ISSN 1744-1986

Technical Report N° 2008/ 15

Using a New Inter-rater Reliability Statistic

Debra Trusso Haley
Pete Thomas
Marian Petre
Anne De Roeck

27th August, 2008

Department of Computing
Faculty of Mathematics, Computing and Technology
The Open University

Walton Hall, Milton Keynes, MK7 6AA
United Kingdom

http://computing.open.ac.uk

Using a New Inter-rater Reliability Statistic

Debra Trusso Haley, Pete Thomas, Marian Petre, Anne De Roeck
The Open University

1.1 Introduction[1]

This paper discusses methods to evaluate Computer Assisted Assessment (CAA) systems, including some commonly used metrics as well as unconventional ones. I found that most of the methods to measure automated assessment reported in the literature were not useful for my purposes. After much research, I found a new metric, the Gwet AC1 inter-rater reliability (IRR) statistic (Gwet, 2001), that is a good solution for evaluating CAAs. Section 1.6 discusses AC1, but first I describe other possible metrics to motivate why I think that AC1 is the best available for evaluating an automated assessment system.

I focus on two types of metrics that I label external and internal metrics. External metrics can be used for reporting and sharing results. Internal metrics are used for comparing results within a research project.

Producers of CAAs need an easily understandable external metric to report results to consumers of CAAs, i.e., those wishing to use a particular system. In addition to reporting results to potential consumers, researchers may wish to share their results with other researchers. Finally, and perhaps most important for this dissertation, producers need an internal metric to quickly compare the results of selecting different parameters of the assessment algorithm. Many choices need to be made when implementing an LSA-based marking system. The LSA literature frequently leaves many of these choices unspecified, including number of dimensions in the reduced matrix, amount and type of training data, types of pre-processing, and weighting functions. The choice of these parameters is an intrinsic aspect of building an LSA marking system. Therefore, researchers need an adequate way to measure and compare the results of the various selections, as I shall explore in this chapter.

[1] This paper is taken from a chapter in my as yet unpublished PhD dissertation.

Section 1.2 describes a simple metric that is often used for external reporting of results. Section 1.3 discusses existing ways to measure the success of LSA-based assessment systems and motivates the need for new metrics. Sections 1.3 and 1.4 discuss several standard statistical tests that could be used to measure the success of automated assessment systems and argue that none of them is suitable for my purposes. Section 1.5 discusses possible metrics that use the distance between two vectors for comparing automated assessment systems - the Manhattan distance (L1) and the Euclidean distance (L2). Finally, Section 1.6 explains and justifies the metric I chose to evaluate EMMA – the Gwet AC1 inter-rater reliability statistic and discusses how it overcomes the flaws of the better-known kappa statistic..

1.2 A simple metric (SM)

A simple success measure is to determine the percentage of marks where two markers give identical marks. However, this simple metric (SM) gives an incomplete picture of the results. Consider the hypothetical case illustrated in Table 1-1. It shows how closely two markers agree with a third marker assumed to be the gold standard. Eighty percent of the answers marked by Marker A agreed exactly with the gold standard, 5% differed by 3 points, and 15% disagreed by

Table 1-1 Hypothetical results for two markers that show the simple metric of the percent of identical scores for a four -point question hides important details

	Marker A	Marker B
Point Difference between Markers and a "Gold Standard"	% of questions	
0	80	75
1	0	25
2	0	0
3	5	0
4	15	0

4 points. Seventy-five percent of the answers marked by Marker B agreed exactly with the gold standard and 25% differed by 1 point. The SM awards 80% to Marking System A and 75% to

Marking System B. Clearly, both markers have a high percentage of agreement, but which is the better marker - A or B? The SM says that A is better than B. However, even though A has a higher percentage of identical answers than does B, the latter has 100% of its marks disagreeing with the human by at most one point while A has only 80% of its marks disagreeing by at most one point and 20% that differ by three or more points. The flaw in the SM is that it gives no indication of the spread, or distribution, of the marks. Perhaps the SM is an acceptable external metric to use for reporting results to consumers, but it is inadequate for internal comparison purposes as I demonstrate in the remainder of this section.

