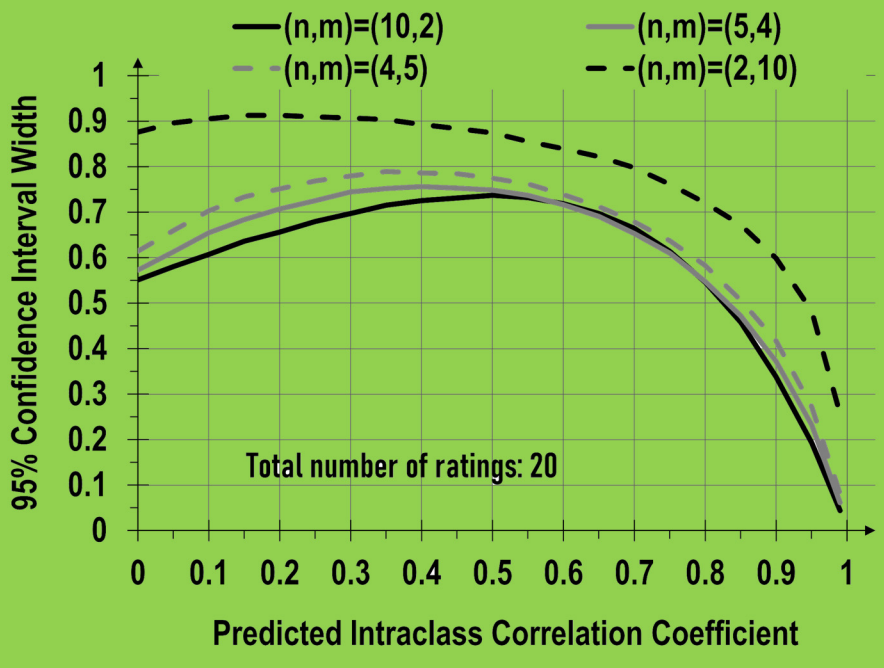
Handbook of Inter-Rater Reliability

Fifth Edition

The Definitive Guide to Measuring the Extent of Agreement Among Raters

Volume 2

Analysis of Quantitative Ratings



Kilem L. Gwet, Ph.D.

HANDBOOK OF INTER-RATER RELIABILITY

Fifth Edition

Volume II

Analysis of Quantitative Ratings

HANDBOOK OF INTER-RATER RELIABILITY

The Definitive Guide to Measuring the Extent of Agreement Among Raters

Fifth Edition

Volume II

Analysis of Quantitative 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 2: Analysis of Quantitative 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 978-1-7923-5464-9

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 2, focuses on ICC methods whereas volume 1 is devoted to CAC 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.

Volume 2 of the Handbook of Inter-Rater Reliability 5th edition contains 2 new chapters not found in the previous editions, and updated versions of 7 chapters taken from the 4th edition. 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 3 is an introductory chapter and an update of chapter 7 in the 4th edition. In addition to providing an overview of the book content similar to that of the 4th edition, this chapter introduces the new multivariate intraclass correlation not covered in previous editions.

- Chapter 4 covers intraclass correlation coefficients in one-factor models and is an update of chapter 8 in the 4th edition. In this 5th edition, there is a separate section devoted to sample size calculations. Two approaches to sample size calculations are now offered: the statistical power approach and the confidence interval approach.

- Chapter 5 covers intraclass correlation coefficients under the random factorial design, which is based on a two-way Analysis of Variance model where the rater and subject factors are both random. This is an update of chapter 9 in the 4th edition. Errors identified in the previous edition have been corrected and section 5.4 on sample size calculations has been expanded substantially. Again, to perform sample size calculations, researchers can now choose between the statistical power approach based on the Minimum Detectable Difference (MDD) and the confidence interval approach based on the target interval length to achieve.

- Chapter 6 covers intraclass correlation coefficients under the mixed factorial design, which is based on a two-way Analysis of Variance model where the rater factor is fixed and the subject factor random. This is an update of chapter 10 in the 4th edition. Most sections have been rewritten to address numerous comments received from readers of the 4th editions, and section 6.4 on sample size calculations has been expanded substantially, with the addition of the statistical power approach.

- Chapter 7 is new and covers Finn's coefficient of reliability. The decision to cover Finn's coefficient is justified by the fact that traditional intraclass correlations may not be applicable if the subject population is homogeneous. ICCs are built upon specific statistical models that may no longer be valid if the sample subjects do not constitute a representative sample from the target population they are supposed to have been selected from.

- Chapter 8 entitled "Measures of Association and Concordance" is an updated version of chapter 12 in the 4th edition. It covers various association and con-

cordance measures often used by researchers. The main novelty here is the discussion of Lin's concordance correlation coefficient and its statistical properties.

- Chapter 9 is new and covers 3 important topics. Section 9.2 discusses the benchmarking of ICC estimates, which consists of qualifying the strength of agreement using a probabilistic-based procedure. Section 9.3 describes a graphical approach for exploring the influence of individual raters in low-agreement inter-rater reliability experiments. The technique can help single out problem raters who may need further training to catch up with the group. Section 9.4 addresses the important problem of multivariate intraclass correlation. This problem must be dealt with whenever you must analyze multiple correlated quantitative variables with the intraclass correlation coefficients.

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.

