CHAPTER 2

Setting Up a Database of Ratings for Analysis

OBJECTIVE

After collecting rating data from subjects, that data 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 when setting up the database, other studies however can be quite challenging to the point where even deciding what the raters and the subjects are and how to prepare the ratings for analysis require substantial effort. Before agreement coefficients can be calculated or statistical models can be built, your rating data must first be adequately organized and the different variables and factors be well defined. You will also see in this chapter that it may sometimes be necessary to analyze more than one aspect of agreement to fully evaluate the extent to which raters agree. The primary objective of this chapter is to provide guidelines to researchers for setting up their datasets of ratings before analysis can begin.

Contents

"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

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. Many problems that researchers encounter with the analysis of their rating data, stem from the difficulty to properly organize their data. Sometimes, the difficulty to adequately define the very notions of subjects, raters or even agreement, is part the problem. This chapter did not exist in the first four editions of this book. However, years of practice in the field of inter-rater reliability have convinced me that researchers needed guidelines to properly organize their inter-rater reliability data and to adequately frame their problem before the analysis itself can begin. Therefore, this chapter explores various scenarios that are expected to create some challenges and discusses ways to overcome them.

This book focuses on categorical ratings, which could be of nominal as well as of ordinal types[1]. There are essentially 4 approaches for organizing such data. Each of these approaches has advantages and disadvantages. I strongly suggest that you set up your dataset using a format that is a variant of one of the following 4 options:

- `Contingency Table` (CROSSTAB)

  The contingency table also referred to as the cross tabulation, the crosstab, or the frequency table, presents data in the form of a matrix showing the distribution of subjects by rater and category.

  For example, Table 2.1 summarizes reliability data produced by two clinicians after they examined 102 patients suffering from spinal pain. Both clinicians classified each of the patients into one of 3 categories defined by the syndrome type they present. The 3 syndrome types are "Derangement", "Dysfunctional" and "Postural". This contingency table indicates that the same 10 individuals that clinician 1 diagnosed with a derangement syndrome were diagnosed with a dysfunctional syndrome by clinician 2.

  The contingency table works well when dealing with 2 raters and when the ratings can be seen as categories into which subjects are classified. If the ratings are in the form of quantitative measurements or if the number of raters is large, then the contingency table as a method of organizing your data must

[1] Note that quantitative ratings of ratio or of interval types whose values are predetermined before the inter-rater reliability experiment is conducted are considered to be of ordinal type.

be ignored. I will also show in subsequent chapters that the use of contingency tables will be challenging if you want to properly handle missing ratings.

Table 2.1: Ratings of Spinal Pain by Clinician and Syndrome Type[a]

	Clinician 2			
Clinician 1	Derangement	Dysfunctional	Postural	Total
Derangement	22	10	2	34
Dysfunctional	6	27	11	44
Postural	2	5	17	24
Total	30	42	30	102

[a]Data initially published by Sim and Wright (2005)

- `Distribution of Raters by subject and category (RDIST)`

  RDIST is another method of presenting your rating data in a table, where each row represents a subject and each column represents one of the categories into which subjects are classified. Each table entry is the number of raters who classified the subject represented by the row, into the category represented by the column. Table 2.2 for example, shows how 6 raters are distributed across 30 subjects and 5 categories. Note that each row sums to 6, which represents the number of raters. This will be case if each rater rates all subjects. Otherwise, some row totals will be smaller than 6.

  Although the RDIST format can accommodate multiple raters, it will only work well when the ratings are in the form of categories[2]. Moreover, the RDIST format conceals the ratings associated with individual raters, making it impossible to use certain agreement coefficients such as one proposed by Conger (1980) to be discussed in subsequent chapters.

[2]It is because the number of categories is generally fixed and applies to all subjects.

Table 2.2: Diagnoses on 30 Subjects by 6 Raters per Subject[a]

	Category				
Subject	Depression ($j = 1$)	Personality Disorder ($j = 2$)	Schizophrenia ($j = 3$)	Neurosis ($j = 4$)	Other ($j = 5$)
1	0	0	0	6	0
2	0	3	0	0	3
3	0	1	4	0	1
4	0	0	0	0	6
⋮	⋮	⋮	⋮	⋮	⋮
28	0	2	0	4	0
29	1	0	5	0	0
30	0	0	0	0	6

[a]This is an extract of a dataset initially published by Fleiss (1971)

- `Wide Data Format (WDF)`

The WDF is a way to organize your data in a table format where each row represents a subject, each column a rater and each data point represents the rating the rater assigned to the subject. As shown in Table 2.3, this format is a listing of all ratings organized by subject and raters. Its main advantage 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.