Table 1-2 uses the SM to evaluate the results of an experiment to determine the optimum amount of training data, which is one of the parameters for calibrating an LSA-based marking system.

Table 1-2 Percentages of Agreement between Human and Computer when varying amount of training data

# of Marked Answers	% Equal Scores	Tutor and Computer differ by ±1mark	Tutor and Computer differ by ± 2 marks	Tutor and Computer differ by ± 3 marks	Tutor and Computer differ by ± 4 marks
10	65	20	13	2	1
20	68	15	13	3	1
30	60	27	11	2	1
40	60	27	11	2	1
50	57	31	10	2	0
60	61	25	11	1	1
70	67	18	12	3	1
80	67	18	11	3	1
90	67	18	11	3	1
100	67	17	11	3	2
200	35	54	7	4	1
300	45	39	12	2	2
400	62	22	12	2	2
500	61	24	12	2	2
600	47	38	12	3	1
627	59	27	11	2	1

Finding the correct answer by studying this table is very difficult. Unfortunately, studying a graph turned out to be no easier. Indeed, this difficulty in interpreting the data was a major motivation for finding an alternative success metric.

Figure 1-1 shows the information from Table 1-2 in graphical form. The figure contains a great deal of information, making it difficult to understand and interpret. How can one determine the best amount of training data by looking at this chart? The y-axis shows the percentage of marks. The x-axis shows the amount of training data. The data points show the percentage of marks where the human and the computer agree, or differ by from zero to four points. The first set of data points (marked by an open 0) indicates the cases where there was zero difference between the tutor mark and the computer mark, i.e., they are identical. The second set of data points (marked by a +) is where there was a difference of plus or minus one point. The fifth set (marked with an open square) indicates those questions with the worst results: either the tutor awarded four points and the computer awarded zero points, or vice versa.