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 limited 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 provides primarily a broad view 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 out 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 presented, and the practical context in which they can be used described. Also discussed in this chapter, are the types of reliability data the researcher may collect, and how they affect the way the notion of agreement is defined. I later insist on the need to analyze inter-rater reliability data according to the principles of statistical inference in order to ensure the findings can be projected beyond the often small samples of subjects and raters that participate in a reliability experiment. Figure 1.5 depicts a flowchart that summarizes the process for 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 participants 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 the same 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 raters are key components 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 "gold standard" measurements that researchers agree to use as reference. I will also discuss the notion of validity in this book.

I am well 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 peculiar 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 un-

related 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 definitions 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 by mixing reliability of categorical ratings and reliability of quantitative ratings.

1.1.1 *Categorical versus Quantitative Ratings*

How an inter-rater reliability analysis is conducted depends profoundly upon the type of rating data that will be collected. Your ratings may be categorical or quantitative. It is absolutely essential to be clear at the planning stage whether you will be dealing with categorical ratings or whether you will be dealing with quantitative ratings. The categorical rating is essentially a label that the rater assigns to each subject to define its membership category. A quantitative rating on the other hand, refers to a numeric value that is created by the rater during the rating phase and assigned to the subject. Specific numeric values, which are predetermined before the rating of subjects begins should be treated as nominal ratings. This book focuses on quantitative ratings. Nevertheless, the distinction between these 2 types of ratings must be clarified to avoid any possible confusion. Readers who are interested in categorical ratings needs to read Gwet (2021)

As an example of quantitative ratings, 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. In other words, would they produce the same measurements on the same shoulders? These measurements are numerical values that can only be determined by the raters themselves after the experiment had been performed. The notion of agreement in this case must well understood. As a matter of fact, 2 medical devices from two manufacturers are unlikely to yield two identical values when used on the same subject, even if both are very accurate. Therefore, there is a 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 a sign of agreement and a large variation a sign of disagreement.

Categorical ratings are mainly used for classification purposes and can be either nominal or ordinal[1]. The reliability of a classification process can be established by asking two individuals referred to as raters, to independently perform this classification with the same set of objects. The extent to which these two categorizations coincide represents what is referred to as inter-rater reliability. For nominal ratings,

[1]Ordinal ratings can be ranked in increasing or decreasing order (e.g. small, medium, large), while nominal ratings cannot (e.g. Red, Blue, Yellow).

only an exact match qualifies as an agreement. For ordinal ratings however, exact and partial matches are often considered. A more detailed discussion of these topics can be found in Gwet (2021).

Whether you are dealing with categorical or quantitative ratings, if inter-rater reliability is high then the raters can be used interchangeably without the researcher having to worry about these ratings being affected by a significant rater bias. Interchangeability of raters is what justifies the importance of inter-rater reliability studies. If interchangeability is guaranteed, then the ratings assigned to subjects can be used with confidence without asking which 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.

1.1.2 *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.

Inter-rater reliability analysis of quantitative ratings has been applied multiple times and successfully to neuroimaging[2]. Chen et al. (2017) provides a detailed review of several such applications. Most neuroimaging data are large and of quantitative nature and are typically analyzed with various forms of Analysis of Variance models similar to those that are extensively discussed in this book. For each individual several three-dimensional scans are often collected, each of which containing brain activity measurements at various spacial locations of the brain known as *voxels*[3]. Bowman et al. (2007) provides an interesting account of the application of statistical methods in functional neuroimaging.

[2]This branch of medical imaging focuses on the brain and is often used for clinical diagnoses and for detecting brain lesions or tumors.

[3]A voxel is small cubic region of the brain, which is only a few cubic millimeters in size.

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

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

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

Jung (2003) summarizes the efforts that have been made in this area.

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

- Different coders evaluate the medical records of patients and assign one or multiple codes from a disease classification system. Unlike the typical inter-rater reliability experiment where a rater assigns each subject to one and only one category, here coders can assign a patient do multiple disease categories. For example, Leone et al. (2006) investigated the extent to which neurologists agree when assigning ICD-9-CM[4] codes to patients who have suffered from stroke. The challenge here is to define the notion of agreement in this situation where one coder assigns 3 codes to a patient, while a second coder assigns a single code to the same patient.

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

  In the context of inter-rater reliability, multiple coders may be asked to independently do the mapping between two systems so that the reliability of the mapping process can be evaluated.

1.1.3 *Objectives of an Inter-Rater Reliability Analysis*

When defining the notion of inter-rater reliability, there will always be a degree of impreciseness in what we really mean by it. This issue is acknowledged by Eckes (2011) when he says "... even if high interrater reliability has been achieved in a given assessment context exactly what such a finding stands for may be far from

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

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

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

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

- Other researchers want to summarize the extent of agreement among raters with a single number that quantifies the extent of agreement among raters. For categorical ratings, possible agreement coefficients include Cohen's kappa, Gwet's AC_1 among others extensively discussed in Gwet (2021). For quantitative ratings on the other hand, which are the main focus of this book, agreement coefficients will essentially be in the form of an intraclass correlation coefficients under a variety of ANOVA models.

  After computing the agreement coefficient, subsequent analyses could be of interest. These 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. These types of analyses will be discussed in this book for quantitative ratings.

When designing an inter-rater reliability study, it is essential to set clear analytical goals. What questions do you need to answer? What statistics do you need to

produce? Again, to quantify inter-rater reliability you may not need a universal and rigourous definition. A broad and clear description of study goals is usually all it takes to select the most appropriate methods.

