Free-Marginal Multirater Kappa (multirater κ_{free}): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa

Justus J. Randolph

University of Joensuu

Presented at the Joensuu Learning and Instruction Symposium 2005

October 14-15, 2005

Joensuu, Finland

Author Note:

Justus J. Randolph, Department of Computer Science, University of Joensuu, Finland.

Correspondence concerning this article should be addressed to Justus J. Randolph, Department of Computer Science, University of Joensuu, PO BOX 111, FIN-80101, Finland. E-mail: justusrandolph@yahoo.com

Acknowledgement: Thanks to Roman Bednarik and Niko Myller, who helped me realize the need for a multirater free-marginal kappa.

Abstract

Fleiss' popular multirater kappa is known to be influenced by prevalence and bias, which can lead to the paradox of high agreement but low kappa. It also assumes that raters are restricted in how they can distribute cases across categories, which is not a typical feature of many agreement studies. In this article, a free-marginal, multirater alternative to Fleiss' multirater kappa is introduced. Free-marginal Multirater Kappa (multirater κ_{free}), like its birater free-marginal counterparts (PABAK, S, RE, and κ_{m},) is not influenced by kappa and is appropriate for the typical agreement study, in which raters' distributions of cases into categories are not restricted. Recommendations for the proper use of multirater κ_{free} are included.

Keywords: Multirater Kappa, Cohen's Kappa, Reliability, Measures of Agreement

Free-Marginal Multirater Kappa (multirater κ_{free}): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa

Fleiss' multirater kappa (1971), which is a chance-adjusted index of agreement for multirater categorization of nominal variables, is often used in the medical and behavioral sciences. It is used in medical research when, for example, a researcher wants to know the chance-adjusted measure of agreement between physicians' nominal diagnoses. It is also widely used in the fields of content analysis and meta-analysis when a researcher wants to determine how well raters agree on the coding of nominal variables. Fleiss' multirater kappa is presented in seminal textbooks, like Siegell and Castellan's (1988) *Nonparametric Statistics for the Behavioral Sciences* and can be computed by a number of online resources, such as the Chang's Statistics Toolbox (n.d.), and by a number of offline statistical programs, such as SPSS (see Nicholls, 1997), STAT, and SAS.

Like its marginally-dependent counterparts - such as Cohen's kappa (1960) and Scott's pi (1955) - Fleiss' multirater kappa is appropriate for fixed-marginal validity studies; however, it is not appropriate for agreement studies that have free-marginal distributions. Throughout this paper, marginals are considered to be *fixed* when raters know a priori the quantity of cases that should be distributed into each category. For example, this might be the case when a rater is free to assign cases to categories as long as there will be a certain, predetermined amount of cases in each category in the end. Marginal distributions are considered to be *free* when raters do not know a priori the

quantities of cases that should be distributed into each category. For example, this is the case when a rater is free to assign cases to categories with no limits on how many cases must go into each category. See Brennan and Prediger, [1981]).

When fixed-marginal varieties of kappa (e.g., Cohen's kappa, Scott's pi, and Fleiss' multirater kappa) are used in free-marginal, agreement studies, the value of kappa can vary significantly when the proportions of overall agreement and the number of raters, categories, and cases are held constant but the marginal distributions are allowed to vary. Byrt, Bishop, and Carlin (1993) estimate that in the birater case, when percent of overall agreement is held constant, the value of fixed-marginal versions of kappa can double or halve depending on the characteristics of marginal distributions. This finding is attributed to two aspects that affect marginal distributions - prevalence, defined as "the [true] proportions of cases of various types in a population" (Banerjee, Capazzoli, McSweeney & Sinha, 1999. p. 6) and bias, defined as "the bias of one rater relative to another" (Banerjee et al., 1999. p. 6) in their assignment of cases (See Banerjee et al, 1999; Brennan & Prediger, 1981; Byrt et al., 1993; Cichetti & Feinstein, 1990a, 1990b; Light, 1971).

