Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

CHAPTER 2

Setting Up a Database of Ratings for Analysis

OBJECTIVE

Rating data collected from subjects must be structured in a way that is suitable for analysis. While many inter-rater reliability studies are simple and do not present any particular challenge for structuring the database before it can be analyzed, other studies however can be quite challenging to the point where even deciding what the raters and the subjects are becomes a difficult task. In this case, organizing your ratings properly for analysis will require some preliminary work. You will also learn in this chapter how to approach studies where the very notion of agreement has multiple dimensions. Some situations may even call for an independent analysis of different aspects of agreement. Knowing how to organize your rating data is a critical step for a successful analysis. Therefore, the primary objective of this chapter is to provide guidelines to researchers for setting up their datasets of ratings before analysis can begin.

Contents

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

"The man who grasps principles can successfully select his own methods. The man who tries methods, ignoring principles, is sure to have trouble."
Ralph Waldo Emmerson (May 25, 1803 - April 27, 1882)

2.1 Introduction

Years of practice in the field of inter-rater reliability have convinced me that researchers needed guidelines to properly structure their inter-rater reliability data and adequately frame their problem before the analysis itself can begin. Therefore, this chapter explores various scenarios that are expected to create challenges, and discusses ways to overcome them. I am assuming here that you have collected quantitative ratings. Nominal ratings in the form of qualitative variables with no ordinal structure are out of the scope of this book. If you have collected quantitative ratings and want to use agreement coefficients such as Cohen's kappa, then you will be better served by the "Handbook of Inter-Rater Reliability, Volume I" by Gwet (2021).

Rating data must be organized in a logical manner before it can be analyzed effectively. This is especially true if you are going to use a software package for analysis. In fact, many problems researchers face while analyzing their rating data are created by the bad layout of their dataset. One negative implication of a poorly-organized rating dataset is the difficulty to clearly identify the 3 key components of the analysis of inter-rater reliability data. These are the "`Raters`", the "`Subjects`" and the "`Ratings`". In a properly structured database, the researcher should always know for each quantitative rating, which subject it is assigned to, and what rater produced it. Not being to answer these questions could be a sign that your data may not be well-structured.

Although there is a wide variety of ways a dataset of ratings can be organized, there are a few guiding principles that should normally be followed. This book assumes that a rectangular structure can be adapted to your dataset. That is, your entire dataset will be in the form of columns data of the same type. Each column will contain several values taken by the same variable, and each row must be associated with one subject or one case and is referred to as a record. Unless your data can be in organized in a rectangular form, applying the methods presented in this book will prove difficult. Several examples of rectangular datasets are discussed in this book.

There are essentially 2 ways for you to organize your quantitative rating data before analysis, each of which presents advantages and disadvantages. Unless you are sure about what you are doing, I would expect your data format to be a variant of one of the following 2 options:

- **Wide Data Format** (WDF)
  The wide data format is one way you can organize your data in a special

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

table format where each row represents a subject, each column a rater and each table entry at the junction of the row and column represents the specific rating the rater assigned to the subject. As shown in Table 2.1, this format is essentially a listing of all ratings organized by subject and raters. Its main advantage over the previous 2 formats is the completeness of the information it presents. With this format, there is no loss of information as it shows what rater rated what subject and the specific rating assigned to every subject. A secondary advantage of this format is its ability to use categorical ratings as well as quantitative measurements.

One may see the WDF format as one which only has advantages. No, it does not only have advantages. Its main two disadvantages are the dependency upon the number of raters and its inability to accommodate more than one study factors. In fact, the number of raters in the study equals the number of columns in the table plus 1 (the "units" column). Consequently, a large number of raters will inevitable results in a larger table with many columns. But the more serious limitation of this data format is its inability to accommodate more than one factor. Assume that the raters must rate a group of subjects on 2 characteristics of factors. Each factor will require a separate WDF-type table for all ratings to be displayed and the number of table will increase with the number factors being analyzed. A more efficient table with a different record layout is necessary in this case.

- **Long Data Format** (LDF)
  The long data format is best described with an example. Table 2.2 shows the ratings assigned to 6 subjects by 3 raters on 3 different factors. As you can see the long format requires more rows than the wide format (hence the name long format). However, it allows for the display of ratings collected on multiple factors and can accommodate categorical as well as quantitative measurements. If the subjects are rated on many factors then this should be your initial format of choice, although the specific software product you want to use will ultimately determine the final format before the analysis begins.