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 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? Or scoring duties were 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.

An adequate presentation of the inter-rater reliability problematic cannot consist of detailed information, and computation procedures alone. It must also provide a proper and global view of the essential nature of the problem as a whole, as depicted in figure 1.1.

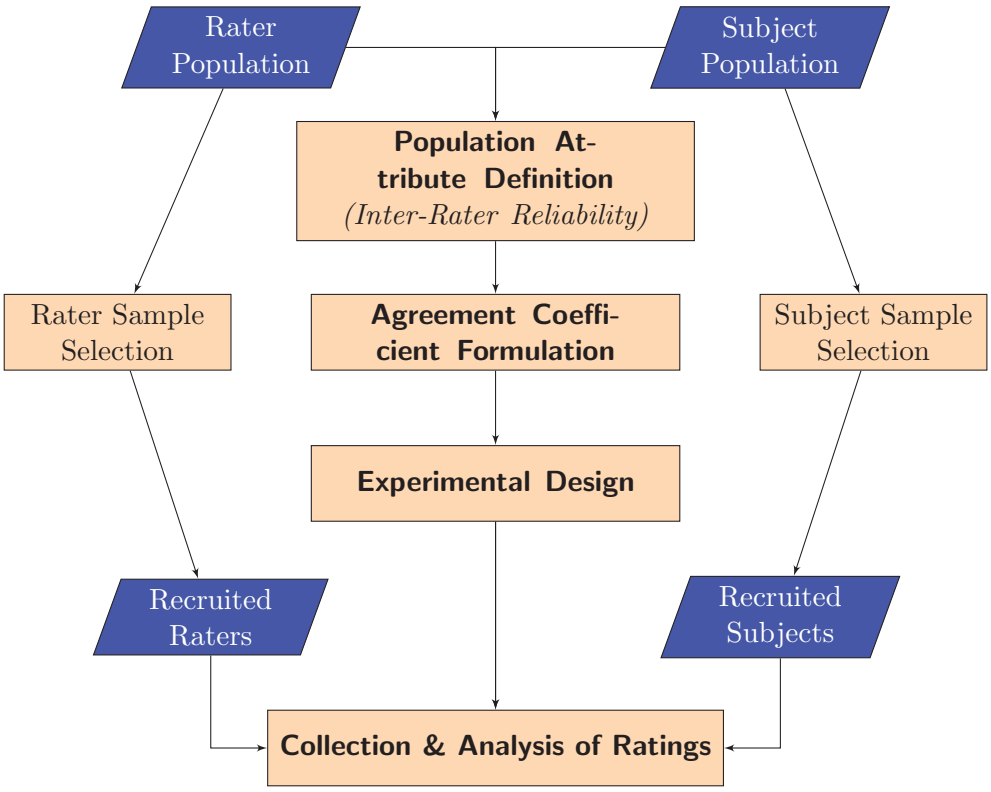


Figure 1.1: Phases of an Inter-Rater Reliability Study Design

1.3 Target Rater & Subject Populations

In a recent past, I 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 by the way? Are they the women who visit a particular obstetric unit? Should other obstetric units be considered as well? Which ones?

- Who are the triage nurses targeted by this study? I am not referring to the triage nurses who may eventually participate in the study. Instead, I am referring to all triage nurses whose lack of proficiency in the triage process may have adverse effects on our predefined target population of pregnant women. They represent our target population of triage nurses. The possibly large number of nurses in this triage nursing population is irrelevant at this point, since we do not yet worry about those who will ultimately be recruited to participate in the study. Recruitment for the study will be addressed at a later time during the experimental design phase.

- In the ideal scenario where each triage nurse in the nursing population was to participate in the prioritization of all pregnant women in the target subject population, we want the extent of agreement among the nurses to be very high. But there is another important outstanding problem we need to address. If the triage patients must be classified into one of 5 possible priority categories, then we need to recognize that even after a maternal and fetal assessments are performed on the patient, two triage nurses may still be uncertain about the correct priority level the patient should be assigned to. This undesirable situation of uncertainty could lead them to assign a priority level 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 badly as we may crave for high agreement among the nurses, this 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 number ultimately produced by the inter-rater reliability experiment.

Proving that the initial formulation of an inter-rater reliability coefficient as a population parameter is useful for studying the population attribute agreed upon can be a delicate task. What we can do in practice is to conduct an experiment and actually compute an estimated agreement coefficient. If the experiment is well designed, this estimated agreement coefficient will be a good approximation of the population inter-rater reliability coefficient. The interpretation of its value in conjunction with the subject-matter knowledge for the tasks the raters are accomplishing will help determine if the way inter-rater reliability is quantified is acceptable. We will further discuss some technical issues pertaining to the formulation of agreement coefficients in section 1.4.

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 to be used as a concrete representation of the inter-rater reliability concept. Quantitative ratings differ from categorical ratings in the way the notion of agreement is handled. Categorical variables can take a predetermined and limited number of values. Therefore, agreement occurs when 2 raters assign the exact same value to the same subject. Quantitative variables on the hand, take their values from a continuum. For all practical purposes it means the different numeric values that a quantitative variable can possibly take are not predetermined before the experiment and are often defined by 2 parameters:

- The range of possible values that can be assigned to the quantitative variable.
- The likelihood for each value to be assigned to the quantitative variable representing the associated law of probability.

