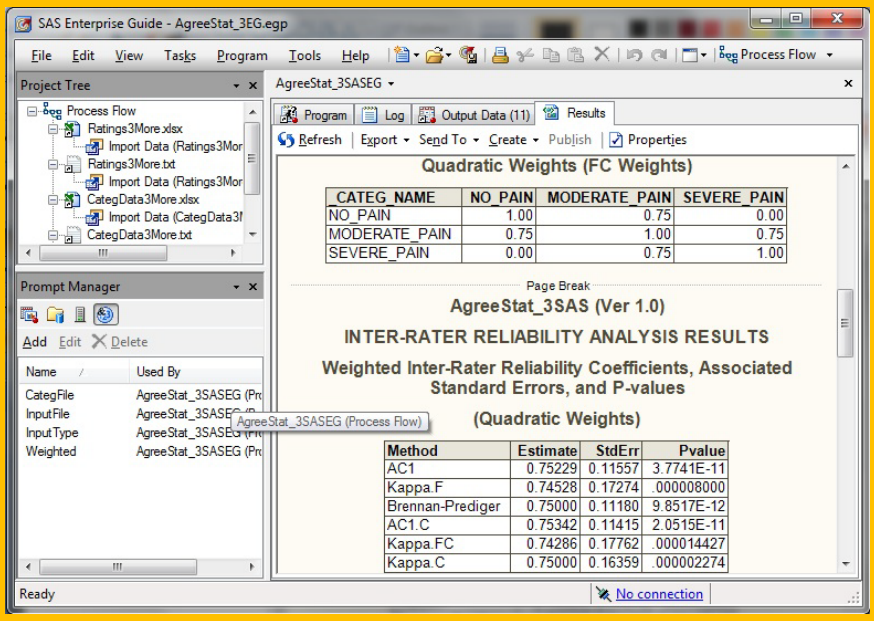
INTER-RATER RELIABILITY USING SAS

A Practical Guide for Nominal, Ordinal, and Interval Data



Kilem L. Gwet, Ph.D.

INTER-RATER RELIABILITY USING SAS

Also by Kilem L. Gwet

- HANDBOOK OF INTER-RATER RELIABILITY (Second Edition): *The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters* (ISBN: 978-0970806246 / 978-0970806222)
- HOW TO COMPUTE INTRACLASS CORRELATION USING MS EXCEL: A Practical Guide to Inter-Rater Reliability Assessment for Quantitative Data. *eBook downloadable at*:

www.agreestat.com

INTER-RATER RELIABILITY USING SAS

A Practical Guide for Nominal, Ordinal, and Interval Data

Kilem Li Gwet, Ph.D.

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

Published by Advanced Analytics, LLC . Printed and bound in the United States of America.

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

Publisher's Cataloguing in Publication Data :

Gwet, Kilem Li
Inter-Reliability with SAS
A Practical Guide for Nominal, Ordinal, and Interval Data/ By Kilem Li Gwet
p. cm.
Includes bibliographical references and index.
1. Biostatistics
2. Statistical Methods
3. Statistics - Study - Learning. I. Title.
ISBN 978-0-9708062-6-0

Contents

Preface

I wrote this book, primarily to assist researchers, and students with the calculation of various inter-rater reliability coefficients for nominal, ordinal, and interval data using the SAS system. The primary focus here is to show practitioners simple step-by-step approaches for organizing their rating data, creating SAS datasets, and using appropriate procedures, or special macro programs to obtain the final coefficients. The agreement coefficients used in this book are first briefly described before being calculated with SAS. I deliberately decided to avoid any formal mathematical presentation of the coefficients under investigation in this book. Instead, these coefficients are introduced using simple numeric examples to show their functionality. This approach has the advantage of presenting the methods in a clear and simple manner, allowing the reader a quick understanding of the mechanics behind the methods. Its biggest disadvantage is perhaps some loss of generality and mathematical rigor, which may not be essential for practitioners less familiar with the field of inter-rater reliability assessment. But advanced readers who want detailed discussion of the different methods may want to see Gwet (2010*a*).