Although though are several versions of kappa that are not influenced by prevalence and bias in the birater case, they have not been applied to the multirater, multicategory case. Therefore, in this paper I introduce a free-marginal, multirater version of kappa, called Free-Marginal Multirater Kappa (multirater κ_{free}), which is not influenced by bias or prevalence and is appropriate for common situations in most reliability/agreement studies. Multirater κ_{free} is an extension of the birater, free-marginal

forms of kappa that use 1/number of categories as the proportion of agreement expected by chance.

In the following sections, I discuss Fleiss' multirater kappa and demonstrate to what degree it is affected by prevalence. I also discuss the birater, free-marginal solution to the prevalence and bias problem. This paper ends with a presentation of multirater κ_{free}, an analysis of its properties, and recommendations for its use.

Fleiss' Multirater Kappa

In 1971, Fleiss introduced a generalization of Cohen's (1960) unweighted kappa for the multirater case. Like all other versions of the kappa statistic, Fleiss' multirater kappa takes the general form presented in Equation 1:

(Insert Equation 1 here, centered) [1]

The logic behind this formula is that a certain proportion of agreement between raters can be expected by chance; the proportion expected by chance is signified by - P_e. By subtracting the P_e from the proportion of overall observed agreement (P_o) and dividing that (i.e., $P_o - P_e$) by the maximum proportion of possible chance-adjusted agreement ($1 - P_e$) yields a statistic, κ, which can take values from 1 to -1. Values between 1 and 0 indicate agreement better than chance, a value of 0 indicates a level of agreement that could have been expected by chance, values between 0 and -1 indicate levels of agreement that are worse than chance.

All forms of kappa are based on Equation 1; however, the various forms of kappa can be distinguished by how P_o and P_e are defined. One class of kappa statistics uses a P_e

that is dependent on marginal distributions; in this article these will be referred to as *fixed-marginal kappas*. Cohen's kappa, Scott's pi, and Fleiss' multirater kappa are fixed-marginal forms of kappa. Other forms of kappa - such as Byrt et al.'s *PABAK* (1993), Brennan and Prediger's κ_m (1981) Maxwells's *RE* (1977), and Bennet, Alpert, & Goldstein's S (1954), are not dependent on their marginal distributions and will be referred to in this article as *free-marginal kappas*.

Coming back to Fleiss' multirater kappa, Fleiss defines P_o as:

Insert Equation 2 here, centered [2]

where *N* is the number of cases, *n* is the number of raters, and *k* is the number of rating categories. Fleiss defines P_e as:

Insert equation 3 here, centered[3]

Table 1, below, is a hypothetical situation in which N = 4, k = 2, and n = 3. The quantity (n_{ij}) in the *yes* and *no* columns is the number of raters who assigned the case to the same category.

[Insert Table 1 here]

From the data in Table 1, Fleiss' P_o should be .67:

$(3^2 + 3^2 + 2^2 + 2^2 + 1^2 + 1^2 + 0^2 + 0^2 – 4*3) / [4*3*(3-1)] = .67$

and Fleiss' P_e should be .50:

$([3 + 2 + 1+ 0] / [4*3])^2 + ([0+1+2+3] / [4*3])^2 = .50$

and Fleiss' multirater kappa should then be .34

$(.67 - .50) / (1-.50) = .34$

The P_o of .67 indicates that raters agreed on 67% of cases. The P_e of .50 indicates that raters would have been expected to agree on 50% of cases purely by chance. The positive value of kappa indicates that agreement is slightly better than what would have been expected by chance.

In Table 1 the proportion of *yes* ratings to the total number of ratings (50%) is equal to the proportion of *no* ratings to the total number of ratings (50%); there is no prevalence of one type of category over another in this data set. If marginal symmetry is defined as the proportion of the number of *yes* ratings to the total number of ratings, where the values of 0 and 1 are indicators of the maximal degree of prevalence and a value of 0.5 indicates that there is perfect symmetry, then the data in Table 1 have perfect symmetry, and thus no prevalence.