Consider for example, a metal working shop where several ovens are used to heat metal specimens. A sample of 3 ovens may be used on 2 metal specimens to produce the following temperatures: 491.50, 498.30, 498.10,493.50, 488.50,471.85. First,

the exact temperature an oven will produced on a given metal specimen cannot be predetermined. Therefore, these are numbers you will only see after the experiment. Second, it is unlikely that 2 ovens will produce the exact same temperature level and this "absolute" agreement cannot be the basis for evaluating the extent agreement among different ovens (or raters in general). Consequently agreement among raters is evaluated for quantitative ratings by comparing the portion of total variance that can be attributed to raters to the portion of total variance attributable to subjects. If the subject variance dominates the raters variance by a significant margin, one concludes that the rater agreement is good.

When developing a statistical procedure, you also need a framework for statistical inference. This framework is expected to provide the law of probability needed to evaluate the statistical precision associated with the agreement coefficients. For quantitative ratings, the general approach that dominates the literature consists of describing rating data with one of the many possible Analysis Of Variance (ANOVA) models. However, sometimes the conditions underlying the ANOVA model may not be satisfied, in which case alternative ad hoc approaches have been considered.

1.4.1 *ANOVA-Based Agreement Coefficients*

Throughout this book, a quantitative rating assigned to subject i by rater j on k^{th} occasion[5] will be denoted by y_{ijk}. Assuming the inter-rater reliability experiment is based on n subjects, r raters and m measurements per subject, the simplest ANOVA model often used to describe this data is the one-way ANOVA model with random subject factor and formulated as follows:

$$y_{ijk} = \mu + s_i + e_{ijk}, \text{ for } \begin{cases} i = 1, \cdots, n, \\ j = 1, \cdots, r, \\ k = 1, \cdots, m. \end{cases} \tag{1.4.1}$$

where μ is the theoretical mean, s_i the subject effect assumed to follow the Normal probability distribution with mean 0 and variance σ_s^2, and e_{ijk} the error effect assumed to follow the Normal probability distribution with mean 0 and variance σ_e^2. Moreover, the 2 random variables s_i and e_{ijk} are assumed to be uncorrelated.

> *The Intraclass Correlation Coefficient (ICC) in this context, is defined as the correlation coefficient between the 2 ratings y_{ijk} and $y_{ij'k}$ assigned to the same subject i on the same occasion k. You will later in subsequent chapters that this correlation coefficient can be formulated as follows:*
>
> $$cov(y_{ijk}, y_{ij'k}) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}.$$

[5] Sometimes, raters have the opportunity to rate the same subject multiple times on different occasions. In this case, 2 indices are needed to label each rating.

Once your experimental data have been collected, computing the ICC under model 1.4.1 amounts to estimating the subject and error variance components. This summarizes the process of formulating agreement coefficients for quantitative ratings Note that model 1.4.1 is only the simplest of many ANOVA models that will be used throughout this book to model quantitative ratings. You will see in subsequent chapters that more elaborate models are recommended in some special situations.

I like to bring your attention a key often overlooked assumption underlying ANOVA models such as that described by equation 1.4.1. The ratings associated with each rater are assumed to represent a random sample from a population that follows the Normal distribution. The main practical implication of this assumption is that some significant subject variation is expected in your data. This will not always be the case in practice. When this assumption is violated by a very homogeneous subject sample, the validity of ICC as a measure of agreement between raters can be compromised. In this case, some alternative approaches may be needed.

1.4.2 *Non-ANOVA Approaches to Agreement Coefficients*

I indicated in the previous section that the extent of agreement among raters evaluated with the Intraclass Correlation Coefficient (ICC) is calculated as the ratio of the subject variance over the total variance, which also includes the rater variance. Since the subject variance is in the numerator and the rater variance in the denominator, ICC will yield a high agreement coefficient only if the subject variance exceeds the rater variance by a significant margin. What if the subject variance is very small due to the subject sample being too homogeneous? Under this scenario, it is near impossible to obtain a high agreement coefficient, because the already small subject variance may never exceed the rater variance by a significant margin.

A homogeneous subject sample is a clear violation of the basic underlying ANOVA assumption of Normality. With this type of data, several authors have recommended alternative ways of formulating agreement coefficients, which will be reviewed in subsequent chapters. These alternative approaches are based on the ratio of the subject variance to its maximum value. Although this formulation is not derived from a theoretical model as the ICC, it appears to provide a valid solution when the ANOVA underlying assumptions are violated.

1.4.3 *Multivariate Inter-Rater Reliability*

In most complex inter-rater reliability experiments, subjects are scored on two factors ore more. As previously discussed, it is very common in the field of neuroimaging that hundreds of brain activity measurements are collected at various locations of the brain for the same patient. These specific brain locations are known as

voxels. When the study involves many patients, it will eventually produce a staggering number of observations. Note that several univariate inter-rater reliability analyzes can always be conducted separately for each variable. However, one cannot ignore the basic fact that all variables associated with the same individual are highly correlated. Ignoring these correlations makes it difficult to develop global measure of reliability as discussed in chapter 9. In fact, averaging individual agreement coefficients in this context will often yield a highly unstable global coefficient. Here is the problem that motivates the need to develop a multivariate approach to inter-rater reliability.