Table 2.3: Classification of 7 subjects by 4 raters into 5 categories.

Units	Rater1	Rater2	Rater3	Rater4
1	a	a		a
2	b	b	c	b
3	c	c	c	c
4	c	c	c	c
5	b	b	b	b
6	a	b	c	d
7	d	d	d	d

- `Long Data Format (LDF)`

  LDF is a format best described with an example. Table 2.4 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. However, if you are dealing with a single factor, I would recommend using the WDF format.

Table 2.4: 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

So far, I presented 4 different ways of organizing rating data, all of which are 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. However, in the real world, inter-rater reliability problems can get quite complex and knowing how to organize your data and how to analyze it can become involved. 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.

2.2 Dealing with the Notions of Subject and Characteristic

As surprising as it may appear, the notions of subject and rater are sometimes fuzzy. Not knowing what defines the subject or what defines the rater makes it impossible to organize your data effectively. Therefore, it is essential to identify the variable or variables that will be used for identifying a subject and a rater before creating your dataset. The subject 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. The 3 versions of the test are labeled as M, VCP and VCE[3] and each has a number of components as shown in Table 2.5. 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 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. Once these ratings are collected, the question becomes "how should they be organized?"

Table 2.5: Components of a linguistic test

Analytic Component	Test Version		
	M	**VCP**	**VCE**
Grammar	Yes	Yes	Yes
Vocabulary	Yes	Yes	Yes
Fluency	Yes	Yes	Yes
Pronunciation	Yes	Yes	Yes
Listening	No	Yes	Yes

An effective way to tackle such a problem is to carefully define the different variables of interest you are dealing with. Each variable will represent a column of data containing the different values a variable can take. Some of these values might need to be repeated to show the relationship among variables. I propose the following two possible lists of variables for this problem before discussing their advantages and disadvantages. The first option defines 7 variables while the second defines 9 variables.

[3]The actual meaning of these acronyms is intentionally omitted as it does not add value to the general understanding of this example.

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 2.7 for option 1 and to Table 2.6 for option 2. The `STUDENT` variable can take the 11 values `Amber, David, Isaac, Jasmine, Lee, Mary, Ricardo, Suzan, Viktor, Yanick, Yin`, while the `VERSION` variable can take the 3 values `M, VCE, VCP`. In Table 2.7 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 2.6 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 similar scoring rubric then any combination of `STUDENT`, `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 is recommended when you see your inter-rater reliability experiment as one that focuses on the rating of a single factor for all subjects. In this particular case, all analytic components whose scoring is assumed to be based on the exact same scoring rubric, can be seen as a single factor that I would name `PROFICIENCY`. If the scoring rubric used differs from one analytic component to another one, then the

ratings cannot be associated with a single factor.

Depending on how the different analytic components are being rated, the rating 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 each of the 5 components `FLUENCY`, `GRAMMAR`, `LISTEN`, `PRONU` and `VOCAB` to represent 5 characteristics involved in this inter-rater reliability study 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. The columns associated with the 5 factors must be analyzed separately, since they represent different factors. However, the `OVERALL` column can be defined if possible in order to conduct a global analysis of proficiency.

With the long-format dataset, each combination of `NAME` and `VERSION` represents a subject. The univariate 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 2.7 without the `COMPONENT` column and with the numbers representing 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, you can still extract the information related to a specific factor, reorganize it in a wide-format before analyzing it.

You may note that the wide-format dataset represented by Table 2.7 has more records than the long-format dataset represented by Table 2.6, which in turn has more variables than the wide-format dataset. In general, when raters' names are listed horizontally, then it is a wide format. But when they are all listed in a single column then it is a long format.

Table 2.6: 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.4 that can be found in Appendix A

Table 2.7: 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.3 that can be found in Appendix A

2.3 Dealing with the Notion of Rater

While the notion of subject can be ill-defined as discussed in section 2.2, the same could happen in some applications with the notion of rater. Therefore, you may need to formally define what a rater is before your database can be properly organized. Two scenarios where the notion of rater is nontrivial are discussed in this section. The first scenario is about the notion of rater in an intra-rater reliability study where multiple ratings are often assigned to one subject. The second scenario explores the notion of rater when there is a close link between the subject and the rater.