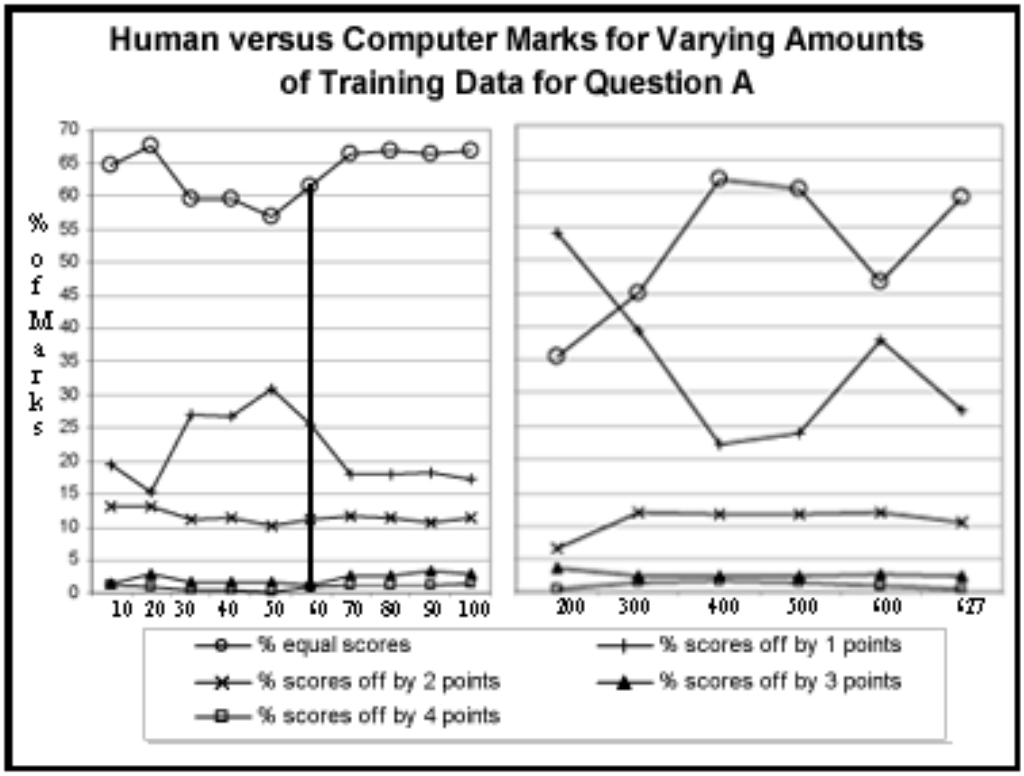

Figure 1-1 Comparison of human and computer marks for various amounts of training data

The legend to the right of the graph shows the correspondence between each set of data points and the amount by which the human and computer scores differ.

The viewer can see all of the results for a particular amount of training data by looking at a vertical slice of the graph. For example, the vertical line in the graph shows that when 60 training examples were used, EMMA matched the human about 61% of the time, differed by one point about 25% of the time, differed by two points about 11% of the time, differed by three points about 1% of the time, and differed by 4 points about 1% of the time.

Looking at a vertical slice of the graph shows the performance of EMMA for a particular amount of training data. Looking at a horizontal set of data points gives another point of view. A data set shows how much the performance varies over different amounts of training data. For example, the set indicated with an x shows that the marks that differed by plus or minus two points ranged from about 7% for 200 training data items to about 13% for 10 and 20 training data items.

I tried several different graphical ways to display the data – Figure 1-1 shows the clearest way I found. Even so, it is difficult to evaluate the overall effectiveness of varying the amount of training data by analysing this figure because it contains a lot of information, all of which is necessary to measure the results. Thus, the SM is not adequate for the internal purpose of evaluating various calibrations of the LSA algorithm. The next two sections discuss several other metrics and explain why I found them, like the SM, to be not useful for my work.

1.3 The inadequacy of existing success measures

The literature offers two widely used techniques to evaluate marking systems – precision and recall, and correlation. The following subsections describe them and suggest why they are inappropriate for evaluating CAAs.

1.3.1 Precision and recall

The first technique is the use of precision and recall; these measures are used widely in LSI and LSA research (Dumais, 1991; Manning & Schütze, 1999; Graesser, Wiemer-Hastings,

Wiemer-Hastings, Harter & The Tutoring Research Group, 2000; Nakov, Valchanova & Angelova, 2003). Precision looks at how relevant the collection of retrieved documents is; it is the ratio of correctly retrieved, i.e. relevant, documents to all retrieved documents. Recall is a measurement of completeness. It is the ratio of correctly retrieved documents to all relevant documents i.e., those that were retrieved plus those that the retrieval system failed to retrieve (Foltz, 1990). As recall goes up, precision tends to go down; in the trivial case, a system achieves 100% recall if all the documents are retrieved, which would give the lowest precision. Information retrieval (IR) researchers plot values of precision for various levels of recall to provide a good picture of the effectiveness of their techniques (Dumais, 2003). The relevance to LSA and marking is that LSA retrieves the marked answers from the training data that are closest to the answer being marked.

It is important to have a good metric to measure success when calibrating a marking system. Dumais, in a widely cited study (1991), used precision and recall to justify the use of log-entropy as the weighting function in the term-frequency matrix. (The decision of a weighting function is a critical choice to be made by LSA researchers.) Nakov, Valchanova & Angelova (2003) used precision and recall figures to argue that the choice of a weighting function is the most crucial of all calibration techniques. Many researchers continue to justify the use of the log-entropy weighting factor (Foltz, Kintsch & Landauer, 1998) by relying on the early work of Dumais (1991). Although log-entropy *may* be the best weighting function, it should be justified for LSA-based assessment systems on research done with LSA-based assessment systems instead of IR systems. Researchers need to remember that Dumais is primarily interested in information retrieval rather than essay assessment.