Liu et al. (2019) discuss an interesting pilot study aimed testing the reliability of a three-dimensional facial camera. The problem is to compare specific facial measurements evaluated clinically to those evaluated by a computerized analysis of images produced by the three-dimensional camera. To conduct this study, 12 edentulous participants (7 women and 5 men) were selected. Digital images for facial reconstruction were captured. The following 8 extraoral soft-tissue facial landmarks were identified:

- Right Outer Canthus (OCR)
- Left Outer Canthus (OCL)
- Right Cheilion (CmR)
- Left Cheilion (CmL)
- Pronasale Nostril Tip (NT)
- Subnasale (SN)
- Philtrum (PT)
- Gnathion (GN)

These landmarks can be identified in Figure 1.2. The measurements of interest are the absolute value of the inter-landmark distances of OCR-OCL, OCR-CmR, OCL-CmL, OCR-GN, OCL-GN, CmR-CmL, NT-GN, and SN-GN. These distances were measured clinically and then on the 3D digital reconstructions. The intraclass correlation coefficient was then applied to evaluate the reliability of digital measurement and inter-examiner reliability.

Figure 1.2: Facial landmarks for testing the reliability of a three-dimensional camera)

For the sake of computing a global inter-rater reliability coefficient for quantitative ratings, I recommend in chapter 9 to first create a composite score using principal component analysis[6](PCA), and to use the most interesting principal component as the composite score. The single composite score associated with each subject will then be used to compute the overall inter-rater reliability coefficient.

1.5 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[7]) 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 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 the clinical and research diagnosis approaches then the two approaches would represent the only two raters of interest. Therefore there would

[6]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 in 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.

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

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 and type 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 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 the *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 seen by the researcher as a "good" representation (with respect to a number of characteristics) of the target subject 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 **??**.

1.5.1 *Sample Selection*

The sample of raters or subjects should ideally be probabilistic. That is the selection of each rater or each subject from their respective target universes must be carried out according to a random process that gives each unit of interest a chance of being chosen for the study. As an example, suppose that 4 of the 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 the two doctors are the only two raters of interest, while the 10 patients make up the target subject universe from which a sample of 4 subjects (i.e patients) must be selected. A simple selection protocol is described in Table 1.1 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 subject 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 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 dramatically increases the likelihood that the inter-rater reliability experiment will be based on a sample of subjects that is representative of the target universe it was selected from.

The patient selection protocol described in Table 1.1 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 in the design of an inter-rater reliability experiment. The same thing that is said about subjects here can be said about 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 1.1: 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	

1.5.2 *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 in 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 scoring duties among multiple raters has some potentially serious impact on the precision with which agreement coefficients are calculated. If scoring duties 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. 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 generate 15 random numbers (3 numbers per subject) as shown in Table 1.2. Then the two smallest numbers determine the 2 raters to be assigned to the subject. Note that using the two largest numbers will work as well.

Table 1.2: 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

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 1.2 has another advantage, which is to remove any possible bias in the process of deciding what rater scores what subject. Removing this bias is essential for ensuring the integrity of the scoring process.

1.6 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 was collected through a series of yes/no questions, analyzing it can still be problematic.

Consider the questionnaire shown in Figure 1.3. 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 and foremost before any analysis can be carried out. I will come back to questionnaire 1.3 later on to show what can be done about it before inter-rater reliability can be computed.

Fortunately, there are simple situations in quantitative research where all raters need is a well-designed scoring rubric before they can observe subjects and assign a numeric code to each of them. For example, consider the rubric shown in Figure 1.4, 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 scores 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 their values.

Researchers in any quantitative field must become familiar with the notion of "variable[8]" without which no coding and therefore no statistical analysis can be done. Let us go back to Figure 1.3 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

[8] 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.

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 1.3. 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 the associated variable is assigned a value of 0, and will take value 1 is 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.

Table 1.3: Scoring Rubric of Newspaper Articles Reporting on Peer-led Sex Education

Claim	Scale				Score
(Variables)	0	1	2	3	Assigned
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

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.

Scoring, coding, or rating is an activity that consists of assigning to a subject

Table 1.4: 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 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 much simpler with a 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 1.4 and aimed at quantifying on a scale from 1 to 4 the extent to which students master a particular concept. While the criteria for assigning a score of 1 are fairly 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 design, and more training for raters may not be sufficient to achieve an acceptable inter-rater reliability. The scoring rubric must also be written in an effective manner.

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 guide-

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 1.3: 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 1.4: Online Community of Inquiry (OCOI) Syllabus Rubric©

lines, 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 the specific group of raters who must 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 ICD coding systems, but is equally important for those performing mapping activities between the two systems.

1.7 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 for quantitative ratings, and as presented in this book is the law of probability that governs the variation of ratings. For any given rater and any given subject, I should be able to evaluate the likelihood that the associated rating exceeds a predetermined threshold. I will usually

hypothesize a statistical model to achieve this objective. 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 unknown 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 initially defined as an attribute of the rater population or a theoretical construct, which in turn should be codified mathematically with respect to both the rater and subject populations. This expression represents the population parameter or the estimand or the inter-rater reliability parameter to be estimated using actual ratings from the reliability experiment. Note that the mathematical expression showing how the ratings are manipulated is called the inter-rater reliability estimator. In sequence we have three things to worry about, the attribute, the estimand, and the estimator. Most published papers on inter-rater reliability tend to limit the discussions to the estimator that produces the agreement coefficient. For the discussion to be complete, it must tie the estimator to the estimand and to the attribute.