Although, the use of SAS here is basic, the intent in this book is not to teach SAS. Nevertheless, if the user has access to SAS and knows how to launch it, this may be all what is needed to compute various inter-rater reliability coefficients with the techniques recommended here. However, some experience with SAS is definitely required to customize my (macro) programs for specific needs. To take full advantage of the techniques I recommend, it is essential

to ensure that the SAS system at your disposal has a license for the IML Procedure also known as the SAS/IML software. IML stands for Interactive Matrix Language, and I used it extensively in some of the solutions I propose. You do not need to know anything about SAS/IML to use my solutions. All you need to know is whether the SAS system you are using can run IML statements.

The FREQ procedure of SAS offers the calculation of Cohen's Kappa as an option, when the number of raters is limited to 2. The introduction of this feature is without doubt a very welcome addition to the system. But I have realized that in addition to offering only Kappa as the only agreement coefficient, the use of FREQ to compute Kappa is full of pitfalls that could easily lead a careless practitioner to wrong results. For example, if one rater does not use one category that another rater has used, SAS does not compute any Kappa at all. I refer to this problem in chapter 1 as the unbalanced-table issue. Even more seriously, if both raters use the same number of different categories, SAS will produce "very wrong" results, because the FREQ procedure will be matching wrong categories to determine agreement. I refer to this issue in chapter 1 as the "Diagonal Issue." There are actually a few other potentially serious problems with weighted Kappa that I have noticed. I have clearly documented all of the problems I was able to identify, and propose a plan for resolving them.

Many analysts are introduced to the SAS system these days through Enterprise Guide. For these analysts I have devoted an entire chapter to the calculation of inter-rater reliability coefficients with SAS Enterprise Guide (EG). Examples of EG project files are presented to show how agreement coefficients would be computed for 2 raters, as well as for 3 raters or more. I am myself an old type PC SAS programmer who at first was reluctant getting into the new SAS point-and-click platform of Enterprise Guide. However, the more I got into it, the more I liked it. EG indeed makes it

easier to implement the solutions I am proposing in this book.

I used SAS version 9.2 when writing this book. If you are using an older version of SAS, you will still find this book useful, except perhaps chapter 6 on Enterprise Guide. There is also a possibility that in future versions of SAS, SAS Institute will correct some of the problems associated with the FREQ Procedure that I documented in this book. In that happens, I will revise this book to reflect those changes.

Kilem Li Gwet, Ph.D.

CHAPTER 1

The SAS Solution and Its Problems

1.1 Overview

My book entitled "Handbook of Inter-Rater Reliability: *The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*," is essentially about methodology, where I discuss at length about the merits of a large number of techniques for evaluating the extent of agreement among raters. In this book, I have decided to shift my focus, from discussing the methods' merits to producing numbers. Because the focus is on production, I confined myself to presenting a non-mathematical overview of the techniques, and to concentrate on one technology, which is SAS. By choosing SAS, I do not really anticipate that a student or a researcher will purchase this software for the sole purpose of computing inter-rater reliability coefficients. SAS is indeed a massive and expensive system typically licensed by institutions. But many students, and professional researchers can access it through their respective institutions. SAS Institute, Inc., the developer of this system offers an inexpensive learning edition, which can only be of limited help for implementing the approaches I recommend in this book. The reason is that many of the solutions I recommend use the SAS/IML software also known as Proc IML, which is not included in the learning edition.

To respond to the growing demand from researchers for software products that can compute inter-rater reliability coefficients, particularly Cohen's Kappa, SAS now includes an option for calculating the Kappa and Weighted Kappa coefficients in the FREQ procedure. Actually the FREQ procedure does not only compute the coefficients, it also computes associated precision measures, such as the standard errors, P-values, and confidence intervals. The community of researchers who already use SAS can now take advantage of these features. However, the solution proposed by SAS with its FREQ Procedure carries a large number of problems. This book is an attempt to clarify and document many of them, and to propose solutions. I am now going to review what I consider as being among the most important of these problems.

