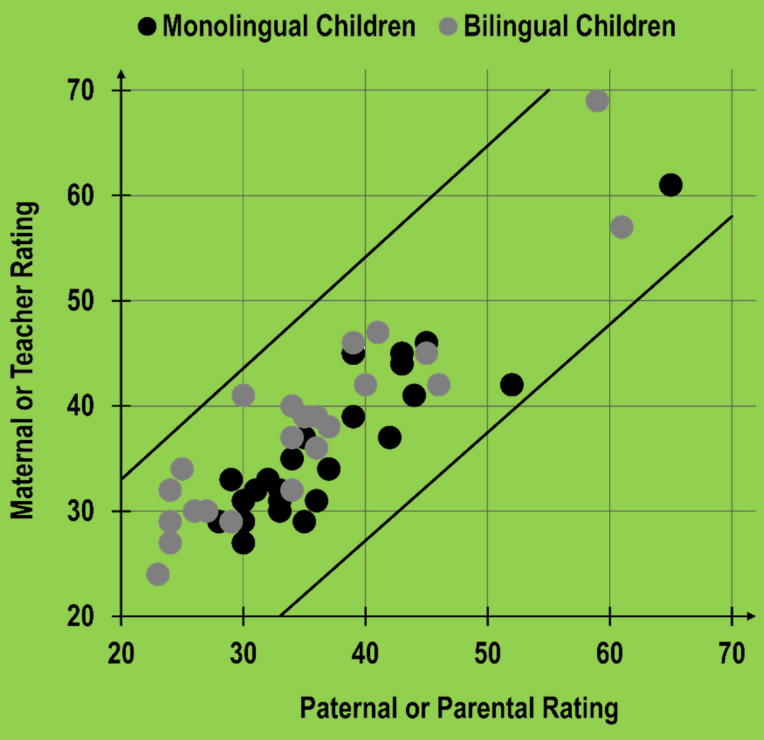
Handbook of Inter-Rater Reliability

Fifth Edition

The Definitive Guide to Measuring the Extent of Agreement Among Raters

Volume 1

Analysis of Categorical Ratings



Kilem L. Gwet, Ph.D.

HANDBOOK OF INTER-RATER RELIABILITY

Fifth Edition

Volume I

Analysis of Categorical Ratings

HANDBOOK OF INTER-RATER RELIABILITY

The Definitive Guide to Measuring the Extent of Agreement Among Raters

Fifth Edition

Volume I

Analysis of Categorical Ratings

Kilem L. Gwet, Ph.D.

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

Published by AgreeStat Analytics, in the United States of America.

AgreeStat Analytics
PO BOX 2696,
Gaithersburg, MD 20886-2696
e-mail: contact@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 - Volume 1: Analysis of Categorical Ratings / By Kilem Li Gwet - 5th ed.
p. cm.
Includes bibliographical references and index.
1. Biostatistics
2. Statistical Methods
3. Statistics - Study - Learning. I. Title.
ISBN 978-1-7923-5463-2

Preface

Ratings that 2 raters independently assign to the same group of subjects may still differ, sometimes substantially. In this case, an observed rating is affected by attributes associated with both the rater and the subject. Other unknown factors could possibly impact rating data, although the rater and the subject are known to be the dominant effects in a well-designed inter-rater reliability experiment. Ratings assigned to subjects are considered reliable if they are solely affected by subject-specific attributes, the rater effect being negligible. Why are reliable ratings important in research? It is because any variation in a reliable rating dataset can be interpreted as valuable information about the subjects under investigation. Improving the quality of rating data by minimizing the rater effect is the primary objective of the study of inter-rater reliability.

Between-rater variation could jeopardize the integrity of scientific inquiries or have dramatic consequences in a clinical setting. As a matter of fact, a wrong drug or wrong dosage of the correct drug may be administered to patients at a hospital due to a poor diagnosis. Likewise, exam grades are considered reliable if they are determined only by the candidate's proficiency level in a particular skill, and not by the examiner's scoring method. The study of inter-rater reliability helps researchers address these issues using an approach that is methodologically sound.

The 4th edition of this book covers Chance-corrected Agreement Coefficients (CAC) for the analysis of categorical ratings, as well as Intraclass Correlation Coefficients (ICC) for the analysis of quantitative ratings. Both topics were discussed in parts II and III of that book, which is divided into 4 parts. The 5th edition however, is released in 2 volumes. The present volume 1, focuses on CAC methods whereas volume 2 is devoted to ICC methods. The decision to release 2 volumes was made at the request of numerous readers of the 4th edition who indicated that they are often interested in either CAC techniques or in ICC techniques, but rarely in both at a given point in time. Moreover, the large number of topics covered in this 5th edition could not be squeezed in a single book, without it becoming voluminous.

Here is a summary of the main changes from the 4th edition that you will find in this book:

- Chapter 2 is new to the 5th edition and covers various ways of setting up your rating dataset before analysis. My decision to add this chapter stems from a large number of questions I received from researchers who wanted to

know how their rating data should be organized. I noticed that sometimes, organizing your data properly will clear the pathway towards resolving most computational problems.

- Chapter 6 entitled "Agreement Coefficients and Statistical Inference" has also been expanded substantially. Section 6.5 on sample size calculation in particular, covers new power calculation methods not discussed in the 4th edition. These are nonparametric methods for computing the optimal number of raters and subjects at the design stage of an inter-rater reliability experiment.

- Chapter 8 on the analysis of agreement coefficients conditionally upon specific categories has been substantially rewritten with more details and added clarity.

- Chapter 9 on the analysis of nominal-scale inter-rater reliability data is new. It addresses several new techniques that were not covered in any of the previous editions of this book. One of these techniques is about the important notion of inter-annotator agreement, which plays a key role in the fields of Natural Language Processing (NLP), computational linguistics or text analytics.

  Another technique discussed in this chapter is the procedure for testing 2 agreement coefficients (correlated and uncorrelated) for statistical significance.

  Also discussed in this chapter is the important problem of measuring the extent of agreement among 3 raters or more, when the same subject cannot be rated by more than 2 raters. I show how such a study can be designed and discuss the statistical implications of such a design.

  The remaining techniques described in this chapter are related to the influence analysis, the intra-rater reliability and to Cronbach's alpha coefficient. Influence analysis is used to detect problem raters in low-agreement studies, whereas Cronbach's alpha coefficient is commonly used in item analysis.