  While LDF is the most general of all data formats presented in this section, using it when subjects are being rated on a single factor may appear as overkill. In reality when dealing with a single factor, there is no need to use a format more general than the WDF.

So far, I presented 2 different ways for you to organize your rating data, all of which being based upon the assumption that the notions of subject and rater are well-defined, that the number of subjects and raters are given and that each rater assigns a single rating to each subject. Unfortunately in the real world, inter-rater

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

reliability problems can get quite complex and knowing how to organize your data and how to analyze it can quickly become a daunting task. In the next few sections, I am going to discuss a few special cases that I have encountered in practice and which require a careful examination.

Table 2.1: Ratings assigned to 5 subjects by 4 raters.

Subject	J1	J2	J3	J4
1	6	1	3	2
1	6.5	3	3	4
1	4	3	5.5	4
5	10	5	6	9
5	9	4.5	5	9
5	9.5	4	6.6	8
4	6	2	4	7
4	7	1	3	6
4	8	2.5	4	5
2	9	2	5	8
2	7	1	2	6
2	8	2	2	7
3	10	5	6	9
3	7	4	6	5
3	8	4	6	8

Table 2.2: Ratings of 6 subjects by 3 raters.

Subject	Rater	factor1	factor2	factor3
1	1	1.753	1.813	1.701
1	2	4.366	4.057	4.504
1	3	2.491	2.238	2.647
2	1	0.801	0.721	0.894
2	2	2.073	2.162	1.881
2	3	2.588	2.996	2.104
3	1	1.563	1.469	1.467
3	2	2.224	2.154	2.029
3	3	1.423	1.466	1.383
4	1	0.791	0.824	0.740
4	2	1.840	1.613	2.083
4	3	2.325	1.959	2.813
5	1	0.798	0.830	0.730
5	2	2.223	2.063	2.591
5	3	2.786	2.484	3.548
6	1	0.965	1.095	0.911
6	2	1.634	1.678	1.395
6	3	1.694	1.532	1.531

2.2 Defining Subjects and Characteristics

As surprising as it may appear, the notions of subject and rater are sometimes very fuzzy. The subject for example, is not always going to be associated with a single well-defined entity to which a rater is expected to assign a rating. Consider for example an inter-rater reliability experiment that takes place in a university setting, where 11 students in the Linguistics department must take 3 versions of the same test. Suppose that the 3 versions of the test are labeled as "M"[1], "VCP" and "VCE." Moreover, the M version of the test has 4 analytic components, which are Grammar, Vocabulary, Fluency and Pronunciation. The VCP and VCE versions of the test however, have each 5 analytical components, which are the 4 components of the M

[1] The actual meaning of these acronyms is irrelevant for us right now.

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

test in addition to the Listening component. Four raters must assign quantitative ratings to these students and some of the raters may not be able to rate all students on all components of each version of the tests. While your ultimate goal is to quantify the extent of agreement among the 4 raters who participated in this inter-rater reliability experiment, several questions remain unanswered for the moment. Should you analyze the rater agreement separately for each version of the test? What about the analytic component of each version? Should you instead focus on a global measure of agreement using all available ratings?

Table 2.3: Description of 3 versions of a linguistics test

Analytic	Test Version		
Component	M	VCP	VCE
Grammar	X	x	X
Vocabulary	X	X	X
Fluency	X	X	X
Pronunciation	x	X	X
Listening		X	X

As you may see, what you are dealing with here is far from a simple experiment involving 2 raters and a few subjects to be rated on a single factor. Unless you take the time to carefully think about the types of analyses you want to perform, you may never find the right approach for organizing these ratings. At first sight, it is unclear what you should consider to be the subject and the subject's attribute (or factor) being rated. Should you consider the student as subject and perhaps the student's proficiency in grammar as an attribute? Perhaps combining the student and the test version being taken could be seen a subject that must be rated on a particular analytic component. How do you set up your database and what ratings do you use in the calculation of what coefficient?

Ultimately you will need to build a dataset containing several columns of data. Each of these data columns is labeled with variable name. Your first task is to identify all of these variables of interest and to identify the values they can take. Some of these values might need to be repeated to show the relationship among the variables. I propose the following two possible lists of variables for this problem (i.e. 2 representations of the same data), before discussing their advantages and disadvantages:

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

Option 1: 7 variables defined

1· The student name referred to as `STUDENT`,
2· The test version named `VERSION`,
3· The test component named `COMPONENT`,
4· Rater1's scores named `RATER1`,
5· Rater2's scores named `RATER2`,
6· Rater3's scores named `RATER3`,
7· Rater4's scores named `RATER4`.

