CHAPTER 9

Intraclass Correlations under the Mixed Factorial Design

OBJECTIVE

This chapter aims at presenting methods for calculating the intraclass correlation for reliability studies where the participating raters are the only raters of interest. The rater factor is considered fixed. Although, the experiments considered are designed in such a way that each rater must rate all subjects, the analysis methods presented in this chapter will properly handle the missing ratings. Only the subjects are assumed to have been randomly selected from a larger subject population. Methods for obtaining confidence intervals and p-values will be discussed as well.

CONTENTS

9.1 The Problem

In the previous chapter, we presented methods for computing intraclass correlation as a measure of inter-rater or intra-rater reliability, under the random factorial design. The random factorial design treats both the subject and the rater effects as random, which is justified only when the subject and rater samples are randomly selected from larger subject and rater universes. However, the rater effect cannot be treated as random in certain types of reliability experiments. For example, the reliability experiment may use a single measuring instrument to score the same subjects on 5 different occasions. This experiment involves a single rater (the measuring instrument), which does not represent any other rater. The rater effect must be considered fixed in such a situation. A fixed rater effect combined with a random subject effect will lead to an experimental design known as the "Mixed factorial design."

There is a fundamental difference between the random and mixed factorial designs regarding the role the rater effect plays in data analysis. The rater effect plays a critical role in data analysis when it is random. When fixed, the rater effect's role in data analysis is minimal. Consequently, we would primarily recommend the mixed factorial design for studies where one expect a minimal systematic rater effect. A systematic rater effect is created in situations where one rater is stringent with all subjects, while another rater is lenient with all subjects. The mixed factorial design is appropriate for measuring whether raters interact the same way with the subjects. We will further discuss about this topic in subsequent sections.

9.2 Intraclass Correlation Coefficient

The mixed factorial design involves a single group of raters as well as a single group of subjects, all of which are rated by each rater. Note that the word raters in this context could designate a group of 5 individuals (for example) operating the same measuring device, or the 5 occasions on which a single individual used the same measuring device to score all subjects. The data produced in both situations can be analyzed with the same methods that will be discussed in this section. The analysis results may be interpreted differently depending on the context. Only the context will tell us whether the intraclass correlation presented in this chapter should be interpreted as inter-rater reliability, or as intra-rater reliability. Therefore, developing an in-depth understanding of the nature of the reliability experiment will be essential to have a valid interpretation of the data analysis.

The scores generated by a mixed design experiment will be described mathematically as,

$$y_{ijk} = \mu + s_i + r_j + (sr)_{ij} + e_{ijk}, \tag{9.1}$$

where, y_{ijk} is the k^{th} replicate measurement[1] that rater j assigned to subject i. The remaining terms in model 9.1 are defined as follows[2]:

- μ is the overall expected value of the y-score for all subjects and raters.
- s_i is the random subject effect, assumed to follow the Normal distribution with 0 mean, and variance σ_s^2.
- r_j is the fixed rater effect, assumed to satisfy the condition,

$$\sum_{j=1}^{r} r_j = 0, \tag{9.2}$$

  where r is the number of raters participating in the experiment.
- $(sr)_{ij}$ is the random subject-rater interaction effect, assumed to follow the Normal distribution with mean 0, and variance σ_{sr}^2, and to satisfy the condition,

$$\sum_{j=1}^{r} (sr)_{ij} = 0, \text{ for any subject } i. \tag{9.3}$$

- e_{ijk} is the random error effect, assumed to follow the Normal distribution with mean 0, and variance σ_e^2.

The subject, interaction, and error effects are considered mutually independent. That is, the magnitude of one effect does not affect another effect. We will also assume that the reliability experiment involves n subjects, r raters, and m measurements per rater and subject.

Model 9.1 is known in the inter-rater reliability literature as Model 3 (see Shrout & Fleiss, 1979 or McGraw & Wong, 1996), and stipulates that under the mixed factorial design, the different effects are additive (i.e. the subject and rater effects must be added to determine their joint impact on the score). From this model we will derive an intraclass correlation coefficient that could be used as a measure of inter-rater reliability or as a measure of intra-rater reliability, depending on how the experiment was set up.

