CHAPTER 7

Benchmarking Inter-Rater Reliability Coefficients

OBJECTIVE

In this chapter, I 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

"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)

7.1 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

G. (1977) provided one of the most widely-used benchmark scales among practitioners, and which will be discussed in section 7.2. 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 of experimental conditions on the magnitude of estimated agreement coefficients will lead to an incomplete and possibly misleading 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 7.2 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. Even if an error-free inter-rater reliability coefficient is 0, its value estimated from small samples of subjects or raters may turn up 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 agreement among raters other than "Poor" would be misleading. The sample-based estimated agreement coefficient which is not statistically significant does not provide a strong enough evidence that the magnitude of the "true" and error-free agreement coefficient is better than 0. Under this condition, the extent of agreement among raters, which by the way is more dependent on the true agreement coefficient than on its estimated value should logically be qualified as poor.*

You need to see the true nature of statistical thinking here. When I claim that the extent of agreement among raters is poor, it does not mean I have proved it. It essentially means the data I have analyzed did not provided sufficient evidence to eliminate poor rater agreement as a real possibility. Can agreement among raters still be excellent even if I qualify it as poor? The answer is yes, although this possibility reduces substantially with a well-designed study.

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 7.2, while a description of the newly-proposed method is described in section 7.3.

7.2 Benchmarking Agreement Coefficients

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. I will identify key factors affecting this interpretation and will demonstrate the need to make them an integral part of any viable benchmarking method.

7.2.1 *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 7.1 describes the benchmark scale that Landis and G. (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% for example 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 subjective nature of their benchmarks, they recommended them as a useful guideline for practitioners. Other authors such as Everitt (1992) have supported this benchmark scale.

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 7.2 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 7.1
Landis and Koch Kappa's Benchmark Scale[a]

Kappa Statistic	Strength of Agreement
−1.0 to 0.0	Poor
0.0 to 0.20	Slight
0.20 to 0.40	Fair
0.40 to 0.60	Moderate
0.60 to 0.80	Substantial
0.80 to 1.00	Almost Perfect

[a]Note that in the original Landis-Koch scale, the first interval is defined as "< 0", while the lower bounds of the remaining intervals are 0.0, 0.21, 0.41, 0.61, and 0.81. The proposed changes are motivated by the fact that most agreement coefficients take values on a continuum between -1 and 1. The original intervals leave some values out of this continuum.

Table 7.2: 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 7.3, and which represents a modified version of the 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 is still being used. Therefore, the argument for supporting the newer Altman's benchmarks remains unclear.

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.

Table 7.3
Altman's Kappa Benchmark Scale[a]

Kappa Statistic	Strength of Agreement
−1.0 to 0.20	Poor
0.20 to 0.40	Fair
0.40 to 0.60	Moderate
0.60 to 0.80	Good
0.80 to 1.00	Very Good

[a]Note that in the original Altman scale, the first interval is defined as "< 0.20", while the lower bounds of the remaining intervals are 0.21, 0.41, 0.61, and 0.81. The proposed changes are motivated by the fact that most agreement coefficients take values on a continuum between -1 and 1. The original intervals leave some values out of this continuum.

A larger number of subjects will generally lead to a more precise agreement coefficient, which is expected to be close to the true value it approximates. Therefore, a straight comparison of 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 7.2.2 aims at demonstrating 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 prove that these factors must be taken into consideration when interpreting agreement coefficients.

7.2.2 *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, I have used the Random Rating (RR) model where raters classify subjects into categories in a purely random manner.

[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