Note that the inter-reliability coefficient generated by the estimator changes each time the raters or subjects who participate in the study change. It is directly affected by the experimental design. The estimand on the other hand solely depends upon both the rater and the subject populations, and are not affected in any way by the experiment. It may change only if you decide to modify the pool of raters and subjects that are targeted by the study. The attribute is the most stable element of all. It can only be affected if the study objective changes. The discrepancy between the estimator and the estimand is what is known as the statistical error. This one can be and should be estimated. It shows how well the experiment was designed. Many different groups of raters and subjects can be formed out of the rater and subject populations. Each of these rater-subject combinations will generate different values for the agreement coefficient. How far you expect any given coefficient to stray away from their average value is measured by the agreement coefficient's standard deviation

Note that the general approach to statistical inference used in this book and based on statistical models is appropriate and widely used in the analysis of quantitative ratings. For the analysis of categorical ratings however, an alternative approach is available. It is based upon the law of probability that governs the selection of raters and subjects from their respective universes. Gwet (2021) provides a more detailed discussion of this approach.

1.8 Book's Structure

This book presents various methods for calculating the extent of agreement among raters for quantitative ratings. To ensure an adequate level of depth in the treatment of this topic, I decided to discuss the precision aspects of the agreement coefficients being calculated, and to expand the methods so that datasets containing missing ratings can be analyzed as well. I always start the presentation of new methods with a simple scenario involving two raters only before extending it to the more general context of multiple raters. The book is divided into four parts:

- Part I has two chapters: the current one and chapter 2, which deals with various ways of organizing rating data before the analysis.
- Part II is made up of five chapters. These are chapters 3 through 7. They deal with various versions of the intraclass correlation coefficient and to Finn's coefficient, which is recommended when the conditions for using the intraclass correlation are not satisfied.
- Part III covers some measures of association and concordance as well as some additional analysis techniques pertaining to the intraclass correlation. This part of the book is made up of the 2 chapters 8 and 9.
- Part IV covers some appendices.

Part II covers 5 of the most important chapters of this book. It starts with chapter 3, which outlines the statistical framework within which various intraclass correlation coefficients (ICC) will be developed. Several Analysis of Variance (ANOVA) models and the underlying assumptions will be reviewed. The ICC as a measure of agreement will later be defined as a function the model parameters. Chapter 4 focuses on simple one-factor ANOVA models. Chapters 5 and 6 discuss the two-factor random and mixed models respectively, where the subject is always random and the rater is random or fixed depending on whether the model is random or mixed.

Occasionally, the conditions underlying the ANOVA models are not met, making the ICC a risky approach for evaluating the extent of agreement among raters. In this case, one would consider the alternative approaches discussed in chapter 7, and which do not rely on ANOVA statistical models.

Part III of this book focuses on the following topics:

- Measures of association and concordance often used by researchers in special situations,
- Special analysis techniques to gain further insight into the intraclass correlation coefficient as a measure of agreement,

- Multivariate intraclass correlation as a method of calculating the global agreement coefficient when subjects are rated on several variables.

This part starts with chapter 8 where several measures of association or concordance are discussed. These measures include the classicsl Pearson's and Spearman's correlation coefficients, Kendall's Tau and Kendall's Coefficient of Concordance (KCC). Also discussed in this chapter is Lin's coefficient, which is a version of the Pearson's correlation coefficient modified to measure agreement more effectively. Chapter 9 is the second and last chapter of part III. In addition to discussing multivariate intraclass correlation coefficients, this chapter also presents ways for obtaining a more insightful interpretation of the univariate intraclass correlation.

1.9 Choosing the Right Method

How your ratings should be analyzed depends heavily on the data type and the ultimate study objectives. I previously indicated that your ratings may be of nominal, ordinal, interval, or ratio types. Figure 1.5 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 1.5 does not identify a specific agreement coefficient that must be used. Instead, it directs you to the chapters that discuss the topics of interest to you. These chapters provide more details that will further help you decide ultimately what coefficients are right for your analysis. However, Figure 1.5 may connect you to Figures 1.6 and 1.7 depending on whether you are interested in measures of association or in measures of agreement. These later figures will direct you to specific sections that treat the agreement coefficient you are interested in.

It follows from Figure 1.5 that chapter 8 is where measures of association are treated. Moreover, Figure 1.6 provides more details regarding the specific section and measures of association that should be used. However, if you are interested in measures of agreement among raters, then you have a number of possibilities.

- If the ratings you can assign to subjects are predetermined, then you need to use a Chance-corrected Agreement Coefficient (CAC) regardless of the data type. CACs are out of the scope of this book. You need to look at volume 1 by Gwet (2021). Ratings are considered predetermined if they are included in a set of values known before the experiment is conducted, and from which all raters must choose the specific score that will be assigned to a subject.

- It is only when the ratings are quantitative and result from a measurement process carried out independently by each rater (e.g. the rating is a human subject's

height, or weight whose values can be determined only after the measurement had been taken), that the intraclass correlation or Finn's coefficient must be considered. Figure 1.7 is more detailed and directs you to specific agreement coefficients to use and their associated sections where they are discussed.

