HANDBOOK

OF

INTER-RATER RELIABILITY

FOURTH EDITION

HANDBOOK OF INTER-RATER RELIABILITY

Fourth Edition

The Definitive Guide to Measuring the Extent of Agreement Among Raters

Kilem Li Gwet, Ph.D.

Advanced Analytics, LLC
P.O. Box 2696
Gaithersburg, MD 20886-2696
USA

Published by Advanced Analytics, LLC; in the United States of America.

Advanced Analytics, LLC
PO BOX 2696,
Gaithersburg, MD 20886-2696
e-mail : gwet@agreestat.com

Publisher's Cataloguing in Publication Data:

Gwet, Kilem Li
Handbook of Inter-Rater Reliability
The Definitive Guide to Measuring the Extent of Agreement Among Raters/ By Kilem Li Gwet - 4th ed.
p. cm.
Includes bibliographical references and index.
1. Biostatistics
2. Statistical Methods
3. Statistics - Study - Learning. I. Title.
ISBN 978-0-9708062-8-4

Preface to the Fourth Edition

The ratings assigned to the very same subjects may differ widely from one rater to another. Many researchers across various fields of research have acknowledged this issue for a long time. The extent of these discrepancies is unacceptable primarily because the analysis of ratings often ignores the presence of the rater-induced source of variation, and assumes most variation in the data to be due to a change in the subject's attribute being investigated. This variation between raters may jeopardize the integrity of scientific inquiries, and could potentially have a dramatic impact on subjects. In fact, a wrong drug or a wrong dosage of the correct drug may be administered to patients at a hospital due to a poor diagnosis, or an exam candidate may see her professional career or even her life plans take a negative turn due to inexplicably large discrepancies in the raters' scoring of her exam. It is the need to resolve these problems that led to the study of inter-rater reliability.

The focus of the previous edition (i.e. third edition) of this *Handbook of Inter-Rater Reliability* is on the presentation of various techniques for analyzing inter-rater reliability data. These techniques include chance-corrected measures, intraclass correlations, and a few others. However, inter-rater reliability studies must be optimally designed before rating data can be collected. Many researchers are often frustrated by the lack of well-documented procedures for calculating the optimal number of subjects and raters that will participate in the inter-rater reliability study. The fourth edition of the *Handbook of Inter-Rater Reliability* will fill this gap. In addition to further refining the presentation of various analysis techniques covered in previous editions, a number of chapters have been expanded considerably in order to address specific issues such as knowing how many subjects and raters to retain for the study, or finding the correct protocol for selecting these subjects and raters to ensure an adequate representation of their respective groups. In particular, the coverage of intra-rater reliability has been expanded and substantially improved.

Unlike the previous editions, this fourth edition discusses the concept of inter-rater reliability first, before presenting the techniques for computing it. The reader who has previously had limited or no exposure to the issues related to inter-rater reliability, will be able to learn about this concept, how it arises in practice, how it affects rating data, and what can be done about it. After a gentle introduction to the field of inter-rater reliability, I proceed in subsequent chapters with the presentation of simple statistical techniques for quantifying the extent of agreement among raters,

and its precision from observed ratings. The reader will quickly notice that this book is detailed. Yes, I wanted it sufficiently detailed so that practitioners can gain considerably more insight into the different topics than would be possible in a book aimed simply at promoting concepts. I wanted a researcher to read this book and be able to implement the proposed solutions without having to figure out hidden steps or expressions not fully fleshed out.

This book is not exhaustive. It does not cover all topics of interest related to the field of inter-rater reliability. I selected topics, which are among the most commonly referenced by researchers in various fields of research. I have also come to the realization that in general, evaluating inter-rater reliability is only one specific task among many others during the conduct of an investigation. Consequently, the time one is willing to allocate to this task may not be sufficient to implement very elaborate techniques that require substantial experience in the use of statistical methods. Therefore, I have made considerable effort to confine myself to techniques that a large number of researchers will feel comfortable implementing. It is also partly because of this that I decided to exclude from the presentation all approaches based on the use of sophisticated theoretical statistical models (e.g. Rasch models, logistic regression models, ...) that generally require considerable time and statistical expertise to be successfully implemented.

I have accumulated considerable experience in the design and analysis of inter-rater reliability studies over the past 15 years, through teaching, writing and consulting. My goal has always been, and remains to gather in one place, detailed, well-organized, and readable materials on inter-rater reliability that are accessible to researchers and students in all fields of research. I expect readers with no background in statistics to be able to read this book. However, the need to provide a detailed account of the techniques has sometimes led me to present a mathematical formulation of certain concepts and approaches. In order to offer further assistance to readers who may not read some equations adequately, I present detailed examples, and provide downloadable Excel spreadsheets that show all the steps for calculating various agreement coefficients, along with their precision measures. Users of the R statistical package will find in appendix B several R functions implementing the techniques discussed in this book, and which can be downloaded for free. I expect the *Handbook of Inter-Rater Reliability* to be an essential reference on inter-rater reliability assessment to all researchers, students, and practitioners in all fields. If you have comments do not hesitate to contact the author.

Kilem Li Gwet, Ph.D.

Preface to the Third Edition

Here, I like to explain why I decided to write the third edition of this book. There are essentially 2 reasons for which I decided to write this third edition:

- The second edition covers various chance-corrected inter-rater reliability coefficients including Cohen's Kappa, Fleiss' Kappa, Brennan-Prediger coefficient, Gwet's AC_1 and many others. However, the treatment of these coefficients is limited to the situation where there is no missing ratings. That is, each rater is assumed to have scored all subjects that participated in the inter-rater reliability experiment. This situation rarely occurs in practice. In fact, most inter-rater reliability studies generate a sizeable number of missing ratings. For various reasons some raters may be unable to rate all subjects, or some reported ratings must be rejected due to coding errors. Therefore, it became necessary to revise the presentation of the various agreement coefficients in order to provide practitioners with clear guidelines regarding the handling of missing ratings during the analysis of inter-rater reliability data.
- Although the second edition offers an extensive account of chance-corrected agreement coefficients, it does not cover 2 important classes of measures of agreement. The first class includes all agreement coefficients in the family of Intraclass Correlation Coefficients (or ICC). The second class of agreement measures omitted in the second edition of this book belong to the family of association measures, whose objective is to quantify the extent of agreement among raters with respect to the ranking of subjects. In this second class of coefficients, one could mention for example the Kendall's coefficient of concordance, Kendall's tau, the Spearman's correlation and the likes. Given the importance of these coefficients for many researchers, there was a need to include them in a new edition.

In addition to expanding the coverage of methods, I have added more clarity into the presentation of many techniques already included in the second edition of this book. Those who read the second edition, will likely find the coverage of weighted agreement coefficients much more readable.

By writing this book, my primary goal was to allow researchers and students in all fields of research to access in one place, detailed, well-organized, and readable materials on inter-rater reliability. Although my background is in statistics, I wanted

to ensure that the content of this book is accessible to readers with no background in statistics. Based the feedback I received about earlier editions of this book, this goal appears to have been achieved to a large extent. I expect the *Handbook of Inter-Rater Reliability* to be an essential reference on inter-rater reliability assessment to all researchers, students, and practitioners in all fields.