Option 2: 9 variables defined

1· The student name referred to as `STUDENT`,
2· The test version named `VERSION`,
3· The rater's name referred to as `RATER`,
4· Student's scores in fluency, named `FLUENCY`,
5· Student's scores in grammar, named `GRAMMAR`,
6· Student's scores in listening, named `LISTEN`,
7· Student's overall test version score, named `OVERALL`,
8· Student's scores in pronunciation, named `PRONU`,
9· Student's scores in vocabulary, named `VOCAB`,

Note that both options only have the `STUDENT` and `VERSION` variables in common and translate to Table 2.4 for option 1 and to Table 2.5 for option 2. The `STUDENT` can take the 11 values `Amber, David, Isaac, Jasmine, Lee, Mary, Ricardo, Suzan, Viktor, Yanick, Yin`, while the `VERSION` variables can take the 3 values `M, VCE, VCP`. In Table 2.4 all ratings associated with one rater are listed in a single column and their analysis across students, test versions and their components are made considerably more convenient. In Table 2.5 on the hand, it is rather the ratings associated with an analytic component of the test that are listed in a single column, those associated with raters being spread across several rows.

If all analytic components are rated using the same or a similar scoring rubric and therefore take similar values, then you may consider "Proficiency" as the single characteristic on which all subjects are being rated. Moreover, any combination of `NAME`, `VERSION` and `COMPONENT` can be seen as a subject (not just the student). The global multiple-rater inter-rater reliability coefficient could then be calculated using all ratings that were collected. Therefore, the wide-format dataset (i.e. Table 2.4) is recommended if the focus of your inter-rater reliability experiment is on the rating of a single factor (i.e. "proficiency") for all subjects.

Depending on how the different analytic components are being rated, the rating

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

of students on GRAMMAR may be done on a scale that is different from the one used to rate the same student on FLUENCY (e.g. one scale may be in numbers and another one in letters). In this case, I would strongly recommend considering proficiency in the 5 components FLUENCY, GRAMMAR, LISTEN, PRONU and VOCAB to represent 5 distinct characteristics and to use the long-format dataset where each column is made up of a set of homogeneous data point that can be analyzed together (see Table 2.5). The columns associated with the 5 factors must be analyzed separately, since they represent different factors. However, an OVERALL column can be defined if needed in order to conduct a global analysis of proficiency.

With the long-format dataset, each combination of NAME and VERSION represents a subject. The analysis of particular factor such as the proficiency in GRAMMAR may require that you extract the variables NAME, VERSION, RATER and GRAMMAR in order to create a wide-format dataset similar to Table 2.4 without the COMPONENT column and with numbers that represent the ratings associated with the students' proficiency in grammar. To recapitulate, the long-format is used to store all of your rating data in a logical way. At the time of analysis, where one factor must be considered at a time, one could use the wide-format for that particular factor to facilitate the analysis.

You may note that the wide-format dataset represented by Table 2.4 is much longer than the long-format dataset represented by Table 2.5, which is wider than the wide-format. As a general rule, when raters' names are listed horizontally, it is a wide format and when they use a single column and repeated several times, it is a long format.

Table 2.4: Eleven Students' Linguistics Test Scores in a Wide Format[a]

STUDENT	VERSION	COMPONENT	RATER1	RATER2	RATER3	RATER4
Suzan	M	Grammar	3		3.5	3
Suzan	M	Vocabulary	3		3.5	3.5
Suzan	M	Fluency	3		3.5	3
Suzan	M	Pronunciation	3		3.5	3
Suzan	M	Overall	3		3.5	3
Mary	M	Grammar	4	4	4	
Mary	M	Vocabulary	3.5	4	3.5	
Mary	M	Fluency	3	3.5	3.5	
Mary	M	Pronunciation	3	4	3.5	
Mary	M	Overall	3.5	3.5	3.5	

[a] This table is an extract of the longer table A.1 that can be found in Appendix A

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

Table 2.5: Eleven Students' Linguistics Test Scores in a Long Format[a]

STUDENT	VERSION	RATER	FLUENCY	GRAMMAR	LISTEN	OVERALL	PRONU	VOCAB
Suzan	M	Rater 1	3	3		3	3	3
Suzan	M	Rater 3	3.5	3.5		3.5	3.5	3.5
Suzan	M	Rater 4	3	3		3	3	3.5
Mary	M	Rater 1	3	4		3.5	3	3.5
Mary	M	Rater 2	3.5	4		3.5	4	4
Mary	M	Rater 3	3.5	4		3.5	3.5	3.5

[a] This table is an extract of the longer table A.2 that can be found in Appendix A

2.3 Defining Raters in Complex Situations

While the majority of inter-rater reliability studies are based on well-defined and identifiable group of raters, this is not always the case. On this section, I am going to discuss 2 examples of rating datasets where knowing what the raters and the subjects are requires some preliminary work. In the first example, there is a need to construct the rater based on existing variables. The second example presents a peculiar situation where the raters and the subjects are closely linked and are not clearly identifiable as raters.