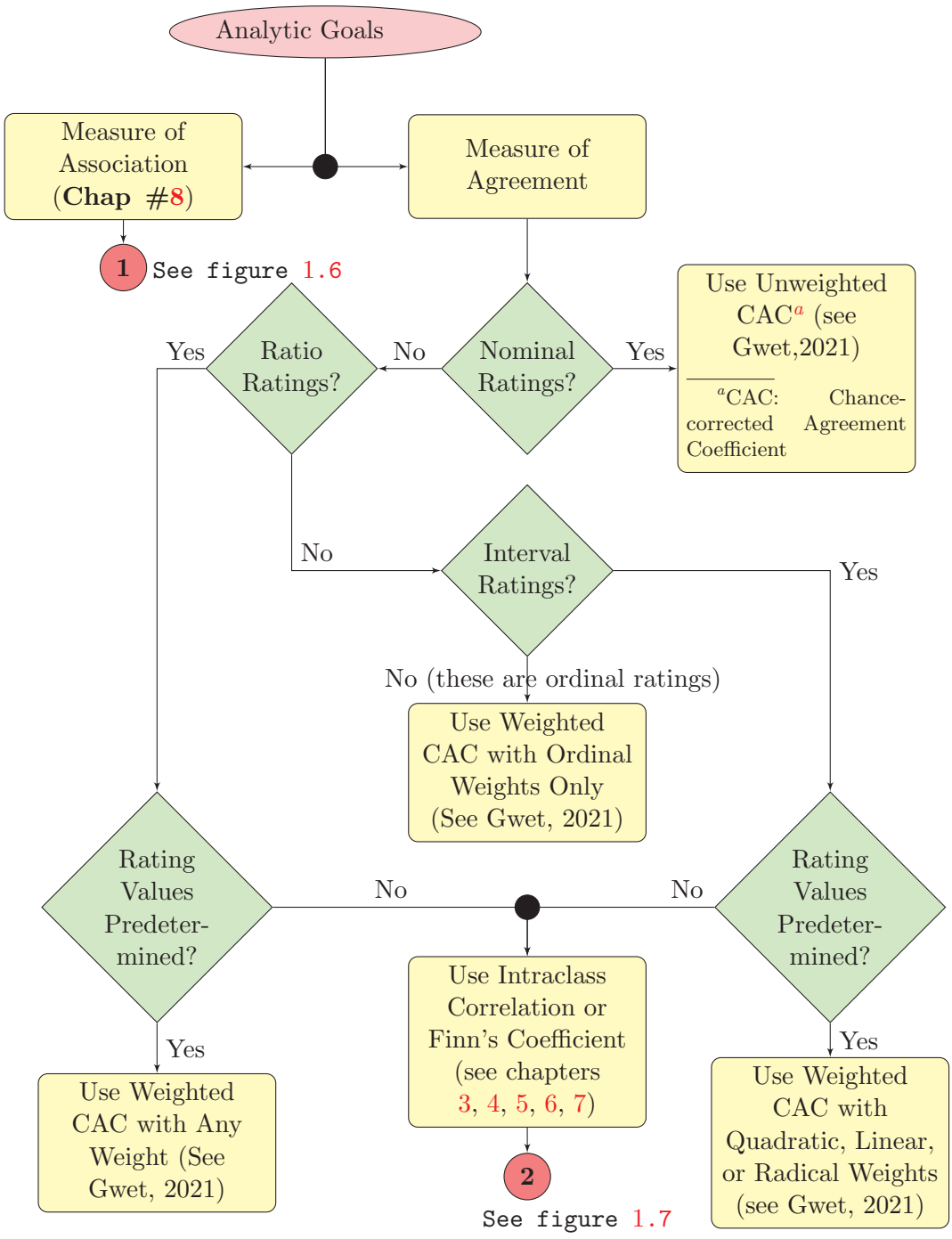


Figure 1.5: Selecting the right agreement coefficient based on analytic goals

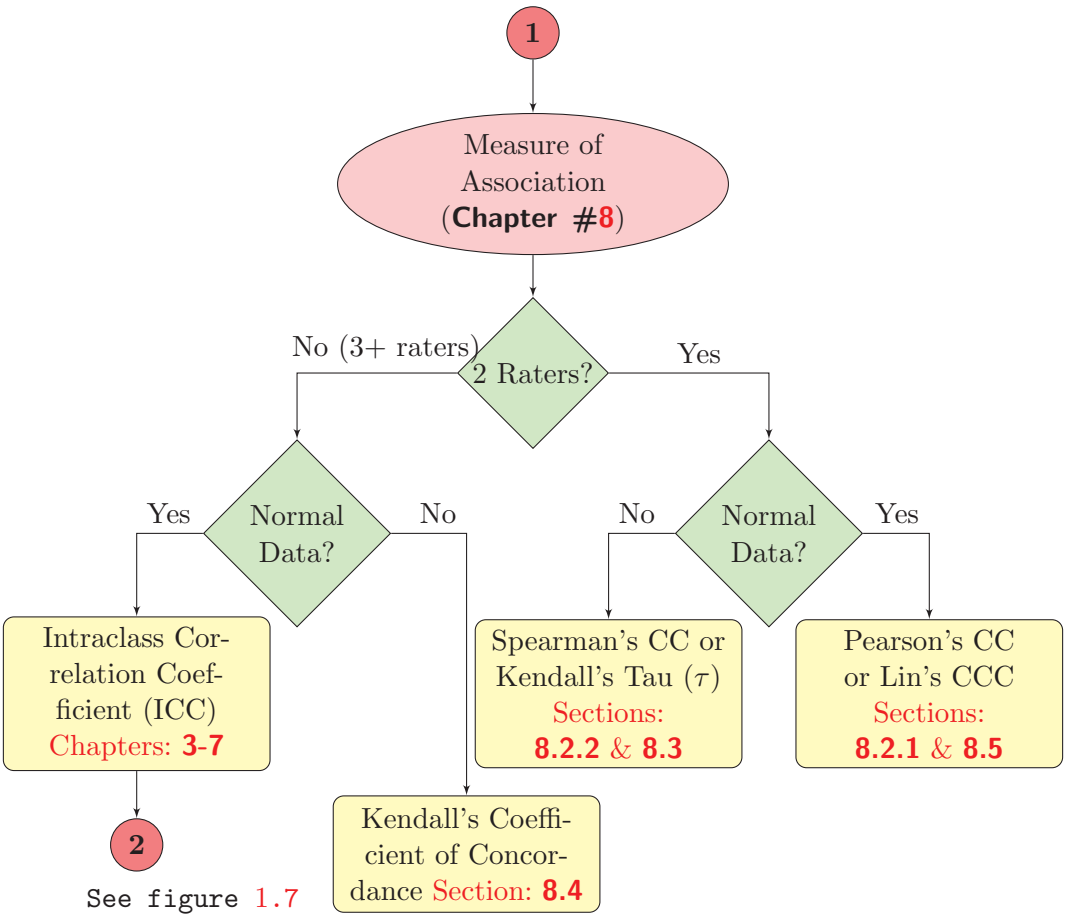