Kilem Li Gwet, Ph.D.

Table of Contents

ACKNOWLEDGMENTS

First and foremost, this book would never have been written without the full support of my wife Suzy, and our three girls Mata, Lelna, and Addia. They have all graciously put up with my insatiable computer habits and so many long workdays, and busy weekends over the past few years. Neither would this work have been completed without my mother inlaw Mathilde, who has always been there to remind me that it was time to have diner, forcing me at last to interrupt my research and writing activities to have a short but quality family time.

I started conducting research on inter-rater reliability in 2001 while on a consulting assignment with Booz Allen & Hamilton Inc., a major private contractor for the US Federal Government headquartered in Tysons Corner, Virginia. The purpose of my consulting assignment was to provide statistical support in a research study investigating the personality dynamics of information technology (IT) professionals and their relationship with IT teams' performance. One aspect of the project focused on evaluating the extent of agreement among interviewers using the Myers-Briggs Type Indicator Assessment, and the Fundamental Interpersonal Relations Orientation-Behavior tools. These are two survey instruments often used by psychologists to measure people's personality types. I certainly owe a debt of gratitude to the Defense Acquisition University (DAU) for sponsoring the research study, and to the Booz Allen & Hamilton's associates and principals who gave me the opportunity to be part of it.

Finally, I like to thank you the reader for buying this book. Please tell me what you think about it, either by e-mail or by writing a review at Amazon.com.

Thank you,

Kilem Li Gwet, Ph.D.

PART I

PRELIMINARIES

CHAPTER 1

Introduction

OBJECTIVE

The primary objective of this chapter is to provide a broad view of the inter-rater reliability concept, and to highlight its importance in scientific inquiries. The difficulties associated with the quantification of inter-rater reliability, and some key factors affecting its magnitude are discussed as well. This chapter stresses out the importance of a clear statement of the study objectives, and a careful design of inter-rater reliability experiments. Different inter-rater reliability types are presented, and the practical context in which they can be used described. Also discussed in this chapter, are the types of reliability data the researcher may collect, and how they affect the way the notion of agreement is defined. I later insist on the need to analyze inter-rater reliability data according to the principles of statistical inference in order to ensure the findings can be projected beyond the often small samples of subjects and raters that participate in a reliability experiment. Figure 1.7.1 represents a flowchart that summarizes the process for identifying the correct type of agreement coefficient to use based on the type of ratings to be collected. The chapters where the recommended agreement coefficients are treated are also identified.

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)

1.1 What is Inter-Rater Reliability ?

The concept of inter-rater reliability has such a wide range of applications across many fields of research that there is no one single definition that could possibly satisfy the specialists in any field. Nevertheless, introducing the general concept is straightforward. During the conduct of a scientific investigation, classifying subjects or objects into predefined classes or categories is a rather common activity. These categories are often values taken by a nominal or an ordinal characteristic. The reliability of this classification process can be established by asking two individuals referred to as raters, to independently perform this classification with the same set of objects. By accomplishing this task, these two individuals will have just participated in what is called an inter-rater reliability experiment expected to produce two categorizations of the same objects. The extent to which these two categorizations coincide represents what is often referred to as inter-rater reliability. If inter-rater reliability is high then both raters can be used interchangeably without the researcher having to worry about the categorization being affected by a significant rater factor. Interchangeability of raters is what justifies the importance of inter-rater reliability. If interchangeability is guaranteed, then the categories into which subjects are classified can be used with confidence without asking what rater produced them. The concept of inter-rater reliability will appeal to all those who are concerned about their data being affected to a large extent by the raters, and not by the subjects who are supposed to be the main focus of the investigation.

Our discussion of the notion of inter-rater reliability in the previous paragraph remains somehow superficial, and vague. Many terms are lousily defined. Although one can easily get a sense of what inter-rater is and how important it is, articulating a universal definition that is applicable in most situations is still problematic. For example the previous paragraph mentions two raters. But can we define inter-rater reliability without being specific about the number of raters ? If we cannot, then how many raters should be considered for this purpose ? What about the number of subjects being rated ? Consider for example faculty members rating the proficiency of nursing students on five aspects of patient care on the following four-point scale: (*i*) None, (*ii*) Basic, (*iii*) Intermediate, (*iv*) Advanced. Here the raters are humans (faculty members), and 4 categories representing an ordinal scale are used. Here, inter-rater (actually inter-faculty) reliability is the extent to which a nursing student is assigned a proficiency level, which is independent of the specific faculty member

who performed the evaluation. The proficiency level should be an attribute of the nursing student and the particular test that is administered, and not an attribute of any particular faculty member. There is no reference to a particular student, nor to a particular faculty member. We do not worry about quantifying inter-rater reliability at the present. Instead, we want to explore the concept only.

So far our discussion has been limited to human raters and to categories as the measurements being produced by the inter-rater reliability experiment. This will not always be the case. Consider a situation where two medical devices designed by two manufacturers to measure the strength of the human shoulder in kilograms. The researcher wants to know whether both medical devices can be interchangeable. Before getting down to the analysis of shoulder strength data, you want to ensure that they are not contaminated by an important medical device factor. It is because what you are studying is the human shoulder, and not the medical device. What is peculiar about this experiment is that the raters are no longer humans, instead they are medical devices. Moreover, the measurements produced by the experiment are no longer categories. Instead, they are numerical values. This changes the notion of agreement entirely, and raises a whole host of new issues. Two medical devices from two manufacturers are unlikely to yield two identical values when used on the same subject. Therefore, we need to have a different way of looking at the closeness of the ratings. This is generally accomplished by looking at the variation in ratings that is due to raters only. A small variation is an indication of ratings that are very close, while a large variation suggests that the raters may have very different opinions. We are implicitly assuming here that isolating the component of the rating variation that is due to the raters alone is feasible.

There are situations where the rater can be seen as an abstract entity to some extent when defining inter-rater reliability, and other situations where the rater must be a concrete entity. For example when discussing about inter-rater reliability of medical devices, unless we clearly identify what medical devices we are referring to, our discussion will carry little interest. Our inter-rater reliability definition will clearly be limited to those devices, and any concrete statistical measure of reliability will directly refer to them. When we explore the inter-rater reliability among faculty members testing the proficiency of nursing students, then it is clearly in our interest not to exclude from consideration any faculty member who is a potential examiner now or in the future. Likewise, we would want to have our sight over all possible nursing students who may have to be evaluated at some point during their program. The general framework retained at this exploratory stage of the investigation will not just help define inter-rater reliability, it will also help to delineate the domain of validity of the concrete agreement measures that will be formulated in the form of inter-rater reliability coefficients.