The reader will notice that this book is very detailed. Yes, I wanted it to be sufficiently detailed for practitioners to gain more insight into the topics, which would not be possible if the book was limited to a high-level coverage of technical concepts. I want the researcher to read this book and be able to implement the proposed solutions without having to figure out hidden steps or unexplained concepts.

This book is not exhaustive. It does not cover all topics of interest related to the field of inter-rater reliability. I selected topics among the most commonly referenced by researchers in various fields of research. Moreover, 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 with advanced statistical techniques. Therefore, I decided to confine myself to techniques

that a large number of researchers will feel comfortable implementing. This is one of the reasons I did not cover any approach that appeals to advanced theoretical statistical models (e.g. Rasch models, logistic regression models, ...) and which generally require considerable time and statistical expertise to be successfully implemented.

I accumulated considerable experience in the design and analysis of inter-rater reliability studies over the past 20 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 less familiar with mathematical equations, I present detailed examples, and provide downloadable Excel spreadsheets that show all the steps for calculating various agreement coefficients, along with their precision measures. 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 of research. If you have comments do not hesitate to contact me at contact@agreestat.com.

Kilem Li Gwet, Ph.D.

Contents

Acknowledgment

First and foremost, this book would never have been written without the full support of my wife Suzy, and our 3 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.

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 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 2 survey instruments that psychologists often use 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 reading this book. Please tell me what you think about it by sending an e-mail to contact@agreestat.com. Alternatively, you may write a review at Amazon.com.

Thank you,

Kilem Li Gwet, Ph.D.

PART I

PRELIMINARIES

List of Part I Chapters

CHAPTER 1

Introduction

OBJECTIVE

This chapter presents an overview of the inter-rater reliability concept and highlights its importance in scientific inquiries. Difficulties associated with the quantification of inter-rater reliability and key factors affecting its magnitude are discussed as well. This chapter stresses the importance of a clear statement of study objectives and a careful design of inter-rater reliability experiments. Different types of inter-rater reliability are discussed 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 1.4 depicts a flowchart that summarizes the process of identifying the correct agreement coefficient to use based on the type of ratings to be collected.

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 specialists in all of these fields. Nevertheless, introducing the general concept is straightforward. During the conduct of a scientific investigation, researchers often gather data that must later be interpreted before inference is made about the issues being investigated. Given the pivotal role of data in scientific inference, it is crucial to ensure that the data production system that may include methods, procedures, equipment and people, is marginally affected by small changes. Broadly speaking, reliability is the extent to which a data production system resists to small changes in its structure. If one piece of equipment is replaced with an alternative but similar one, will the system produce the same data? If some of humans participating in the data collection are replaced with others, will it affect the data? To what extent? If some procedures are changed will you still get valid data? Data reliability is more than just inter-rater reliability. Inter-rater reliability refers to the portion of data reliability that is affected by the specific components of the data production system that you call raters. If the raters are a key component of that system, then inter-rater reliability may well be all you need to quantify. But if there are other important components in the production system, you may need to investigate them as well.

This book is primarily concerned about reliability examined from the viewpoint of data reproducibility. Consequently, inter-rater reliability will be evaluated by the extent of agreement among raters. It is in that sense that the term "inter-rater reliability" and "inter-rater agreement" will often be used interchangeably throughout this text. This approach is more consistent with the concept of reliability in measurement theory. Reliable data ensure reproducibility or consistency of repeated measurements. It does not by any means ensure validity, which refers to consistency with a "gold standard" measurements that researchers agree to use as reference. I will also discuss the notion of validity in this book.

I am however fully aware that some authors have attempted to make a clear distinction between the 2 notions of reliability and agreement. In psychometric theory for example, Tinsley and Weiss (1975) and Tinsley and Weiss (2000) introduced a special notion of reliability that considers as reliable, series of data from different raters, which are similar when expressed in the form of deviations from their overall mean. These authors argued that this notion of reliability is different from that of agreement, which requires raters to generate the exact same ratings. Their notion

of reliability is unrelated to data reproducibility, which is the only problem I have worked on. Other authors such Krippendorff (2011) or Kottner and Streiner (2011) have discussed these issues. While Krippendorff (2011) provides an instructive review of different conceptions of reliability in various fields of research, the argument made by Kottner and Streiner (2011) does not clarify the issue much and may even have confused it further.

Let us turn to the study of reliability in the context of categorical ratings. 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 reliability 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 level 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 can be 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 this moment. 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 are 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 numeric 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 inter-rater reliability among faculty members testing the proficiency level 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 evaluate 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 a systematic examination of the message content. For example, health information specialists must often read general information regarding a patient's condition and the treatment received, before assigning and International Classification of Disease code needed for billing. A poor inter-coder 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 et al. (2013).

In the fields of linguistic analysis, computational linguistics, or text analytics, annotation is a common activity. 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.

Many 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 inter-coder reliability. Medical coding is a specialty in the medical field, which has specific challenges for inter-rater reliability assessment. The need to evaluate inter-coder agreement generally occurs in one of the following two situations:

- Different coders evaluate the patients' medical records 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 a situation where

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

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 their respective primary diagnosis codes (possibly different) fall into the same group of codes. Alternatively, one may use both primary and secondary diagnosis codes to define agreement as being reached when some codes from both raters are included in a predefined group of "similar codes."

- 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 beneficiaries of public health plans using codes in 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 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. Eckes (2011, page 24) acknowledged this issue when he said "... even if high inter-rater 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

if categories are defined on an ordinal scale such as "none", "basic", "intermediate", "Advanced" and "Expert" then the 2 categories "Advanced" and "Expert" represent a disagreement. However, these 2 categories are often seen as representing a "partial agreement," which can be justified when compared to the 2 extreme categories "none" and "expert" that are expected to represent total disagreement. 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). Nevertheless, very few statistical concepts have been applied more widely than this one.

Ratings collected from a reliability experiment are generally presented in the form of a data table where the first column contains subject identifiers and the subsequent columns contain the ratings that each rater assigned to these subjects. Two types of analyzes can then be performed on such data:

- Some researchers are primarily interested in studying the different factors that affect the rating magnitude. 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.

