KappaAcc: Deciding Whether Kappa is Big Enough by Estimating Observer Accuracy

Technical Report 28

Roger Bakeman

Developmental Laboratory
Department of Psychology
Georgia State University
Atlanta, GA 30303

bakeman@gsu.edu

Summary:

Kappa, a statistic that gauges inter-observer agreement corrected for chance, is generally regarded as superior to the percentage agreement sometimes reported. But when is kappa big enough? No one value of kappa can be regarded as universally acceptable. Any published article that claims otherwise is misinformed. Whether a given value of kappa is acceptable depends on the percentage accuracy researchers require of their observers. How accurate observers need to be to achieve a particular value of kappa depends on several factors, the most important of which is the number of codes or ratings observers are asked to assign to events. Described here and available for download at no charge is a computer program, **KappaAcc**, that computes various kappa statistics and estimates observer percentage accuracy. Whenever a kappa is reported, estimated observer accuracy should be reported as well, as a way to judge the adequacy of the computed value.

Support for the development of KappaAcc was provided by the National Institutes of Health, NICHD (R01-HD035612).

KappaAcc: Deciding Whether Kappa is Big Enough by Estimating Observer Accuracy

1 Omnibus and Weighted Kappa

Researchers who ask observers to code or rate behavior typically gauge inter-observer agreement with kappa (Cohen, 1960). The kappa statistic is useful both when training observers and later when published reports seek to convince others that observers were not just making idiosyncratic judgments. Typically researchers ask two observers to code (nominal scale) or rate (ordinal scale) the same sequence of events (or time intervals, or sessions), applying *K* codes or ratings to *N* events. Their *N* pairs of judgments are then tallied in a *K* × *K* kappa table (or agreement matrix); rows represent one observer, columns the other observer, and rows and columns are labeled with the *K* codes or ratings.

Cohen's kappa is an omnibus statistic, a single number that summarizes the agreement evidenced by the kappa table. The standard formula is:

$\kappa = \frac{P_O - P_C}{1 - P_C}$, where $P_O = \sum_{i=1}^{K} p_{ii}$ and $P_C = \sum_{i=1}^{K} p_{+i} p_{i+}$

P_O represents observed agreement (the sum of the probabilities on the upper-left to lower-right diagonal), P_C represents chance agreement (the sum of the corresponding row and column probability products), and the formula emphasizes that kappa gauges observer agreement corrected for chance.

The formula for weighted kappa (Cohen, 1968) is more general:

$$\kappa_{wt} = 1 - \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} x_{ij}}{\sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} e_{ij}}$$

Each observed value (x_{ij}) and each expected value (e_{ij}) is multiplied by the corresponding cell in an array of weights (w_{ij}). (Note: The formula for weighted kappa in Bakeman and Quera, 2011, p. 82, contains a typo; the "1 –" before the fraction was inadvertently omitted.) When *K* = 5, standard weights would be as shown below.

In an array of weights, 0s indicate agreements. Here, the 0s on the diagonal indicate times when both observers made the same judgment. The 1s indicate disagreements. With standard weights—all off-diagonal cells set to 1—all disagreements are weighted equally and both the usual and the weighted kappa formulas yield identical results.

	A	B	C	D	E
A	0	1	1	1	1
B	1	0	1	1	1
C	1	1	0	1	1
D	1	1	1	0	1
E	1	1	1	1	0

standard

If observers agreed for all events, the sum of the $w_{ij}x_{ij}$ products would be 0, the fraction after "1 –"would be 0, and so κ and κ_{wt} would equal 1, indicating perfect agreement.

When codes are nominal, it usually makes the most sense to weight all disagreements equally. But when ratings are ordinal, other arrays of weights could make sense. Here are three possibilities that could be used when observers are asked to rate events (or time intervals or sessions) 1 to 5:

	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

linear

	1	2	3	4	5
1	0	0	1	1	1
2	0	0	0	1	1
3	1	0	0	0	1
4	1	1	0	0	0
5	1	1	1	0	0

w/1 standard

	1	2	3	4	5
1	0	0	1	2	3
2	0	0	0	1	2
3	1	0	0	0	1
4	2	1	0	0	0
5	3	2	1	0	0

w/1 linear

The **linear** array weights more extreme disagreements more highly (e.g., weighting a 1–3 disagreement 2 but a 1–5 disagreement 4). The **w/1 standard** array regards disagreements within one scale point as agreements and weights them 0 (e.g., weighting a 1–2 disagreement 0 but a 1–3 disagreement 1). And the **w/1 linear** array likewise regards disagreements within one scale point as agreements but weights more extreme disagreements more highly (e.g., weighting a 1–3 disagreement 1 but a 1-4 disagreement 2). The KappaAcc program lets users select the weighting scheme, including any custom scheme they might wish to define, that makes sense to them.

2 No One Value of Kappa is Universally Acceptable

Once paired observer judgments are tallied in a kappa table and kappa or weighted kappa computed, researchers want to know whether the computed value is big enough. Does its value indicate adequate observer agreement? It is possible to find articles in the literature that assign categorical terms like excellent to specific values of kappa. Such articles are misinformed, offer no reasoned arguments for their judgements, and to avoid providing them any traction should, in my opinion, not be cited.

Observer accuracy provides a reasoned way for determining whether a particular value of kappa is adequate. In the real world, the true value of observer accuracy is unknowable. Its computation requires that we know the true state of affairs. But in an ideal world of simulated observers, we can specify the true state of affairs, specifically: (a) the number of codes or ratings and their simple probabilities; (b) observer accuracy—e.g., the probability that an observer will assign code A when the event is truly an A; and (c) observer inaccuracy—e.g., the probability that an observer will assign code B when the event is actually an A.

Gardner (1995) has shown us how to model observer decision making in the ideal world. His equations let us determine the value of kappa that would result if two observers of specified accuracy and inaccuracy were asked to assign *N* codes or ratings to events of specified probability (see also Bakeman, Quera, McArthur, & Robinson, 1997; Bakeman

& Quera, 2011). The inference is, if simulated observers of known accuracy achieve a value of kappa just slightly greater than the value achieved by observers in the real world, it is reasonable to assume that the actual observers are at least as accurate as the simulated ones.

Using Gardner's equations, the KappaAcc program determines the percentage accuracy required of simulated observers to achieve the magnitude of the kappa observed by the actual observers. Parameters for the simulated observers are set as follows.

First, KappaAcc sets the simple probabilities as the means of the two observers' corresponding probabilities. The model requires estimates of the simple probabilities; estimating them with averages makes the estimates reflect the observed variability—important because we know from earlier work (Bakeman et al., 1997; Bakeman & Quera, 2011) that the estimated value of kappa declines somewhat when simple probabilities are more variable.

Second, KappaAcc sets accuracy—the probability that each observer will correctly code or rate each event—to a given value, e.g. 90%. The same value is set for each code or rating primarily because this is the simplest assumption and because rationalizing different accuracies for different codes seems somewhere between challenging to impossible. Likely, in the real world observers are not equally accurate across codes or ratings, which only means that the estimated percentage accuracy produced by KappaAcc should be interpreted as an average accuracy for the given codes or rating. Similarly, the same values are set for both observers, again because this is the simplest assumption. Likely in the real world observers are not equally accurate, but again the estimated percentage accuracy should be interpreted as an average value for the two observers.

Third, KappaAcc sets inaccuracy—the probability that each observer will incorrectly code or rate each event—to a given value, e.g. 2.5% when $K = 5$. With $K = 5$, there is one way observers can be accurate (coding an A an A) but four ways they could be inaccurate (coding an A as a B or C or D or E). Again, the same value is set for all errors for each code, and for both observers, for the reasons given for accuracy.

In sum, given a particular kappa table, with its row and column marginals (i.e., each observer's simple probabilities), KappaAcc determines a percentage accuracy that, given Gardner's equations, will yield an estimated value of kappa for simulated observers that is just slightly less than the actual computed value of kappa. The inference is that the observers must have been at least this accurate, on average, to produce the observed value of kappa. Researchers need to commit to a percentage accuracy they find acceptable, but this shouldn't prove difficult. As Bakeman and Quera (2011) wrote:

> Especially when training and checking observers, our main concern should not be the magnitude of kappa, but the level of observer accuracy we regard as acceptable. … Any judgment is arbitrary, but 80% is a good candidate for a

> minimum level of acceptability. Gardner (1995) characterized 80% as discouragingly low "but possibly representative of the accuracy of classification for some social behaviors or expressions of affect" (p. 347). It seems reasonable to expect better, and—although 100% accuracy will likely elude us—85%, 90%, or even 95% accuracy may represent reasonable goals. (p. 65)

3 Gardner's Model

Gardner (1995) modeled the decision making of each observer with an array of conditional probabilities, labeled *rho* (ρ) for the first observer and *sigma* (σ) for the second. Rows represent the observers' decision—the code or rating they assigned—and columns represent the true state of affairs. Thus cells on the diagonal indicate the probability that an observer will code or rate an event correctly.

	A	B	C	D	E
A	.90	.025	.025	.025	.025
B	.025	.90	.025	.025	.025
C	.025	.025	.90	.025	.025
D	.025	.025	.025	.90	.025
E	.025	.025	.025	.025	.90

rho (ρ)

If observers are 90% accurate and are equally accurate for all codes or ratings, then with $i = 1$ to K, ρ_{ii} and $\sigma_{ii} = .90$. And if observers are equally inaccurate for all codes or ratings, then when $i \neq j$ with $i = 1$ to K, $j = 1$ to K, ρ_{ij} and $\sigma_{ij} = (1 - .90)/(K - 1)$. The arrays to the right reflect these circumstances.

	A	B	C	D	E
A	.90	.025	.025	.025	.025
B	.025	.90	.025	.025	.025
C	.025	.025	.90	.025	.025
D	.025	.025	.025	.90	.025
E	.025	.025	.025	.025	.90

sigma (σ)

The simple probabilities for each code or rating constitute the third array— labeled pi (π)—used by Garner's model. If codes are equiprobable, then with $i = 1$ to K, $\pi_i = 1/K$, as shown here for $K = 5$.

A	B	C	D	E
.20	.20	.20	.20	.20

pi (π)

As noted in Bakeman et al. (1997):

> Now we can compute the expected unconditional probabilities for the cells of the agreement matrix given fallible observers. The formula is
>
> $$u_{ij} = \sum_{k=1}^{K} \rho_{i|k}\sigma_{j|k}\pi_k$$
>
> where u_{ij} represents a cell in the $K \times K$ agreement matrix. Each u_{ij} is the sum of K terms, where each term represents the probability that the first observer will code an event C_i and the second observer will code it C_j given a C_k event. Per basic probability theory, the probability of the joint event that constitutes each term is a product (AND), whereas the probability of any of these joint events occurring is a sum (OR). The terms in each series exhaust the possible ways the first observer might code an event C_i when the second observer codes it C_j. (p. 359)

Applying Gardner's formula for expected unconditional probabilities to the arrays for rho, sigma, and pi that assume 90% observer accuracy and equiprobable codes gives the values shown in the array to the right (expressed as percentages).

	A	B	C	D	E
A	16.250	0.938	0.938	0.938	0.938
B	0.938	16.250	0.938	0.938	0.938
C	0.938	0.938	16.250	0.938	0.938
D	0.938	0.938	0.938	16.250	0.938
E	0.938	0.938	0.938	0.938	16.250

expected probabilities (*u*)

Rounded to two significant digits, the value of kappa for this table is .77. The inference is that when a computed value of kappa is .77, given five equiprobable codes or ratings and parameter settings—observers equally accurate, errors equally distributed—observers can be regarded as at least 90% accurate, on average. Moreover, as we have shown previously, slightly lower values of kappa are acceptable when simple probabilities are more variable and when the number of codes or ratings is fewer (Bakeman et al., 1997; Bakeman & Quera, 2011, pp. 65–68, 165-166).

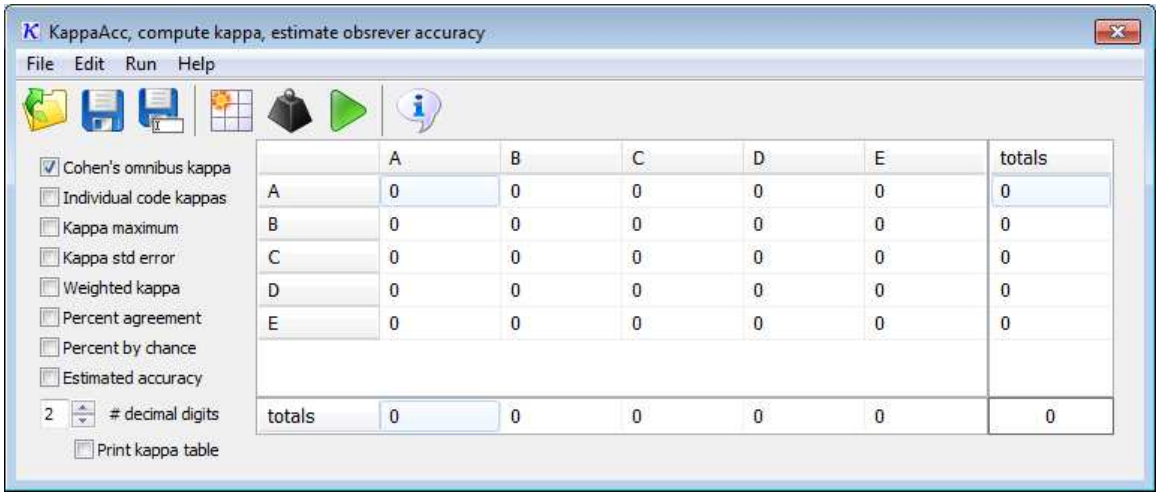
Here is another example. Observers were asked to rate 28 sessions 1 to 7; the kappa table that resulted is shown to the right. When computing kappa, disagreements within 1 scale point were counted as agreements (0) and other disagreements were weighted 1 (the w/1 scheme shown earlier).

	1	2	3	4	5	6	7	total
1	2	5						7
2	1	3	1					5
3	1	2	1					4
4			5	1				6
5			3	1	1			5
6					1			1
7								
total	4	10	10	2	2			28
M p	.1964	.2679	.2500	.1429	.1250	.0179	0	

The weighted kappa for this table was .70 (rounded to two digits; .6989 rounded to four digits). One observer did not rate any sessions 6 or 7 and the other observer did not rate any sessions 7. As a result the effective number of codes (*K*) for this table is 6, which the kappa computations take into account. Of the 28 paired judgements, only 4 differed by more than 1 scale point. KappaAcc estimated that, given the mean probabilities for each code (*M p* in the table), estimated probabilities by observers who were 89% accurate would yield a weighted kappa of .6987 and .7225 by observers who were 90% accurate (rounded to four digits). Hence, it seems reasonable to claim that a kappa of this magnitude requires observers who are at least 89% accurate on average.

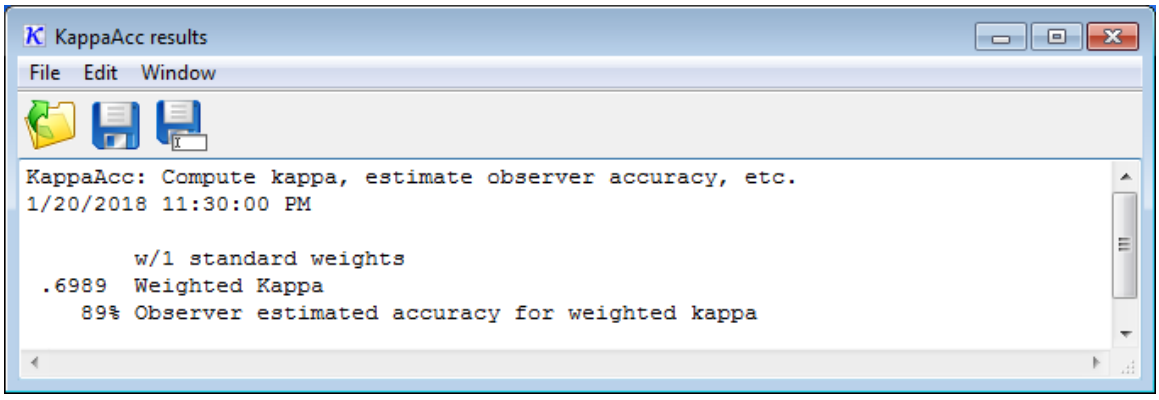
4 The KappaAcc Program

KappaAcc is an evolution of our initial ComKappa program (Robinson & Bakeman, 1997). It is essentially the same as ComKappa3, except for some minor bug fixes and, most notably, the additional capability to compute estimated observer accuracy. When initially invoked, a main window and a results window open. Here is the main window:



Selecting the table icon  or *Run > Define a new table* lets you define the number of codes or ratings and provide labels for them. You can then enter the values for the kappa table directly in the window or copy-and-paste the values from a spread sheet (with right clicks). If you want other than the standard weights, select the weight icon  or *Run > Specify weights* to select the weights you want. Next check the statistics you want computed and then select the compute icon  or *Run > Compute stats*.

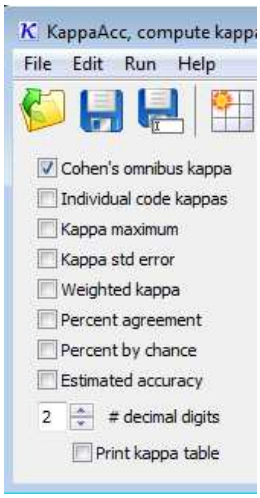
Here is the results window after entering the kappa table for the 1–7 ratings just given, selecting *w/1 standard weights* (agree within 1 point=0, otherwise 1), checking *Weighted kappa* and *Estimated accuracy*, and specifying 4 decimal digits for the output.



The contents of the results window can be saved to a text file (*File > Save*) or copy-and-pasted into a spread sheet (with right clicks).

Program details. KappaAcc is programmed in Pascal and compiled using Embarcadero® Delphi 10 Seattle. It will run on Windows computers or on Apple computers with a Windows simulator. It is contained in a single executable file, KappaAcc.exe, and once placed in a folder on your computer can be invoked, like any other program, with a double click. You could also create a shortcut and place it on your desktop. If your computer's security measures block running of "unknown" executable files, you may need help from your local IT people. This write-up as a PDF file and KappaAcc.exe are contained in the KappaAcc.zip file which can be downloaded at no charge from http://www2.gsu.edu/~psyrab/BakemanPrograms.htm.

5 Statistics Computed by KappaAcc



Before selecting the compute icon  check the appropriate boxes for the statistics you want computed. Possibilities are:

Cohen's omnibus kappa. Kappa as described earlier using standard weights.

Individual code kappas. As noted earlier, kappa is an omnibus statistic; it summarizes agreement for a set of mutually exclusive and exhaustive codes. Computing a separate kappa for each code (forming a 2 x 2 table for each code and computing its kappa) can be informative because it identifies particularly problematic codes.

Kappa maximum. In theory, values of kappa can vary from –1 to +1, where 1 represents perfect agreement. Negative values are rare and indicate greater than chance disagreement. But kappa can equal 1 only when the tallies for the corresponding rows and columns are the same —that is, when the simple probabilities for each code are the same for both observers. If not, the value of kappa can be no higher than kappa maximum.

Kappa std error. For completeness, CompKappa4 computes kappa's standard error. However, as noted in Bakeman and Gottman (1997):

> True, the standard error of kappa has been described (… Bakeman & Gottman, 1997, p. 65)—which means that a standardized kappa could easily be computed. However, statistical significance for kappa is rarely reported; as Bakeman and Gottman note, even relatively low values of kappa can still be significantly different from zero, but not of sufficient magnitude to satisfy investigators. (p. 63)

Weighted kappa. Especially useful for ordinal scales, as described earlier.

Percent agreement. Observed agreement—the sum of the probabilities on the upper-left to lower-right diagonal; P_O in the standard kappa formula.

Percent by chance. Chance agreement—the sum of the corresponding row and column probability products; P_C in the standard kappa formula.

Estimated accuracy. As described earlier.

6 Recommendation

To indicate that inter-observer agreement is adequate, journal articles commonly give the value of kappa obtained for sessions independently coded by two observers. I recommend such articles also give the estimated observer accuracy as computed by KappaAcc. For the example just given this could be stated as follows:

When two observers independently rated 28 sessions, the value of weighted kappa was .70 (disagreements within one scale point, counted as agreements, were weighted 0 and other disagreements were weighted 1). To produce a kappa of this magnitude simulated observers would need to be at least 89% accurate on average (computed by the KappaAcc program, Bakeman, 2018).

To cite this report:

Bakeman, R. (2018). KappaAcc: Deciding Whether Kappa is Big Enough by Estimating Observer Accuracy (Technical Report 28). Retrieved from Georgia State University website: http://bakeman.gsucreate.org/DevLabTechReport28.pdf

7 References

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, UK: Cambridge University Press.

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences.* Cambridge, UK: Cambridge University Press.

Bakeman, R., Quera, V., McArthur, D., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 2*, 357–370.

Gardner, W. (1995). On the reliability of sequential data: Measurement, meaning, and correction. In J. M. Gottman (Ed.), *The analysis of change* (pp. 339–359). Hillsdale, NJ: Erlbaum.

Robinson, B. F., & Bakeman, R. (1998). ComKappa: A Windows 95 program for calculating kappa and related statistics. *Behavior Research Methods, Instruments, and Computers, 30*, 731–732.