In the inter-rater reliability literature, it is rather common to encounter other notions such as that of intra-rater reliability, or test-retest reliability. While inter-rater reliability is concerned about the reproducibility of measurements by different raters, intra-rater reliability on the other hand is concerned about self-reproducibility. It can be seen as a special case of inter-rater reliability. Instead of having several raters rate the same subject as in the case of inter-rater reliability, you would have the same rater rating the same subject on several occasions, also known as trials or replicates. In other words, intra-rater reliability can be seen as inter-trial or inter-replicate reliability. It does not raise any new challenges. Instead it requires an adaptation of existing ideas and approaches initially developed to assess inter-rater reliability. In addition to intra-rater reliability, the inter-rater reliability has several other branches that will be explored at a later time when the context is appropriate.

SOME APPLICATIONS OF INTER-RATER RELIABILITY

There is little doubt that it is in the medical field that inter-rater reliability has enjoyed an exceptionally high popularity. Perhaps this is due to medical errors having direct and possibly lethal consequences on human subjects. We all know stories of patients who have received the wrong medication or the right medication at a wrong dosage because the wrong illness was diagnosed by a medical personnel with insufficient training in the administration of a particular test. Therefore, improving the quality of medical tests was probably far more urgent than improving for example the quality of a video game. Patient care for example in the field of nursing is another highly sensitive area where inter-rater reliability has found a fertile ground. Chart abstractors in a neonatal intensive care unit for example play a pivotal role in the care given to newborn babies who present a potentially serious medical problem. Ensuring that the charts are abstracted in a consistent manner is essential for the reliability of diagnoses and other quality care indicators.

The field of psychometrics, which is concerned with the measurement of knowledge, abilities, attitudes, personality traits, and educational attainment, has also seen a widespread use of inter-rater reliability techniques. The use of inter-rater reliability is justified here by the constant need to validate various measurement instruments such as questionnaires, tests, and personality assessments. A popular personality test is the Myers-Briggs Type Indicator (MBTI) assessment, which is often used to categorize individuals according to their personality type (e.g. Introversion, Extraversion, Intuition, Sensing, Perception, ...). These classifications often help managers match job applicants to different job types, and build project teams. Being able to test the reliability of such a test is essential for their effective use. When used by different examiners, a reliable psychometric test is expected to produce the same categorization of the same human subjects. Eckes (2011) discusses eloquently

the inter-rater reliability issues pertaining to the area of performance assessment.

Content analysis is another research field where inter-rater reliability has found numerous applications. One of the pioneering works on inter-rater reliability by Scott (1955) was published in this field. Experts in content analysis often use the terminology "inter-coder reliability." It is because raters in this field must evaluate the characteristics of a message or an artifact and assign to it a code that determines its membership in a particular category. In many applications, human coders use a codebook to guide the systematic examination of the message content. For example, health information specialists must often read general information regarding a patient's condition, the treatment received before assigning one of the numerous International Classification of Disease codes so that the medical treatment administered by doctors can be processed for payment. A poor intercoder reliability in this context would result in payment errors and possibly large financial losses. More information regarding the application of inter-reliability in content analysis can be found in Krippendorff (2012), or Zhao, Liu, and Deng (2013).

In the fields of linguistic analysis, computational linguistics, or text analytics, annotation is a big thing. Linguistic annotations can be used by subsequent applications such as a text-to-speech application with a speech synthesizer. There could be human annotators, or different annotation tools. Experts in this field are often concerned about different annotators or annotation techniques not being in agreement. This justifies the need to evaluate inter-rater reliability, generally referred to in this field of study as inter-annotator reliability. Carletta (1996) discusses some of the issues that are specific to the application of inter-rater reliability in computational linguistics. Even in the area of software testing or software process assessment, there have been some successful applications of inter-rater reliability. Software assessment is a complex activity where several process attributes are evaluated with respect to the capability levels that are reached. Inter-rater reliability, also known in this field as inter-assessor reliability is essential to ensure the integrity of the testing procedures. Jung (2003) summarizes the efforts that have been made in this area.

Numerous researchers have also used the concept of inter-rater reliability in the field of medical coding, involving the use of one or multiple systems of classification of diseases. The terminology used most often by practitioners in this field is intercoder reliability. Medical coding is a specialty in the medical field, which has specific challenges posed by inter-rater reliability assessment. The need to evaluate intercoder agreement generally occurs in one of the following two situations:

- Different coders evaluate the medical records of patients and assign one or multiple codes from a disease classification system. Unlike the typical inter-rater reliability experiment where a rater assigns each subject to one and only one category, here coders can assign a patient do multiple disease categories.

For example, Leone et al. (2006) investigated the extent to which neurologists agree when assigning ICD-9-CM[1] codes to patients who have suffered from stroke. The challenge here is to define the notion of agreement in this situation where one coder assigns 3 codes to a patient, while a second coder assigns a single code to the same patient.

Several approaches are possible depending on the study objective. One approach is to define agreement with respect to the primary diagnostic code only. They have to be identical for the coders to be in agreement. A second approach is to create groups of codes and to consider that two coders have agreed if they respective primary diagnosis codes (possibly different) fall into the same group of codes. Alternatively one may use both primary and secondary diagnosis codes provided group membership can be well defined.

- The concept of inter-rater reliability has also been successfully used in the field of medical coding to evaluate the reliability of mapping between two coding systems. Mapping between two coding systems is an essential activity for various reasons. For example behavioral health practitioners consider the Diagnostic and Statistical Manual (DSM) of Mental Disorders to be their nomenclature. However, the US federal government pays claims from the beneficiaries of public health plans using codes from the International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM). Likewise, the Systematic Nomenclature of Medicine-Clinical Terms (SNOMED CT) was developed to be used in Electronic Health Records (EHR) for data entry and retrieval and is optimized for clinical decision support and data analysis.

  In the context of inter-rater reliability, multiple coders may be asked to independently do the mapping between two systems so that the reliability of the mapping process can be evaluated. All raters take each code from one system, and map it to one or several codes from the second system. This data is generally analyzed as follows:

  ⇒ Suppose that a SNOMED code such as **238916002** is mapped to a single ICD-9-CM **60789** by coder 1, and to the three ICD-9-CM codes **60789**, **37454**, and **7041** by coder 2. The analysis of this data is made easier if the coders assign multiple codes by order of priority. One may consider one of the following two options for organizing this data:

  OPTION 1

  In option 1, all ICD-9-CM codes from each rater are displayed vertically following the priority order given to them. Each row of Table 1.1 is treated

[1] ICD-9-CM: International Classification of Diseases 9th Revision - Clinical Modification

as a separate subject that was coded independently from the others. The bullet point indicates that coder 1 did not code subjects 2 and 3.

Table 1.1: Option 1

Subject	SNOMED	Coder 1	Coder 2
1	238916002	60789	60789
2	238916002	•	37454
3	238916002	•	7041

This option more or less ignores subjects 2 and 3 in the calculation of agreement. It has nevertheless been used by some authors (c.f. Stein et al. - 2005).