Although precision and recall are useful for evaluating IR techniques, I believe that using them to measure automated marking systems is irrelevant. Recall is not important – it makes no difference how many documents are returned because the marking system looks at only a pre-determined number that are the closest matches to the document being marked. Precision, on the other hand, is very important – the documents judged by the marking system to be relevant must actually be relevant. Precision, however, is a binary measure. It assumes that the documents are

relevant or not. EMMA uses the cosine similarity measure to rank the documents in terms of how similar they are to the answer being marked. It then awards a mark by calculating the weighted average (using the cosine measure) of the five most similar answers. This feature of LSA provides a finer-grained measure than the technique of using precision and recall, which is better suited to information retrieval.

1.3.2 Correlation

The second technique to evaluate marking systems is statistical correlation, which is used by many researchers (Wiemer-Hastings, 1999; Foltz, Gilliam & Kendall, 2000; Perez, Gliozzo, Strapparava, Alfonseca, Rodriquez & Magnini, 2005). The most widely known correlation measures are Pearson's r, Spearman's rho, and Kendall's tau_b (Dancey & Reidy, 2002). The formulas for the Pearson and Spearman measures given by Daniel (1977) are shown below. Spearman's rho calculates the correlation by calculating the difference between each pair of data points, or ranks. Daniel gives a correction if there are many tied ranks (1977 p. 364).

Equation 1-1 The Pearson correlation coefficient

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

X and Y are the two variables being correlated and n is the number of cases.

Equation 1-2 The Spearman rank correlation coefficient

$$r_s = \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

n is the number of cases, d is the difference between the ranks

Kendall's tau is based on concordant and discordant pairs (Stegmann & Lucking, 2005). Equation 1-3 gives the formula. Given 2 observations: (x_i, y_i) and $\left(x_j, y_j\right)$ they are:

concordant if when $x_j > x_i$ then $y_j > y_i$

discordant if when $x_j > x_i$ then $y_j < y_i$

tied if $x_i = x_j$ and/or $y_i = y_j$

Equation 1-3 Kendall's tau

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

where n_c = number of concordant pairs

and n_d = number of discordant pairs

Pearson's r is used when data are normally distributed. For many marking schemes, the marks are negatively skewed, i.e., the tail on the distribution graph goes to the left (Rowntree, 2004 p. 59). This trait occurs because markers tend to give high, rather than evenly distributed, marks. Figure 1-2 gives an example of a skewed distribution. The data are taken from the corpus and are typical of all of the questions I have examined. The figure shows that the data are non-normally distributed and thus a non-parametric test (e.g., Spearman's rho or Kendall's tau_b) is the appropriate choice (Dancey & Reidy, 2002; Rowntree, 2004 p. 125).

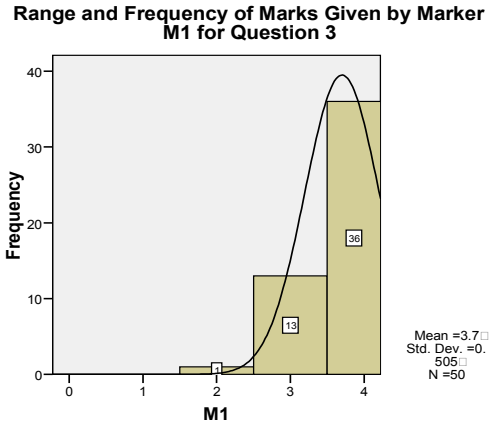

Figure 1-2 Histogram showing that marks are non-normally distributed

Correlation statistics indicate how well one variable can be used to predict another variable. If human-assigned marks and computer-assigned marks agree, the correlation would be perfect. Even if the human marks were always twice the computer marks, the correlation would once again be perfect. In this case, a good correlation would not mean a good marking system.