2.3.1 *Intra-rater Reliability*

In this section, I want to describe experiments with an unusual type of raters. While inter-rater reliability quantifies the extent of agreement among raters, intra-rater reliability quantifies the reproducibility of ratings by the same raters. Intra-rater reliability refers to agreement of raters with themselves on different occasions. Consider a situation where one rater must rate 5 subjects on two occasions. The outcome of such a study would be described as in Table (a) of Figure 2.1. One convenient way to analyze Table (a) data, is to consider each of the 2 occasions as a "virtual" rater and to display the rating data as shown in Table (b) of Figure 2.1, before computing an ordinary inter-rater reliability coefficient between the 2 virtual raters Vrater1 and Vrater2. If the number of ratings per subject is 3, then 3 virtual raters will be necessary to compute the intra-rater reliability coefficient.

Table (a)

Subject	Rater
1	a
1	a
2	b
2	c
3	c
3	c
4	a
4	a
5	b
5	b

Table (b)

Subject	Vrater1	Vrater2
1	a	a
2	b	c
3	c	c
4	a	a
5	b	b

Figure 2.1: Ratings of 5 subjects on 2 occasions by the same rater in an intra-rater reliability study

Table (a)

Subject	Rater1	Rater2	Rater3
1	a	a	a
1	a	b	a
2	b	b	b
2	c	c	c
3	c	c	b
3	c	c	c
4	a	a	a
4	a	a	a
5	b	a	b
5	b	b	b

Table (b)

Vsubject	Vrater1	Vrater2
1	a	a
2	a	b
3	a	a
4	b	c
5	b	c
6	b	c
7	c	c
8	c	c
9	b	c
10	a	a
11	a	a
12	a	a
13	b	b
14	a	b
15	b	b

Figure 2.2: Ratings of 5 subjects on 2 occasions by 3 raters in an intra-rater reliability study.

The intra-rater reliability problem discussed in the previous paragraph involves a single rater and 2 ratings per subject. Consider now, an intra-rater reliability study with 3 raters who must rate each of 5 subjects on 2 occasions. This data would naturally be organized in a wide format as shown in Table (a) of Figure 2.2 for the purpose of evaluating inter-rater reliability. For evaluating intra-rater reliability

however, the more effective way to set up your dataset is to create 2 "virtual raters" (since the number of ratings per subject is 2)[4] and 15 "virtual subjects" as shown in Table (b) of Figure 2.2. Each "actual subject" is rated on 2 occasions by 3 "actual raters." For the purpose of calculating intra-rater reliability, each occasion is seen as a "virtual rater" who rates a total of 15 "virtual subjects".

2.3.2 *Rating of subjects by different groups of raters*

In section 2.3.1, I discussed about intra-rater reliability experiments where you needed to create fictitious raters so as to quantify the reproducibility of ratings. In this section, I want to present another scenario involving unusual raters. It is an experiment where subjects are rated by different raters who are closely linked to the subjects they are rating. Consider Table 2.8, which shows rating data from members of 7 families who evaluated their own neuroticism[5]. The first column of this table contains a family identifier `FamID`, whereas the remaining columns contain family ratings assigned by their own members `FM1`, `FM2`, `FM3` and `FM4`.

The family is the subject and its members are the only raters to rate it. 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 2.8, 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 2.8 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.

Before closing this section, one needs to stress out that the number of subjects and number of raters participating in an inter-rater reliability must be determined at the design stage, and not be dictated by the outcome of the experiment. If these numbers are determined after the experiment had taken place, they become random variables that will increase the magnitude of the agreement coefficient variance. Therefore, in

[4] If there were 3 replicates then you would need 3 virtual raters. As a general rule, the number of virtual raters is determined by the number of ratings per subject and per rater.

[5] Neuroticism is known in psychology as a personality trait, which is typically defined as a tendency toward anxiety, depression, self-doubt, and other negative feelings.

the study of family neuroticism discussed in this section, a proper design will set the number of family members allowed to be part of the study. The actual participating family members could be selected randomly if necessary.

Table 2.8: 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 Dealing with the Notion of Agreement in a Multiple-Level Process

In the field of Natural Language Processing (NLP), annotation is a critical and often complex endeavour. Given the importance of having a reliable annotation scheme, Inter-Annotator Agreement (IAA) is often calculated and is expected to be high. What is peculiar about the field of NLP however, is that even in a simple annotation project, evaluating the IAA is almost going to be anything but a regular inter-rater reliability experiment. Calculating the IAA often requires an in-depth analysis of a multistage process leading to an annotation. Even the very notion of agreement so trivial in most inter-rater reliability studies must be carefully defined as agreement and disagreement may occur at any stage in the complex annotation process. When can we then claim that two annotators are in agreement about the annotation of a particular linguistic artifact?