- Other researchers want to quantify the extent of agreement among raters with a single summary statistics (e.g. kappa, intraclass correlation, Spearman correlation, etc...). Subsequent analyses that include identifying problem raters, comparing agreement coefficients obtained on different occasions or from different subject groups, or testing hypotheses about the magnitude of an agreement coefficient are also often of interest. These are the types of analyses that are primarily addressed in this book. You will see in the next section that a proper implementation of this approach requires a careful specification of the experimental design parameters. The problems associated with the formulation of agreement coefficients will be addressed in subsequent chapters.

1.2 Scope and Design of Inter-Rater Reliability Experiments

Many articles on inter-rater reliability assessment are limited to a description of the experiment that produced the ratings and to the method used for analyzing those

ratings. Oftentimes little space is devoted to discussing the strength and validity of the information collected. The researcher who obtains a high inter-rater reliability coefficient of 0.95 for example may conclude that the extent of agreement among raters is very high and therefore the raters are interchangeable. But what raters exactly are interchangeable? Are we just referring to the two raters who participated in the reliability experiment? Can we extrapolate these findings 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 will still agree at that same level when rating other subjects? What subject population are we allowed to infer to? Were all subjects rated by the same pair of raters? Were scoring duties distributed among several pairs of raters? Most inter-rater reliability studies published in the literature do not address these critical questions. This deficiency makes it difficult to have an accurate interpretation of many published studies.

In order to facilitate the interpretation of study findings, it is essential to start the development of a new inter-rater reliability experiment by clarifying the scope of the investigation and by providing a detailed description of the experimental design. The scope of the investigation will help articulate an abstract definition of inter-rater reliability separated from the calculation procedure while the experimental design will help specify all calculation procedures. I will show one way to approach this process in the next few paragraphs.

1.2.1 *Scope of the Investigation*

I once participated 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 visiting 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 the 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? 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 to participate in the study. Recruitment for the study will be addressed at a later time during the experimental design phase.

- In the ideal situation 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 triage nurses to be very high. But there is an important outstanding problem we need to address. If the patients must be classified into one of 5 possible priority categories, then we need to recognize that even after a maternal and fetal assessments are performed on the patient, a triage nurse may still be uncertain about the correct priority level the patient should be assigned to. This undesirable situation of uncertainty could lead to a priority level being assigned that does not reflect the patient's specific condition. An agreement among nurses reached under uncertainty is known in the inter-rater reliability literature as *Chance Agreement*. As desirable as a high agreement among nurses may be, chance agreement is not the type of agreement that we want. Instead, we want to prevent chance 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 representation of the inter-rater reliability produced by the experiment. An adequate presentation of the inter-rater reliability problem 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, as depicted in figure 1.1.

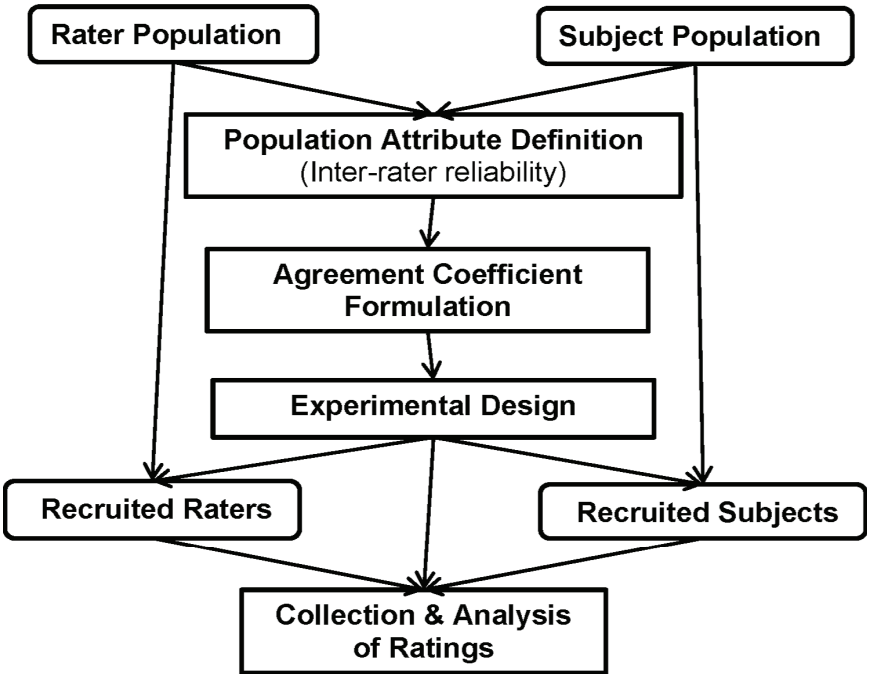


Figure 1.1: Phases of an Inter-Rater Reliability Study

1.2.2 *Experimental Design*

An inter-rater reliability experiment must be carefully designed. An experiment is said to be well designed if it produces accurate agreement coefficients (i.e. coefficients with a small standard error[2]) given the often limited resources available. Experimental design involves determining how many raters and subjects should be recruited, what protocol should be retained for selecting them, how raters should be assigned to subjects, how scoring should be performed. Some very simple studies

[2]The standard error is a statistical measure that tells us how far any given agreement coefficient strays away from its average value.

may only require a few of these activities to be performed.

Some inter-rater reliability studies are based on a fixed and predetermined number of raters. For example, if the purpose of the study is to investigate the extent of concordance between clinical assessment and research methods in the assessment of diagnosis, then the 2 approaches would be the only raters of interest. Therefore there would be no need to worry about other raters not being part of the experiment. If the study aims at investigating the extent of concordance among chart abstractors or medical coders then deciding about the number of raters to include in the study becomes an important issue to be addressed. A common mistake made in many inter-rater reliability studies is to start by recruiting a few raters prior to specifying the entire universe of raters concerned about the scoring of subjects. The correct approach consists of specifying the target rater universe (or population) first, then a subset of this population should be selected for participation in the study according to a predetermined and rigorous random selection protocol. This process that consists of specifying the target rater universe, calculating the required number of raters and performing the selection of a smaller subset of raters is referred to as the *Rater Sampling*, or the *Sampling of the Rater Population*. This sampling process will lead to a *Rater Sample* or *Sample of Raters* supposed to be a good representation of the entire rater universe.

