turning this problem around a bit and focusing on "judges" or "raters" and what happens to the intraclass correlation estimate of interrater reliability if the judges are not using the rating response scale the same way.

Problem in Context

- Ratings are any kind of coding (in our case ordinal rating) made concerning attitudes, cognitions, or behaviors. In our case we are interested in the kinds of ratings made by third-parties of a particular individual's attitudes, behaviors, knowledge, aptitude, suitability, cognition, etc.. That is, judges or raters, rating others.
- The principle goal is to determine and quantify the degree of agreement among raters when using a particular rating scheme.
- There are many different methods for quantifying the level of inter-rater agreement but we will focus on the intraclass correlation because it is widely used with rating scale data ….. And ordinal ratings at that.

Problem in Context

- There are several nice review papers on rater reliability but we will follow the Shrout & Fleiss (1979, *Psych. Bulletin*) framework and the intraclass correlation (ICC) and its corresponding variance decomposition strategy.
- Imagine, we have four judges rating some competency, behavior, or attitude. Likewise, imagine that the ratings are on a Likert scale.
- The question we are asking in this paper is what happens to the ICC estimate if the raters are (a) using an ordinal response scale, and (b) if one or more of the judges use the rating scale differently than the other judges.

Methodology

- We are examining a case similar to Table 2 in Shrout and Fleiss wherein in we have 4 judges rating.
- We are studying the case of model 2 which assumes that raters are randomly selected from some population of raters and each rater rates all "objects or patients or students" on the variable.
- This is a two-way random model.
- We are also investigating the ICC for a single rater, ICC(2,1), as well as for an average of the raters, ICC(2,4) in the Shrout and Fleiss notation.

Methodology

- We simulated the ratings based on a population correlation matrix and then we made transformations on those random variables to mimick the rating process. Except for the focus on raters, the methodology is the same as in the opening paper of this symposium (Zumbo & Ochieng, 2003).
- We also investigated the effect of rater severity (i.e., rater 'toughness'). The idea being that a rater who is using the scale differently than the others may also be the one that is more severe. (Recall that the type of differential scale use has the mis-responder being unable to distinguish between the top two scale points and using the lower of the two points.
- Recall that we are simulated the process of judges responding to rating scales.

Methodology

- Our simulation design has
  - Number of rating scale points from 3 to 9
  - Number of judges using the scale differently (i.e., mis-responding to the scale) 0, 1, 2 out of 4 judges
  - Number of tough judges i.e., the number of judges with latent variable means one standard deviation below the the other judges on the unobserved (latent) judge scale: 0, 1, 2, out of 4 judges.
  - So, completely cross 7 x 3 x 3 simulation experiment.
  - Note that in our simulation design the judge that misresponds is, where applicable, also the judge who is more severe.

Methodology Continued

- Reminder: the type of mis-responding involves not being able to differentiate between the top two response scale options and hence responding with the lesser of the two.
- We used the following correlation matrix modeled from Shrout & Fleiss

ICC(2,1)= .7575

ICC(2,4)=.9259

Correlations

		J1	J2	J3	J4
J1	Pearson Correlation	1	.742**	.722**	.747**
	Sig. (2-tailed)	.	.000	.000	.000
	N	100000	100000	100000	100000
J2	Pearson Correlation	.742**	1	.893**	.727**
	Sig. (2-tailed)	.000	.	.000	.000
	N	100000	100000	100000	100000
J3	Pearson Correlation	.722**	.893**	1	.715**
	Sig. (2-tailed)	.000	.000	.	.000
	N	100000	100000	100000	100000
J4	Pearson Correlation	.747**	.727**	.715**	1
	Sig. (2-tailed)	.000	.000	.000	.
	N	100000	100000	100000	100000