To fix ideas, let us consider a simple example from the medical field. Two annotators are to annotate clinical documents typed by a physician during patient consultations. The annotators are looking for specific segments of text that can be seen as describing a medical condition, so that the identified text segments can be marked with the condition name and a tag indicating whether the condition is present or absent. For example, the physician may write in a note that "The patient had a mild sore throat without fever." In this case, the text segment *"mild sore throat"* would be annotated as *"Sore Throat - Present"*, while the text segment *"without fever"* would be annotated as *"Fever-Absent"*. I am going to review several of the issues that must be resolved here before an IAA can be adequately quantified:

- Since annotation takes place at the text segment level, should agreement between annotators be evaluated at that level? In other words, should you con-

sider the text segment to be the subject that must be rated by all annotators? Although most researchers in the NLP field have a tendency to consider text segments as subjects (see Savkov et al., 2016), this approach presents many challenges that could be avoided. The subject to be rated must be uniquely defined and identical for all annotators. If annotators already disagree about the text segments that must be annotated, then using these same text segments as basis for evaluating agreement will inevitably lead to more challenges. For the example presented in this section, using the clinical note as basis for evaluating is more appropriate. This issue is further discussed in section 9.2 of chapter 9.

- Figure 2.3 shows an extract of a clinical note, which was annotated by 2 annotators. Each annotator needed to perform the following 3 tasks:

  1) A specific text segment must be identified.
  2) A clinical finding (e.g. "Sore throat") must be associated with the identified text segment.
  3) The clinical finding further be marked as present or absent.

  Note that, the first annotator could mark 3 text segments while the second only marks 2 with a possible overlap between some of the text segments.

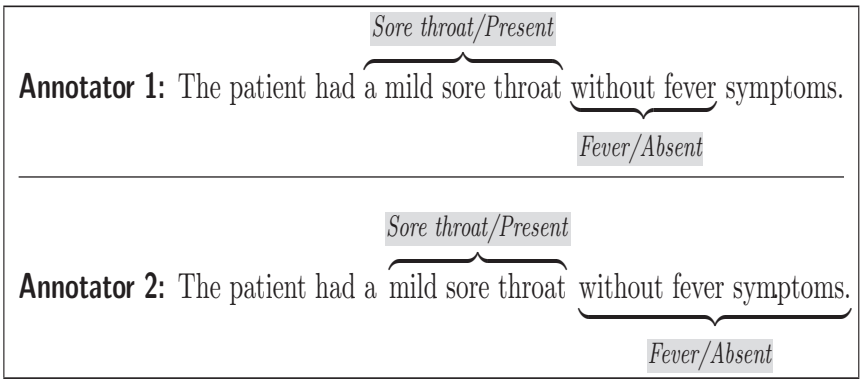

Figure 2.3: Annotation of a sentence by 2 annotators 1 & 2.

In Figure 2.3, both annotators highlighted 2 text segments. However, the 2 highlighted segments are not the same, although the clinical finding and their presence indicators are the same. Do these annotators agree? About what? To what extent? Here is a situation where the often trivial notion of agreement becomes complex.

Many previous attempts to apply traditional inter-rater reliability coefficients to the field of NLP resulted in a failure. This failure is often explained by the large number of text segments that are not annotated for not carrying any information of interest and which make the magnitude of various proportions negligible. Consequently, some alternative more traditional information retrieval metrics have been suggested by Hripcsak and Rothschild (2005)

Researchers in the field of NLP often consider the text segment as the fundamental unit of analysis. This choice is highly questionable. As a matter of fact, it is up to each annotator to identify the specific text segments that will be marked and this disqualifies the text segment as an ideal unit of analysis. The choice of segments is an integral part of the annotation process. A different approach would be to consider the clinical note as the basic unit of analysis. That is, all annotated text segments related to a particular clinical note will be used to determine the extent to which two annotators agree at the clinical note level[6]. The smallest entity for which agreement is evaluated must be defined before the start of the annotation work and cannot be a product this work. Further technical details on this issue are presented in section 9.2 of chapter 9.

Figure 2.3 depicts a typical example of inter-rater reliability study where the absence of a clear-cut definition of the notions of subject and agreement could make it difficult to organize rating data and to analyze it. If the critical phase of concepts definition is overlooked or neglected, then the entire inter-rater reliability experiment can become questionable.

