# CHAPTER 6

# Intraclass Correlations under the Mixed Factorial Design

## OBJECTIVE

This chapter aims at presenting methods for analyzing intraclass correlation coefficients for reliability studies based on a random sample of subjects and a fixed group of raters. Therefore, the rater factor is considered fixed, while the subject factor is considered random. The intraclass correlation coefficient (ICC) used for quantifying intra-rater reliability is a valid measure of reproducibility. However, the ICC used for quantifying inter-rater reliability is a valid measure of consistency and would be a valid measure of agreement only if there is no systematic bias in the rating process from one rater. This chapter also discusses methods for obtaining confidence intervals and $p$-values for all types of ICCs under model 3, in addition to presenting a detailed account of the methods used to determine the optimal number of subjects during the experimental design.

## Contents

## 6.1 The Problem

In chapter 5, I presented methods for computing the intraclass correlation coefficient as a measure of inter-rater or intra-rater reliability, under the pure 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 two measuring instruments (an existing one and a new one) to take measurements on subjects. This experiment involves two raters (i.e. the measuring instruments expected to produce comparable measurements) and no other rater is under consideration. 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. Unlike the random factorial model of the previous chapter, which proceeds with a direct evaluation of the rater variance, the mixed factorial model is essentially based on the analysis of the interaction between raters and subjects. The raters alone do not represent a source of uncertainty to be analyzed since their effect is fixed. The only source of uncertainty involving the raters relies on the subject-rater interaction. A large subject-rater interaction has a negative effect on inter-rater reliability; but may have a positive effect on intra-rater reliability for a given total variation in the ratings. Note that the subject-rater interaction is high if the score difference between raters varies considerably from subject to subject and is low otherwise. I will discuss these relationships further in subsequent sections.

If the raters strongly agree then the subject-rater interaction will be small and the mixed factorial design will rightfully yield a high intraclass correlation coefficient. However, a small subject-rater interaction or its complete absence in a well-designed experiment could well yield a high intraclass correlation without the raters being in high agreement. This typically occurs in situations where there is a large and systematic gap (across subjects) between ratings from two raters. Although these situations are unusual in practice, especially with raters having basic training, researchers may want to take a precautionary measure by testing that ratings from all raters come from distributions with a common expected value. If this hypothesis is rejected, then I will not recommend the use of a mixed factorial design for the purpose of analyzing inter-rater reliability. Nevertheless, using a mixed factorial design for the purpose of studying inter-rater reliability is generally an effective approach for quantifying the extent of agreement among raters when the rater effect is fixed. This is due to us not having to worry about the rater variance component that represents raters that did not participate in the reliability experiment.

For the purpose of analyzing intra-rater reliability, the mixed factorial design works well when the rater effect is fixed. For a given total variation in the ratings, having a high subject-rater interaction may be beneficial since it could lead to a higher coefficient of intra-rater reliability. In fact, the error variance under the mixed factorial design includes not only the variance due to random experimental errors, but also the variance due to replication, which actually measures reproducibility. Therefore, a small error variance is an indication of small variance due to replication as well, which in turn is an indication of high reproducibility. Consequently, a high subject-rater interaction for a given total variation is not problematic and may even be a sign of a relatively small error variance and high reproducibility. The relationship between the subject-rater interaction, inter-rater and intra-rater reliability will be further discussed in subsequent sections.

## 6.2 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 operating the same measuring device, or could also designate 5 measuring devices that the same individual operates to score all subjects. The data produced in both situations can be analyzed with the same methods that are discussed in this section. The analysis results may be interpreted differently depending on the context. Therefore, developing an in-depth understanding of the nature of the reliability experiment is essential to properly interpret the data analysis.

The abstract representation of ratings under the mixed factorial design is given by,

$$y_{ijk} = \mu + s_i + r_j + (sr)_{ij} + e_{ijk}, \qquad (6.2.1)$$

where $y_{ijk}$ is the $k^{th}$ replicate score[1] that rater $j$ assigned to subject $i$. The remaining terms in model 6.2.1 are defined as follows[2]:

- $\mu$ is the expected value of the $y$-score.
- $s_i$ is the random subject effect, assumed to follow the Normal distribution with 0 mean and variance $\sigma_s^2$.