Fleiss' Multirater Kappa, Prevalence, and Symmetry

It can be shown that Fleiss' multirater kappa, like its birater fixed-marginal counterparts, is affected by prevalence. Table 2 shows a hypothetical multirater situation where the number of raters (3), categories (2), cases (4), and percent of overall agreement

(.67) is the same as in Table 1; however, in Table 2 the marginal distributions are not symmetrical; the proportion of *yes* ratings (83%) is not equal to the proportion of *no* ratings (17%). Despite Table 2's matrix having all of the same characteristics as Table 1's matrix, except for the symmetry of the marginal distributions, Fleiss' multirater kappa for the data in Table 2 is -.2, whereas in Table 1 kappa it is .34

[Insert Table 2 here]

Figure 1 plots the kappa of a data set with the same general parameters as Table 1 and Table 2 as it takes on each of the 9 values of symmetry (i.e., .83, .75, .65, .58, .50, .42, .25, and .17) that are possible for a data set with 3 raters, 4 cases, and 2 categories. Figure 1 shows that for Fleiss' multirater kappa there is a strong quadratic relationship between symmetry (i.e., the number of *yes* ratings / the number of total ratings) and the value of kappa when the number of raters, categories, cases, and percent of overall agreement is held constant.

According to Brennan and Prediger (1981), the dependency of fixed-marginal kappas on their marginal distributions make them desirable for validity studies or studies where marginal distributions are known to raters beforehand. For example, it would be desirable to use a fixed-marginal kappa when (a) one wishes to know the chance-adjusted level of agreement of a trainee-rater's ratings with 'true' ratings and (b) the trainee-rater was informed of the number, but not which, of the cases should be assigned to each case – that is, when marginals are fixed. However, this same dependency is an undesirable characteristic for reliability/agreement studies when the (a) level of agreement of interest is between two or more raters with each other or (b) the number of cases that should be assigned to each category is not known by the raters a priori. Brennan and Prediger argue

that having fixed marginals, in both rows and columns, is an assumption that should be met before marginally-dependent kappa statistics like Cohen's kappa, Scott's pi, and Fleiss' multirater kappa can be used.

Free-Marginal Alternatives

One popular solution to the prevalence paradox in an agreement/reliability study with two raters is to assume that marginals are free and set P_e equal to $1/k$, where k is the number of rating categories. For example, if there are three rating categories, under the null hypothesis any two raters would be expected to agree on 1/3 of the cases. The free-marginal solution in the birater case has been suggested by many and has been shown to avoid the prevalence and bias paradox of high agreement but low kappa (Bennet et al. 1954; Brennan & Prediger, 1981; Byrt et al., 1954; Lawlis & Lu, and Maxwell, 1977) .

Free-marginal versions of kappa are recommended when raters are not restricted in the number of cases that can be assigned to each category, which is often the case in the typical agreement study. Cohen notes that in the typical reliability/agreement study "there is no criterion for the 'correctness' of judgments, and the judges are a priori deemed equally competent to make judgments. Also, there is *no restriction* [italics added] on the distribution of judgments over categories for either judge" (as cited in Brennan and Prediger, p. 692).

Although free-marginal solutions avoid the prevalence and bias paradoxes associated with fixed-marginal version of kappa, they are not without their faults. Concerning Bennet's S, a free-marginal version of kappa, Scott (1955) argues that,

> As the number of categories increases, S increases, for a fixed P_o. And herein lies a spurious effect . . . The index is based on the assumption that all categories in

> the dimension have equal probability of use 1/k [where k is number of categories]. This is an unwarranted assumption for most behavioral and attitudinal research. (p. 322)

Others, like Lawlis and Lu (1972), argue that "every judgment has the same probability of occurring under the hypothesis that the judges have no understanding of the scales being applied and their ratings are purely random" (p. 17). Applying the logic of hypothesis testing to kappa, I am led to believe that Lawlis and Lu's assumption is correct because kappa is, or should be, interpreted as index of how likely a particular level of agreement is, given chance, not how likely chance is, given a particular level of agreement.

Lawlis and Lu (1972) do concede that when there are categories that are intended to be avoided, the probability of agreement by chance is higher than 1 / *k* categories. Brennan and Prediger suggest that the empty or avoided category problem can be remedied by doing pilot studies and collapsing or revising categories so that they are functional. I, on the other hand, suggest that the answer is not to avoid using empty or low-frequency categories altogether, but to have a strong theoretical justification for including low-frequency or empty categories. If the population parameter is such that there is a low instance of one or more theoretically justified categories, then it does not follow that raters should be automatically *punished* because of natural population parameters.