OPTION 2

A better approach may be option 2 of Table 1.2, where the bullet points are replaced by the Coder 1's code with the lowest priority level. Now there is a "Weight" column that determines what weight (between 0 and 1) will be assigned to the disagreement. The use of weights in inter-rater reliability is discussed more thoroughly in the next few chapters.

Table 1.2: Option 2

Subject	SNOMED	Coder 1	Coder 2	Weight
1	238916002	60789	60789	1
2	238916002	60789	37454	0.75
3	238916002	60789	7041	0

⇒ Once an option for organizing rating data is retained, then one may use one of the many standard computation methods that will be discussed in the next few chapters.

THE STUDY OF INTER-RATER RELIABILITY

When defining the notion of inter-rater reliability, there will always be a degree of impreciseness in what we really mean by it. This issue is acknowledged by Eckes (2011, p. 24) when he says "... even if high interrater reliability has been achieved in a given assessment context exactly what such a finding stands for may be far from clear. One reason for this is that there is no commonly accepted definition of inter-rater reliability." Even the notion of agreement can sometimes be fuzzy. For example

when categories represent an ordinal scale such as "none", "basic", "intermediate", "Advanced", and "Expert," it is not difficult to see to see that although "Advanced", and "Expert" represent a disagreement, these two categories are often justifiably seen as a "partial agreement", especially when compared to two categories such as "none" and "expert." Nevertheless, there is no doubt that the concept of inter-rater reliability is of great importance in all fields of research. Therefore, it is justified for us to turn to the question of which methods are best for studying it. Many ill-defined scientific concepts have been thoroughly investigated in the history of science, primarily because their existence and importance raise no doubt. For example the notion of probability has never been thoroughly defined as indicated by Kolmogorov (1999). However, very few statistical concepts have been applied more widely than this one.

1.2 Defining Experimental Parameters

Many articles about inter-rater reliability are characterized by a description of an experiment that produces ratings, and the method for analyzing those ratings. But these articles devote little space to discuss about the strength of the evidence presented, and its validity. The researcher who obtains a high inter-rater reliability coefficient of 0.95 may conclude that the extent of agreement among raters is very good and therefore the raters are interchangeable. But what raters exactly are interchangeable ? Are we just referring to those two raters who participated in the reliability experiment ? Can we extrapolate to other similar raters who may not have participated in the study ? If the two participating raters agreed very well on the specific subjects that were rated, can we conclude that they still agree at the same level when rating other subjects ? What subject population are we allowed to infer to ? Generally many inter-rater reliability studies published in the literature do not address these critical questions. One of the reasons these issues are not addressed is that inter-rater reliability studies are not based on a precise definition of what should be quantified. Although starting with a general theoretical definition of inter-rater reliability may not connect directly with your specific application, we can still do a good job clarifying the scope of our investigation, and providing a detailed description of our ultimate goal. I will show in the next few paragraphs how this problem could be approached.

In a recent past, I was involved in the design of an inter-rater reliability study aimed at evaluating the extent to which triage nurses agree when assigning priority levels for care to pregnant women who present themselves in an obstetric unit with a health problem. If different triage nurses were to assign different priority levels to the same patients then one can see the potential dangers to which such disagreements may expose future mothers and their fetuses. Rather than rushing into the collection

of priority data with a few triage nurses and a handful of mothers-to-be who happen to be available, it is essential to take the time to carefully articulate the ultimate goal of this study. Here are a few goals to consider:

- The concern here is to ensure that the extent of agreement among triage nurses is high in order to improve patient-centered care for the population of pregnant women.
- But what is exactly that population of pregnant women we are servicing by the way ? Are they the women who visit a particular obstetric unit ? Should other obstetric units be considered as well ? Which ones ?
- Who are the triage nurses targeted by this study ? I am not referring to the triage nurses who may eventually participate in the study. Instead, I am referring to all triage nurses whose lack of proficiency in the triage process may have adverse effects on our predefined target population of pregnant women. They represent our target population of triage nurses. The possibly large number of nurses in this triage nursing population is irrelevant at this point, since we do not yet worry about those who will ultimately be recruited. Recruitment for the study will be addressed at a later time.
- In the ideal scenario where each triage nurse in the nursing population was to participate in the prioritization of all pregnant women in the target subject population, we want the extent of agreement among the nurses to be very high. But there is another important outstanding problem we need to address. If the triage patients must be classified into one of 5 possible priority categories, then we need to recognize that even after a maternal and fetal assessment are performed on the patient, two triage nurses may still be uncertain about the correct priority level the patient should be assigned to. This undesirable situation could lead them to assign a priority level that is not directly dictated by the patient's specific condition. An agreement among nurses that results from such an unpredictable situation is known in the inter-rater reliability literature as *Chance Agreement*. As badly as we may crave for high agreement among the nurses, this is not the type of agreements that we want. We want to prevent these types of agreement from giving us a false sense of security.

  All the issues raised above could lead to the following definition of inter-rater reliability for this triage study:

  > *Inter-rater reliability is defined as the propensity for any two triage nurses taken from the target triage nursing population, to assign the same priority level to any given pregnant woman chosen from the target women population, chance agreement having been removed from consideration.*

The above definition of inter-rater reliability does not provide a blueprint for calculating it. But that was not its intended purpose either. Instead, its purpose is to allow the management team to agree on a particular attribute of the nursing population that should be explored. Once this phase is finalized, the next step would be for the scientists to derive a formal mathematical expression to be associated with the attribute agreed upon, under the hypothetical situation where both target populations (raters and subjects) are available. This expression would then be the population parameter or coefficient (also known as inter-rater reliability coefficient) associated with the concept of inter-rater reliability. Now comes the experimental phase where a subset of raters and a subset of subjects are selected to derive an estimated inter-rater reliability coefficient, which is the concrete number ultimately produced by the inter-rater reliability experiment. An adequate presentation of the inter-rater reliability problematic cannot consist of detailed information, and computation procedures alone. It must also provide a proper and global view of the essential nature of the problem as a whole.