An inter-rater reliability study will not just involve rater sampling. It will often require *Subject Sampling* as well, unless the researcher decides that the study findings must apply solely to the specific group of subjects that participated in the experiment. If sampling the subject universe is necessary then that universe must be specified so that the scope of the experiment and the subjects to which the study findings apply are known. After calculating the number of subjects required to achieve the study accuracy goals, the researcher can proceed with the actual selection of subjects that will make up the *Subject Sample*. The subject sample is essentially a list of subject names that the researcher sees as a "good" representation (with respect to a number of characteristics) of the target patient population. These are the subjects that are contacted and eventually recruited to participate in the study. Since the task of recruiting a subject is not always successful, it is recommended to create a subject sample containing slightly more subjects than needed. The number of subjects in the subject sample will often be referred to as the subject sample size. Issues related to the determination of this sample size are discussed in chapter 6.

Sample Selection

The sample of raters or subjects should ideally be probabilistic. That is the selection of each rater and each subject from their respective target universes must be carried out following a random process that gives each unit of interest a chance of being chosen for the study. As an example, suppose that 4 of 10 patients hospitalized

in a psychiatric hospital must be evaluated by two doctors as part of an inter-rater reliability experiment. For the sake of simplicity, I assume that both doctors are the only raters of interest, while the 10 patients make up the target patient universe from which a sample of 4 subjects (i.e patients) must be selected. A simple selection protocol is described in Table 1.3 and is carried out by first assigning a selection probability of 0.4 (obtained by dividing 4 by 10) to each of the 10 patients in the target patient universe. The next step consists of assigning an arbitrary random number between 0 and 1 to each of the 10 patients in the universe (see "Random Number" column). Only the 4 patients (2, 3, 5, and 7) associated with the 4 smallest random numbers[3] make it to the subject sample of 4. By giving an equal chance of selection to each of the 10 patients in the target universe, this selection procedure increases the likelihood of obtaining a representative sample of the target universe.

The patient selection protocol described in Table 1.3 represents what is known as the sampling plan. It formally links the subjects that participate in the experiment to their home universe. This link ties the agreement coefficient produced by the experiment to the target subject universe that served as basis for articulating the population attribute (or construct). This example shows the importance of the sampling plan when designing an inter-rater reliability experiment. A sampling plan might also be needed for selecting the raters as well if there is a need to sample raters from a target rater population. These design issues will be further discussed in subsequent chapters.

Table 1.3: Sampling of the Patient Universe

Patient	Selection Probability	Random Number	Patient Sample
1	0.4	0.42838	
2	0.4	0.41048	X
3	0.4	0.12451	X
4	0.4	0.97345	
5	0.4	0.15262	X
6	0.4	0.98749	
7	0.4	0.15323	X
8	0.4	0.79993	
9	0.4	0.81326	
10	0.4	0.52606	

[3]Using the 4 largest random numbers instead of the smallest ones will still lead to a valid procedure.

Assignment of Raters to Subjects

In some inter-rater reliability experiments, each rater must score all subjects. However, the scoring as indicated by Axelson and Kreiter (2009) "... is typically a labor-intensive process, scoring duties are often distributed across multiple judges." Even when the scoring of subjects is not labor-intensive, distributing this task across multiple raters is sometimes recommended as a way to minimize costs. If the subjects to be rated are laboratories spread over a vast geographic area, then it is more cost effective to assign a small number of raters (2 or 3) to each laboratory rather than ask each rater to visit all of them. The problem here is that the decision to require each rater to score all subjects or to distribute the scoring duty among multiple raters has some potentially serious effects on the agreement coefficients' precision. If the scoring duty must be distributed across multiple raters then how should that distribution be done? How many raters should be assigned to one subject? What impact would it have on the accuracy of the agreement coefficients?

As a matter of fact, requiring each rater to score all subjects is the most effective design in terms of minimizing the agreement coefficient standard error. That is, for the same number of subjects and raters, it will yield more accurate agreement coefficients than alternative designs that distribute scoring duties among multiple raters. This is due to the fact that assigning different raters to different subjects is a process that can be done in many different ways and therefore creates a new source of variation that can only increase the standard error of an agreement coefficient. In order to streamline this process and be able to perform a statistical evaluation of its impact on the precision of agreement coefficients, the assignment of raters to subjects must be done randomly. Consider an inter-rater reliability experiment where 3 raters must score 5 subjects with the same subject being scored by no more than 2 raters. A convenient and replicable way to implement this is to generate 15 random numbers (3 numbers per subject) to populate Table 1.4. Then the two smallest numbers in each row determine the 2 raters to be assigned to the subject. Note that using the two largest numbers will work as well.

It follows from the above table that raters 2 and 3 will score subject 1, while raters 1 and 3 are assigned to subject 2. A random assignment of raters to subjects as described in Table 1.4 has another advantage, which is to remove any possible bias in the process of deciding which rater scores which subject. Removing this bias is essential for ensuring the integrity of the scoring process.

Table 1.4: Random Assignment of Raters to Subjects

Subject	Rater 1	Rater 2	Rater 3
S1	0.9553	0.8098	0.0209
S2	0.5808	0.7961	0.5669
S3	0.6780	0.0778	0.6728
S4	0.4596	0.1434	0.0758
S5	0.7650	0.3401	0.9624

1.3 Scoring of Subjects/Items

The interaction between raters and subjects produces information about subjects in the form of completed questionnaires, annotated texts or medical records. In general, this raw primary information collected by raters cannot be analyzed when it is presented in the form of narrative text. Even when the information about subjects is collected through a series of yes/no questions, analyzing it can still be problematic.

Consider the questionnaire shown in Figure 1.2. It was designed by a PhD student and aimed at gathering information about newspaper articles that reported on peer-led sex education. Raters were to use it to rate several newspaper articles. The question now is to know how these completed questionnaires can be used to quantify the extent of agreement among raters. As a matter of fact, computing an inter-rater reliability from a batch of completed questionnaires such as this one is an impossible task. It is because that data must be properly coded first. Coding can be a complex and slow activity. However, it must be done first before any analysis can be carried out. I will come back to questionnaire 1.2 later to show what can be done about it before inter-rater reliability can be computed.

