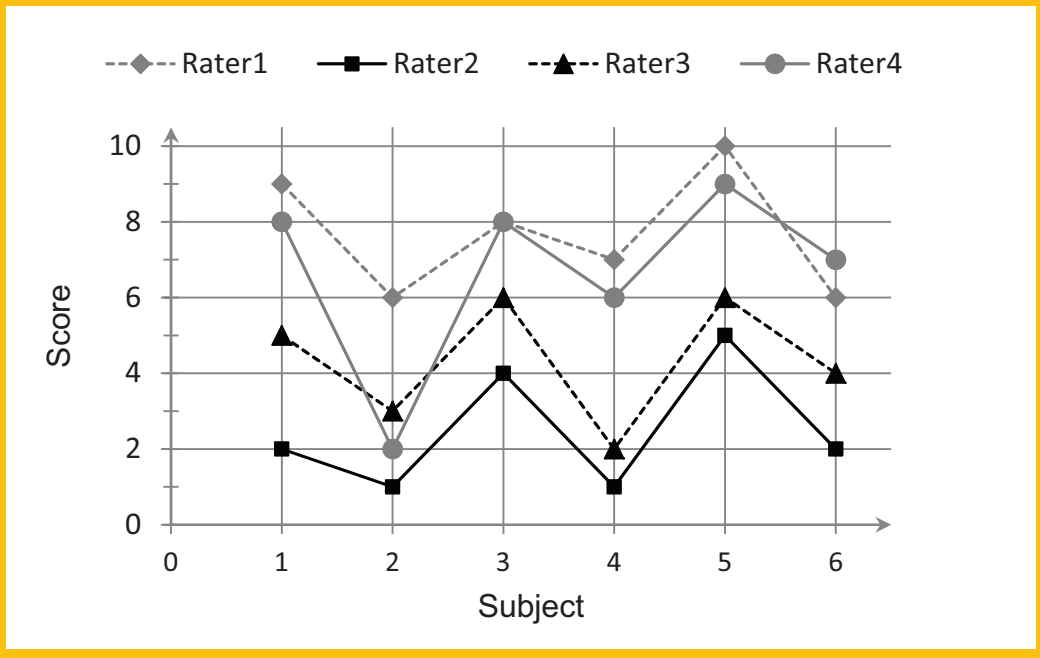
HANDBOOK OF INTER-RATER RELIABILITY

THIRD EDITION

The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters

A Handbook for Researchers, Practitioners, Teachers & Students



Kilem L. Gwet, Ph.D.

HANDBOOK

OF

INTER-RATER RELIABILITY

THIRD EDITION

HANDBOOK OF INTER-RATER RELIABILITY

Third Edition

The Definitive Guide to Measuring the Extent of Agreement Among Raters

Kilem Li Gwet, Ph.D.

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

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

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

Publisher's Cataloguing in Publication Data:

Gwet, Kilem Li
Handbook of Inter-Rater Reliability
The Definitive Guide to Measuring the Extent of Agreement Among Raters/ By Kilem Li Gwet - 3nd ed.
p. cm.
Includes bibliographical references and index.
1. Biostatistics
2. Statistical Methods
3. Statistics - Study - Learning. I. Title.
ISBN 978-0-9708062-7-7

Preface to the Third Edition

The preface to the second edition of the *Handbook of Inter-Rater Reliability* can be found on the next page. Here, I like to explain why I decided to write the third edition of this book. There are essentially 2 reasons for which I decided to write this third edition:

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

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

By writing this book, my primary goal was to allow researchers and students in

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

Kilem Li Gwet, Ph.D.

Preface to the Second Edition

Professional researchers or graduate students who report their research findings are often required to include inter-rater reliability statistics into their analysis. These statistics are quality indicators of the measurement reproducibility. Two raters scoring the same subjects under the same conditions are expected to achieve a high level of consistency in their scores. Otherwise, they will be a source of variation in research data if they are allowed to score different subjects independently. In the later case, the variation associated with the measurements will be attributable to both the raters and the subjects, making it impossible to study the subjects alone. This situation will ultimately lead to the collapse of the whole research project, since its main purpose is precisely the study of subjects. A single rater cannot carried out a massive collection of research data within a reasonable timeframe. Assigning more raters to this task creates the need to minimize the extra variation in the data that multiple raters will add. This is achieved by conducting a special study where selected raters must score the same group of subjects. This experiment will provide the data needed for quantifying the extent to which the raters agree. The resulting measure is referred to as inter-rater reliability. A low inter-rater reliability indicates a possible need for additional training to the raters. After achieving an acceptable level of agreement, they can conduct data collection activities independently.

The early sixties saw the development of various measures for quantifying consistency in the scores that different observers also known as raters assign to the same subjects. The raters could be two physicians examining the same group of patients in a medical facility. While our judgement reflects our thoughts, the lack of transparency of our cognitive processes makes it difficult for others to always agree with us when observing the same phenomenon. The fact that each score reflects the rater's personal perception of the classification process can be detrimental to the credibility of scientific research where high agreement is required. This book summarizes the various inter-rater agreement analysis techniques proposed in the literature, and discusses the contributions of scientists such as Fleiss, Cohen, Everitt, Kraemer, and others whose pioneering work broke the ground for this development. Also extensively discussed is my own contribution to the inter-rater reliability literature.

The scores assigned to subjects can either be qualitative (also known as discrete or nominal) or quantitative. I chose to focus on the treatment of qualitative, and enumerable quantitative scores (i.e. ordinal and interval), and to model-free methods

similar to the Kappa coefficient initially proposed by Cohen (1960). The analysis of continuous quantitative ratings is not covered, primarily because this field is already treated within a solid theoretical framework originally developed in other areas of the statistical science. The classical theory of reliability, the ANOVA (Analysis Of Variance), and loglinear regression techniques widely used in statistical science have provided an adequate framework for studying continuous ratings. The absence of such a framework for analyzing nominal scores provides a fertile ground where researchers can explore new procedures. Consequently, a plethora of procedures has submerged the literature with no common framework to evaluate their merit. I felt the need to review existing practices and concepts with the objective of describing their purpose as well as showing their limitations, all that within a single framework of statistical inference. This is the primary motivation for writing this book.

Initially developed and mostly used in the social and medical sciences, inter-rater reliability assessment is gaining ground in other areas such as software development or linguistics. Inter-rater reliability testing is required nowadays in many research studies, not only those conducted by experienced researchers and scientists, but also those students conduct as part of their master's or doctorate dissertations. One goal this book aims at, is the presentation in one place, of all contributions of notice to the literature where practitioners can start their inquiries, and be exposed to the main problems and issues that have been studied in the past.

This text is intended to general practitioners, researchers, students with general analytical background. Being able to read basic mathematical expressions will ease the reading without it being a prerequisite for accessing the material. The key concepts and main approaches are explained in plain language independently of the mathematical formulas. The book is full of numerical examples to show how the different techniques are implemented in practice. To facilitate the use of the techniques presented in this book, I developed a user-friendly point-and-click Excel VBA program called AgreeStat, which can be downloaded from the website www.agreestat.com. This program can handle a large number of response categories. It can calculate various agreement coefficients available in the literature for 2 raters or more, along with their standard errors. Conditional analysis on specific categories has been implemented as well.

Kilem Li Gwet, Ph.D.

Table of Contents

ACKNOWLEDGMENTS

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

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

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

Thank you,

Kilem Li Gwet, Ph.D.