CHAPTER 6

Benchmarking Inter-Rater Reliability Coefficients

OBJECTIVE

In this chapter, I will discuss about several ways in which the extent of agreement among raters can be interpreted once it has been quantified with one of the agreement coefficients discussed in the past few chapters. Given the agreement coefficient's magnitude, should you conclude that the extent of agreement among raters is "Excellent", "Good", or "Poor ?" To answer this question, I will review some benchmark scales proposed in the literature, will discuss their weaknesses, and will recommend an alternative benchmarking model that accounts for the precision with which the agreement coefficient has been estimated. I argue that the magnitude of the agreement coefficient alone is insufficient to qualify the extent of agreement among raters. It is because accurate numbers based on a well-designed experiment must lead to a stronger statement than inaccurate numbers based on a limited and ill-designed experiment.

CONTENTS

6.1 Overview **164**

6.2 Benchmarking the Agreement Coefficient **165**
6.2.1 Existing Benchmarks **166**
6.2.2 Agreement Coefficient's Sources of Variation **168**

6.3 The Proposed Benchmarking Method **173**
6.3.1 The Method **174**
6.3.2 The Benchmark Probabilities and the Interpretation of the New Method **177**

6.4 Concluding Remarks **180**

"Concrete measures can determine progress, but they do not really measure values."

Peter Block : "The Answer to How Is Yes : Acting on What Matters"
(Berrett-Koehler, 2002)

6.1 Overview

"Extent of agreement among raters" is often a vague notion in our imagination. The inter-rater reliability coefficient codifies it in a logical way, allowing researchers to have a common and concrete representation of an abstract concept. The many different logics used in this codification led to various forms of the agreement coefficient. However, for an inter-rater reliability coefficient to be useful, researchers must be able to interpret its magnitude. Although concrete agreement coefficients determine the extent to which raters agree among themselves, these measures do not tell researchers how valuable that information is. Should an agreement coefficient of 0.5 for example be considered good, fair, or bad ? Should it be considered acceptable ? What are the practical implications for implementing a classification system that is backed up with a 0.50 inter-rater reliability coefficient ? These are some of the questions that are addressed in this chapter.

In the course of the development of inter-rater reliability coefficients, it appeared early that a rule of thumb was needed to help researchers relate the magnitude of the estimated inter-rater reliability coefficient to the notion of extent of agreement. Practitioners wanted a threshold for Kappa, beyond which the extent of agreement will be considered "good." The process of comparing estimated inter-rater reliability coefficients to a predetermined threshold before deciding whether the extent of agreement is good or bad is called *Benchmarking*, and the thresholds used to make the comparison are the *Benchmarks*.

Many scientific fields use standards of quality to distinguish the acceptable from the unacceptable. These standards are expected to vary from one field to another one. Regarding inter-rater reliability coefficients, the following two questions should be answered:

- What makes a good extent of agreement good ?
- How high should the inter-rater reliability coefficient be for the extent of agreement as a construct to be considered good ?

Accumulated experience in a particular discipline have generally provided the answer to these two questions as far as the use of Kappa is concerned. Landis and Koch (1977) provided one of the most widely-used benchmark scales among practitioners, and which will be discussed in section 6.2. Researchers having used the

Kappa statistic over a long period have found the proposed benchmark scale useful.

While the use of accumulated experience for benchmarking has undeniable merits, ignoring the influence that experimental conditions have on the magnitude of estimated agreement coefficients will lead to an incomplete interpretation of their significance. I demonstrate in the next few sections that a benchmarking model that does not account for the number of subjects and raters that participated in the reliability experiment, as well as the number of response categories could validate an agreement coefficient, which carries a large error margin. An agreement coefficient of 0.50 for example, is labeled as "moderate" according to all benchmark scales known in the literature. While this may be acceptable in a study involving 25 subjects, 3 raters and 4 response categories, I show in section 6.2 that an agreement coefficient of this magnitude is not even *statistically significant* if the study is based on 10 subjects, 2 raters and 2 response categories. The lack of statistical significance indicates that the "true" value of the coefficient (i.e. free of sampling errors) could well be as small as 0. In the absence of the "true" agreement coefficient, the error margin associated with the estimated agreement coefficient becomes informative ; because it provides the only description of the neighborhood where the truth is situated. If an error-free inter-rater reliability coefficient is 0, its value estimated from small samples of subjects or raters may appear as high as 0.5 or even higher due to sampling errors alone.

