Statistics Corner:

How do we calculate rater/coder agreement and Cohen's Kappa?

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

I am working on a study in which two raters coded answers to 6 questions about study abroad attitudes/experience for 170 Japanese university students. The coding was done according to a rubric in which there were 4 – 8 possible responses per question. Since most—if not all—of the data is categorical, I have heard that Cohen's Kappa is the most common way of ascertaining inter-rater agreement. What is the best way to actually calculate that? Since more and more people are moving away from single-rater assessments to multi-rater assessments, this question should be relevant to *Shiken Research Bulletin* readers.

Answer:

In order to address your question, I will have to describe the *agreement coefficient* as well as the *Kappa coefficient*. I will do so with a simple example, then with the more complex data that you have in your study.

Simple agreement coefficient example

In the realm of ratings or codings (hereafter simply called codings) of various categories, an *agreement coefficient* can be used to estimate the proportion of codings assigned by two raters or coders (hereafter simply called coders) that coincide. In the simplest scenario, let's say that two coders listen to the interview data of 120 students who were interviewed just after returning from a study abroad experience. After listening to each interview, each of two coders is required to decide if the student was generally positive about the living abroad experience or generally negative. In other words, the coders are required to code each student's experience as positive or negative. Figure 1 illustrates how we need to lay out the results for the two coders in order to calculate an agreement coefficient.

In some cases, the codings agree between the two coders. When the two assigned codings for a student are both positive, that student is counted in cell A; when the two assigned codings for a student are both negative, that student is counted in cell B. The other cells indicate that the two coders disagreed in their codings (i.e., Coder A assigned a positive coding, but Coder B assigned a negative one, or vice versa). Notice that the row totals Row1 and Row2 are given to the right of Figure 1, and column totals Col1 and Col2 are given at the bottom. Notice also that the grand total (also affectionately known as *N*) is shown at the bottom right.

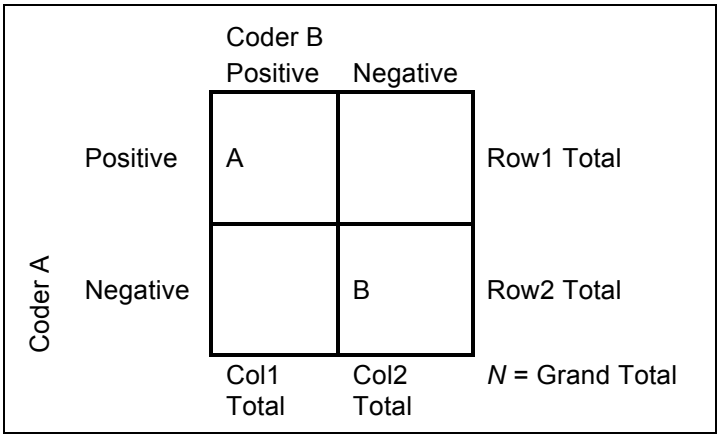

Figure 1. **Layout for positive and negative coding data for two coders.**

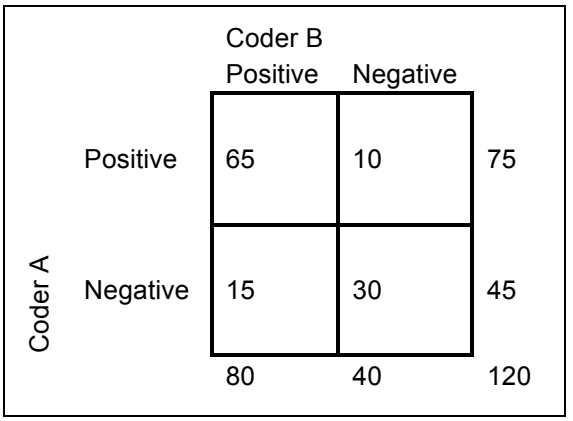

Figure 2. Sample data for positive and negative coding data for two coders.

Coming back to our imaginary scenario, notice in Figure 2 that 65 out of the 120 students were classified as positive by both coders, while 30 others were classified as negative by both coders. In addition, 25 students (10 + 15 = 25 students) are classified differently by the two coders.

With this information in hand, the following formula (Equation 1) can be used to calculate the agreement coefficient:

$$p_o = \frac{A + B}{N} \tag{1}$$

where:

p_o = agreement coefficient (or proportion observed)
A = number of agreed codings in cell A
B = number of agreed codings in cell B
N = total number of codings

Substituting the values found in Figure 2 into the equation, we get (Equation 2):

$$p_o = \frac{65 + 30}{120} = \frac{95}{120} = .7916666 \approx .79 \quad (2)$$

Thus, the agreement coefficient in this case is about .79, which means that the coders classified the students in the same way about 79% of the time. Note that by extension, the coders disagreed 21% of the time (100% - 79% = 21%).