2.3.1 *Myocardial Blood Flow Measurements*

In this section, I review an example of rating data where it may be necessary to combine 2 variables to define a rater or a subject. Table 2.6 shown Myocardial Blood Flow[2] (MBF) measurements often used to diagnose coronary artery disease (CAD). These measurements are obtained based on a computer-assisted analysis of heart scans of CAD patients. The first column `VSOFT` represents the software package used in the analysis, `VID` is the patient identification number, `VMOD` is the particular data analysis model that is implemented in the software, `R/S` determines whether the measurements were taken when the heart was at rest ("`R`") and during physical stress ("`S`"). The remaining variables `GLOBAL`, `LAD` and `RCA` represent different parts of the heart where the measurement was taken.

Typically, researchers conduct these studies to determine the extent to which different software packages agree with each other. This dataset can be analyzed in

[2] Myocardial blood flow can be defined as the volume of blood transiting through tissue at a certain rate.

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

many different ways. Regardless of how the data will be analyzed, look at how it is organized in Table 2.6.

- Each column represents one variable, which gives you direct access to its content. VSOFT tells you which software package produced the data. Note that there are only 2 software packages used in this dataset (1 and 3). Unlike software 1, which only implements model 2, software package 3 implements both models 2 and 4 as indicated by variable VMOD.

- This dataset is organized logically in the sense that it clearly establishes the relationships between the MBF variables being measured and the corresponding conditions under which they were measured. The long format is used here to describe the data and access is made easy using any software package of choice.

Note that there is a natural way to define subjects that consists of using the values of variable VID. These are natural subjects since they are expected to theoretically represent the primary source of variability in your dataset. Any other source of variation in your data is undesired and would ideally be removable. Sources of variation that are undesirable include statistical noise due to measurement errors, discrepancies among analytical models or software implementation. A subtle issue to consider is whether or not a heart being in rest versus stress modes represents a natural source of variation. If it is expected that rest and stress measurements should be different then a combination of VID and R/S should define the subject and all rows of data must be used in the analysis. Otherwise, the R/S variable should be considered as a factor to be investigated.

Regarding the definition of raters, your options must be carefully examined. Here are a few possibilities to consider:

- You may decide that it is the extent of agreement between the two models 1 and 2 that must be investigated. In this case, Models 1 and 2 are the 2 raters that will drive your inter-rater reliability analysis along with the patient factor defined by the VMOD variable. The software and R/S effects will be confounding[3] and not explicitly part of the statistical model. One disadvantage of this framework is due to the confounding variables VMOD and R/S status increasing the variance of the intraclass correlation.

- Another possibility of interest would be to investigate the extent of agreement between the two software packages. The software and the patient will then be the 2 main factors used to explain variation in the MBF measurements. The model and R/S will be confounding factors that will impact MBF measurements through the software and the patient.

[3] Confounding factors or variables are variables with a hidden effect on the MBF measurements.

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

Table 2.6: Myocardial Blood Flow measurements

VSOFT	VID	VMOD	R/S	GLOBAL	LAD	LCx	RCA
1	1	2	R	4.366	4.057	4.504	5.275
1	1	2	S	2.491	2.238	2.647	3.044
1	2	2	R	2.073	2.162	1.881	2.062
1	2	2	S	2.588	2.996	2.104	2.313
1	3	2	R	2.224	2.154	2.029	2.810
1	3	2	S	1.423	1.466	1.383	1.397
1	4	2	R	1.840	1.613	2.083	2.287
1	4	2	S	2.325	1.959	2.813	2.894
1	5	2	R	2.223	2.063	2.591	2.862
1	5	2	S	2.786	2.484	3.548	3.233
3	1	2	R	4.840	4.450	4.530	5.670
3	1	4	R	5.360	6.100	6.780	2.920
3	1	2	S	2.674	2.418	2.559	3.115
3	1	4	S	2.721	2.891	3.626	1.553
3	2	2	R	2.070	2.230	1.930	1.970
3	2	4	R	2.530	3.420	2.410	1.440
3	2	2	S	2.202	2.506	2.010	1.970
3	2	4	S	2.144	2.571	1.617	2.149
3	3	2	R	2.590	2.520	2.190	3.050
3	3	4	R	1.650	1.960	1.700	1.160
3	3	2	S	1.962	2.191	1.698	1.943
3	3	4	S	1.250	1.704	1.318	0.739
3	4	2	R	2.700	2.620	2.770	2.720
3	4	4	R	1.900	1.960	2.170	1.570
3	4	2	S	2.647	2.339	2.916	2.833
3	4	4	S	1.979	1.782	2.009	2.415
3	5	2	R	2.760	2.190	3.050	3.280
3	5	4	R	1.999	1.871	2.165	2.024
3	5	2	S	3.247	2.433	3.910	3.814
3	5	4	S	2.186	1.713	2.198	3.368