Another problem with the correlation statistic is that it would be low in the case where computer marks are off by plus-or-minus one point. In this situation, the computer mark could not be used to predict the human mark even though I argue that the overall results of the marking system would be very good if all the contradictory marks differ by only one point in either direction. For these reasons, the standard correlation statistics may not be useful for evaluating automated marking systems.

Table 1-3 shows various correlation coefficients for the case where two markers have 96% identical marks, and 4% where they differ by two points. For one mark, marker 1 scored higher than marker 2 and in the other case, scored lower than marker 2. Intuition would lead one to expect a high correlation between these two markers because they agree exactly on 96% of the marks, but SPSS calculates essentially zero correlation between the two markers. This example

Table 1-3 Output from SPSS that shows no correlation for two markers who have 96% identical answers

m1 * m2 Crosstabulation

Count

		m2		
		2	4	Total
m1	2	0	1	1
	4	1	48	49
Total		1	49	50

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	-.020	.014	-.722	.470
	Kendall's tau-c	-.002	.002	-.722	.470
	Spearman Correlation	-.020	.014	-.141	.888[c]
Interval by Interval	Pearson's R	-.020	.014	-.141	.888[c]
N of Valid Cases		50			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

shows that the traditional correlation statistics fail the common sense test and are not applicable to the problem of comparing the similarity between human and computer markers.

Section 1.4 looks at a traditional statistical test (and its non-parametric variations) and explains why it, also, failed to help evaluate my automated marking system.

1.4 Problems with the traditional t-test

Having considered and rejected the metrics described in Section 1.3, I turn to the traditional t-test as a candidate for evaluating automated marking systems. It comes in parametric and non-parametric versions. The parametric tests are more powerful than the non-parametric versions but the data must meet three assumptions to use them: normally distributed populations, approximately equal variations of the populations, and no extreme scores.

The t-test compares the means of two groups. For marking systems, one group is the human-assigned scores and the other group is the computer-assigned scores. When all participants take place in both conditions (short answer marked by tutor and short answer marked by EMMA), the study design is known as within-participants (also called repeated measures or related design) and the appropriate parametric statistical test for comparing the groups is the t-test (Dancey & Reidy, 2002). The SPSS output of the t-test includes the mean scores for each group, the difference between them, and the standard deviations. With these values, one can compute the effect size, which is the difference of the means divided by the mean of the standard deviations. Confidence intervals around the effect sizes are an additional tool for evaluating results (Aberson, 2002). Therefore, if the data meet the three assumptions for using parametric tests, employing effect sizes with confidence intervals could be a good way to evaluate automated marking systems.

Unfortunately, I cannot use the t-test and effect sizes with confidence intervals because my data are not normally distributed, as suggested by Figure 1-2. The marks given by tutors are highly negatively skewed because the marks tend to cluster towards the high end of the marking scale. If the marks are not normally distributed, the effect sizes and confidence intervals will be incorrect (Thompson, 2002) and the t-test is not applicable. I can, however, use the Wilcoxon

signed ranks test, which is the non-parametric version of the t-test. This test statistic is calculated by ranking the differences between the two scores. But the scores with zero difference are ignored because "they do not give us any information" (Dancey & Reidy, 2002).

There are two problems with the Wilcoxon t-test. The first problem is the elimination of those cases where the difference is zero. Dancey & Reidy (2002) claim that these cases do not give us any information, which may be true when trying to establish that there *is* a difference between two groups. However, when evaluating marking systems, I want to establish that there is *no* difference between two groups or that the difference is *very small*. If, for example, a marking system produces marks that agree with the human 95% of the time, that figure is informative, contradicting one of the assumptions of the Wilcoxon test. I need a test statistic that takes into account the number of cases where the difference between two marks is zero.