Multirater Free-Marginal Kappa: Multirater κ_{free}

Multirater κ_{free} is a version of kappa that can be used in the multirater case. Basically, it uses the same P_o as does Fleiss' multirater kappa, but the P_e is

(Insert Equation 4 here, centered)[4]

Since multirater κ_{free} uses Fliess' P_o and the P_e in Equation 4, multirater κ_{free} in its entirety is:

[Insert Equation 5 here]

Coming back to Tables 1 and 2, using the multirater κ_{free} formula, the kappa for both tables is .33. In contrast, when using Fleiss' multirater formula, the kappa for Table 1, which has perfect symmetry, is .34 and for Table 2, which has a high degree of asymmetry, the kappa is -.02.

Figure 2, which is a slight variation of Figure 1, illustrates how values of multirater κ_{free} and Fleiss' multirater kappa compare over different levels of symmetry when all other variables – number of raters, cases, categories, and percent of overall agreement- are held constant. Figure 2 shows that the values of multirater κ_{free} are constant over varying levels of symmetry and that Fleiss' multirater kappa varies. Multirater κ_{free} and Fleiss' multirater kappa converge when there is perfect symmetry (i.e., when there are an equal number of cases in each category.)

While Fleiss' multirater kappa varies as a function of symmetry of marginal distributions, multirater κ_{free} varies as a function of the number of categories. Figure 3 illustrates the relationship between values of multirater κ_{free} and number of rating categories (from 2 to 10) when the parameters of the data sets, besides the number of

rating categories, in Tables 1 and 2 are held constant. As shown in Figure 3, the value of multirater κ_{free} raises nonlinearly as the number of categories increase. The difference in multirater κ_{free} between *k* and *k+1* decreases as k increases.

Recommendations for the Use of Multirater κ_{free}

Both multirater κ_{free} are Fleiss' multirater kappa can be used as agreement indices when there are more than two raters assigning cases to nominal categories; however, the decision of whether multirater κ_{free} or Fleiss' multirater kappa should be used is determined by whether all marginals are fixed or whether one or more marginals are free. Multirater κ_{free} is appropriate when one or more marginals are not fixed. If all marginals are fixed, then Fleiss' multirater kappa is the appropriate index. In validity studies when an index of agreement beyond chance is desired and when the marginals of the test ratings are fixed *and* proportional to the marginals of the criterion ratings, then it is appropriate to use Fleiss' multirater kappa; otherwise multirater κ_{free} should be used.

If one decides to use multirater κ_{free}, the number of categories should be carefully considered. I suggest using as few categories as possible, each of which must have a strong theoretical or empirical justification. Using more categories than are theoretically justified will spuriously inflate the value of multirater κ_{free}.

Summary

In summary, kappa statistics can classified according to the number of raters that are used (birater or multirater) and whether they assume fixed or free-marginal distributions. Table 3 shows how the common kappa statistics, including multirater κ_{free}, fall into those classifications.

[Insert Table 3 here.]

As is illustrated in Table 3, Fleiss' multirater kappa, the most widely-used multirater index of interrater agreement for variables with nominal categories, is not appropriate for situations in which marginals are not fixed. Although there are many versions of kappa that are suitable for situations in which marginals are not fixed, these versions have only been explicated in the birater case. In this article, I introduced a multirater version of kappa (multirater κ_{free}) that *is* appropriate when marginals are not fixed, which is the case in the typical agreement study. Unlike Fleiss' multirater kappa, values of multirater κ_{free} do not vary as a function of the symmetry of marginal distributions; they vary as a function of the number of rating categories used. Multirater κ_{free} is recommended as an alternative to Fleiss' multirater kappa when raters do not know a priori how cases are to be distributed in categories. Since multirater κ_{free} can be inflated by adding superfluous categories, it is recommended that as few rating categories be used as are theoretically or empirically justifiable. Information on calculating multirater κ_{free} with SPSS can be found in Randolph (2005)

References

Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999). Beyond kappa: A review of agreement measures. *Canadian Journal of Statistics (27)*1, 3-23.