Fortunately, there are simple situations in quantitative research where a well-designed scoring rubric is all the raters need before they can observe subjects and assign a numeric code to each of them. For example, consider the rubric shown in Figure 1.3 and which was designed to rate online course syllabi with respect the course potential to create an educational community of inquiry (COI). Note that this rubric is used to score each syllabus on 5 attributes named "Instructional Design for Cognitive Presence," "Technology Tools for COI," "COI Loop for Social Presence," "Support for Learner Characteristics," and "Instructor Feedback for Teaching Presence." For each of these attributes, the rubric describes the conditions that must be met before a specific score shown in the first column can be assigned to a syllabus.

Each of the 5 attributes of this rubric is a variable. The rubric is essentially a tool that provides a detailed description of the relationship between a set of variables retained in a study and a set of values associated with the content of each variable.

Researchers in any quantitative field must become familiar with the notion of "variable[4]" without which no coding and therefore no statistical analysis can be done. Let us go back to Figure 1.2 for a moment. A questionnaire such as this one does not provide the researcher with well defined variables with their respective values that can be used for creating a coding rubric. The variables and their values will have to be defined so the newspaper articles can be assigned a numeric code that can be used for analysis. Several claims can be checked, and for each checked claim additional information may be provided. One possible way to resolve this problem is to use the rubric shown in Table 1.5. It follows from this table that each claim is seen as a different variable that can take 4 values 0, 1, 2, and 3. If a particular claim is unchecked then the associated variable is assigned a 0 value, and will take value 1 if the claim is checked but no evidence was presented to support. The variable takes values 2 or 3 depending on whether the evidence presented is anecdotal or research-based. The rightmost column contains the score assigned to a particular newspaper. The "Total" row of the table contains in its rightmost cell, the sum of all scores, and will represent the overall newspaper score.

Inter-rater reliability can be calculated separately for each variable, although a global inter-rater reliability can also be calculated based on the total score assigned to one newspaper. The most important idea to remember is the need to identify your subjects, and to define variables that can only take one value per subject before a coding rubric can be defined.

The purpose of coding is to transform raw information to numbers, or to well-defined categories into which subjects can be classified. Note that the field of coding is vast, very diverse and well beyond the scope of this book. In the medical field for example, there are professional coders who are trained to assign specific codes to medical conditions. The techniques presented in this book assume that the researcher is able to assign unique codes to each subject under investigation.

[4]A variable is the opposite of a constant and represents a characteristic or an attribute of interest in a research study associated with a subject and which can take values that vary from subject to subject.

Table 1.5: Scoring Rubric of Newspaper Articles Reporting on Peer-led Sex Education

	Scale				
Claim *(Variables)*	0	1	2	3	Score
Cost	Unchecked	Checked, NoEv[a]	Checked, Anec[b]	Checked, Re[c]	_____
Credibility	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Empowerment	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Naturalism	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Efficacy	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Modelling	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Educator Benefit	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Acceptability	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Outreach	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Reinforcement	Unchecked	Checked, NoEv	Checked, Anec	Checked, Re	_____
Total					_____

[a]NoEv = No Evidence, [b]Anec = Anecdotal Evidence, [c]Res = Research Evidence

Scoring, coding, or rating is an activity that consists of assigning to a subject a label (or score), which is later used to determine what action should be taken about the subject. A score in medical diagnosis for example, is either used to identify appropriate treatment services for the patient or to allow service providers to get paid. Given the immense implication a score has, its accuracy is of critical importance and often a source of controversies. This explains why research in the field of inter-rater reliability has grown considerably in the past few years.

The clarity of a scoring rubric will make the training of raters simpler with direct impact on inter-rater reliability. However, the complexity of scoring rubrics may vary considerably from one application to another. Consider a simple and not so well-designed rubric shown in Table 1.6, aimed at quantifying on a scale of 1 to 4 the extent to which students master a particular concept. While the criteria for assigning a score of 1 are clear, the top 2 scores of 4 and 5 however are potentially a source of disagreement among raters. What is the difference between the two statements "... understanding of concept is clearly evident" and "... understanding of concept is evident"? When something is evident it is evident. Moreover, whether using "logical thinking to arrive at conclusion" is more brilliant than showing "thinking skills to arrive at conclusion" is anybody's guess. This example shows that a better study

POSITIVE CLAIMS What claims does the source make for peer-led sex education? (Select as many as apply, a blank space has been provided for those not fitting the criteria)		**EVIDENCE** Does the source cite evidence to support the claim?		**EVIDENCE SOURCE** What form of evidence does the source cite to support the claim?	
		Yes The source cites evidence to support this claim	**No** The source does not cite evidence to support this claim	**Research** The source cites a form of research e.g. a study to support this claim	**Anecdote** The source cites anecdotal evidence e.g. opinion or experience to support this claim
Cost (Positive) Peer education is cost-effective					
Credibility (Positive) Peer educators have credibility with the target population					
Empowerment (Positive) Peer education is empowering					
Naturalism (Positive) Peer education uses pre-established means of communication					
Efficacy (Positive) Peer educators are more successful than professionals					
Modelling (Positive) Peer educators are positive role models					
Educator Benefit (Positive) Peer education is beneficial to peer educators					
Acceptability (Positive) Peer education is acceptable when other education is not					
Outreach (Positive) Peer education can be used to educate those who are 'hard to reach'					
Reinforcement (Positive) Peers can reinforce learning through ongoing social contact					

Figure 1.2: Questionnaire to Support the Content Analysis of Newspaper Articles Reporting on Peer-led Sex Education.