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

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

- $r_j$ is the fixed rater effect, assumed to satisfy the condition,

$$\sum_{j=1}^{r} r_j = 0, \tag{6.2.2}$$

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 $\sigma^2_{sr}$, and to satisfy the condition,

$$\sum_{j=1}^{r} (sr)_{ij} = 0, \text{ for any subject } i. \tag{6.2.3}$$

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

The subject, interaction and error effects are considered mutually independent. That is, the magnitude of one effect does not affect that of another effect. Now, suppose that the reliability experiment involves $n$ subjects, $r$ raters and $m$ measurements per rater and subject. Later in this section, I will consider the more general situation where the number of measurements $m$ varies by rater and by subject.

Model 6.2.1 is known in the inter-rater reliability literature as Model 3 (see Shrout and Fleiss, 1979; McGraw and 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 I will derive two intraclass correlation coefficients. One coefficient will be a measure of inter-rater reliability, while the other will be used as a measure of intra-rater reliability.

### 6.2.1 *Defining Inter-Rater Reliability*

The inter-rater reliability based on model 6.2.1, is by definition the correlation coefficient between the scores $y_{ijk}$ and $y_{ij'k}$ associated with two raters $j$ and $j'$, the same subject $i$ and the same trial number $k$ (if any). It follows from equation 6.2.1 that this inter-rater reliability (denoted by $\rho$) is defined[3] as,

$$\boxed{\rho = \frac{\sigma^2_s - \sigma^2_{sr}/(r-1)}{\sigma^2_s + \sigma^2_{sr} + \sigma^2_e}} \tag{6.2.4}$$

[3] Note that $\rho = \text{Corr}(y_{ijk}, y_{ij'k}) = \text{Cov}(y_{ijk}, y_{ij'k})/\left[\sqrt{\text{Var}(y_{ijk})}\sqrt{\text{Var}(y_{ij'k})}\right]$. However, the covariance term can be re-written as, $\text{Cov}(y_{ijk}, y_{ij'k}) = \sigma^2_s + \text{Cov}[(sr)_{ij}, (sr)_{ij'}]$. By taking the variance of both sides of equation 6.2.3, one can prove that $\text{Cov}\big[(sr)_{ij}, (sr)_{ij'}\big] = -\sigma^2_{sr}/(r-1)$.

Equation 6.2.4 provides the definitional expression of inter-rater reliability based on the idealized model of equation 6.2.1. This coefficient belongs to the family of intraclass correlation coefficient (ICC) and the next step in its exploitation is to specify the procedure for calculating it from experimental data. But at this stage, I first need to see whether equation 6.2.4 actually measures the extent of agreement among the $r$ raters that participated in the experiment.

A careful examination of expression 6.2.4 suggests that $\rho$ varies from 0 to 1 and takes a high value closer to 1 only when the subject variance $\sigma_s^2$ exceeds the combined variance $\sigma_{sr}^2 + \sigma_e^2$ by a wide margin. That is, $\rho$ will be high when the error and interaction variances are both relatively small. Consequently, you will obtain a high $\rho$ value if the following 3 conditions are satisfied:

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

(*b*) The subject-rater interaction is limited (i.e. $\sigma_{sr}^2$). 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 combined.

If the raters are in agreement and the experiment well planned, these three conditions will be met. However, there is a problem that stems from the fact that these three conditions could be met without the raters being in agreement. Consider the reliability data of Table 6.1 and the associated graph in Figure 6.1. 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 two graphs associated with raters 1 and 2 remains constant across subjects. This data will nevertheless yield a high $\rho$ value, despite the fact that raters 1 and 2 clearly disagree about the scoring of subjects across the board. This reality has led authors such as Bartko (1976), or McGraw and Wong (1996) to consider $\rho$ of equation 6.2.4 as a measure of consistency and not as a measure of agreement.