The second problem with the Wilcoxon t-test is that it shows whether two groups are different but not by how much. To solve that problem, I can look at the mean difference given by the descriptive statistics – no difference or very small differences would allow me to conclude that there is no significant difference between two groups. However, as mentioned earlier, calibrating an LSA-based marking system is critical. How should I compare the results of calibrating the system? I cannot use mean differences by themselves; I must consider the standard deviations. This requirement leads me back to effect sizes, but the results will be invalid because my data are not normally distributed.

For the reasons given above, I cannot use correlation statistics, t-tests, or effect sizes with confidence intervals. Section 1.5 presents possible alternative metrics and provides an example of using them to evaluate test results.

1.5 Success metrics using the distance between two vectors

The inability to locate an appropriate metric, as described in previous sections, combined with the difficulty in interpreting Figure 1-1 and Table 1-2, led me to investigate two metrics from the field of vector space theory. The Manhattan Distance measure (L1) and the Euclidean Distance measure (L2) are two metrics used to calculate the distance between two vectors

(Gerald & Wheatley, 1970). Their application to marking exams is as follows. One vector is the list of scores given to a question by one marker; the other vector is the list of scores assigned by another marker. If the vectors are identical, the distance between the vectors is zero and the two markers would agree perfectly.

The measures are calculated using the well-known formulas (Gerald & Wheatley, 1970) shown below. These formulas compute the distance between the vectors in slightly different ways. The L1 computes the sum of the differences between each point in the two vectors; the L2 computes the square root of the sum of the squares of the differences.

Equation 1-4 The Manhattan Distance (1 norm, or L1):

$$\mathrm{M(X,Y)} = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Equation 1-5 The Euclidean Distance (2-norm, or L2):

$$M(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where X = (x_1, x2, xn) and Y = (y_1, y2, yn) are two n-dimensional vectors

The L1 and L2 metrics provide a figure that could be used to assess the results of the different experiments quickly. Table 1-4 shows the same information as Table 1-2 except that it includes L1 and L2 and is sorted by L2. This table shows that 50 is the amount of training data that corresponds to the best outcome for this question using either L1 or L2 as the metric. This evaluation agreed with a careful hand analysis, taking into consideration the number of marks that differ by 1 or more points, of all of the numbers given in Table 1-2.

Table 1-4 Result of varying the amount of training data - sorted from best to worst

# of Marked Answers	% Equal Scores	Tutor and Computer differ by ±1	Tutor and Computer differ by ± 2	Tutor and Computer differ by ± 3	Tutor and Computer differ by ± 4	Manhattan Distance L1	Euclidean Distance L2
50	57	31	10	2	0	93	9.6
60	61	25	11	1	1	95	9.7
30	60	27	11	2	1	97	9.8
40	60	27	11	2	1	98	9.9
627	59	27	11	2	1	100	10.0
10	65	20	13	2	1	105	10.2
80	67	18	11	3	1	107	10.3
70	67	18	12	3	1	108	10.4
20	68	15	13	3	1	109	10.5
90	67	18	11	3	1	110	10.5
100	67	17	11	3	2	114	10.7
500	61	24	12	2	2	116	10.8
400	62	22	12	2	2	119	10.9
200	35	54	7	4	1	122	11.1
600	47	38	12	3	1	124	11.1
300	45	39	12	2	2	133	11.5

The Manhattan and Euclidean distance measures at first seemed to be promising tools for automated marking researchers to evaluate their systems. Unlike the simple metric (SM) these two metrics take into consideration the values of agreement between human and computer over the whole range of possibilities. That is, they evaluate the results where human and computer marks are identical, where they are off by plus or minus one point, plus or minus two points, and so on until the worst result which is where the human and computer differ by the maximum point value of the question. The SM uses just the value where the human and computer marks are identical and can lead to ambiguity, as demonstrated in Table 1-1. L1 and L2 give a richer picture of the effectiveness of an automated marking system than the SM and are no more difficult to analyse than the SM. There are, however, two problems with them. The first problem is that they are not widely used for evaluating CAAs and no agreed upon cut-off levels exist. The second and more serious problem arises when comparing answers with differing point values. The problem arises because neither L1 nor L2 is normalised, so that the distance between vectors cannot be compared in a sensible manner.