Low potential for building a community of inquiry	1-9 points
Moderate potential for building a community of inquiry	10-17 points
High potential for building a community of inquiry	18-25 points

Online Community of Inquiry Syllabus Rubric© (Rogers & Van Haneghan, 2016)

Scale	Instructional Design for Cognitive Presence	Technology Tools for COI	COI Loop for Social Presence	Support for Learner Characteristics	Instructor Feedback for Teaching Presence
Low (1 point each)	Instructional design offers **limited** cognitive activities (e.g., no exchange of ideas).	**Limited** technology offering to facilitate a COI (e.g., email & assignment tool).	Communication actions are **limited** to S-T interactions only. No open communication planned.	Learner support and available resources are not identified or **limited**.	Syllabus provides **no** information on format for obtaining instructor feedback. No direct instruction (focusing discussion) mentioned. Instructor offers face-to-face office hours only.
Basic (2 points each)	Instructional design offers **minimum** cognitive activities. Exploration (exchange of ideas) is the only one present. This is at the knowledge level of inquiry.	Technology could **minimally** facilitate a COI (e.g., email, assignment tool, & a forum tool).	Open communication actions provide for **minimum** student-teacher (S-T) and student-student (S-S) interactions.	**Minimum** learner support and available resources are identified (e.g., disability services).	Syllabus provides **minimum** information on format for obtaining instructor feedback. No direct instruction mentioned. Instructor offers face-to-face office hours.
Moderate (3 points each)	Instructional design offers **adequate** cognitive activities such as exploration and integration (connecting ideas). This is at the comprehension level of inquiry.	Technology could **adequately** facilitate a COI (e.g., email, assignment, forum, & collaborative tools for individual or group project sharing with other students).	Open communication actions provide for **adequate** S-T and S-S interactions. Collaboration is encouraged to build group cohesion through words, a point-system, or by example.	**Adequate** learner support and available resources are identified (e.g., disability & remedial services).	Syllabus provides **adequate** information on feedback format. Text-based direct instruction is mentioned (or live lecture for blended course). Instructor offers online office hours.
Above Average (4 points each)	Instructional design offers **ample** cognitive activities such as exploration, integration, and resolution (applying new ideas). This is at the application level of inquiry.	Technology could **amply** facilitate a COI (e.g., email, assignment, forums, collaborative tools, & synchronous meeting tools).	Open communication actions provide for **ample** S-T and S-S interactions and opportunities for student-led moderation of forums. Collaboration is required to build group cohesion and a rubric and guidelines are provided.	**Ample** learner support and available resources are identified and offered (e.g., disability, remedial services, & strategies).	Syllabus provides **ample** information on feedback format with prompt turnaround time. Multi-modal direct instruction is mentioned (e.g., narrated PowerPoint, video tutorial, or podcasts). Instructor offers online office hours.
Exemplary (5 points each)	Instructional design offers **extensive** cognitive activities such as exploration, integration, resolution, and triggering events (analysis, synthesis, evaluation).	Technology could **extensively** facilitate a COI (e.g., email, assignment, forum, collaborative tools, & synchronous meeting tools) in innovative ways.	Open communication actions provide for **extensive** S-T, S-S, and student-participant/expert (S-P/E) interactions and opportunities for student-led moderation of forums. Collaboration is required to build group cohesion and a rubric and guidelines are provided.	**Extensive** learner support and available resources are identified (e.g., disability, remedial services, strategies, scaffolding of assignments, or lab component).	Syllabus provides **extensive** information on feedback format and prompt turnaround time. Multi-modal direct instruction is mentioned. Instructor offers online office hours and social media venues for classroom interactions.
Subtotal	_____Points	_____Points	_____Points	_____Points	_____Points
Total	_____Points	**Directions:** This is a 5-point rubric with the following scales: low, basic, moderate, above average, and exemplary. The points awarded determine the course's potential of developing an online community of inquiry (COI).			

Figure 1.3: Online Community of Inquiry (OCOI) Syllabus Rubric©

design and more rater training may not be sufficient to achieve an acceptable inter-rater reliability. The scoring rubric must also be well written.

Table 1.6: Example of Scoring Rubric

Score	Description
4	• Student's understanding of the concept is clearly evident • Student uses effective strategies to get accurate results • Student uses logical thinking to arrive at conclusion
3	• Student's understanding of the concept is evident • Student uses appropriate strategies to get accurate results • Student shows thinking skills to arrive at conclusion
2	• Student has limited understanding of the concept • Student uses strategies that are ineffective • Student attempts to show thinking skills
1	• Student has a complete lack of understanding of the concept • Student makes no attempt to use a strategy • Student shows no understand

A far more complex "scoring rubric" is the Diagnostic and Statistical Manual of Mental Disorders (DSM), the standard classification of mental disorders used by mental health professionals in the United States. It is a very elaborate set of guidelines, which includes diagnostic criteria indicating symptoms that must be present to qualify for a particular diagnosis. On these criteria, the American Psychiatric Association issued the following warrant:

> *While these criteria help increase diagnostic reliability (i.e., the likelihood that two doctors would come up with the same diagnosis when using DSM to assess a patient), it is important to remember that these criteria are meant to be used by trained professionals using clinical judgment; they are not meant to be used by the general public in a cookbook fashion.*

Although this particular scoring rubric was drafted by a plethora of internationally-known experts, its magnitude and complexity still require a formal inter-rater reliability experiment to be conducted with a representative sample of the group of professionals expected to use it.

The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT®) is a clinical terminology developed by the College of American Pathologists (CAP)

to effectively classify electronic health records. This coding system allows clinical information to be recorded using identifiers that refer to concepts, and covers a wide range of clinical specialties, disciplines and requirements, and is now owned and maintained by the International Health Terminology Standards Development Organisation (IHTSDO). It contains over 100,000 diagnosis concepts and requires considerable training to be used effectively. One may visit the webpage https://www.snomed.org/ to get a sense of the magnitude of this gigantic scoring rubric. Another challenging activity related to SNOMED is the often needed mapping between SNOMED and DSM, which is based on the International Classification of Diseases (ICD) and used for health management, reimbursement and resource allocation decision-making. Inter-rater reliability is not just important for coders using SNOMED-CT or the ICD coding systems, but is equally important for those performing mapping activities between the two systems.

1.4 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 section 1.2.1, and assume that 75 triage nurses have been identified in the target nursing population (in a more formal way, I would say $R = 75$), and 1,000 patients in the patient population (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 score all 1,000 patients by assigning one of 5 priority levels to each of them.

Suppose for simplicity 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 the 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 priority k to patient i 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 $P_{a|i,k}$ over 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 formulated mathematically as follows:

$$P_a = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{5}\frac{R_{ik}(R_{ik}-1)}{R(R-1)}. \tag{1.4.1}$$

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 is logical. The ratings that are collected from experiment can only take 5 discrete values. Therefore the notion of agreement is straightforward and is defined by the assignment of the same priority level by 2 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 another example where a psychiatrist classifying his patients into one of five categories named "Depression", "Personal Disorder", "Schizophrenia", "Neurosis", and "Other.[5]" This five-item scale is *Nominal* since no meaningful ordering of these categories 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.