**. Correlation is significant at the 0.01 level (2-tailed).

ICC for a Single Rater

Judge "impact"				number of judges who are using the response scale differently; judge using different thresholds for the rating response		
				.0000	1.0000	2.0000
All of the judges being equally "tough" on their targets	number of Likert scale points	3.0000	single rater ICC	.5787	.5029	.4367
		4.0000	single rater ICC	.6375	.6176	.6003
		5.0000	single rater ICC	.6739	.6667	.6601
		6.0000	single rater ICC	.6972	.6940	.6908
		7.0000	single rater ICC	.7120	.7103	.7084
		8.0000	single rater ICC	.7204	.7195	.7186
		9.0000	single rater ICC	.7286	.7280	.7273
Three judges equally tough (i.e., equal unobserved variable means) and one judge being a bit tougher (i.e., with an unobserved score mean that is one standard deviation lower than the others.	number of Likert scale points	3.0000	single rater ICC	.4726	.4539	.3871
		4.0000	single rater ICC	.5230	.5190	.4993
		5.0000	single rater ICC	.5490	.5476	.5395
		6.0000	single rater ICC	.5650	.5644	.5603
		7.0000	single rater ICC	.5743	.5739	.5714
		8.0000	single rater ICC	.5794	.5794	.5780
		9.0000	single rater ICC	.5860	.5859	.5848
Two judges equally tough (i.e., equal unobserved variable means) and two judges being a bit tougher (i.e., with an unobserved score mean that is one standard deviation lower than the others.	number of Likert scale points	3.0000	single rater ICC	.4277	.4111	.3961
		4.0000	single rater ICC	.4849	.4813	.4779
		5.0000	single rater ICC	.5118	.5106	.5094
		6.0000	single rater ICC	.5268	.5263	.5259
		7.0000	single rater ICC	.5351	.5348	.5345
		8.0000	single rater ICC	.5399	.5400	.5400
		9.0000	single rater ICC	.5464	.5464	.5463