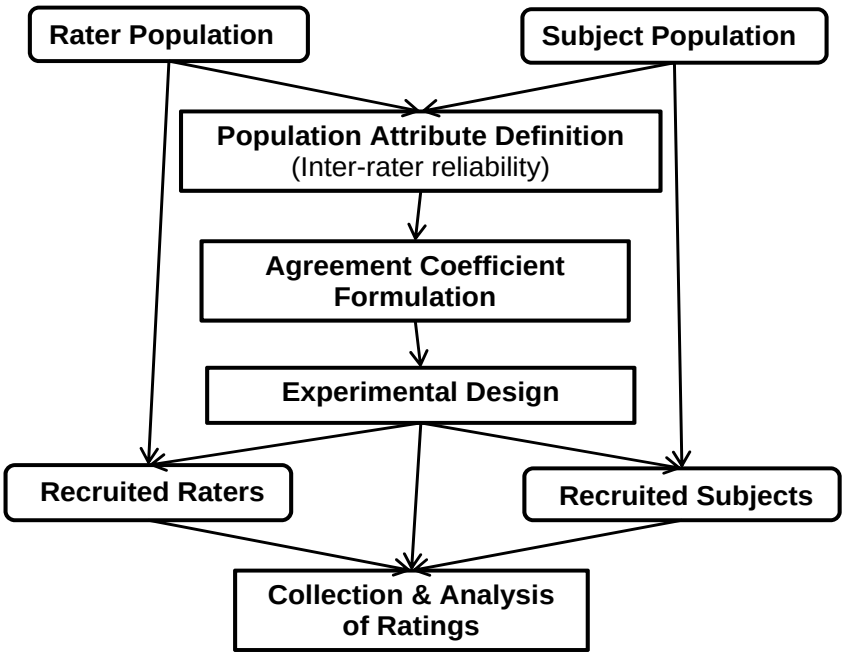

Figure 1.2.1: Phases of an Inter-Rater Reliability Study

Proving that the initial formulation of an inter-rater reliability coefficient as a population parameter is useful for studying the population attribute agreed upon can be a delicate task. What we can do in practice is to conduct an experiment and actually compute an estimated agreement coefficient. If the experiment is well designed,

this estimated agreement coefficient will be a good approximation of the population inter-rater reliability coefficient. The interpretation of its value in conjunction with the subject-matter knowledge of the tasks the raters are accomplishing will help determine if the way inter-rater reliability is quantified is acceptable. We will further discuss some technical issues pertaining to the formulation of agreement coefficients in the next section.

An inter-rater reliability experiment must be carefully designed. This design involves determining how many raters and subjects should be recruited, and what protocol should be retained for selecting them. Since only a subset of the rater population is generally retained for the experiment, it is essential to have a formal link between the participating raters and their home universe to ensure that the agreement coefficient that will ultimately be calculated will have a connection to the target population of raters around which the population attribute agreed upon was articulated. The same thing can be said about the subjects that will be recruited to participate in the inter-rater reliability experiment. Although a complete discussion of these design issues will take place in subsequent chapters, we will further explore some of these issues in the next few sections.

Ratings collected from a reliability experiment are generally presented in the form of a data table where the first column represents the subjects, and the subsequent columns representing the raters and the different ratings they assigned to these subjects. Two types of analyzes can then be performed on such data:

- Some researchers simply want to summarize the extent of agreement among raters with a single number that quantifies the extent of agreement among raters (e.g. kappa, intraclass correlation, Spearman correlation, etc...). These agreement coefficients may be formulated in many different ways depending on the study objectives as will be seen later.

- Other researchers are primarily interested in studying the different factors that affect the magnitude of the ratings. This task is often accomplished by developing statistical models that describe several aspects pertaining to the rating process. These statistical models, which are often described in the form of logit or log-linear models are not covered in this book. Interested readers may want to read Agresti (1988), Tanner, and Young (1985), Eckes (2011), or Schuster and von Eye (2001) among others.

1.3 Formulation of Agreement Coefficients

I indicated in the previous section that after defining the population attribute considered to represent inter-rater reliability, the next step is to formulate the agreement coefficient that will quantify it. This formulation takes the form of an algebraic expression that shows how the ratings will be manipulated to produce a number representing the inter-rater reliability coefficient. Let us consider the maternal fetal triage study discussed in the previous, and assume that 75 triage nurses have been identified in the target nursing population (let us say in a more formal way that $R = 75$), and 1,000 patients in the patient population affected by the triage processes being investigated (that is $N = 1,000$). Although the inter-rater reliability experiment will likely not involve all 75 raters and all 1,000 potential patients, I still want to formulate the agreement coefficient under the ideal scenario where each of the 75 triage nurses prioritizes all 1,000 patients by assigning one of 5 priority levels to them.

Suppose for simplicity sake that you are only interested in the "*the propensity for any two triage nurses taken from the target triage nursing population, to assign the same priority level to any given pregnant woman chosen from the target women population.*" Assuming we do not have to worry about the notion of chance agreement, this population attribute can be quantified by the relative number of pairs of triage nurses who assign a patient to the same priority level, averaged over all patients in the patient population. Let R_{ik} designate the number of nurses who assign patient i to priority level k. The total number of pairs of raters that can be formed out of R nurses in the population is $R(R-1)/2$. Likewise, the number of pairs of nurses that can be formed out of those R_{ik} who assigned patient i to priority k is $R_{ik}(R_{ik}-1)/2$. Now the relative number of pairs of nurses who assign the exact same priority level k to patient i is $P_{a|i,k} = R_{ik}(R_{ik}-1)/R(R-1)$. This means the relative number of pairs of nurses who assign any of the same priority level to patient i is obtained by summing the values $P_{a|i,k}$ for all 5 priority levels 1, 2, 3, 4, and 5. That is, $P_{a|i} = P_{a|i,1} + P_{a|i,2} + P_{a|i,3} + P_{a|i,4} + P_{a|i,5}$. Averaging all these values $P_{a|i}$ over all patients in the patient population will yield the agreement coefficient P_a we are looking for. All these operations can be codified mathematically as follows:

$$P_a = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{5}\frac{R_{ik}(R_{ik}-1)}{R(R-1)}. \tag{1.3.1}$$

This quantity becomes the estimand that will later be approximated using actual ratings from the reliability experiment.

The formulation of the agreement coefficient I just discussed appears reasonable for a simple reason. The ratings that are collected from the experiment can only

take five discrete values. Therefore the notion of agreement is straightforward and corresponds to an assignment of the same priority level by two raters. These ratings belong to the group of data known to be of nominal type. However, agreement coefficients recommended for nominal scales will be inefficient for ordinal, interval or ratio scales. And vice-versa, agreement coefficients suitable for the analysis of ratio data may not be indicated for analyzing nominal data.

Consider for example a psychiatrist classifying his patients into one of five categories named "Depression", "Personal Disorder", "Schizophrenia", "Neurosis", and "Other.[2]" This five-item scale is *Nominal* since no meaningful ordering of items is possible (i.e. no category can be considered closer to one category than to another one). On the other hand, patients classified as "Certain," "Probable," "Possible," or "Doubtful" after being tested for Multiple Sclerosis, are said to be rated on an *Ordinal Scale*. The "Certain" category is closer to the "Probable" category than it is to the "Doubtful" category. Consequently, disagreements on an ordinal scale should be treated differently from disagreements on a nominal scale. This is a situation where the type of rating data (Nominal or ordinal) will have a direct impact on the way the data is being analyzed. Some inter-rater reliability studies assign continuous scores such as the blood pressure level to subjects. The data scale in this example is a continuum. As result, it is unreasonable to require agreement between two raters to represent an assignment of the exact same score to a subject. Agreement in this context is often measured by the within-subject variation of scores. With a different notion of agreement comes different ways of formulating agreement coefficients.