If all students were coded exactly the same way by both coders, the coefficient would be 1.00 [e.g., $(A + B) / N = (70 + 50) / 120 = 1.00$], so 1.00 is the maximum value the agreement coefficient can have. However, unlike the reliability coefficients (for more on this concept, see Brown, 1997, 2002) that researchers often use, the agreement coefficient can't be lower than what would result by chance. Put another way, with 120 students, we might reasonably expect 30 students per cell by chance alone. This would result in a coefficient of .50 [$(A + B) / N = (30 + 30)/120 = 60/120 = .50$]. Thus, no agreement coefficient can logically be any lower than what would occur by chance alone—in this case, no lower than .50. This is very different from reliability estimates, which can go as low as .00.

Simple Kappa coefficient example

The *Kappa coefficient* (κ) arose (due to Cohen, 1960) to adjust for this chance-lower-limit problem by providing an estimate of the proportion of agreement in classifications beyond what would be expected to occur by chance alone. The adjustment is given in the following formula (Equation 3):

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \quad (3)$$

where:

κ = the Kappa coefficient
p_o = agreement coefficient (or proportion observed)
p_e = proportion classification agreement that could occur by chance alone,
In this case: $p_e = (Row1 \times Col1) + (Row2 \times Col2) / N^2$ [in this case]

Before calculating κ, a researcher must calculate p_o and p_e for the particular data involved. We have already calculated $p_o = .7916666$ for the data in Figure 2; the calculations for p_e for the same data are as follows (Equation 4):

$$p_e = \frac{(Row1 \times Col1) + (Row2 \times Col2)}{N^2} = \frac{(75 \times 80) + (45 \times 40)}{120^2} = \frac{6000 + 1800}{14400} = \frac{7800}{14400} = .5416666 \quad (4)$$

Notice that we are calculating p_e by (a) multiplying the total for the row in which cell A is found by the column for cell A (Row1 × Col1), (b) multiplying the row total for cell B times the column total for cell B (Row2 × Col2), then (c) adding the two results together [(Row1 × Col1) + (Row2 × Col2)], and (d) dividing the whole thing by N^2. In doing so, we are calculating the proportion of the expected frequencies for cells A and B.

Since the p_o for the data in Figure 2 was .7916666 and now we know that p_e is .5416666[1], Kappa turns out to be (Equation 5):

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} = \frac{(.7916666 - .5416666)}{(1 - .5416666)} = \frac{(.25)}{(.4583334)} = .5454544 \approx .55 \quad (5)$$

This Kappa coefficient is an estimate of the coding agreement that occurred beyond what would be expected to occur by chance alone. It can also be interpreted as a proportion of agreement, .55 in this case, or as a percentage of agreement, 55% in this case. Unlike the agreement coefficient, Kappa represents the percentage of classification agreement beyond chance, so it is not surprising that Kappa is usually lower than the agreement coefficient. Like the agreement coefficient, Kappa has an upper limit of 1.00, but unlike the agreement coefficient, Kappa has the more familiar .00 lower limit.

Your more complex agreement coefficient example

In the case of your more complex coding criteria, your data should be laid out as shown in Figure 3 for your Question 1. Figure 4 shows the labels and examples you sent me for your Question 1 rubric. Notice that instead of cells A and B (as in the simpler example), we are focused here on agreements in cells A, B, C, D, E, and F.

		Coder B						
		5	4	3	2	1	0	
Coder A	5	A						Row1
	4		B					Row2
	3			C				Row3
	2				D			Row4
	1					E		Row5
	0						F	Row6
		Col1	Col2	Col3	Col4	Col5	Col6	N

Figure 3. **Layout for categories 5, 4, 3, 2, 1, and 0 data for two coders for question 1 in your study.**

Your actual data are displayed in Figure 5. Notice that the agreements in cells A, B, C, D, E, and F (at 74, 21, 1, 9, 20, and 25, respectively) are fairly high relative to the other cells. That is an indication that the two raters are tending to agree with each other in their ratings.

[1] Note that, for the sake of accuracy, I keep all the decimal places that my 100-Yen-shop calculator gives me until the very last step where I round the result off to two-places in order to be consistent with APA formatting.

5 = A strong interest in SA	Ex: I'm very interested in abroad study because I want to speak English fluently. [T1]
4 = A mild interest in SA	Ex: A little. But, I like to go abroad. [K4]
3 = Neutral and/or ambivalent	Ex: It's so-so. I want to study abroad, but I don't have money. And you? [T80]
2 = Little interest in SA	Ex: I have just little interested about study abroad. [K28]
1 = A strong disinterest in SA	Ex: I'm nothing. I like Japan school. And you? [T90]
0 = No response	

Figure 4. **Coder options for question 1 in your study.**

		Coder B						
		5	4	3	2	1	0	
Coder A	5	74	0	0	0	0	0	74
	4	0	21	0	1	1	0	23
	3	1	5	1	3	0	0	8
	2	0	2	0	9	0	0	11
	1	0	0	0	2	20	0	22
	0	3	1	0	1	0	25	30
		78	78	29	10	16	21	25

Figure 5. **The actual data for categories 5, 4, 3, 2, 1, and 0 for two coders for question 1 in your study[2].**

With Figures 3 and 5 in hand, the calculation of the agreement coefficient is simple with the following equation (6):

$$p_o = \frac{A + B + C + D + E + F}{N} \tag{6}$$

where:

p_o = agreement coefficient (or proportion observed)
A to F= number of agreed codings in cells A to F
N = total number of codings

[2] You asked me in an aside if this could be a Likert item (see Brown, 2000, 2011). I'm inclined to say "no" because you really are interested in how consistently the coders judged data to be in these different categories (including "no response). If you had been asking the students themselves to rate their experience using this scale, then I would say that it should be analyzed as a Likert item. Even if all the above were not true, your other questions are all clearly nominal in nature, so it is probably best if you use agreement and Kappa to consistently analyze all of your questions the same way.