Table 6.1: 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 |

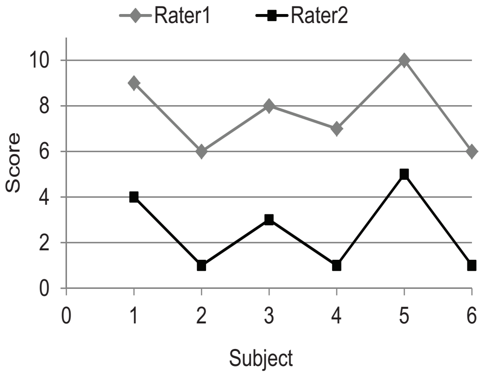

Figure 6.1: Rating data from Table 6.1

Table 6.2: 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 |

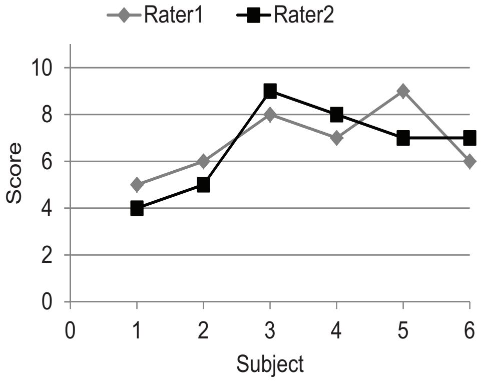

Figure 6.2: Rating data from Table 6.2

Table 6.2 ratings and the associated graph of Figure 6.2 show a reasonably good agreement among raters. Because of the (small) subject-rater interaction depicted in this figure (i.e. the gap between the two curves changes from subject to subject) the $\rho$ value of equation 6.2.4 associated with this data will be smaller than that of Table 6.1. Although agreement in Figure 6.2 is higher than in Figure 6.1, equation 6.2.4 tends to favor Figure 6.1. This is due to the observed subject-rater interaction, which penalizes Figure 6.2.

The problem I just described regarding Tables 6.1 and 6.2 stems from the systematic rater effect observed in Figure 6.1, 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. *There should not be any* *"substantial" rater effect that is independent of the subjects. The subject-rater interaction and the experimental errors are expected to be the sole sources of disagreement among raters.* 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 6.2.4 will adequately measure the extent of agreement among raters, as long as the raters are not heavily biased when rating subjects. When using the mixed factorial design to analyze inter-rater reliability, researchers may want to first test that the ratings from all raters are distributed around a common expected value. For two raters only, one may use the Mann-Whitney test proposed by Mann and Whitney (1947) or the Wilcoxon rank sum test recommended by Wilcoxon (1949) . For three raters or more, one may use the Kruskal-Wallis test proposed by Kruskal (1952) and Kruskal and Wallis (1952). If these statistical tests reject the hypothesis that the ratings are distributed around the same expected value, then I would not use the mixed factorial design for analyzing inter-rater reliability.

If the experiment uses a single rater who rates each subject several times, then there will be no rater bias issue and equation 6.2.4 may 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, because of the absence of an explicit rater variance component.

### 6.2.2 *Defining Intra-Rater Reliability*

The mixed factorial design where the rater effect is fixed can be used for analyzing the inter-rater reliability as discussed in the previous section. But you may use it to analyze intra-rater reliability as well, provided the experiment produces repeated measurements on the same subjects. Usually two trials are sufficient to evaluate intra-rater reliability. As usual, I define the intra-rater reliability coefficient as the correlation coefficient between two ratings $y_{ijk}$ and $y_{ijk'}$ associated with the same rater, the same subject and two trials $k$ and $k'$. Using the statistical model described with equation 6.2.1, one can establish that the intra-rater reliability coefficient is defined by,

$$\gamma = \frac{\sigma_s^2 + \sigma_{sr}^2}{\sigma_s^2 + \sigma_{sr}^2 + \sigma_e^2}, \tag{6.2.5}$$

where $\sigma_s^2$, $\sigma_{sr}^2$ and $\sigma_e^2$ are the variance components due to the subject, subject-rater interaction and error effects. It follows from this expression that the magnitude of the intra-rater reliability coefficient is essentially determined by the ratio of the error variance to the total of the subject and subject-rater interaction variances.

Consequently, if the error variation exceeds the total subject and interaction variation by a wide margin, then the intra-rater reliability coefficient will be small. However, a relatively small error variance will lead to a high intra-rater reliability coefficient. This is what you would normally expect since the error variation is induced by both the replication and of the experimental error. Hence a small error variation implies a small replication variation, which guarantees reproducibility.