NOMINAL RATINGS

With a nominal scale, two raters agree when their respective ratings assigned to a subject are identical, and are in disagreement otherwise. In this context, agreement and disagreement are two distinct and opposite notions, and the relative number of times agreement occurs would normally be sufficient to determine the extent of agreement among raters. Unfortunately the limited number of values that raters can assign to subjects increases the possibility of an agreement happening by pure chance. Intuitively, the smaller the number of categories the higher the likelihood of chance agreement. Consequently, our initial intuition that the relative number of agreement occurrences can be used as an inter-rater reliability measure is unsatisfactory and must be adjusted for chance agreement. A key motivation behind the development of the well-know Kappa coefficient by Cohen (1960) was to propose an agreement coefficient that will be corrected for chance agreement.

The notion of agreement sometimes appears in the form of internal consistency

[2]Although the same patient may present multiple symptoms, we assume that the rating will be determined by the most visible symptom.

in scale development. When a set of questions are asked to a group of participating subjects in order to measure a specific construct, the scale developer expects all of the questions to show (internal) consistency towards the measurement of a unique latent construct. High internal consistency is an indication of a high degree of agreement among the questions (called items in the jargon of item response theory) with respect to the construct associated with the subjects. One of the best known measures of internal consistency is Cronbach's alpha coefficient (Cronbach, 1951) discussed in Part IV of the book.

ORDINAL RATINGS

When the measurement is ordinal, agreement and disagreement are no longer two distinct notions. Two raters A and B who rate the same patient as "Certain Multiple Sclerosis" and "Probable Multiple Sclerosis" are not quite in total agreement for sure. But are they in disagreement ? Maybe to some extent only. That is, with ordinal scales, a disagreement is sometimes seen as a different degree of agreement, a *Partial Agreement*. An ordinal scale being nominal and ordered, the chance-agreement problem discussed previously remains present and becomes more complex with a changing notion of disagreement. This problem has been addressed in the literature by assigning weights to different degrees of disagreement as shown by Cohen (1968) among others.

With ordinal ratings, there is another kind of agreement that may be of interest to some researchers. It is the agreement among raters with respect to the ranking of subjects. Since the subjects participating in the reliability experiment can be ranked with respect to the scores assigned by one rater, a researcher may want to know whether all raters agree on which subject is the best, which one is the second best, and so on. For government examiners who score proposals submitted by potential contractors, the actual score may not matter as much as the ranking of the proposals. In this case, the most appropriate agreement coefficients would belong to the family of measures of concordance, or association as will be seen later in this book.

INTERVAL AND RATIO RATINGS

The distinction between the notions of interval and ratio data is not as important in the field of inter-rater reliability assessment, as it would in other fields. Nevertheless knowing that distinction will help researchers make a better choice of agreement coefficients. An example of interval data, which is not of ratio type, is the temperature expressed either in degree Celsius or in degree Fahrenheit. The difference between 35^0F and 70^0F is 35^0F, which represents a drastic change in the intensity of heat we feel. However, only a comparison between 2 temperature values can give meaning to

each of them. An isolated value such as 35^0F does not represent a concrete measure in the absence of a natural origin, making it meaningless to apply certain arithmetic operations such as the multiplication or the division[3]. Ratio data on the other hand, such as the weight, the height, or the body mass index possess all the properties of the nominal, ordinal, and interval data, in addition to allowing for the use of all arithmetic operations, including the multiplication and the division.

Why should we care about rating data being of interval or ratio type ? It is because interval/ratio-type ratings would require special methods for evaluating the extent of agreement among raters. The very notion of agreement must be revised. Given the large number of different values a score may take, the likelihood of two raters assigning the exact same score to a subject is slim. Consequently, the extent of agreement among raters is best evaluated by comparing the variation in ratings caused by the raters to the variation in ratings caused by random errors.

1.4 Different Reliability Types

The inter-rater reliability literature is full of various notions of reliability. Different terms are used to designate similar notions (e.g. intra-rater reliability, and test-retest reliability), while the word inter-rater reliability has been used with different meanings in different contexts. There is also an important distinction to be made between validity and reliability. While reliability is necessary (although insufficient) to ensure validity, validity is unnecessary for a system to be reliable. This section reviews some uncommon reliability types often encountered in the literature, and will discuss the relationship between reliability and validity.

1.4.1 *Undefined Raters and Subjects*

There are studies where identifying what entities represent subjects and raters can be unclear. As an example consider a reliability study discussed by Light (1971) where 150 mother-father pairs are asked a single 3-item multiple choice question. The ratings obtained from this experiment are reported in Table 1.3. The problem is to evaluate the extent to which mothers and fathers agree on a single issue, which could be related to children education for example. Instead of having two raters (one mother and one father) rate 150 subjects as is often the case, this special expe-

[3]For example, if you write $70^0\text{F} = 2 \times 35^0\text{F}$ then you will be giving the false impression that at 70^0F the heat intensity is twice higher than at 35^0F. The only thing we really know for sure is that the intensity of the heat at 70^0F is substantially higher than at 35^0F. By how much ? Twice ? Three times ? We cannot really say it in any meaningful way.

riment involves 150 raters rating a single subject[4]. This could nevertheless be seen as a classical inter-rater reliability as long as 75 raters of one type (e.g. fathers) are paired with 75 raters of a different type (e.g. mothers). However, it is unwise to treat these raters as raters. Instead, you should see inter-rater reliability in this context as being calculated not between two human raters, but rather between two types of raters: "Mothers," and "Fathers." The different mother-father pairs can be seen as distinct "subjects." "Mothers," and "Fathers" are virtual raters, whose ratings come from specific mother-father pairs.

Table 1.3: Distribution of Mother-Father Pairs by Response Category

	Mothers			
Fathers	1	2	3	Total
1	40	5	5	50
2	8	42	0	50
3	2	3	45	50
Total	50	50	50	150

The main advantage of this design lies in the possibility it offers to extrapolate the calculated extent of agreement to a population of mothers and fathers larger than the 150 study participants.

1.4.2 *Conditional Reliability*

When the extent of agreement among raters on a nominal or ordinal scale is unexpectedly low, it is common for researchers to want to identify the specific category or categories on which raters have difficulties agreeing. This effort aims at finding some of the root causes of the weak rater agreement. The method used consists of calculating the extent of agreement among raters based on the pool of subjects known to have been classified by a given rater into the category to investigate. The resulting agreement coefficient is constrained by the requirement (or condition) to use only subjects whose membership in one category was determined by a given rater, and is known as the conditional agreement coefficient.

The reference rater whose ratings are used to select the subjects for conditional analysis, could be chosen in a number of ways. In a two-rater reliability experiment for example, the reference rater will necessarily be one of the two participants. In

[4]Rating a subject in this context amounts to one mother-father pair providing a personal opinion on a social issue.

a multiple-rater reliability experiment however, the reference rater may be chosen arbitrarily, or may represent the most experienced of all raters whose ratings may be considered as the gold standard. Fleiss (1971), or Light (1971) among others studied such conditional analyzes.

