CHAPTER 6

Benchmarking Inter-Rater Reliability Coefficients

OBJECTIVE

In this chapter, we will learn how the extent of agreement among raters should 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 we conclude that the extent of agreement among raters is "Excellent", "Good", or "Poor?" To answer this question, we 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.

CONTENTS

"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 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. We 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, we prove 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.*

We 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 deferred until 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

some of their limitations. We will identify key factors affecting the meaning of the estimated inter-rater reliability, and will demonstrate the need to make them an integral part of any viable benchmarking method.

6.2.1 *Existing Benchmarks*

Benchmarking is essential for communicating the results of a reliability study to a wide audience, in addition to providing guidelines to help practitioners with the use of agreement statistics. Three benchmarking models proposed in the literature will be reviewed in this section. Although most of these models were developed to be used with the Kappa coefficient, they are often used in practice with other agreement coefficients as well.

Table 6.1 describes the benchmark scale that Landis and Koch (1977) proposed. It follows from this table that the extent of agreement can be qualified as "Poor", "Slight", "Fair", "Moderate", "Substantial", and "Almost Perfect" depending on the magnitude of Kappa. A Kappa value between 40% and 60% indicates a moderate agreement level, while ranges of values ($60\% - 80\%$), and ($80\% - 100\%$) indicate substantial and almost perfect agreement levels respectively. Although the authors acknowledge the arbitrary nature of their benchmarks, they recommended them as a useful guideline for practitioners. Other authors such as Everitt (1992) have supported this benchmark scale.

Table 6.1:
Landis and Koch-Kappa's Benchmark Scale

Kappa Statistic	Strength of Agreement
< 0.0	Poor
0.0 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

Fleiss (1981) proposed another benchmark scale where the first three value ranges of the Landis-Koch benchmark are collapsed into a single range "0.40 or less" labeled as "Poor." Table 6.2 shows the 3 ranges of values that make up Fleiss' benchmarking scale. Kappa values in the $40\% - 75\%$ range for example represent an "Intermediate to Good" extent of agreement, while all Kappa values in the $75\% - 100\%$ range indicate an "Excellent" extent of agreement. This scale has the advantage of having a small

number of categories while presenting the middle category as that of acceptable values, the low-end category as that of the unacceptable, and the high-end category as that of excellence.

Table 6.2: Fleiss' Kappa Benchmark Scale

Kappa Statistic	Strength of Agreement
< 0.40	Poor
0.40 to 0.75	Intermediate to Good
More than 0.75	Excellent

Altman (1991) proposed his benchmark scale summarized in Table 6.3, and which represents a modified version of Landis-Koch's proposal. The only noticeable difference is the first two ranges of values of Landis-Koch's proposal that Altman collapsed into a single category labeled as "Poor." Landis-Koch's proposed benchmarking method was published several years before Alman's, and are still being used. Therefore, the argument for supporting the newer Altman's benchmarks remains unclear.

Table 6.3: Altman's Kappa Benchmark Scale

Kappa Statistic	Strength of Agreement
< 0.20	Poor
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Good
0.81 to 1.00	Very Good

Our objective in this chapter is not to recommend the use of a specific Benchmark scale. Practitioners who want to use these scales could choose one that meets their analytical goals. For example, those who only want to know whether the extent of agreement is excellent or poor will want to use Fleiss' benchmarks, whereas researchers with a desire for a finer categorization may prefer either Landis-Koch or Altman proposals. The more critical issue we must address is that of the error margin associated with agreement coefficients. The magnitude of the error margin alone may lead to a spurious characterization of the agreement coefficient.

A large number of subjects will generally lead to a precise agreement coefficient, which is expected to be close to the true value of the inter-rater reliability it approximates. Therefore, a straight comparison of such a precise coefficient with existing benchmarks will yield conclusions that will apply to the true parameter of interest

as well. A small number of subjects on the other hand, reduces the precision of the inter-rater reliability coefficient, in addition to exposing that precision to further degradation due possibly to a small number of raters, or a small number of response categories. Comparing an agreement coefficient loaded with errors to a predetermined quality benchmark can only produce a questionable characterization of the extent of agreement among raters. Consequently, accounting for the number of subjects, raters, and response categories becomes critical when interpreting the magnitude of sample-based agreement coefficients. Section 6.2.2 aims to demonstrate that the number of subjects (n), raters (r), and categories (q) can substantially affect the agreement coefficient probability distribution[1], and therefore its error margin. This will be a demonstration that these factors must be taken into consideration in the way agreement coefficients are interpreted.

6.2.2 *Agreement Coefficient's Sources of Variation*

To demonstrate how the number of subjects (n), raters (r), and categories (q) affect the agreement coefficient's distribution, a well-established statistical approach is to simulate an inter-rater reliability study with a computer, and to repeat the experiment many times in order to obtain the agreement coefficient's probability distribution. For this particular problem, we have used the Random Rating (RR) model where raters classify subjects into categories in a purely random manner. More specifically, we want to study how the quantities n, r, and q affect the 95^{th} percentile[2] of an agreement coefficient. For this purpose, we used Kappa and Pi as examples of coefficients for illustration purposes. Note that any estimated agreement coefficient that exceeds the RR-based 95^{th} percentile (also called *Critical Value*) will be inconsistent with random rating. Therefore, the observed ratings are unlikely to have been done randomly. On the other hand, an agreement coefficient smaller than the critical value will be considered consistent with random rating. An agreement coefficient smaller than the critical value raises serious doubt about the very existence of intrinsic agreement among raters. An agreement coefficient that exceeds the critical value is perceived as being consistent with the existence of an intrinsic agreement among raters.