Note that a large error variance will automatically lead to a small intra-rater reliability coefficient whether the ratings are actually reproducible or not. This reminds us that intra-rater reliability experiments must be carefully conducted to minimize possible experimental errors, whose presence make any reproducibility of measurements or rater agreement undetectable.

### 6.2.3 *Calculating Inter-Rater and Intra-Rater Reliability Coefficients*

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 regular 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 and Fleiss (1979) as well as McGraw and Wong (1996) described the simpler procedure for complete data and without replication (i.e. $m = 1$). I will first present this procedure (slightly adapted to accommodate replicates), before discussing the more general procedure for missing scores.

#### SPECIAL METHOD FOR BALANCED DATA

If your data is complete and two measurements or more are taken on each subject (i.e. $m \geq 2$), then the inter-rater reliability coefficient $\rho$ of equation 6.2.4 can be estimated[4] by,

$$\mathrm{ICC}(3,1) = \frac{(\mathrm{MSS}-\mathrm{MSI})-(\mathrm{MSI}-\mathrm{MSE})/(r-1)}{\mathrm{MSS}+r(\mathrm{MSI}-\mathrm{MSE})+(rm-1)\mathrm{MSE}}, \tag{6.2.6}$$

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:

[4] The estimated value of $\rho$ based on experimental data is designated by ICC(3,1) to keep the notation used in the literature, where number 3 refers to model 3 and number 1 refers to the fact that the unit of analysis is 1 basic measurement as opposed to an average of several measurements.

- Mean Squares for Subjects (**MSS**)
  The MSS is calculated by summing the squared differences $(\overline{y}_{i\cdot\cdot} - \overline{y})^2$ over all $n$ subjects and multiplying the summation by $rm/(n-1)$. Note that $\overline{y}$ is the overall score average, while $\overline{y}_{i\cdot\cdot}$ is the average of all measurements associated with subject $i$.

- Mean Squares for Interaction (**MSI**)
  The MSI is calculated by summing the squared differences $(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y})^2$ over all $n$ subjects and $r$ raters and by multiplying the summation by $m/[(r-1)(n-1)]$. The term $\overline{y}_{ij\cdot}$ represents the average of all measurements associated with rater $j$ and subject $i$. The term $\overline{y}_{\cdot j\cdot}$ on the other hand, is the average of all measurements associated with rater $j$.

- Mean Squares for Errors (**MSE**)
  The MSE is calculated by summing the squared differences $(y_{ijk} - \overline{y}_{ij\cdot})^2$ over all $rnm$ measurements and by dividing the summation by $rn(m-1)$. Note that this mean of squares can be calculated only if there are 2 measurements or more taken on each subject (i.e. $m \geq 2$). In case the experiment is conducted without replication (i.e. $m = 1$), do not compute this quantity. Instead, compute MSI and rename it as MSE since interaction and error effects will be blended together.

Using observed ratings, the intra-rater reliability coefficient $\gamma$ of equation 6.2.5 can be estimated data as follows:

$$\text{ICC}_a(3,1) = \frac{\text{MSS} + r\text{MSI} - (r+1)\text{MSE}}{\text{MSS} + r\text{MSI} + (rm - r - 1)\text{MSE}}, \tag{6.2.7}$$

where the different means of squares are defined as above.

If there is a single measurement taken on each subject (i.e. $m = 1$ or no replication) then you cannot evaluate intra-rater reliability, although inter-rater reliability can be calculated. To compute the inter-rater reliability coefficient with a single measurement per subject, you need to assume a statistical model similar to that of equation 6.2.1 but without the interaction term[5]. Under this restricted model, the inter-rater reliability coefficient is estimated as follows:

$$\text{ICC}(3,1) = \frac{\text{MSS} - \text{MSE}}{\text{MSS} + (r-1)\text{MSE}}, \tag{6.2.8}$$

where the mean of squares for subjects (MSS) and the mean of squares for errors (MSE) are calculated as follows:

[5] Assuming a model without interaction does not mean that you refute the existence of a possible interaction. Instead, it means you will treat the interaction and error effects as a single effect.