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

1.4.1 *Nominal Ratings*

With nominal scales, 2 raters agree when the ratings they assign to the same subject are identical and disagree 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 small 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 of 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 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 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. One of the best known measures of internal consistency is Cronbach's alpha coefficient (Cronbach, 1951) discussed in Part III of the book.

1.4.2 *Ordinal Ratings*

When categories are 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 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 discussed later in this book.

1.4.3 *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[6]. 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 due to raters to the variation in ratings due to random errors.

1.5 Different Reliability Types

The inter-rater reliability literature is full of various notions of reliability. Different terms are used to designate similar concepts (e.g. intra-rater reliability and test-retest reliability), and the same 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

[6]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 know is that the intensity of the heat is substantially higher at 70^0F than at 35^0F. By how much? Twice? Three times? We cannot say it with certainty.

section reviews some uncommon reliability types often encountered in the literature, and discusses the relationship between reliability and validity.

1.5.1 *Undefined Raters and Subjects*

In some studies, identifying which entity represents a subject and which one represents a rater may not be straightforward. 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 1.7. 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 experiment involves 150 raters rating a single subject[7]. 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 1.7: 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

What we should learn from this example is that, sometimes the researcher needs to construct the notions of rater and subject, before the study is designed.

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

1.5.2 *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 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 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 seen as the gold standard. Fleiss (1971), or Light (1971) among others studied such conditional analyzes.

1.5.3 *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 subjects. This topic is discussed in part III of this book.

In this book, the discussion on internal consistency evaluation 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).

1.5.4 *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

these "true" scores. Do raters who achieve high inter-rater reliability also agree on the correct category, when it exists? Or do they often agree on the wrong 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 validity. Validity is measured with special validity coefficients to be discussed in chapter 8.

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 considered 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.

1.5.5 *Multivariate Inter-Rater Reliability*

In many inter-rater reliability experiments, subjects are scored on two factors ore more. Consider for example a study where 3 psychiatrists interview a sample of patients and rate them on a five-point scale on each of the following 3 factors:

- $x_1 = lack\ of\ concentration;$
- $x_2 = despondency;$
- $x_3 = anxiety.$

With rating data available for 3 factors, the researcher has the option to evaluate the extent of agreement among psychiatrists separately for each factor, or evaluate an overall agreement among psychiatrists using all ratings from all 3 factors. The question now is how to compute the global agreement coefficient when ratings are available on several factors, which may even be based on different scales. The general approach I would recommend depends on the type of rating data that you have. For nominal and ordinal scales, I would recommend computing an agreement coefficient for each individual factor first, before averaging them over all factors. Ideally, you would avoid averaging agreement coefficients that are associated with highly correlated factors. Because, such an operation will likely result in an overall agreement

coefficient with high variance.

For quantitative type ratings (interval and ratio), I would recommend creating a composite score based on principal components analysis[8](PCA) as a first step. The next step is to use the first principal component as the composite score, as it is the one that carries the most information about your dataset. The single composite score associated with each subject will then be used to compute the Intraclass Correlation Coefficient (ICC). Note that the treatment of ICCs is out of the scope of this book, but is covered in Gwet (2021) for those interested in pursuing this topic.

1.6 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 the context of inter-rater reliability and as presented in this book is the law of probability that governs the selection of raters from the target rater population and the selection of subjects from the target subject population. I expect to be able to pick any rater from the rater population (or any 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 formulated mathematically with respect to both the rater and subject populations. This mathematical expression represents the population parameter or the estimand or the inter-rater reliability parameter to approximate using actual ratings from the reliability experiment. Note that the expression showing how 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,

[8]The Principal Component Analysis is a dimension-reduction statistical technique that reduces a large set of variables to a smaller set of variables called the "Principal Components" that still contains most of the information of the large set. Because the first principal component accounts for the largest portion of the rating variation, I will retain it as the composite score.

the estimand, and the estimator. Most published papers on inter-rater reliability tend to limit the discussions to the estimator that generates 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 produced by the estimator changes each time the raters or subjects who participate in the study change. 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 evaluated. 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 6 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.

1.7 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 present agreement coefficients along with their associated standard errors and to generalize them in such a way that datasets with missing ratings can be analyzed. I always start the presentation of new methods with a simple scenario involving 2 raters and a two-level nominal scale, before expanding it to the more general context of 3 raters or more, and to ordinal, interval, or ratio ratings. This book is divided into 4 parts:

- Part I has 2 chapters: the current introductory chapter 1 and chapter 2, which presents various ways of organizing rating data before analysis.
- Part II is made up of 5 chapters, from chapter 3 through chapter 7. Chance-

corrected Agreement Coefficients (CAC) are discussed in these chapters for nominal, ordinal, interval and ratio ratings. Also discussed in Part II chapters, are the theoretical foundation of many agreement coefficients in chapter 5, the framework for statistical inference in chapter 6 and the benchmarking methods for qualifying the magnitude of agreement coefficients in chapter 7.

- Part III covers important miscellaneous topics in 2 chapters. Chapter 8 is devoted to the study of agreement coefficients conditionally upon specific categories as well as agreement with the gold standard. Chapter 9 on the other hand, presents special agreement coefficients such as the inter-annotator agreement, and covers various additional methods that enhance the analysis inter-rater reliability data in many different ways.

- Part IV of the book includes appendices A, B, and C. 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. Appendix C contains some datasets used for modeling agreement coefficient variances in chapter 6 to determine the optimal number of subjects when planning an inter-rater reliability experiment.

Part II of this book starts in chapter 3 with a critical review of several agreement coefficients proposed in the literature for analyzing nominal ratings. This review includes Cohen's kappa coefficient, its generalized versions to multiple raters, Gwet's AC_1, Krippendorff's alpha, or Brennan-Prediger coefficient among others. In chapter 4, I show that with the use of proper weights, the agreement coefficients discussed in chapter 3 can be adapted to produce a more efficient analysis of ordinal and interval ratings. In chapter 5, I use the AC_1 coefficient proposed by Gwet (2008a), and Aickin's alpha coefficient of Aickin (1990) as examples to show how agreement coefficients can be constructed to achieve specific analytic goals, 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 6, I introduce the basic principles of statistical inference in the context of inter-rater reliability assessment. I stress out the importance of defining the target population of raters and the target population of subjects prior to selecting the subjects and raters that will be part of the inter-rater reliability experiment.