Bennet, E. M., Alpert, R., & Goldstein, A.C. (1954). Communications through limited response questioning. *Public Opinion Quarterly (18),* 303-308.

Brennan, R. L. & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement (41),* 687-699.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology* (46), 423-429.

Chang, A. (n.d.). *Statistics toolbox*. Retrieved October 13, 2005 from http://department.obg.cuhk.edu.hk/researchsupport/statmenu.asp

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement (1),* 37-46.

Feinstein, A. R. & Cicchetti, D. V. (1990a). High agreement but low kappa: I. The problems of the two paradoxes. *Journal of Clinical Epidemiology (43),* 543- 549.

Feinstein, A. R. & Cicchetti, D. V. (1990b). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology (43),* 551-558.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin (76),* 378-382.

Lawlis, G. F. & Lu, E. (1972). Judgements of counselling process: Reliability, agreement and error. *Psychological Bulletin (78),* 17-20.

Light, R. J. (1971). Measures of response agreement for qualitative data. Some generalizations and alternatives. *Psychological Bulletin (76),* 365-377.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry (130),* 79-83.

Nichols, D. (1997). *MKAPPASC.SPS/MKAPPASC.TXT (SPSS Macro/SPSS read me), 1997*. Retrieved September 12, 2005 from ftp://ftp.spss.com/pub/spss/statistics/nichols/macros/.

Randolph, J. J. (2005). SPSS *macro for computing multirater free-marginal kappa*. Retrieved October 17, 2005 from http://www.geocities.com/justusrandolph/mrak_macro

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly (19),* 321-325.

Siegel, S. & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad [1]$$

$$P_o = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij^2} - Nn \right) \quad [2]$$

$$P_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 \quad [3]$$

$$P_e = \frac{1}{k} \quad [4]$$

$$\kappa_{free} = \frac{\left[\frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij^2} - Nn \right) \right] - \left[\frac{1}{k} \right]}{1 - \left[\frac{1}{k} \right]} \quad [5]$$

Table 1

Hypothetical Data Set 1

Case	*Category*		*Sum*
	Yes	*No*	
1	3	0	3
2	2	1	3
3	1	2	3
4	0	3	3
Total	6	6	12

Table 2

Hypothetical Data Set 2

Case	*Category*		*Sum*
	Yes	*No*	
1	3	0	3
2	2	1	3
3	2	1	3
4	3	0	3
Total	10	2	12

Table 3

A Rater and Marginal Framework for Classifying Kappa Statistics

	Birater	*Multirater*
Assumes fixed-marginal	Cohen's κ Scott's π Bennet's *S*	Fleiss' Multirater κ
Assumes free-marginal	Brennan and Prediger's κ_m Maxwell's *RE* Byrt et al's PABAK	Multirater κ_{free}