In the next section we will start by investigating intraclass correlation when the unit of analysis is the individual score. The individual score represents the first raw score that one individual rater assigns to the subject. Afterwards, we will investigate intraclass correlation when the unit of analysis is a mean of k scores.

[1] Many reliability experiments only involves one replication (the first one)

[2] These are standard conditions used in all ANOVA models

9.2.1 *The Individual Score as Unit of Analysis*

The inter-rater reliability based on model 9.1, is by definition the correlation coefficient between the scores y_{ijk} and $y_{ij'k}$ associated with 2 raters j and j', the same subject i, and the same replicate number k (if any). It follows from equation 9.1 that this inter-rater reliability (denoted by ρ) is defined[3] as,

$$\boxed{\rho = \frac{\sigma_s^2 - \sigma_{sr}^2/(r-1)}{\sigma_s^2 + \sigma_{sr}^2 + \sigma_e^2}} \tag{9.4}$$

Equation 9.4 provides the definitional expression of inter-rater reliability based on the idealized model of equation 9.1. This coefficient belongs to the family of intraclass correlation coefficient (ICC), and the next step in its exploitation will be to specify the procedure for calculating it from experimental data. But at this stage, we first need to see whether equation 9.4 actually measures the extent of agreement among r raters that participated in the experiment.

A careful examination of expression 9.4 suggests that ρ varies from 0 to 1, and takes a high value closer to 1 only when the subject variance σ_s^2 exceeds the combined variance $\sigma_{sr}^2 + \sigma_e^2$ by a wide margin. That is, ρ will be high when the error and interaction variances are both reasonably small. Consequently, you will obtain a high ρ value if the following 3 conditions are satisfied:

(a) The experiment is sufficiently well designed to keep the experimental error low (i.e. σ_e^2 is small),

(b) The subject-rater interaction is limited (i.e. σ_{sr}^2). This variance component particularly, may have a dramatic impact on the intraclass correlation.

(c) The subject variance is substantially larger than the error and interaction variances.

The problem is that these 3 conditions could be met without the raters being in agreement at all. Consider the reliability data of Table 9.1 and the associated graph in Figure 9.1. This is a typical example of ratings characterized by total absence of any subject-rater interaction effect (i.e. $\sigma_{st}^2 = 0$). That is the gap between the 2 graphs associated with raters 1 and 2 remains constant across subjects. This data will nevertheless yield a high ρ value, despite the fact that raters 1 and 2 clearly

[3]Note that $\rho = \mathsf{Corr}(y_{ijk}, y_{ij'k}) = \mathsf{Cov}(y_{ijk}, y_{ij'k}) / \big[\sqrt{\mathsf{Var}(y_{ijk})}\sqrt{\mathsf{Var}(y_{ij'k})}\big]$. However, the covariance term can be re-written as, $\mathsf{Cov}(y_{ijk}, y_{ij'k}) = \sigma_s^2 + \mathsf{Cov}[(sr)_{ij}, (sr)_{ij'}]$. By taking the variance of both sides of equation 9.3, one can prove that $\mathsf{Cov}\big[(sr)_{ij}, (sr)_{ij'}\big] = -\sigma_{sr}^2/(r-1)$.

disagree about the scoring of subjects across the board. This reality has led authors such as Bartko (1976), or McGraw & Wong (1996) to consider ρ of equation 9.4 as a measure of consistency, and not as a measure of agreement.

More problematic may be the data in Table 9.2, and the associated graph in Figure 9.2, which indicates a reasonably good agreement among raters. Because of the (small) subject-rater interaction depicted in this figure (i.e. the gap between the 2 curves changes from subject to subject) the ρ value of equation 9.4 associated with this data will be smaller than that of Table 9.1. Although agreement in Figure 9.2 is higher than in Figure 9.1, equation 9.4 tends to favor[4] Figure 9.1. This is due to the observed subject-rater interaction, which penalizes Figure 9.2.