- MSS is the summation of all squared differences $(\overline{y}_{i\cdot} - \overline{y})^2$ that is multiplied by $r/(n-1)$. Moreover, $\overline{y}_{i\cdot}$ is the average of all scores associated with subject $i$ and $\overline{y}$ the overall mean score.

- MSE on the other hand, is the summation of all squared differences $(y_{ij} - \overline{y}_{i\cdot} - \overline{y}_{\cdot j} + \overline{y})^2$ that is divided by $(r-1)(n-1)$, where $\overline{y}_{\cdot j}$ is the average of all scores associated with rater $j$.

The following example illustrates the calculation of the inter-rater and intra-rater reliability of equations 6.2.6 and 6.2.7.

### Example 6.1

To illustrate the calculation of ICC(3, 1) of equation 6.2.6 when the data is complete, let us consider the data shown in Table 6.3. Four chiropractors evaluated twice the distance along the spine of a particular condition on 16 patients.

Therefore $n = 16$ (number of subjects), $r = 4$ (number of raters or chiropractors) and $m = 2$ (the number of replicates). It follows from equation 6.2.6 that we need to compute the 3 means of squares MSS, MSI and MSE. These means of squares may be computed following the definitions given above. For more details, you may want to look at the Excel worksheet named "Example 6.1" contained in the downloadable Excel workbook "`www.agreestat.com/books/icc5/chapter6/chapter6examples.xlsx`."

The three means of squares needed here are given by,

$$\bullet\ \text{MSS} = 15,961.333 \quad \bullet\ \text{MSI} = 1,852.558 \quad \bullet\ \text{MSE} = 1,771.555.$$

Consequently, the intraclass correlation of equation 6.2.6 associated with inter-rater reliability is calculated as follows:

$$\begin{aligned}\text{ICC}(3,1) &= \frac{(15,961.333 - 1,852.558) - (1,852.558 - 1,771.555)/(4-1)}{15,961.333 + 4 \times (1,852.558 - 1,771.555) + (4 \times 2 - 1) \times 1,771.555},\\ &= 14,081.77/28,686.23 = 0.4909.\end{aligned}$$

As for the intra-rater reliability, the associated intraclass correlation coefficient of equation 6.2.7 is given by,

$$\begin{aligned}\text{ICC}_a(3,1) &= \frac{15,961.33 + 4 \times 1,852.56 - (4+1) \times 1,771.55}{15,961.33 + 4 \times 1,852.56 + (4 \times 2 - 4 - 1) \times 1,771.555},\\ &= 14,513.82/28,686.24 = 0.5059.\end{aligned}$$

Table 6.3: Chiropractic Assessment[a] of 16 Patients by 4 Chiropractors

| | Chiropractor ($j$) | | | | | Chiropractor ($j$) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Patient | CC | PK | JA | LM | Patient | CC | PK | JA | LM |
| 1 | 115 | 132 | 22 | 33 | 9 | 140 | 74 | 65 | 66 |
| 1 | 45 | 34 | 243 | 10 | 9 | 114 | 101 | 185 | 86 |
| 2 | 191 | 191 | 216 | 193 | 10 | 122 | 56 | 60 | 50 |
| 2 | 197 | 196 | 223 | 208 | 10 | 98 | 70 | 64 | 53 |
| 3 | 50 | 29 | 26 | 25 | 11 | 100 | 72 | 122 | 190 |
| 3 | 27 | 23 | 31 | 26 | 11 | 56 | 114 | 124 | 51 |
| 4 | 63 | 175 | 29 | 189 | 12 | 120 | 110 | 103 | 32 |
| 4 | 52 | 93 | 65 | 92 | 12 | 97 | 127 | 38 | 124 |
| 5 | 195 | 149 | 170 | 155 | 13 | 125 | 86 | 12 | 123 |
| 5 | 166 | 142 | 164 | 180 | 13 | 131 | 102 | 82 | 91 |
| 6 | 67 | 160 | 35 | 33 | 14 | 100 | 29 | 36 | 23 |
| 6 | 170 | 41 | 22 | 159 | 14 | 53 | 35 | 22 | 32 |
| 7 | 192 | 140 | 138 | 184 | 15 | 42 | 39 | 84 | 50 |
| 7 | 72 | 120 | 143 | 127 | 15 | 38 | 51 | 59 | 40 |
| 8 | 153 | 61 | 77 | 172 | 16 | 18 | 120 | 18 | 93 |
| 8 | 170 | 61 | 72 | 52 | 16 | 54 | 115 | 22 | 24 |