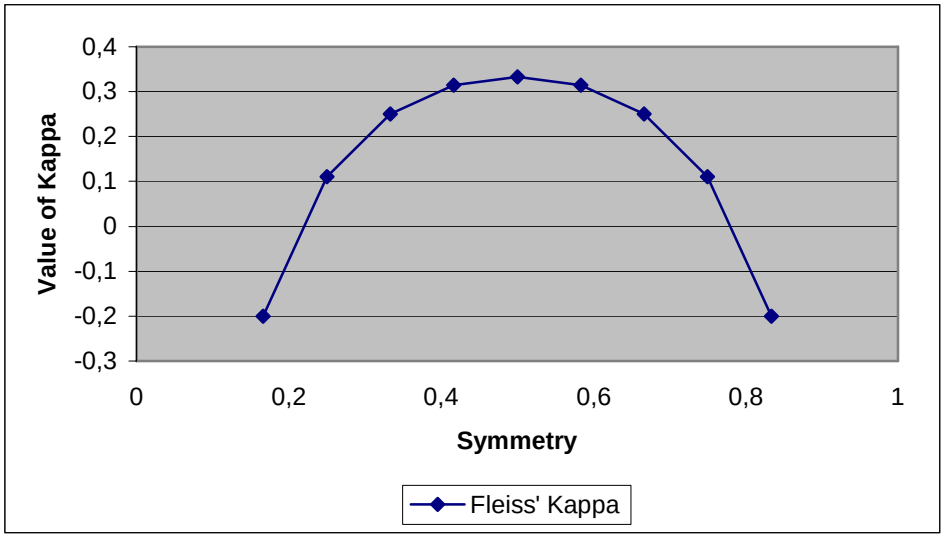

Figure 1. The Relationship between Symmetry and Kappa - Fleiss' Kappa

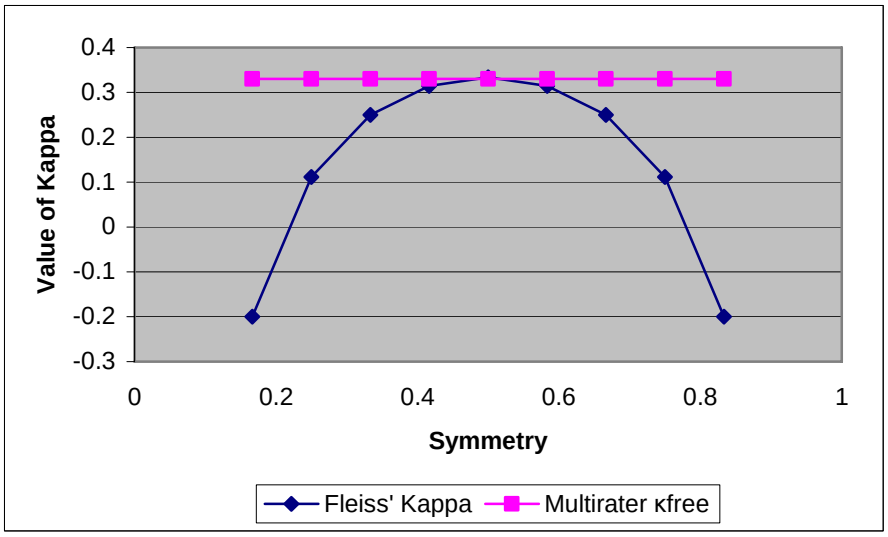

Figure 2. The Relationship between Symmetry and Kappa - Fleiss' Kappa and MFK

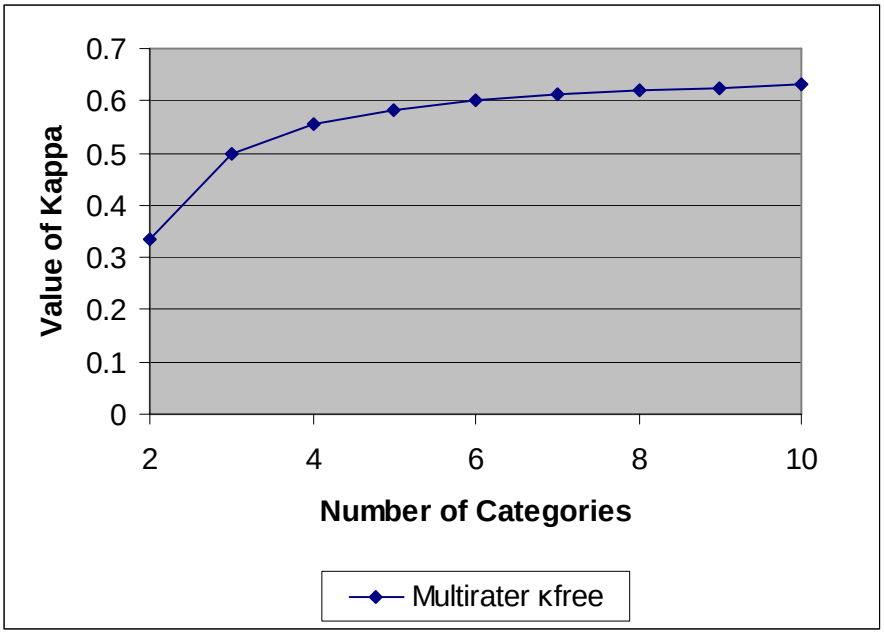

Figure 3. The Relationship between multirater κ_{free} and Number of Rating Categories