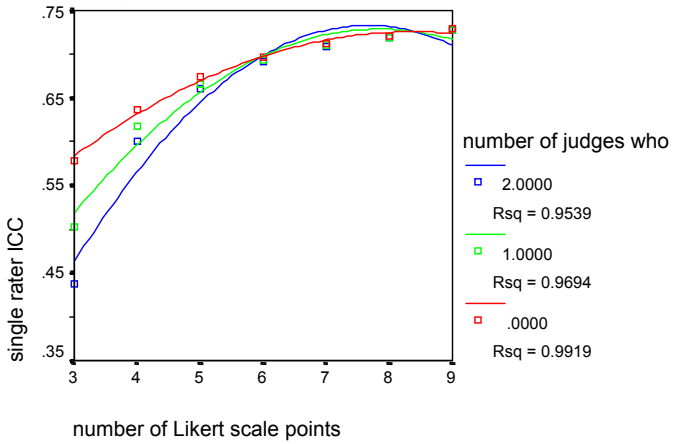
ICC for continuous variate 0.7575

ICC for continuous variate 0.6050

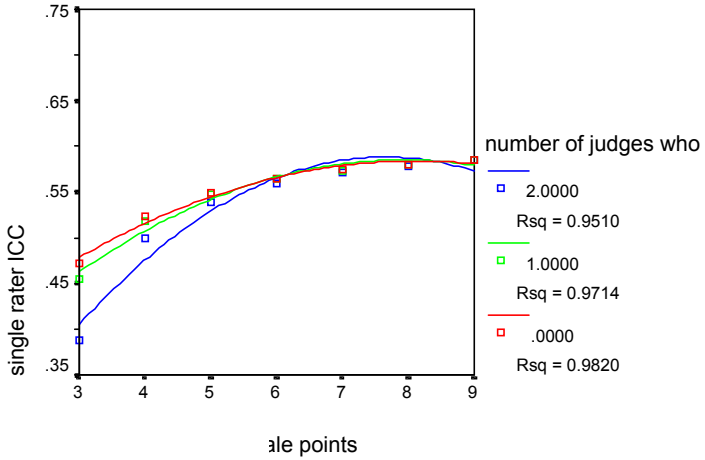
ICC for continuous variate 0.5664

Graphical Depiction of the Information on the Table

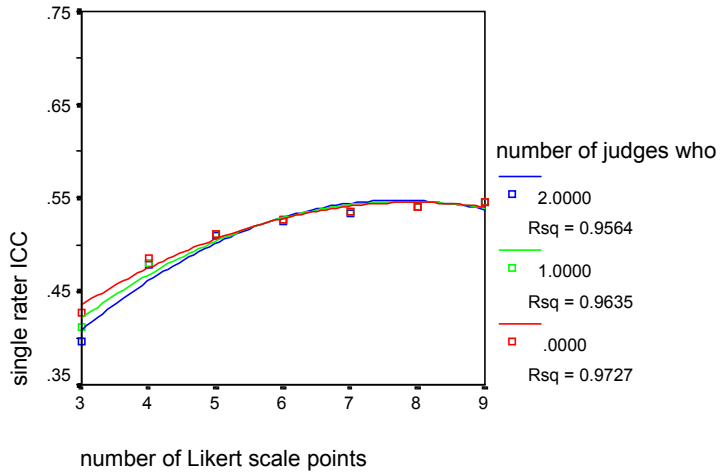
All judges equally tough on their targets



One judge is inherently tougher than the other three



Two judges are inherently tougher than the other two



Statistical Modeling of the simulation outcomes

- We fit a linear additive statistical model to the simulation outcomes for the single rater ICC.
- As expected the number of judges rating more severely than other judges resulted in a statistically significant effect, $F(1, 58)=271.9$, $p=.0001$, with more judges rating severely resulted in a lower ICC.
- There was a quadratic effect of the number of scale points: Linear $F(1, 58)=59.74$, $p=.0001$, Quadratic $F(1, 58)=35.7$, $p=.0001$.
- No statistical effect of mis-responding.

ICC for the average of the raters

All of the judges being equally "tough" on their targets

ICC for continuous variate .9259

			number of judges who are using the response scale differently; judge using different thresholds for the rating response		
			.0000	1.0000	2.0000
number of Likert scale points	3.0000	average of raters ICC	.8460	.8019	.7562
	4.0000	average of raters ICC	.8755	.8660	.8573
	5.0000	average of raters ICC	.8921	.8889	.8859
	6.0000	average of raters ICC	.9021	.9007	.8994
	7.0000	average of raters ICC	.9082	.9075	.9067
	8.0000	average of raters ICC	.9115	.9112	.9108
	9.0000	average of raters ICC	.9148	.9146	.9143

ICC for the average of the raters

One judge inherently tougher than the other three

ICC for continuous variate .8597

			number of judges who are using the response scale differently; judge using different thresholds for the rating response		
			.0000	1.0000	2.0000
number of Likert scale points	3.0000	average of raters ICC	.7819	.7687	.7164
	4.0000	average of raters ICC	.8143	.8119	.7996
	5.0000	average of raters ICC	.8296	.8288	.8241
	6.0000	average of raters ICC	.8386	.8383	.8360
	7.0000	average of raters ICC	.8436	.8435	.8421
	8.0000	average of raters ICC	.8464	.8464	.8456
	9.0000	average of raters ICC	.8499	.8499	.8493

ICC for the average of the raters

Two judges inherently tougher than the other two

ICC for continuous variate .8393

			number of judges who are using the response scale differently; judge using different thresholds for the rating response		
			.0000	1.0000	2.0000
number of Likert scale points	3.0000	average of raters ICC	.7493	.7363	.7240
	4.0000	average of raters ICC	.7902	.7877	.7855
	5.0000	average of raters ICC	.8075	.8067	.8060
	6.0000	average of raters ICC	.8166	.8163	.8161
	7.0000	average of raters ICC	.8215	.8214	.8212
	8.0000	average of raters ICC	.8244	.8244	.8244
	9.0000	average of raters ICC	.8282	.8281	.8281

Statistical Modeling of the simulation outcomes

- We fit a linear additive statistical model to the simulation outcomes for an average of the raters ICC.
- As expected the number of judges rating more severely than other judges resulted in a statistically significant effect, $F(1, 58)=267.9$, $p=.0001$, with more judges rating severely resulted in a lower ICC.
- There was a quadratic effect of the number of scale points: Linear $F(1, 58)=84.5$, $p=.0001$, Quadratic $F(1, 58)=53.0$, $p=.0001$.
- A small statistical effect of mis-responding, $F(1, 58)=5.4$, $p=.023$, with the mis-responding attentuating the ICC.

Closing remarks

- We want to be clear that we are not suggesting using the ICC when the ratings are on a 3 or 4 point scale, but rather documenting what happens when you use it.
- There needs to be further study of the mis-rating (i.e., misresponding or differential responding) but the type of misratings we studied seem to have little effect on the ICC.