1.4.3 *Reliability as Internal Consistency*

In the social sciences, survey questionnaires often contain groups of questions aimed at collecting the same information from different perspectives. If a specific set of questions provides highly correlated information from different respondents in a consistent manner, it is considered to be reliable. This reliability is known as *Internal Consistency Reliability*. There are numerous situations in practice that lead to special forms of the internal consistency reliability. Internal consistency does not deal with raters creating scores. Instead, it deals with item questions used to create summary scores (also known as scales) based on information collected from the subjects. This topic is discussed in part IV of this book.

The discussion on internal consistency evaluation in this book will focus on Cronbach's alpha proposed by Cronbach (1951). Additional information on this topic could be found in other textbooks on social research methods such as those of Carmines and Zeller (1979), or Traub (1994).

1.4.4 *Reliability versus validity*

Reliability coefficients quantify the extent to which measurements are reproducible. However, the existence of a " true" score associated with each subject or object raises the question as to whether the scores that the raters agreed upon match the "true" scores. Do raters who achieve high inter-rater reliability also agree on the correct category, when it exists ? Or do they simply agree on any category ? These are some important questions a researcher may want to consider in order to have an adequate interpretation of the magnitude of agreement coefficients.

If two raters agree frequently, then the scores they assign to subjects are considered reliable. If both raters agree on the subject's "true" score, then these scores are considered valid. Valid scores are scores that are both reliable and match the reference score, also known as the "Gold Standard." Classical inter-rater reliability coefficients will generally not measure the validity of scores. Validity is measured with special validity coefficients to be discussed in chapter 11.

As seen earlier in this chapter, the "true" score does not always exist. The scoring of the quality of customer service in a department store for example reflects the rater's personal taste or opinion. No score in this case can a priori be conside-

red standard or true, although if customer service consistently receives low ratings, it could reasonably be considered to be of poor quality. This may still provide valuable information to managers, primarily because it shows that the raters (i.e. the store customers) agree on something that has the potential to affect the business profitability.

1.5 Statistical Inference

The analysis of ratings often leads researchers to draw conclusions that go beyond the specific raters and subjects that participated in the experiment. This process of deducing from hard facts is known as inference. However, I recommend this inference to be statistical. Before enumerating some of the benefits of statistical inference, I must stress out that what distinguishes statistical inference from any other type of inference is its probabilistic nature. The foundation of statistical inference as it applies in this context of inter-rater reliability, and as presented in this book is the law of probability that governs the selection of raters from the rater population and the selection of subjects from the subject population. I expect to be able to pick any rater from the rater population (or one subject from the subject population) and tell precisely the likelihood that it will be recruited to participate in the experiment. These laws of probabilities tie the set of recruited raters and subjects to their respective populations. These links will make it possible to evaluate the chance for our calculated agreement coefficient to have the desired proximity with its population-based estimand. Here is where you find one of the main benefits of statistical inference.

In the past few sections, I indicated that before an inter-rater reliability study is formally designed the target rater and subject populations must be carefully defined first. Then inter-rater reliability is defined as an attribute of the rater population, which in turn should be codified mathematically with respect to both the rater and subject populations. This expression represents the population parameter or the estimand or the inter-rater reliability parameter to be estimated using actual ratings from the reliability experiment. Note that the expression showing how the ratings produced by the experiment are manipulated is called the inter-rater reliability estimator. In sequence we have three things to worry about, the attribute, the estimand, and the estimator. Most published papers on inter-rater reliability tend to limit the discussions to the estimator that produces the numbers. For the discussion to be complete, it must tie the estimator to the estimand and to the attribute.

Note that the inter-reliability coefficient generated by the estimator changes each time the raters or subjects who participate in the study change. It is directly affected by the experimental design. The estimand on the other hand solely depends upon both the rater and the subject populations, and are not affected in any way by

the experiment. It may change only if you decide to modify the pool of raters and subjects that are targeted by the study. The attribute is the most stable element of all. It can only be affected if the study objective changes. The discrepancy between the estimator and the estimand is what is known as the statistical error. This one can be and should be estimated. It shows how well the experiment was designed. Many different groups of raters and subjects can be formed out of the rater and subject populations. Each of these rater-subject combinations will generate different values for the agreement coefficient. How far you expect any given coefficient to stray away from their average value is measured by the agreement coefficient's standard deviation

Chapter 5 is entirely devoted to the treatment of this important topic. Although I have decided to use the laws of probability governing the selection of raters and subjects as the foundation of statistical inference, this is not to claim that it is the only possible foundation that is available. Researchers who based these analyzes on theoretical statistical models may decide to use the hypothetical laws of probability that come with these models as their foundation. This alternative approach for inference is not considered in this book.

1.6 Book's Structure

This book presents various methods for calculating the extent of agreement among raters for different types of ratings. Although some of the methods were initially developed for nominal ratings only, they have been extended in this book to handle ordinal, interval, and ratio scales as well. To ensure an adequate level of depth in the treatment of this topic, I decided to discuss the precision aspects of the agreement coefficients being calculated, and to expand the methods so that datasets containing missing ratings can be analyzed as well. I always start the presentation of new methods with a simple scenario involving two raters only before extending it to the more general context of multiple raters, and nominal, ordinal, interval, or ratio ratings. The book is divided into five parts:

- Part I has a single chapter, which is the current one.
- Part II is made up of five chapters. These are chapters 2, 3, 4, 5, and 6. They deal with many different types, and aspects of the chance-corrected agreement coefficients (CAC). Agreement coefficients are discussed in these chapters for ratings that are of nominal, ordinal and interval types. Also discussed in part II chapters are the benchmarking methods for qualifying the magnitude of the agreement coefficients as well as the techniques for evaluating associated statistical errors.
- Part III covers the Intraclass Correlation Coefficients (ICC), which are recommended for use with quantitative measurements. This part of the book is made

up of the four chapters 7 to 10.

- Part IV is essentially about miscellaneous topics, which include the study of agreement coefficients conditionally upon specific categories, as well as a discussion of various measures of association or measures of agreement based on ranks. The two chapters 11 and 12 are included in this part of the book.
- Part V of the book includes appendices A and B. Appendix A contains a number of datasets that the reader may use for practice, while appendix B discusses a number of software options that may be considered for analyzing inter-rater reliability data.

Part II of this book starts in chapter 2 with a detailed critical review of several agreement coefficients as proposed in the literature for the analysis of nominal ratings. This review includes Cohen's Kappa coefficient, its generalized versions to multiple raters, Gwet AC_1, Krippendorff's alpha, or Brennan-Prediger coefficients among others. In chapter 3, I show that with the use of proper weights, the agreement coefficients discussed in chapter 2 can be adapted to produce a more efficient analysis of ordinal and interval ratings. In chapter 4, I use the AC_1 coefficient proposed by Gwet (2008*a*), and Aickin's alpha coefficient of Aickin (1990) as examples to show how agreement coefficients can be constructed from scratch, and why these two particular coefficients are expected to yield valid measures of the extent of agreement among raters. The theory underlying these two coefficients is also discussed in details. In chapter 5, I introduce the basic principles of statistical inference in the context of inter-rater reliability assessment. I stress the importance of defining the population of raters and the population of subjects that are targeted by the inter-rater reliability study. I did not use the model-based approach to statistical inference of Kraemer et al. (2002) and others. Instead, I decided to introduce for the first time the design-based approach to statistical inference in the field of inter-rater reliability assessment. The design-based approach to statistical inference is widely used in sample surveys and relies on the random selection of the sample of subjects from a well-defined finite population, as well as from the random selection of raters from a well-defined population of raters. Chapter 6 addresses the important problem of benchmarking the inter-rater reliability coefficient. The problem consists of determining different thresholds (or benchmarks) that could be used to interpret inter-rater reliability as poor, good, or excellent. We review different benchmarks that have been proposed in the literature before proposing a new benchmarking model specific to each inter-rater reliability coefficient.