1.6 The inter-rater reliability statistics

This section discusses two statistics used for inter-rater reliability – Cohen's kappa and Gwet's AC1. Kappa is the better known of the two. AC1, first introduced in 2001 (Gwet), corrects some of the deficiencies of kappa.

1.6.1 The problem with the kappa inter-rater reliability statistic

Cohen's kappa statistic is used for inter-rater reliability (Cohen, 1960). I tried this measure and then discarded it because it gave me non-sensible results. I had instances where EMMA and the human raters agreed by as much as 97% but the kappa statistic was close to zero, indicating no correspondence. I was reassured to find a paper by two researchers that gave an example of what they called an "absurd kappa value" (Stegmann & Lucking, 2005). Table 1-5 shows the example they used – it was taken from another paper (DiEugenio & Glass, 2004). In each of the experiments illustrated, the observers agreed in 90% of the cases but one case showed a high

Table 1-5 Tables illustrating a balanced (left) and a skewed (right) distribution

Observer B	Observer A		Total
	1	2	
1	45	5	50
2	5	45	50
Total	50	50	100

kappa = .8

Observer B	Observer A		Total
	1	2	
1	90	5	95
2	5	0	5
Total	95	5	100

kappa = an "absurd" -0.0526

kappa figure of 0.8 while the other showed essentially no agreement at kappa = -0.0526. The problem with the kappa statistic suggested by Table 1-5 was first documented by Feinstein and Cicchetti (1990) as summarised in their abstract:

> In a fourfold table showing binary agreement of two observers, the observed proportion of agreement, P0 can be paradoxically altered by the chance-corrected ratio that creates κ as an index of concordance. In one paradox, a high value of P0 can be drastically lowered by a substantial imbalance in the table's marginal totals either vertically or horizontally. In the second pardox, (sic) κ will be higher with an asymmetrical rather than

> symmetrical imbalance in marginal totals, and with imperfect rather than perfect symmetry in the imbalance. An adjustment that substitutes Kmax for κ does not repair either problem, and seems to make the second one worse.

DiEugenio & Glass (2004) explain the problem in more accessible language: "κ is affected by skewed distributions of categories (the **prevalence problem**) and by the degree to which the coders disagree (the **bias problem**)."

Researchers in several disciplines have noted the problems with the kappa statistic and have begun to use the AC1 statistic. Chan, in a statistics tutorial for the medical profession (2003), provides an example where kappa gives strange results because one of the rater categories has a small percentage. He recommends the use of AC1. Several researchers in the field of software process improvement (Huo, Zhang & Jeffrey, 2006) suggest the use of AC1. Two computational linguists interested in the automatic classification of documents (Purpura & Hillard, 2006) suggest the use of AC1. Another group of computational linguists interested in classifying documents (Yang, Callan & Shulman, 2006) use AC1. Two researchers at the Dartmouth Medicine School (Blood & Spratt, 2007) recommend the use of AC1 and have created and made freely available a macro for the statistical package SAS. However, they caution that AC1 is still a new statistic:

> "Although the AC1 and AC2 statistics are about five years old now, they remain infants in the statistical world, especially since so few people have been exposed to them. With greater usage will come greater scrutiny, and with greater scrutiny may come identification of problems inherent in these statistics. Therefore, as is always the case with new statistics, caution should be exercised in their use and further examination should occur before they are adopted as the standard."

The next subsection describes the Gwet AC1 statistic.