> *If an inter-rater reliability coefficient is not "Statistically significant," then any characterization of the agreement among raters other than "Poor" would be misleading. The sample-based estimated agreement coefficient which is not statistically significant does not provide strong enough evidence that the "true" magnitude of the agreement coefficient (i.e. free of sampling errors) is better than 0. Under this circumstance, the extent of agreement among raters, which is more dependent on the true agreement coefficient than on its estimated value is logically expected to be poor.*

I propose in this chapter, a new approach for interpreting the inter-rater reliability coefficient that uses existing benchmark scales as well as actual experimental parameters such as the number of subjects, raters, and response categories. Moreover, different benchmarking models are proposed for different agreement coefficients. The current approach to benchmarking is reviewed in section 6.2, while a description of the newly-proposed method is described in section 6.3.

6.2 Benchmarking the Agreement Coefficient

This section's objective is to review various benchmark scales proposed in the literature for interpreting the magnitude of the Kappa statistic, and to discuss

6.4 Concluding Remarks

The primary objective of this chapter was to present an alternative benchmarking model for interpreting the extent of agreement among raters based on the magnitude of the calculated agreement coefficient. The approach currently advocated in the inter-rater reliability literature is based upon a straight comparison between the calculated agreement coefficient and a number of benchmarks proposed by various authors. Using a Monte-Carlo experiment, I demonstrated that this classical approach tends to provide an overly optimistic characterization of the extent of agreement among raters, ignoring the adverse effects that a small number of subjects or raters can have on the agreement coefficient precision. The Monte-Carlo experiment has proved that the classical benchmarking model would characterize the extent of agreement among raters as "Excellent" even when the ratings are obtained through a purely random process. A situation where no intrinsic agreement is expected to occur among raters. This problem is created by estimated agreement coefficients that are sometimes artificially inflated by errors due to the sampling of subjects, or that of raters. The experiment has also demonstrated that a small number of categories will increase the magnitude of theses errors.

In order to provide a fair comparison between agreement coefficients obtained from different studies based on different designs, I have recommended a new benchmarking process that is probabilistic. That is each benchmark range of values is assigned a membership probability. This probability represents the likelihood that the estimand of a particular agreement coefficient falls into the benchmark range of values. After computing these benchmark probabilities, one option would be to simply present them and leave it up to others to decide whether they want to characterize the extent of agreement as very good, intermediate or poor. They will still be able to use the benchmark probabilities to justify their decisions. Instead, I have decided to recommend a rule for characterizing the extent of agreement, which is to select the highest benchmark level that is associated with the smallest cumulative probability that exceeds 95%. The 95% cut-off point is a standard of acceptability in statistical science. Practitioners may decrease or increase that cut-off point if deemed necessary.

I believe that the choice of benchmark scale is less important than the way it is used for characterizing the extent of agreement among raters. Having said that, I do believe that Fleiss' benchmark scale presented in Table 6.2 is bad. It is because of the unduly large width of its benchmark intervals. For example the Intermediate-to-Good range of values goes from 0.4 to 0.75, and is too broad to be very helpful in practice. Moreover, the two words "Intermediate" and "Good" have meanings that are too different for them to be lumped into a single category. Intermediate generally

means it could get much better, while good is always considered satisfactory. If and inter-rater reliability of 0.75 may be deemed acceptable, very few people will admit an inter-rater reliability of 0.4 as being acceptable. However the Landis-Koch and Altman's benchmark scales are both acceptable.

Unlike the classical benchmarking model that is applied uniformly to all agreement coefficients, the newly-proposed model is tailored to each agreement coefficient. The standard error of the estimated agreement coefficient plays a pivotal role in this new process. The standard error quantifies the quality of the study design, will reward well-designed studies with higher benchmark probabilities, while penalizing poorly designed studies. It prevents poorly-designed inter-rater reliability studies from producing an "Excellent" extent of agreement among raters based solely on an imprecise estimated agreement coefficient.