2.5 Dealing with Multiple Ratings per Rater and per Subject

A typical inter-rater reliability study requires each rater to assign a single rating to the subject being rated. However, this will not always be the case. In the field of medical coding for example, mapping two coding systems is often necessary for various reasons. For example, the coding system used for billing may not meet the health facility clinical needs and may be inadequate for medical research. To conduct their activities, medical researchers often need to translate medical codes created by billers into a new nomenclature system more appropriate for research. Consequently, being able to map one coding system into another one in a reliable way is essential to ensure the integrity of medical research.

Stein et al. (2005) describe an experiment aimed at testing the reliability of mapping between the Patient Condition (PC) coding system used by the US Department of Defense and the International Classification of Disease, 9^{th} revision (ICD-9-M).

[6] If the clinical document is very long with a large number of annotated text segments, one may consider using sections of the document as unit of analysis.

For each PC code, medical coders must associate all ICD-9 codes deemed to be related to the PC code. The PC code can be seen as a subject that must be classified into one or many categories defined by the ICD-9 codes. Here lies the peculiar nature of this reliability experiment, characterized by the possibility to assign many categories to a single subject. Such rating data would be best organized in the wide format as shown in Table 2.9.

Table 2.9: Mapping between PC and ICD-9-M Coding Systems[a]

PC	Coder A	Coder B	Coder C
0001	x	800	x
0001	x	801	x
0001	x	802	802
0001	803	803	803
0001	805	804	x
0001	850	850	850

[a]This table is an extract (slightly modified) of Table 1 of Stein et al. (2005).

It follows from Table 2.9 that each PC code must often be repeated in many rows to match the largest number of different ICD-9 codes assigned to it by one coder. Coder B in this case, assigned the maximum of 6 ICD-9 codes to the PC code 0001. This explains why the PC code 0001 was repeated in 6 different rows. All ICD-9 codes used by 2 coders or more must be reported in the same row and ICD-9 codes used by a single coder (e.g. ICD-9 code 805 is used by coder A only) can be reported in any of the 6 rows. Stein et al. (2005) has suggested a data structure, where only identical ICD-9 codes are reported in the same row. This approach will likely overstate the extent of agreement among raters, since disagreement among coders never appears in the same row[7].

Regardless of how Table 2.9 data is organized, there will still be the problem of not having a fixed number of subjects if each row is to be treated as a subject. Moreover, the number of ICD-9 codes considered in the experiment is known only after the experiment is completed. These 2 problems make it difficult to evaluate the precision of agreement coefficients. I would recommend the following approach for addressing these problems:

1) Since the mapping is done from PC codes to ICD-9 codes, it is indicated to

[7]You should note that all rows of data with a single ICD-9 code reported, are always excluded from the calculation of the percent agreement.

consider the initial list of PC codes as your subject sample. Each subject is then assigned one or multiple ICD-9 codes.

2) To properly design such an inter-rater reliability experiment, it is desirable to have for each PC code i, a predetermined list of q_i acceptable ICD-9 codes. Any ICD-9 code not on the list should be considered invalid. Otherwise, the number of categories would be sample-dependent and expected to increase the agreement coefficient variance.

3) In this context, the number of raters r_{ik} who classified subject i into category k must be 0 for all ICD-9 codes k not on the short list of q_i codes associated with PCD code i. For those ICD-9 codes among q_i, r_{ik} would be calculated as usual.

2.6 Concluding Remarks

The primary objective of this chapter was to review various types of rating data encountered in practice and the best way to organize them in databases before analysis. This chapter starts with a general review of the cross tabulation, the distribution of raters by subject and category, the wide and long format for raw ratings. For categorical ratings, these are the 4 most widely-used ways of organizing rating data. Although all 4 options have been used in practice, most software packages cannot handle rating data in the form of a distribution of raters by subject and category.

In some applications, organizing your rating data in a logical way is made difficult when subjects and raters are not readily identifiable. Only a careful examination of the analytic goals can help decide how subjects and raters should be defined. The difficulty of knowing what a subject is was illustrated with a linguistic test containing several analytic components and administered to a group of students. I also showed with an intra-rater reliability example how analytic goals can help to properly define what should be considered as a rater.

Inter-annotator agreement often calculated in the field of Natural Language Processing (NLP) requires that the notion of agreement be properly defined. As a matter of fact, defining the notion of agreement is the biggest challenge one has to face when computing the inter-annotator agreement. This issue is further discussed in section 9.2 of chapter 9. Finally, I discussed the case of multiple ratings assigned to one subject. A priori this should not be problem except when each subject can only be assigned ratings from a specific set of labels, and the number of labels or categories are unknown at the time of study design.