[a] *The numbers represent the distance in millimeters along the spine, of a particular condition.*

## GENERAL METHOD FOR UNBALANCED DATA

It is very common in inter-rater and intra-rater reliability studies, for one rater to generate more or less ratings than another rater. This situation may occur because one rater for whatever reason failed to rate one subject that other raters have rated. Even if all raters rate the same number of subjects, one rater may still take more measurements on some subjects than other raters. In both situations, we are dealing with the common problem of missing ratings. Instead of ignoring it as is often the case in the inter-rater reliability literature, I would like to address it adequately by adapting some general statistical techniques previously proposed by Searle (1997, page 474).

For data containing missing ratings, the inter-rater reliability coefficient ICC$(3, 1)$, and the intra-rater reliability coefficient $\text{ICC}_a(3,1)$ are calculated as follows:

$$\text{ICC}(3,1) = \frac{\widehat{\sigma}_s^2 - \widehat{\sigma}_{sr}^2/(r-1)}{\widehat{\sigma}_s^2 + \widehat{\sigma}_{sr}^2 + \widehat{\sigma}_e^2}, \tag{6.2.9}$$

$$\text{ICC}_a(3,1) = \frac{\widehat{\sigma}_s^2 + \widehat{\sigma}_{sr}^2}{\widehat{\sigma}_s^2 + \widehat{\sigma}_{sr}^2 + \widehat{\sigma}_e^2}, \tag{6.2.10}$$

where $\widehat{\sigma}_e^2$, $\widehat{\sigma}_s^2$, $\widehat{\sigma}_{sr}^2$ are the estimated values of the error variance, the subject variance and the subject-rater interaction variance calculated from experimental data. Note that the literature offers different methods for computing these variance components. If some ratings are missing, then these methods will yield different estimates. The method used in this section is referred to in the literature as *Henderson Method III*. This method is not expected to be superior to the alternatives in any way. It was chosen by me for being more convenient to implement in practice than its competitors.

The methods for calculating these variance components are first presented for the case of a mixed linear model with subject-rater interaction, which requires some replication (i.e. multiple measurements are taken on some subjects). Next, I will discuss the case of a mixed model without subject-rater interaction, which also covers the situation where there is no replication (i.e. only one rating per subject and per rater).

- If the analysis is to be done under the mixed model with subject-rater interaction, then some subjects must have been rated multiple times. That is the number of measurements $m_{ij}$ associated with subject $i$ and rater $j$ must be 2 or more for some subjects. The variance components are then calculated as follows:

$$\widehat{\sigma}_e^2 = (T_{2y} - T_{2sr})/(M - \lambda_0), \tag{6.2.11}$$

$$\widehat{\sigma}_{sr}^2 = \big[T_{2sr} - \text{RSS} - (\lambda_0 - n - r + 1)\widehat{\sigma}_e^2\big]/h_6 \tag{6.2.12}$$

$$\widehat{\sigma}_s^2 = \big[T_{2sr} - T_{2r} - (\lambda_0 - r)\widehat{\sigma}_e^2\big]/(M - k_4) - (r-1)\widehat{\sigma}_{sr}^2/r. \tag{6.2.13}$$

  where $M$ is the total number of measurements, $\lambda_0$ is the number of non-empty subject-rater cells[6] $(i,j)$, while $k_4$ is calculated by summing the quantities $m_{ij}^2/m_{\cdot j}$ over all subjects and raters (note that $m_{ij}$ is the number of measurements associated with subject $i$ and rater $j$ and $m_{\cdot j}$ the number of measurements associated with rater $j$).

[6] If the reliability experiment involves 2 raters and 3 subjects for example, then the total number of subject-rater cells would be $3\times 2 = 6$. Suppose all subjects are rated by both raters, except subject #3 that is rated by rater #1 only. Therefore, the cell associated with subject #3 and rater #2 will be empty and does not count. The number of non-empty cells will then be $\lambda_0 = 5$ (i.e. $6-1=5$).