- You may also consider the software package and the specific model it implements as being one rater. In this case, Table 2.6 suggests the presence of 3 raters. These are 1-1, 3-1 and 3-2 where s-m refers to software s and model m. Inter-rater reliability can then be studied separately for “Rest” and “Stress”

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

measurements, or includes all measurements and consider `R/S` to be a confounding factor.

2.3.2 *Self-Rating of Neuroticism by Family Members*

A scenario in which you may encounter unusual raters is one where subjects are rated by different raters closely linked to the subjects being rated. As an example, consider Table 2.7, which shows rating data from members of 7 families who evaluated their own neuroticism. The family is the subject and its members represent the raters. The goal here is to see if members of the same family agree among themselves on what would be perceived as the family neuroticism. What is peculiar about this example is the strong subject-rater relationship. The raters differ from subject to subject and are even an integral part of the subjects they are rating.

Looking at Table 2.7, there is nothing a priori that prevents us from considering FM1, FM2, FM3 and FM4 as 4 fixed raters that rated 7 independent subjects. For the purpose of computing inter-rater reliability, FM1, FM2, FM3 and FM4 have to be seen as 4 virtual raters, some of whom may have rated fewer subjects. When interpreting the magnitude of the inter-rater reliability coefficient, you will have to stray away from the notion of virtual rater and see a high coefficient as a sign of agreement among family members around a common perception of neuroticism. The missing ratings in Table 2.7 that are due to some families being larger than others, do not pose any particular problem as one can see their occurrence as a random phenomenon associated with the 4 virtual raters.

Table 2.7: Self-rating of neuroticism by members in 7 families

FamID	FM1	FM2	FM3	FM4
1	0.79	0.51	0.60	
2	1.09	1.30		
3	1.26	1.43	0.40	0.53
4	0.49	0.64		
5	0.98	0.68	0.53	
6	1.34	0.45		
7	1.25	1.47	2.19	0.85

2.4 Concluding Remarks

It would never have been possible to cover in a single chapter, all possible challenges practitioners may in encounter in practice when organizing their ratings in datasets. However, I did attempt to cover the most common situations I myself

Get the entire ebook for $19.95 using the link: https://sites.fastspring.com/agreestat/instant/icc5ed978_1_7923_5464_9e

have been exposed to. Organizing your rating data properly gives you the first clear view of what data your inter-rater reliability experiment has produced. It is the preliminary assessment of your ratings, which gives you some insight into the way the raters interacted with the subjects they have rated.

Unfortunately, there is no one single general method for organizing rating data that can systematically be applied to all inter-rater reliability experiments. Therefore, practitioners and researchers must always take time to carefully look at the type of data that was collected and to determine a meaningful way to organize it. In this chapter I have discussed various format types, their advantages and disadvantages. These discussions should guide you in your search for an effective data structure for your ratings.

I have essentially discussed 2 general approaches for organizing your rating data. These are the wide and long data formats (`WDF` and `LDF`). The main advantage of these 2 data formats lies in their rectangular layout. All collected ratings are laid out in a way that shows what raters produced them and which subjects they are assigned to. As a general principle, you would use the `WDF` format when the raters rate subjects on a single factor. If the subjects are rated on 2 factors or more, then the `LDF` format is recommended.

I have also discussed situations where the notions of subject and rater are fuzzy. Researchers are advised not jump into analyzing their data prior to clarifying these concepts. Is the rater defined by a single variable? Should 2 or more variables be used to define a rater? Which variable identifies the subject? How many ratings did each rater assign to each subject? You should be able to answer all of these questions before considering the analysis. No software package can do this preliminary work for you.