1.2 Number of Raters Limited to 2

The implementation of Kappa in the FREQ procedure is almost entirely based on the book that Fleiss and others wrote in 2003[1], and is limited to 2 raters only. Many inter-rater reliability experiments involve 3 raters or more. Although several agreement coefficients have been proposed in the literature for multiple raters, none is yet implemented in SAS. However, the SAS Institute's support group has developed the magree.sas macro program that could be downloaded at:

http://support.sas.com/kb/25/addl/fusion25006_1_magree.sas.txt

This macro program is also based entirely on Fleiss et al. (2003) recommendations, many of which could be highly questionable, particularly those related to the standard errors associated with the coefficients. I have developed another SAS macro program that implements in addition to the Fleiss' Kappa, several other multiple-

[1] Actually it is the book that Fleiss wrote in 1981 that was revised by others in 2003

rater agreement coefficients that were proposed in the literature.

I must say that, I did like very much the way SAS implemented the calculation of the Cicchetti-Allison, and Fleiss-Cohen weights for computing the weighted Kappa coefficient. What I mean specifically, is that if you define the categories as character-type values, then the weights are calculated based on sequential integer values from 1 to the number of categories. However, if your categories are numeric, then these numeric values are automatically used for defining the weights. This implementation gives the user considerable latitude for customizing the weights. Some researchers have contacted me about the availability of software products that would allow them to use their own set of weights. Although you can always use your weights, I do not recommend that approach because of the possible abuses that may result from such anarchy. Instead, I recommend that practitioners use numeric categories and change their values to reflect the way they view the seriousness of some disagreements. I will further discuss about this issue later in this book. While the FREQ procedure does not output the weights it has used[2], the solutions I propose always output these weights. I do believe that researchers should see the weights that were used for calculating weighted agreement coefficients.

1.3 Agreement Coefficient Options Limited to Kappa

Although several agreement coefficients have been proposed in the literature, SAS has only implemented Kappa in the FREQ procedure. I understand very well that the widespread use of Kappa must have played an important role in the decision that SAS developers made. But any researcher with some experience with the use of Kappa must know by now that there are situations in practice where Kappa will produce rather strange results. These

[2]Maybe I should say I haven't found a way to print them out from SAS

problems, largely known in the literature as the Kappa paradoxes have led several researchers to look for alternative coefficients. The Brennan-Prediger coefficient is one possible option that I like (see Brennan & Prediger (1981)). I also proposed a few years ago, the AC_1, which I considered to be a refinement of the Brennan-Prediger proposal (see Gwet (2008*a*)).

In the case of multiple raters (3 or more), researchers may want to explore alternative generalized coefficients due to Conger (1980), Fleiss (1971), Brennan-Prediger (1980), Gwet (2008*a*). SAS does not offer any of these options presently. I will present in chapters 5 and 6, a SAS macro that implements all of these agreement coefficients, and a few more, and will show you step by step how it can be used. From organizing input data to specifying the parameters, PC SAS, and SAS Enterprise Guide users will learn to compute various agreement coefficients and their associated precision measures within a short period of time.

1.4 The Diagonal Problem

The FREQ procedure of SAS was developed many years before the AGREE and KAPPA options were added to it. But the addition of these options did not change the broad way the FREQ procedure looks at a contingency table. To be concrete, let us consider the following contingency table:

Table 1.1: Distribution of 100 Patients by Rater and Level of Pain

		Rater 1			
		Moderate	No	Severe	Total
Rater 2	Mild	25	5	7	37
	No	6	24	4	34
	Severe	11	1	17	29
	Total	42	30	28	100

At first sight, Table 1.1 resembles any ordinary frequency table showing the distribution of 100 patients by rater and pain level. But the raters had to score the patients on a 4-level scale. Each rater only used 3 of the 4 levels, with the 3 levels used not being the same. Although the information reported in Table 1.1 is accurate, the diagonal in this case does not represent agreement as we normally expect it in the area of inter-rater reliability. The AGREE and KAPPA options of the FREQ procedure however, will treat all diagonal elements of Table 1.1 as representing agreement, leading to wrong results. In other words, Table 1.1 is a contingency table, but it is not an agreement table. It is the agreement table that you need to compute inter-rater reliability adequately. The correct agreement table associated with Table 1.1 is the following:

Table 1.2: Distribution of 100 Patients by Rater and Level of Pain

		Rater 1				
		Mild	Moderate	No	Severe	Total
Rater 2	Mild	0	25	5	7	37
	Moderate	0	0	0	0	0
	No	0	6	24	4	34
	Severe	0	11	1	17	29
	Total	0	42	30	28	100

In Table 1.2, all diagonal elements now represent agreement. Therefore, when using the FREQ procedure of SAS, you need to ensure that the contingency table being used is in fact an agreement table. This issue is discussed in great details in chapters 2 and 3, and simple solutions are proposed.

1.5 The Unbalanced-Table Problem

Consider Table 1.3 representing the distribution of 100 elderly patients by rater and level of function. Once again, this table is a normal contingency table in a traditional sense. One thing ho-

wever stands out, which is that rater 1 has used one function scale level more than rater 2. This situation led to the unbalanced table we have with 3 columns and only 2 rows. Therefore our contingency table does not have a diagonal. With no diagonal, the FREQ procedure of SAS cannot compute the Kappa coefficient nor any other related statistics.

Table 1.3: Distribution of 100 Patients by Rater & Functional Level

		Rater 1			
		Independent	Assistance	Dependent	**Total**
	Independent	25	5	7	37
Rater 2	Dependent	17	25	21	63
	Total	42	30	28	100

We have another case here where the contingency table is not an agreement table. With no diagonal in the table, SAS will simply ignore our request to compute Kappa. Most mainstream statistical techniques such as the Chi-square test that have historically been implemented in the FREQ procedure can work just fine with any contingency table. Moreover, the contingency tables typically analyzed in statistics involve 2 variables with different levels, and only the broad table configuration really matters. But the agreement table is a very special contingency table, which must be handled with care. I refer to this issue as the unbalanced-table issue, and further discuss it in chapters 2, and 3. I also propose simple solutions for resolving it.

The agreement table based on Table 1.3 data is given by Table 1.4. Now we have a balanced table with identical categories placed in the same order rowwise and columnwise. An agreement table ready for inter-rater reliability analysis.

Table 1.4: Distribution of 100 Patients by Rater & Functional Level

		Rater 1			
		Independent	Assistance	Dependent	**Total**
Rater 2	Independent	25	5	7	37
	Assistance	0	0	0	0
	Dependent	17	25	21	63
	Total	42	30	28	100

1.6 Then Ordinal Data Problem

Several authors in the literature have advocated the use weighted agreement coefficients so that disagreements, which are more serious and those which are less serious do not receive the same treatment. It follows from table 1.4 that Rater 1 has considered dependent 7 patients who rater 2 considered independent. Here is a disagreement with obviously far more serious consequences that the one that occurred on the 5 patients raters 1 and 2 classified into the assistance and independent groups respectively. The more serious disagreements generally receive a smaller weight than the less serious ones.

By default, SAS sorts categories in alphabetical order for treatment. It means that Table 1.4 data for example will be organized as shown in Table 1.5 below.

Table 1.5: Distribution of 100 Patients by Rater & Functional Level

		Rater 1			
		Assistance	Dependent	Independent	**Total**
Rater 2	Assistance	0	0	0	0
	Dependent	25	21	17	63
	Independent	5	7	25	37
	Total	42	30	28	100

Both tables should a priori produce the same results. That is only

true if you do not care about the weighted Kappa. If you do, then you will want to know that when computing the weighted Kappa from Table 1.5, SAS considers the Assistance-Independent disagreement to be more serious than the Dependent-Independent disagreement by assigning the smallest weight (typically a 0 weight) to the former. It is actually the Dependent-Independent disagreement that must be zero-weighted, as it is the one that should not be given any credit towards agreement. Therefore, when comes to agreement coefficient, even the way categories are labeled matters. I discuss the weighting further in chapter 4.