For any particular set of values associated with n, r, and q, the RR-based 95^{th} percentile is obtained by repeating the simulation of the Random Rating model a large number of times[3]. Each iteration of the experiment yields one agreement coeffi-

[1] For all practical purposes, the agreement coefficient distribution in this context essentially refers to what can be expected from an agreement coefficient in terms of magnitude and variation

[2] The 95^{th} percentile is a threshold that an agreement coefficient can exceed only with a small chance below 5%.

[3] Randomness and its relationship with games of chance popular in the city of Monte-Carlo (Monaco) led Metropolis and Ulam (1949) to refer to this method as the Monte-Carlo method

cient. After several iterations, a series of values are generated providing a probability distribution of the agreement coefficient of interest. Repeating the Monte-Carlo experiment for different values of n, q and r allows us to evaluate their impact on the inter-rater reliability distribution. The Monte-Carlo experiment is implemented as described in the following steps:

The Monte-Carlo Experiment

1. *Assign specific values to n, r, and q. For example, one may decide to simulate a reliability study based on random rating, with $n = 8$ subjects, $r = 3$ raters, and $q = 5$ level Likert scale.*
2. *Generate a table of ratings similar to Table 6.4, where the table entries are computer-generated random integers between 1 and 5.*
3. *Use Table 6.4 data to compute all agreement coefficients of interest and record them.*
4. *Steps 2 and 3 are repeated 100,000 times. However, the number of iterations will not exceed the number of rating tables that can possibly be created.*

Table 6.4: Raw Ratings from a Monte-Carlo Experiment

Rater 1	Rater 2	Rater 3
5	5	4
2	3	1
3	3	3
5	5	5
4	3	5
5	5	4
1	1	2
3	2	2

This Monte-Carlo experiment was carried out several times with n taking the values {2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60}, q taking the values {2, 3, 4, 5}, and r the values {2, 3, 4, 5, 10, 20}. The results obtained are depicted in Figures 6.1, 6.2, and 6.3.

Figure 6.1 shows the variation of the 95^{th} percentile of Kappa for 2 raters only, as a function of the number of subjects (n) and the number of categories (q) under the RR model. It follows from this graph that if the number of subjects is 10 and the number of categories is 2, then the 95^{th} percentile of Kappa will be greater than 0.5. This example shows that 2 raters scoring 10 subjects in a pure random

manner, making no effort to categorize them in any logical way can still achieve a Kappa above 0.5 more than 5% of the times. Why is this possible ? It is because any agreement coefficient based on only 10 subjects, 2 raters, and 2 categories is exposed to substantial error margin and cannot always have a value in the neighborhood of 0 where it is supposed to be under the RR model. The number of subjects in this case - limited to 10 - is too small for the concrete value of Kappa or any other agreement coefficient to be consistent with its conceptual definition and theoretical properties.

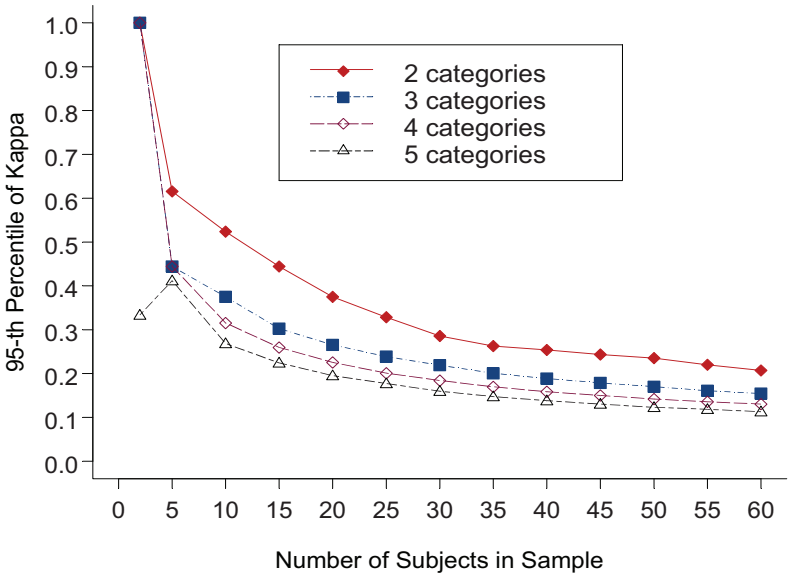

Figure 6.1. Kappa Coefficient By Subject Sample Size and Number of Response Categories Under Random Rating

Figure 6.1 also shows that Kappa's critical value decreases as the number of subjects or the number of categories increases. As a result, more participating subjects or more categories in a reliability study make Kappa more accurate. That the precision of Kappa is impacted by the number of categories may come as surprise. In fact the form of Kappa does not suggest any natural dependency upon the number of categories.

Following the analysis of Figure 6.1, it is natural to ask whether the observed relationship linking the critical value to the number of subjects, raters, and categories will still hold with agreement coefficients other than Kappa. The answer to this question if yes. Figure 6.2, shows the 95^{th} percentile of Scott's Pi statistic as a function of n and q when the number of raters is limited to 2. We can still observe a decrease in the magnitude of the critical value following an increase in number of subjects or response categories.