Table 9.1: Ratings without Subject-Rater Interaction

Subject	Rater (j)	
(i)	Rater1	Rater2
1	9	4
2	6	1
3	8	3
4	7	1
5	10	5
6	6	1

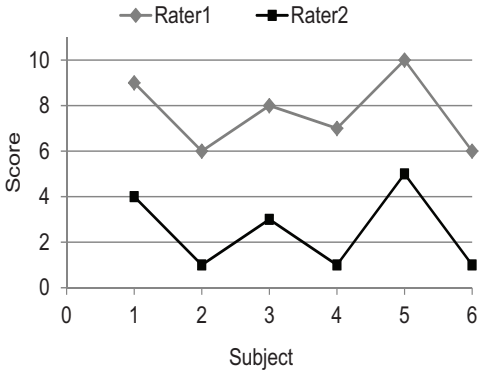

Figure 9.1 : Table 9.1 Rating Data

Table 9.2: Ratings with Subject-Rater Interaction

Subject	Rater (j)	
(i)	Rater1	Rater2
1	5	4
2	6	5
3	8	9
4	7	8
5	9	7
6	6	7

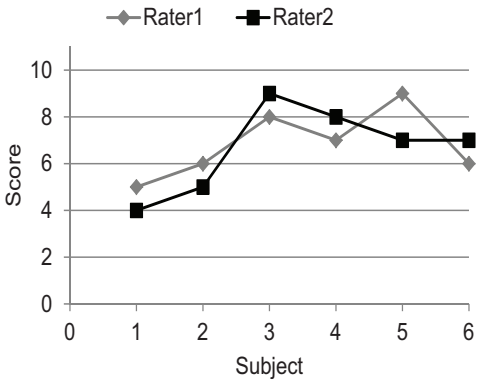

Figure 9.2 : Table 9.2 Rating Data

The problem that we just described regarding Tables 9.1 and 9.2 stems from the

[4]This will become more evident in the next section after the computation procedures are fully specified.

systematic rater effect observed in Figure 9.1, with rater 1 consistently scoring higher than rater 2. The intraclass correlation coefficient based on a mixed factorial design will be appropriate if the experiment is set up in such a way that some raters do not exhibit any systematic bias towards stringency or leniency. Generally, specific scoring instructions given to raters will minimize the systematic bias in their scoring process. The variation in their scores will depend more on how they interact with subjects.

Equation 9.4 will measure the extent of agreement among raters, as long as the raters are not biased when rating subjects. If the experiment uses a single rater who rates each subject on a number of occasions, then there will be no rater bias issue, and equation 9.4 may well be used. *Note that if you can use the mixed factorial design, you should always do so. It will generally yield a higher (sometimes much higher) intraclass correlation than the random factorial design of the previous chapter.*

In the next section, we will present procedures for calculating ρ. We will use the notation of Shrout & Fleiss (1979), who labeled the calculated value of ρ as $\mathsf{ICC}(3,1)$.

9.2.2 *Calculating Inter-Rater Reliability with Individual Score as Unit of Analysis*

When your data contains no missing score (i.e. each rater has scored all subjects), then the ICC can be calculated using a simple procedure based on standard means of squares. In practice however, many experimental datasets contain some missing scores, which makes it necessary to have a computational procedure that can handle them properly. The more general procedure that can handle missing scores will work with complete data as well, and involves complex calculations. Shrout & Fleiss (1979) as well as McGraw & Wong (1996) have described the simpler procedure for complete data, and without replication (i.e. $m = 1$). We will first present this procedure (slightly adapted to accommodate replicates), before discussing the more general and complex procedure for handling missing scores.

Calculations with Complete Experimental Data

If your data is complete, and 2 measurements or more are taken on each subject (i.e. $m \geq 2$), then the ICC associated with Model 3 of equation 9.1 is given by,

$$\mathsf{ICC(3,1)} = \frac{(\mathsf{MSS} - \mathsf{MSI}) - (\mathsf{MSI} - \mathsf{MSE})/(r-1)}{\mathsf{MSS} + r(\mathsf{MSI} - \mathsf{MSE}) + (rm-1)\mathsf{MSE}}, \tag{9.5}$$

where, MSS, MSI, and MSE represent respectively the mean of squares for subjects, the mean of squares for the subject-rater interaction, and the mean of squares for errors. These means of squares are calculated as follows: