Scoring Long-form Constructed Response:
Statistical Challenges in Model Validation

Harry A. Layman
September 19, 2014

a

Abstract:

The increasing desire for more authentic assessment, and to the assessment of higher order cognitive abilities, is leading to an increased focus on performance assessment and the measurement of problem-solving skills, among other changes, in large scale educational assessment.

Present practice in production scoring for constructed response assessment items where student responses of one to several paragraphs are evaluated on well-defined rubrics by distributed teams of human scorers currently yields – in many cases -- results that are barely acceptable even for course-grained, single-dimension metrics. That is, even when scoring essays on a single, four to six point scale (as for example was done in the ASAP competition for automated essay scoring on Kaggle[1]), human scorer inter-rater reliability is marginal (or at least, less reliable than might be expected), in the sense that inter-rater agreement rates ranged from 28 to 78%, with associated quadratic-weighted Kappas ranging from .62 to .85.

Said another way, about half the time, two raters, or human (averaged) scores and AI scoring engines, will yield the same result for a simple measure, while the rest of the time, the variation can be all over map. Kappa doesn't really tell us very much about this variation, which is a concern, because "better" (higher) Kappas might also mask abnormal or biased relationships, where models with slightly lower Kappas might, on examination, provide an intuitively more appealing result. And for scoring solutions that seek to use more detailed scoring rubrics, and to provide sub-scores and more nuanced feedback, while still solving for reliability and validity in overall scoring, the challenge of finding the "best" model for a given dataset will be even greater.

This paper focuses solely on the problem of evaluating models that attempt to mimic human scores that are provided under the best of conditions (e.g. expert scorers not impacted by timing constraints), and addresses the question of how to define the "best" performing models. The aforementioned Kaggle competition chose Quadratic Weighted Kappa as a means of measuring the conformance of scores reported by a model and scores assigned by human scorers. Other Kaggle competitions routinely use other metrics as well[2], while some critics of the ASAP

[1] See https://www.kaggle.com/c/asap-aes
[2] See https://www.kaggle.com/wiki/Metrics

competition in particular, and of the use of AES technology in general, have argued that other model performance metrics might be more appropriate[3].

As a single descriptive statistic, QWK has inherent limits in describing the differences between two populations of results. Accordingly, this short note will present an example to illustrate the extent of these limitations. In short I think that – at least for a two-way comparison between a set of results from human scorers and a set of results from a trained machine learning model trying to emulate human scorers, the basic "confusion matrix" that shows a two dimensional grid with exact match results on a diagonal, and non-exact matches as outliers provides an unbeatable visualization of just how random, or not, the results of using a model might look against a set of "expert" measures.

Future efforts will considers suggested alternatives to QWK, hopefully leading to some more usable and "better" criterion for particular use cases and suggestions for further research.

I. Quadratic-Weighted Kappa (QWK)

Kappa is a statistical property designed to describe the "inter-rater reliability" between scores assigned by two different graders of the same set evaluations. In its basic form, Kappa considers all exact matches as equivalent and all non-matches as simply a "no match". In cases where the assigned values are not simply "categories" but have some ordered relationship between them, a form of Kappa can be calculated that recognizes the "degree of difference" between two assigned values. This recognition can be in the form of a simple arithmetic difference (e.g. linear weighted) or, as with the way "variance" is determined in a population by using the square of the differences, can be done in the form of a "square of the difference", yielding the "quadratic weighted kappa".

QWK is used in model evaluation as a way of characterizing the degree to which one set of numeric values conforms to another set of values – that is, the degree to which a model "accurately" predicts the actual output of a function or process.

Two helpful web pages that can do quick Kappa calculations on data already prepared in the form of a Confusion Matrix can be found at http://www.marcovanetti.com/pages/cfmatrix and even more helpfully (this one includes quadratic-weighted kappa, and a host of other related calculations) at http://vassarstats.net/kappa.html.

Excellent and thorough definitions of Kappa, and its relevance for use in comparing two sets of outcomes for inter-rater reliability, can be found in many places. These range from simple, mechanical and statistical definitions (and some implicit assertions or assumptions that might be worth examination) to detailed examinations of various forms of Kappa (meaning also the linear and quadratic and other acknowledgements of the relationship between classification labels or classifications) and specifically the "chance-corrected" aspect of the calculation, independence assumptions and other factors that give more, or less, weight to the idea that QWK is or is not a

[3] See particularly section entitled "Flawed Experimental Design I". from Les C. Perlman's paper at http://journalofwritingassessment.org/article.php?article=69

good measure for what we are trying to get at – the degree of fidelity between model outputs for a trained "scoring engine" and actual outputs from human judgment. These texts often include commentary about other related or similar statistical properties (e.g. Interclass Correlation coefficient, Pearson's r, etc.) which provide additional context. See for example see:

- https://www.kaggle.com/c/asap-aes/details/evaluation
- http://standardwisdom.com/softwarejournal/2011/12/confusion-matrix-another-single-value-metric-kappa-statistic/
- http://www.john-uebersax.com/stat/kappa.htm,
- http://kappa.chez-alice.fr/ and
- http://www.medcalc.org/manual/kappa.php.

Also worthwhile are the notes and discussion on the two Kappa calculation web pages / sites noted above.

In general, when the categories being compared have an ordered relationship – whether as a measure of the degree of coherence of an argument, for example, or the degree of effectiveness of a drug – weighted Kappas are generally preferred. In many cases, again for ordered measures, quadratic weighted Kappas are therefore not unusual.

II. QWK: Intuitive Pros and Cons for Constructed Response Scoring

Given a set of problems scored on some ordered scale between say 1 and 6, or 0 and 9, or what have you, it makes sense that a discrepancy between two scorers is of greater concern as the size of that difference is greater. The degree of subjectivity involved in the assessment leads to this notion, and so also to the idea that small differences can often be ignored -- and indeed, it is only by considering plus or minus 1 or so called "adjacent" measures to be equivalent that many constructed response rubrics and scoring systems achieve sufficient statistical reliability. It similarly makes sense that greater differences should result in even greater reason to question the validity and equivalence of a set of measures – many scores being far apart might suggest more of a random relationship between the two sets of values than some reasonably predictive model playing a role.

So while it is intuitively reasonable that some sort of weighting be factored into the discrepancies between predicted and actual scores when evaluating a scoring "model". Some models for weighted Kappas – used in evaluating medical treatment outcomes[4], for example, indeed weight differences between outcomes by assigning a specific weight to each pair of possible outcomes – perhaps reflecting for specific situations the relative benefit or damage embodied in the two different conditions.

My primary concern with QWK as a model validation "threshold" indicator is that this one single value masks many possible very different results. Meaning that two models of near equal "accuracy" as indicated by QWK might differ markedly in their specific tendencies – with one, for example, consistently over-predicting while another might have differences from human scorers that are neither more or less likely to be above the human score. Additional distribution

[4] See, again, http://www.medcalc.org/manual/kappa.php

characteristics and bias might also be masked or simply missed by this one simple metric approach, which I will illustrate with a simple example scenario.

III. Quadratic-weighted Kappa: Limitations on use in Model Validation

Or perhaps I should stress "limitations on over-reliance in comparative model evaluation" – since while QWK might provide one "threshold test" of many possible acceptance criteria for a scoring model, this example will illustrate how it can also mask significant difference.

Scenario A: A "good" Model

In this scenario, we present a "confusion matrix" for a 1 to 5 scoring result for an essay evaluation on a simple, "how coherent is this essay" kind of scale (or 'how good is the writing", or "how correct is the answer", etc.) that largely seems to reflect a "pretty good" scoring model as illustrated by this confusion matrix A:

A. Model A data

Scenario A2

	Model Output					
True Values	**1**	**2**	**3**	**4**	**5**	Counts
1	7	4	3	2	1	17
2	4	13	6	3	1	27
3	6	12	31	14	4	67
4	2	6	16	28	13	65
5	0	1	3	8	11	23

0.452261307

					Total Ovservations:		199
Counts	19	36	59	55	30	199	
accuracy:	**45.23%**		**QWK;**	**0.5235**			
kappa:	**0.2791**		err/min/max		0.0822	0.3624	0.6846

Note that, to keep this example simple, the data set has only 199 members, and many of the projected values match the expert scorer's values, or are within one point of the "true" score.

For this model the various statistical properties are listed below the matrix (and in the pages at the end). But note that the observed Kappa with Quadratic Weighting for this model is 0.5235; note also the pattern of "non-matches" around the diagonal.

Scenario B & C: A good model that always "over-scores"

B. Model B data

I have proposed here two models which provide very similar output to model A, but the output is such that it is never lower than the actual human score provided for any data point. This "systematic" too-generous scoring in this case is the result of a model that performs almost the same as in scenario A, but in every case that yielded a below-actual score, this model produced score that had the same "accuracy", except that the "wrong" values were always "wrong" in the over-scoring direction. So any essay scored below its true score was assigned a value above the true score by as much, so that accuracy (or inaccuracy) was preserved. Since the data in this scenario include 11 of 199 essays with the "top" score, and hence cannot be over-scored, I illustrate the possible consequences two ways: with model B and model C output.

For model B, only 11 essays with a true score of 5, or top of scale, are included in the data. The other 12 top-rated outcome data points in this example are now essays with lower true scores, and are "over-scored" by the same amounts that these original 11 top-scored data points were previously underscored. This is designed to yield a model that has identical accuracy to the original, and to see, despite the strong and consistent "over-scoring" bias, what happens to the Kappa. We see that the Kappa actually "improves", despite what is clearly an over-scoring bias.

The new confusion matrix and associated descriptive statistics are presented below. Note that the observed Kappa with Quadratic Weighting for this model is 0.5607.

Similar true values, very similar kappa, clear bias / model always over predicts

Scenario B2

	Model Output					
True Valu	1	2	3	4	5	Counts
1	7	4	3	4	1	19
2	0	13	10	9	2	34
3	0	0	31	26	13	70
4	0	0	0	28	37	65
5	0	0	0	0	11	11

0.452261

				Total Ovservations:		199
Counts	7	17	44	67	64	199
accuracy:	**45.23%**		**QWK:**	**0.5607**		
kappa:	**0.2946**		err/min/max		0.0721	0.4194	0.702

C. Model C Data:

In another variation on the "over-scoring biased model", Model C shows a confusion matrix (below) that reflects an algorithm that is just as good as the one in Model A, with all the same input, but here the 23 essays with top scores all get those tops scores by the model, true to its

tendency to never under-score an output. With now 12 additional "exact match" scores from the model, the accuracy and correlation between the two data sets is significantly better, and as expected the Kappa in this instance is now much higher. Note that the observed Kappa with Quadratic Weighting for this model is now 0.67 (and accuracy has improved from 45.23 to 51.26%)..

same true values, different but equal or higher scores

Scenario C

	Model Output					
True Valu	1	2	3	4	5	Counts
1	2	1	1	2	1	7
2	0	5	3	2	1	11
3	0	0	8	4	2	14
4	0	0	0	4	4	8
5	0	0	0	0	5	5

					Total Ovservations:		45
Counts	2	6	12	12	13	45	
accuracy:	**53.33%**		**QWK:**	0.5121			
kappa:	**0.415**				0.1539	0.2105	0.8137

The goal here was to show a deteriorating set of model outputs – outcomes that were less and less representative of human scoring behavior, that nonetheless also showed improving measures of quadratic-weighted kappa (QWK).

IV. Conclusions

The visualization of the "Confusion Matrix" brings to direct attention the degree to which the non-match outcomes from a scoring model adhere to a pattern that clusters closely to the diagonal line, and the degree of symmetry with respect to expected outcomes that the model yields. Perhaps some additional, single-value descriptive statistics can be selected to describe this degree of "symmetry", as well as the overall degree of "variance" (and it's dispersion). The combination of these values, taken together and weighted appropriately, might provide an algorithmic approach to the intrinsic human judgment that, for now, is more of a "visualization" than a quantified notion of "how good" a model is.

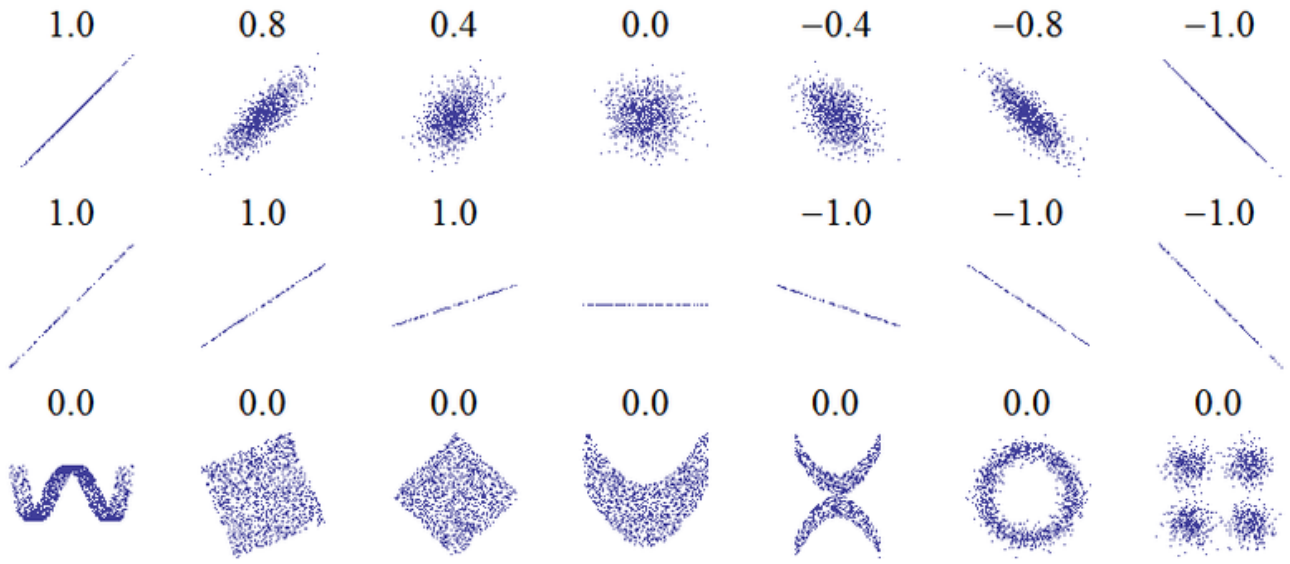
One direction would be to work with some real models, and real data, and see just what combination of descriptive statistics around model performance most closely match the intuitive

sense provided by visualization as to how to describe the "best" model for the purposes of automated essay scoring.

Additional note:

The challenge of represented the "degree of fit" between the model-predicted outcomes and the actual human scoring outcomes across a distribution of data points is the same challenge as encountered when representing the correlation between two variables with a single descriptive statistic. In particular the "Pearson r" or Pearson product-moment correlation coefficient, which is defined in Wikipedia[5] as "a measure of the linear correlation (dependence) between two variables X and Y, expressed as a value between +1 and -1, where 1 is total positive correlation, 0 is no correlation and -1 is total negative correlation."

The challenge here is similar, and the difficulties alluded to in this paper are similar to the issues raised by reviewing this broad set of scatter diagrams, and their associated (and overly simplified) Pearson r values, as presented below.[6] Distributions in the bottom row begin to get at the sorts of challenges raised by assuming that a QWK of say .45 is from a "better" model than one with a QWK of 4.04, in the absence of knowledge of the total relationship between the underlying data sets.



Note: An example of the correlation of x and y for various distributions of (x,y) pairs. Code and references per note 6 below.

[5] See http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient
[6] See http://commons.wikimedia.org/wiki/File:Correlation_examples.png

Kappa as a Measure of Concordance in Categorical Sorting

Kappa provides a measure of the degree to which two judges, A and B, concur in their respective sortings of N items into k mutually exclusive categories. A 'judge' in this context can be an individual human being, a set of individuals who sort the N items collectively, or some non-human agency, such as a computer program or diagnostic test, that performs a sorting on the basis of specified criteria. [Click here for an explanation of the conceptual and computational details of kappa.]

		B			
		1	**2**	**3**	**Total**
A	**1**	**44**	5	1	50
	2	7	**20**	3	30
	3	9	5	**6**	20
Total		60	30	10	100

k = 3
N = 100

To begin, select the number of categories by clicking the appropriate button below; then enter your data into the appropriate cells of the data-entry matrix. After all data have been entered, click the «Calculate» button. To perform a new analysis, click the «Reset» button and start over. The analysis assumes that each entered value is an integer equal to or greater than zero.

Note that measures of weighted kappa are meaningful only if the categories are ordinal and if the weightings ascribed to the categories faithfully reflect the reality of the situation. The weightings in this case are determined by the imputed relative distances between successive ordinal categories. By default, each of these distances is set at '1'. You are free to change any or all of these distances, though I recommend you do so only if you have good reason for it.

The author is grateful to César Roberto de Souza for detecting an error in the original programming for this module and suggesting the appropriate correction.

Select the number of categories:	2	3	4	5	6	7	8
Number selected = 5							

Basis for weighting: imputed relative distances between ordinal categories

1~2	2~3	3~4	4~5	5~6	6~7	7~8	⇐ successive ordinal categories
1	1	1	1	---	---	---	⇐ imputed relative distances

Data Entry

		B								
		1	2	3	4	5	6	7	8	Totals
A	1	7	4	3	2	1	----	----	----	17
	2	4	13	6	3	1	----	----	----	27
	3	6	12	31	14	4	----	----	----	67
	4	2	6	16	28	13	----	----	----	65
	5	0	1	3	8	11	----	----	----	23
	6	----	----	----	----	----	----	----	----	----
	7	----	----	----	----	----	----	----	----	----
	8	----	----	----	----	----	----	----	----	----
Totals		19	36	59	55	30	----	----	----	199

Reset Calculate

The designation "nc" appearing in any of the following cells means "this quantity cannot be calculated." This will typically occur only when your data entries in the above table include a substantial proportion of zeros.

Unweighted Kappa

Observed Kappa	Standard Error	.95 Confidence Interval	
0.2791		Lower Limit	Upper Limit
Method 1	0.0464	0.1881	0.3701
Method 2	0.0464	0.1882	0.37

0.8809	maximum possible unweighted kappa, given the observed marginal frequencies

0.3168	observed as proportion of maximum possible

Kappa with Linear Weighting

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.4071	0.0467	0.3156	0.4986

0.9085	maximum possible linear-weighted kappa, given the observed marginal frequencies
0.4481	observed as proportion of maximum possible

Kappa with Quadratic Weighting

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.5235	0.0822	0.3624	0.6846

0.9363	maximum possible quadratic-weighted kappa, given the observed marginal frequencies
0.5591	observed as proportion of maximum possible

Frequencies of Agreement

Category	Maximum Possible	Chance Expected	Observed
1	17	1.62	7
2	27	4.88	13
3	59	19.86	31
4	55	17.96	28
5	23	3.47	11
6	---	---	---
7			

8			
Total	181	47.79	90

Proportions of Agreement				.95 CI of Observed	
Category	Maximum Possible	Chance Expected	Observed	Lower Limit	Upper Limit
1	0.8947	0.0472	0.2414	0.1102	0.4393
2	0.75	0.084	0.26	0.1508	0.4061
3	0.8806	0.1872	0.3263	0.2357	0.4312
4	0.8462	0.1761	0.3043	0.215	0.4103
5	0.7667	0.07	0.2619	0.1439	0.4232
6	---	---	---	---	---
7					
8					
Composite	0.9095	0.2402	0.4523	0.3822	0.5242

Confidence intervals for proportions are calculated according to the Wilson efficient-score method, corrected for continuity.

Home	Click this link **only** if you did not arrive here via the VassarStats main page.

Kappa as a Measure of Concordance in Categorical Sorting

Kappa provides a measure of the degree to which two judges, A and B, concur in their respective sortings of N items into k mutually exclusive categories. A 'judge' in this context can be an individual human being, a set of individuals who sort the N items collectively, or some non-human agency, such as a computer program or diagnostic test, that performs a sorting on the basis of specified criteria. [Click here for an explanation of the conceptual and computational details of kappa.]

		B			
		1	2	3	Total
A	1	44	5	1	50
	2	7	20	3	30
	3	9	5	6	20
Total		60	30	10	100

k = 3
N = 100

To begin, select the number of categories by clicking the appropriate button below; then enter your data into the appropriate cells of the data-entry matrix. After all data have been entered, click the «Calculate» button. To perform a new analysis, click the «Reset» button and start over. The analysis assumes that each entered value is an integer equal to or greater than zero.

Note that measures of weighted kappa are meaningful only if the categories are ordinal and if the weightings ascribed to the categories faithfully reflect the reality of the situation. The weightings in this case are determined by the imputed relative distances between successive ordinal categories. By default, each of these distances is set at '1'. You are free to change any or all of these distances, though I recommend you do so only if you have good reason for it.

The author is grateful to César Roberto de Souza for detecting an error in the original programming for this module and suggesting the appropriate correction.

Select the number of categories:	2	3	4	5	6	7	8
Number selected = 5							

Basis for weighting: imputed relative distances between ordinal categories

1~2	2~3	3~4	4~5	5~6	6~7	7~8	⇐ successive ordinal categories
1	1	1	1	---	---	---	⇐ imputed relative distances

Data Entry

		B								
		1	2	3	4	5	6	7	8	Totals
A	1	7	4	3	4	1	----	----	----	19
	2	0	13	10	9	2	----	----	----	34
	3	0	0	31	26	13	----	----	----	70
	4	0	0	0	28	37	----	----	----	65
	5	0	0	0	0	11	----	----	----	11
	6	----	----	----	----	----	----	----	----	----
	7	----	----	----	----	----	----	----	----	----
	8	----	----	----	----	----	----	----	----	----
Totals		7	17	44	67	64	----	----	----	199

Reset Calculate

The designation "nc" appearing in any of the following cells means "this quantity cannot be calculated." This will typically occur only when your data entries in the above table include a substantial proportion of zeros.

Unweighted Kappa

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.2946			
Method 1	0.0454	0.2055	0.3837
Method 2	0.0435	0.2094	0.3798

0.6441	maximum possible unweighted kappa, given the observed marginal frequencies

0.4574	observed as proportion of maximum possible

Kappa with Linear Weighting

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.4351	0.0407	0.3554	0.5148

0.4351	maximum possible linear-weighted kappa, given the observed marginal frequencies
1	observed as proportion of maximum possible

Kappa with Quadratic Weighting

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.5607	0.0721	0.4194	0.702

nc	maximum possible quadratic-weighted kappa, given the observed marginal frequencies
nc	observed as proportion of maximum possible

Frequencies of Agreement

Category	Maximum Possible	Chance Expected	Observed
1	7	0.67	7
2	17	2.9	13
3	44	15.48	31
4	65	21.88	28
5	11	3.54	11
6	---	---	---
7			

8			
Total	144	44.47	90

Proportions of Agreement				.95 CI of Observed	
Category	Maximum Possible	Chance Expected	Observed	Lower Limit	Upper Limit
1	0.3684	0.0264	0.3684	0.1723	0.6137
2	0.5	0.0604	0.3421	0.2014	0.5142
3	0.6286	0.1571	0.3735	0.2718	0.487
4	0.9701	0.1987	0.2692	0.1892	0.3667
5	0.1719	0.0495	0.1719	0.0929	0.291
6	---	---	---	---	---
7					
8					
Composite	0.7236	0.2235	0.4523	0.3822	0.5242

Confidence intervals for proportions are calculated according to the Wilson efficient-score method, corrected for continuity.

Home	Click this link **only** if you did not arrive here via the VassarStats main page.

Kappa as a Measure of Concordance in Categorical Sorting

Kappa provides a measure of the degree to which two judges, A and B, concur in their respective sortings of N items into k mutually exclusive categories. A 'judge' in this context can be an individual human being, a set of individuals who sort the N items collectively, or some non-human agency, such as a computer program or diagnostic test, that performs a sorting on the basis of specified criteria. [Click here for an explanation of the conceptual and computational details of kappa.]

		B			
		1	2	3	Total
A	1	44	5	1	50
	2	7	20	3	30
	3	9	5	6	20
Total		60	30	10	100

k = 3
N = 100

To begin, select the number of categories by clicking the appropriate button below; then enter your data into the appropriate cells of the data-entry matrix. After all data have been entered, click the «Calculate» button. To perform a new analysis, click the «Reset» button and start over. The analysis assumes that each entered value is an integer equal to or greater than zero.

Note that measures of weighted kappa are meaningful only if the categories are ordinal and if the weightings ascribed to the categories faithfully reflect the reality of the situation. The weightings in this case are determined by the imputed relative distances between successive ordinal categories. By default, each of these distances is set at '1'. You are free to change any or all of these distances, though I recommend you do so only if you have good reason for it.

The author is grateful to César Roberto de Souza for detecting an error in the original programming for this module and suggesting the appropriate correction.

Select the number of categories:	2	3	4	5	6	7	8
Number selected = 5							

Basis for weighting: imputed relative distances between ordinal categories

1~2	2~3	3~4	4~5	5~6	6~7	7~8	⇐ successive ordinal categories
1	1	1	1	---	---	---	⇐ imputed relative distances

Data Entry

		B								
		1	2	3	4	5	6	7	8	Totals
A	1	7	4	3	2	1	----	----	----	17
	2	0	13	10	3	1	----	----	----	27
	3	0	0	31	26	10	----	----	----	67
	4	0	0	0	28	37	----	----	----	65
	5	0	0	0	0	23	----	----	----	23
	6	----	----	----	----	----	----	----	----	----
	7	----	----	----	----	----	----	----	----	----
	8	----	----	----	----	----	----	----	----	----
Totals		7	17	44	59	72	----	----	----	199

Reset Calculate

The designation "nc" appearing in any of the following cells means "this quantity cannot be calculated." This will typically occur only when your data entries in the above table include a substantial proportion of zeros.

Unweighted Kappa

Observed Kappa		.95 Confidence Interval	
0.3689	Standard Error	Lower Limit	Upper Limit
Method 1	0.0459	0.279	0.4588
Method 2	0.0448	0.2811	0.4567

0.6812	maximum possible unweighted kappa, given the observed marginal frequencies

0.5415	observed as proportion of maximum possible

Kappa with Linear Weighting

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.5338	0.0396	0.4563	0.6113

0.5338	maximum possible linear-weighted kappa, given the observed marginal frequencies
1	observed as proportion of maximum possible

Kappa with Quadratic Weighting

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.67	0.0646	0.5433	0.7967

nc	maximum possible quadratic-weighted kappa, given the observed marginal frequencies
nc	observed as proportion of maximum possible

Frequencies of Agreement

Category	Maximum Possible	Chance Expected	Observed
1	7	0.6	7
2	17	2.31	13
3	44	14.81	31
4	59	19.27	28
5	23	8.32	23
6	---	---	---
7			

8			
Total	150	45.309999	102

Proportions of Agreement				.95 CI of Observed	
Category	Maximum Possible	Chance Expected	Observed	Lower Limit	Upper Limit
1	0.4118	0.0256	0.4118	0.1943	0.6655
2	0.6296	0.0553	0.4194	0.2507	0.6074
3	0.6567	0.154	0.3875	0.2826	0.5033
4	0.9077	0.184	0.2917	0.2056	0.3946
5	0.3194	0.096	0.3194	0.2172	0.4411
6	---	---	---	---	---
7					
8					
Composite	0.7538	0.2277	0.5126	0.4411	0.5836

Confidence intervals for proportions are calculated according to the Wilson efficient-score method, corrected for continuity.

Home	Click this link **only** if you did not arrive here via the VassarStats main page.