I did not use the model-based approach to statistical inference of Kraemer et al. (2002) and others. Instead, I used the design-based approach to statistical inference, which I introduced in the field of inter-rater reliability assessment 2 decades ago. The design-based approach to statistical inference is widely used in sample surveys and relies on the random selection of subjects and raters from their respective target populations. It is the randomization of the subject selection process that forms the

basis for statistical inference. No other assumption is made regarding the rating process.

Chapter 7 addresses the important problem of benchmarking inter-rater reliability coefficients. The problem consists of determining different thresholds (or benchmarks) that could be used to interpret inter-rater reliability as poor, good, or excellent. I review different benchmark scales proposed in the literature before describing a more efficient benchmarking model initially introduced in the 4^{th} edition of this book and which is specific to each inter-rater reliability coefficient.

Part III of this book focuses on miscellaneous topics and starts with chapter 8, which focuses on the analysis of inter-rater reliability coefficients conditionally upon the subject membership to specific categories. Chapter 8 also treats validity coefficients, which quantify the extent of agreement with the gold standard when available. 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 the traditional chance-corrected measures. For example, the extent of agreement among annotators in the fields of Natural Language Processing (NLP) or Computational Linguistics. This problem is addressed in section 9.2 of chapter 9. Section 9.3 of chapter 9 deals with the problem of testing 2 agreement coefficients (correlated or uncorrelated) for statistical significance. This problem is often important when comparing 2 inter-rater reliability studies. In section 9.4, I address the problem of evaluating the extent of agreement among 3 raters or more when the same subject can only be rated twice. Section 9.5 shows you how to quantify the impact (or influence) of individual raters on an agreement coefficient. Section 9.6 deals with the notion of intra-rater reliability, which measures raters' ability to reproduce their own ratings. A popular measure of association in the field of item analysis and known as Cronbach's alpha is discussed in section 9.7.

1.8 Choosing the Right Method

How your ratings should be analyzed depends on the type of data you have collected, and on the ultimate objectives of your analysis. I previously indicated that your ratings may be of nominal, ordinal, interval, or ratio type. Figure 1.4 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 1.4 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 in-

terest 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 III of this book. You may review the content of these chapters if your analysis needs are out of the ordinary.

Figure 1.4 indicates that if you are dealing with ratio or interval ratings, then you can use one of the chance-corrected agreement coefficients of chapters 3 or 4 only if these ratings are predetermined before the experiment is conducted. Otherwise, you will need the intraclass correlation coefficients, which are discussed in another book by Gwet (2021). Ratings are predetermined if the researcher knows all the values that can be assigned to a subject before the beginning of the experiment. These predetermined ratings are an integral part of the experimental design. However, if the rating is the subject's height, or weight whose values can be determined only after the measurements are taken, then the intraclass correlation is what I recommend.

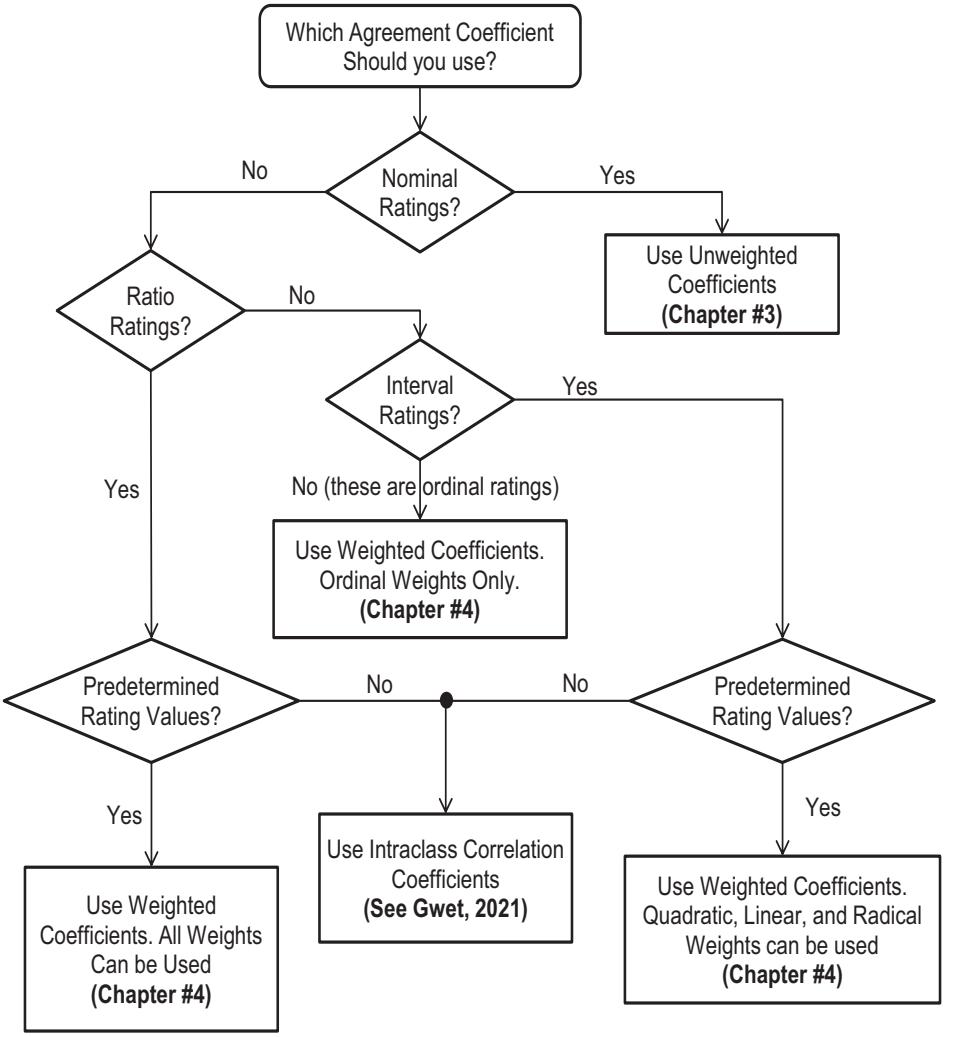


Figure 1.4: Choosing an Agreement Coefficient