Substituting the values found in Figure 5 into the equation, we get (Equation 7):

$$p_o = \frac{74 + 21 + 1 + 9 + 20 + 25}{170} = \frac{150}{170} = .88235294 \approx .88 \quad (7)$$

Thus, the agreement coefficient is about .88, which indicates that the coders agreed in their classifications of students about 88% of the time (and that they disagreed 12% of the time).

Your more complex Kappa coefficient example

Recall that the Kappa coefficient arose to adjust for this chance-lower-limit problem by providing the proportion of consistency in classifications beyond what would be expected to occur by chance alone and that the adjustment is given in the following formula (Equation 8):

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \quad (8)$$

where:

κ = the Kappa coefficient
p_o = agreement coefficient (or proportion observed)
p_e = proportion classification agreement that could occur by chance alone
In this case:
p_e = [(*Row*1 × *Col*1) + (*Row*2 × *Col*2) + (*Row*3 × *Col*3) + (Row4 × Col4) + (*Row*5 × *Col*5) + (*Row*6 x *Col*6)] / N^2

Before calculating κ, we must calculate p_o for the particular data table involved. For the data in Figure 5, the calculations would be as follows (Equation 9):

$$p_e = \frac{(Row1 \times Col1) + (Row2 \times Col2) + (Row3 \times Col3) + (Row4 \times Col4) + (Row5 \times Col5) + (Row6 \times Col6)}{N^2} \quad (9)$$

Notice that this time we are calculating p_e by multiplying the row and column totals for cells A, B, C, D, E, and F and adding them up before dividing by N^2. In doing so, we are calculating the proportion of the expected frequencies for cells A through F. Substituting in the values from Figure 5, we get (Equation 10):

$$\begin{aligned} p_e &= \frac{(74 \times 78) + (23 \times 29) + (8 \times 10) + (11 \times 16) + (22 \times 21) + (30 \times 25)}{170^2} \\ &= \frac{5772 + 667 + 80 + 176 + 462 + 750}{28900} = \frac{7907}{28900} = .2735986 \end{aligned} \quad (10)$$

Given that p_o for the data in Figure 5 is .88235294 and p_e is .2735986, Kappa turns out to be (Equation 11):

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} = \frac{(.88235294 - .2735986)}{(1 - .2735986)} = \frac{(.6087543)}{(.7264014)} = .8380411 \approx .84 \quad (11)$$

This Kappa coefficient shows that the proportion of coding agreement that occurred beyond what we would expect by chance alone is .84, so the percentage of agreement is a respectable 84%, and we know that it could have fallen anywhere between .00 to 1.00.

Conclusion

This column described how to calculate rater/coder agreement and Cohen's Kappa. I have shown here how to lay out the data and calculate agreement and Kappa coefficients for a simple set of data based on binary decisions by two coders, as well as for the data generated in your study for six-category decisions by two coders. I hope you see how you need to arrange your data for Questions 2 to 6 in order to calculate p_o, and more importantly, how to calculate p_e from the row and column totals associated with each agreement cell (regardless of the number of decisions involved) in the process of then calculating agreement and Kappa coefficients for each question in your study. Please note that I would generally report both the agreement and Kappa coefficients because they provide different types of information, both of which may be interesting to some readers.

References

Brown, J. D. (1997). Statistics Corner: Questions and answers about language testing statistics: Reliability of surveys. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1*(2), 17-19. Retrieved from http://www.jalt.org/test/bro_2.htm

Brown, J. D. (2000). Statistics Corner. Questions and answers about language testing statistics: What issues affect Likert-scale questionnaire formats? *Shiken: JALT Testing & Evaluation SIG Newsletter, 4*(1), 18-21. Retrieved from http://www.jalt.org/test/bro_7.htm

Brown, J. D. (2002). Statistics Corner. Questions and answers about language testing statistics: The Cronbach alpha reliability estimate. *Shiken: JALT Testing & Evaluation SIG Newsletter, 6*(1), 14-16. Retrieved from http://www.jalt.org/test/bro_13.htm

Brown, J. D. (2011). Statistics Corner. Questions and answers about language testing statistics: Likert items and scales of measurement? *Shiken: JALT Testing & Evaluation SIG Newsletter, 15*(1), 10-14. Retrieved from http://www.jalt.org/test/bro_34.htm

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Where to Submit Questions:

Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu

JD Brown
Department of Second Language Studies
University of Hawai'i at Mānoa
1890 East-West Road
Honolulu, HI 96822
USA

Your question can remain anonymous if you so desire.