Figure 1.6: Selecting the right measure of association

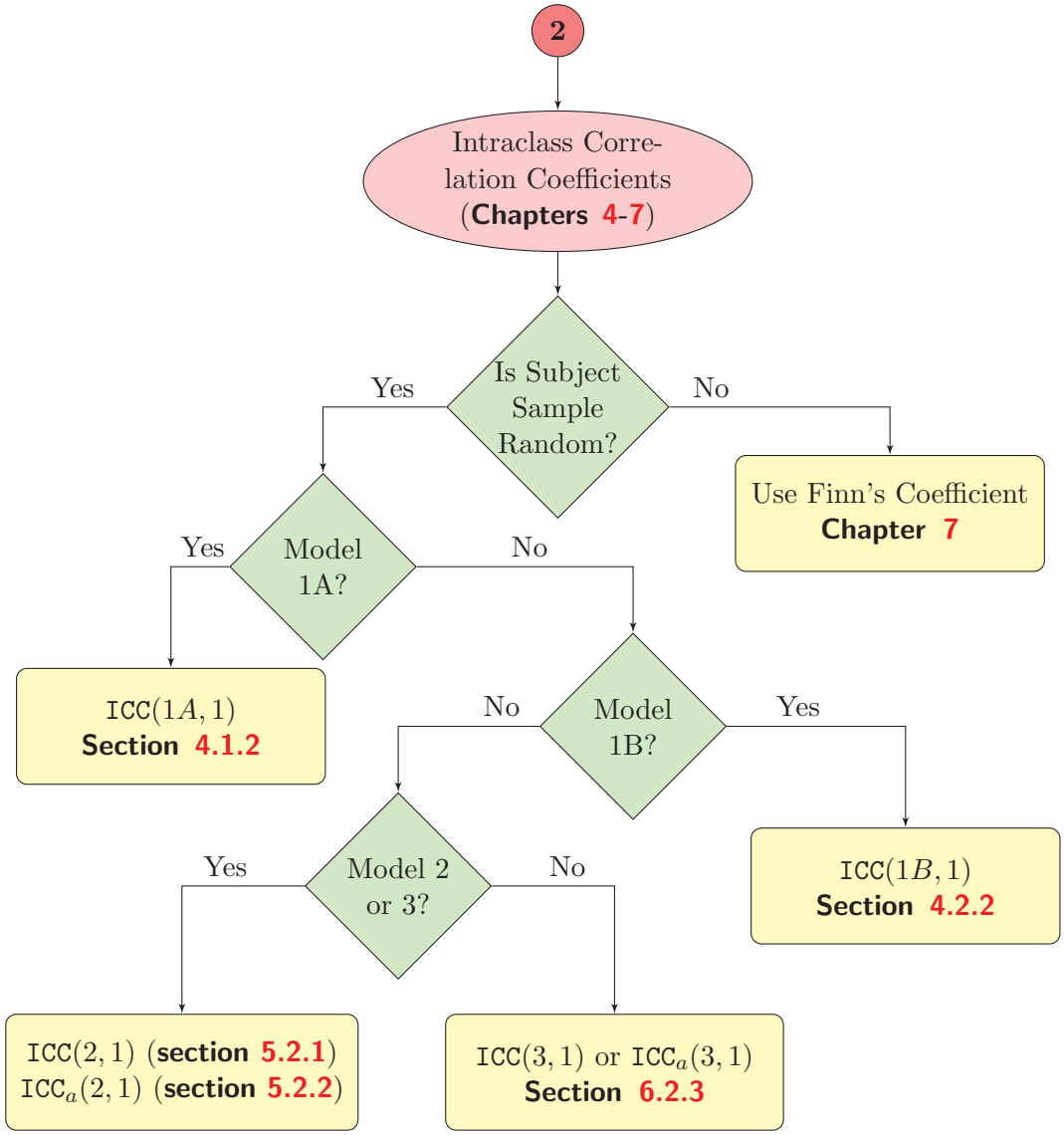


Figure 1.7: Selecting the right intraclass correlation coefficient