Part III of this book starts with chapter 7, which provides a general introduction to the problem of analyzing continuous or quantitative ratings. In chapter 8, I discuss the intraclass correlation in one-factor studies. In these studies, the rater's quantitative score is investigated as a function of the rater effect alone, or as a function of the

subject effect alone. These are typically studies where each subject may be rated by a different group of raters, or where each rater may rate a different group of subjects. Chapter 9 covers the intraclass correlation in two-factor studies based on a random factorial design. In this chapter, the rater's quantitative score is investigated as a function of both the rater and the subject effect. The group of subjects is assumed to be representative of the larger population of subjects it was randomly selected from. Likewise, the group of raters is assumed to be representative of the larger group it was selected from. Chapter 10 is the last chapter of part III of the book, and deals with the intraclass correlation in two-factor studies based on a the mixed factorial design. That is, while the group of subjects is assumed to be representative of a larger group of subjects, the participating raters on the other hand are the only raters of interest.

Part IV of this book focuses on miscellaneous topics, and starts with chapter 11 on the conditional analysis of inter-rater reliability. Validity coefficients are discussed in this chapter, as well as agreement coefficients conditionally upon specific categories. The conditioning is done on the "true" category when one exists, or on the category chosen by a given rater otherwise. This chapter explores additional methods for enhancing the analysis of inter-rater reliability data beyond a single statistic. Chapter 12 is devoted to the study of various measures of association or concordance, in addition to discussing Cronbach's alpha, a popular statistic in the field of item analysis.

1.7 Choosing the Right Method

How your ratings should be analyzed depends heavily on the type of data they represent, and on the ultimate objectives of your analysis. I previously indicated that your ratings may be of nominal, ordinal, interval, or ratio types. Figure 1.7.1 is a flowchart that shows what types of agreement coefficients should be used and the chapters where they are discussed, depending on the rating data type. Note that this chart describes my recommendations, which should not preclude you from treating ordinal ratings for example as if they were nominal, ignoring their ordinal nature if deemed more appropriate.

Figure 1.7.1 does not identify a specific agreement coefficient that must be used. Instead, it directs you to the chapters that discuss the topics that must be of interest to you. These chapters provide more details that will further help you decide ultimately what coefficients are right for your analysis. You will also notice that this chart does not include the special topics that are addressed in part IV of this book. You may review the content of these chapters if your analysis needs are out of the ordinary.

Figure 1.7.1 indicates that if you are dealing with ratio or interval ratings, then you can use one of the chance-corrected agreement coefficients of chapters 2 or 3 only if these ratings are predetermined before the experiment is conducted. Otherwise, you will need the intraclass correlation coefficients of chapters 7 to 10. The ratings are predetermined if before the experiment the researcher knows that each subject can for example only be assigned one of the values 1.35, 3.7, 4.12, and 6.78. However, if the rating is the subject's height, or weight whose values can be determined only after the measurement had been taken, then the intraclass correlation is what I recommend.

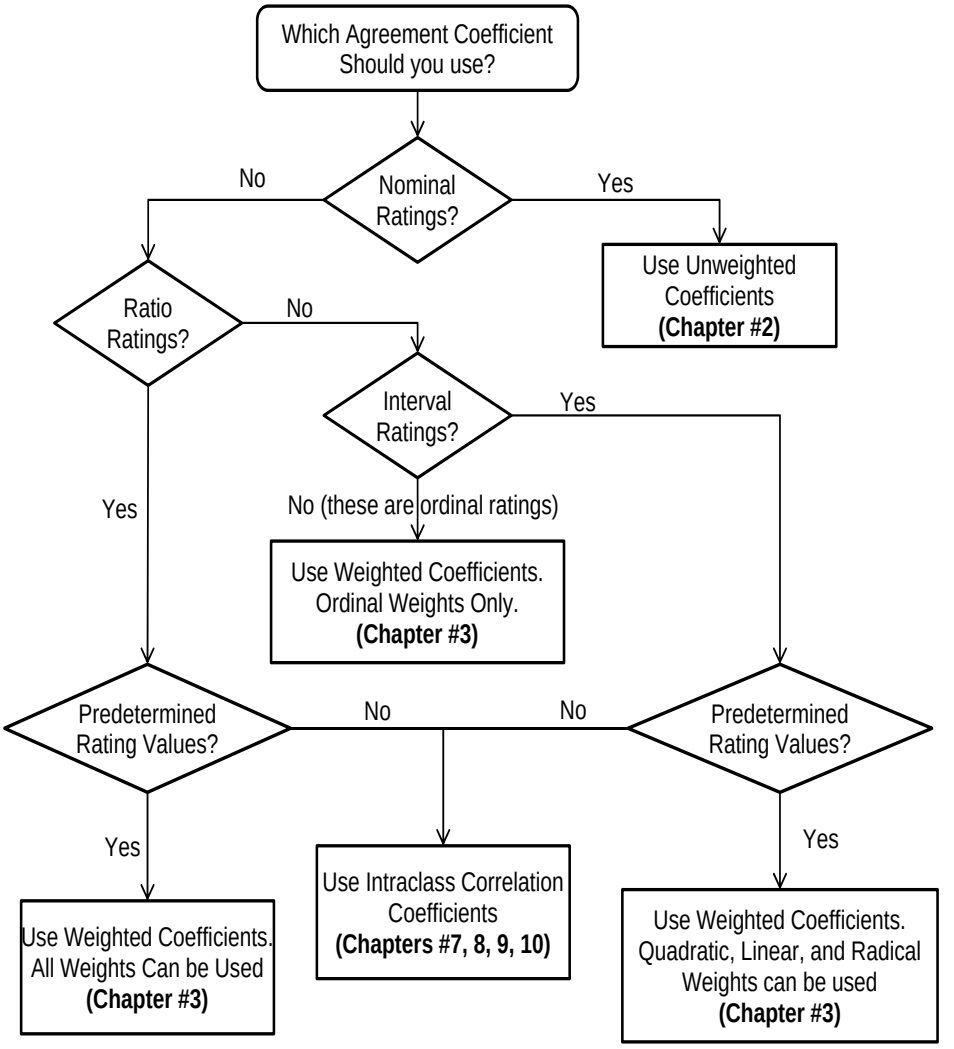

Figure 1.7.1: Choosing an Agreement Coefficient