1.6.2 The Gwet AC1 inter-rater reliability statistic

Kilmer Gwet has written extensively about the problems of the kappa statistic and has proposed AC1 (2001; 2002a; 2002b), which he claims overcomes the problems with kappa. What follows, except where noted, comes from (Gwet, 2002a). The explanation is for the simplified case of two raters and two categories of ratings. Gwet uses the following table in his formulas. "A" is the number of times both raters gave a rating of "1". "B" is the number of times rater A gave a "2" when rater B gave a "1". "A1" is the total number of times Rater A gave a "1" and "A2" is the total number of times Rater A gave a "2". N is the total number of observations.

Table 1-6 Distribution of subjects by rater and response category

	Rater A		
Rater B	**1**	**2**	**Total**
1	A	B	**B**1=A+B
2	C	D	**B**2=C+D
Total	**A**1=A+C	**A**2=B+D	N

According to Gwet (2001), the kappa formula takes the form of Equation 1-6. It is equivalent to the SM discussed in Section 1.2 corrected by the probability of chance agreement. He shows an example similar to Table 1-5 and claims that kappa can be misleading, "especially when the sum of the marginal probabilities is very different from 1". He claims that kappa incorrectly overstates the correction due to chance agreement.

Equation 1-6 The kappa formula

$$kappa = \frac{p - e(\kappa)}{1 - e(\kappa)}$$

where p = the overall agreement = $\frac{A + D}{N}$

and e(κ) = the chance agreement probability = $\frac{A1}{N} * \frac{B1}{N} + \frac{A2}{N} * \frac{B2}{N}$

and A, A1, A2, B1, B2, D, and N are the figures in Table 1-6.

Equation 1-7 shows Gwet's AC1 statistic.

Equation 1-7 Gwet's AC1

$$AC1 = \frac{p - e(\gamma)}{1 - e(\gamma)}$$

where $e(\gamma) = 2P_1(1 - P_1)$

$$P_1 = \frac{(A1 + B1)/2}{N}$$

and $p = \frac{A + D}{N}$

and

AC1 = the first order agreement coefficient

e (γ) = the chance-agreement probability

P_1 = the approximate chance that a rater classifies a subject into category 1

A1 = the number of times a rater, A, classifies a subject into category 1

A = number of times both raters classifies a subject into category 1

D = number of times both raters classifies a subject into category 2

p = the overall agreement

Equation 1-7 shows how to calculate AC1 for the simple case of two raters and two rating categories. Blood & Spratt (2007) give the formula for the general case. I implemented this formula in Java and used it to evaluate the results I have from EMMA.

Equation 1-8 The AC1 formula for the general case

$$AC1 = \frac{p_a - p_{e\gamma}}{1 - p_{e\gamma}}$$

where $p_a = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{q=1}^{Q}\frac{r_{iq}\left(r_{iq}-1\right)}{r\left(r-1\right)}\right\}$

and $p_{e\gamma} = \frac{1}{Q-1}\sum_{q=1}^{Q}\pi_q\left(1-\pi_q\right)$

and $\pi_q = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{iq}}{r}$

P_a = the overall agreement probability

$P_{e\gamma}$ = the chance-agreement probability

r_{iq} = the number of raters who classified the ith object into the qth category. The index *i range* from 1 to n and q ranges from 1 to Q

n = the number of objects rated

Q = the number of categories in the rating scale

r = the total number of raters

π_q = the probability that a rater classifies an object into category q

1.7 A worked example

I demonstrate the use of the kappa and AC1 formulas first with the balanced example and then with the skewed example given in Table 1-5. Table 1-7 summarises the results.

1.7.1 Balanced distribution

$$p = \frac{A+D}{N} = \frac{45+45}{100} = .9$$

$$e(\kappa) = \frac{A1}{N} * \frac{B1}{N} + \frac{A2}{N} * \frac{B2}{N} = \frac{50 * 50}{100 * 100} + \frac{50 * 50}{100 * 100} = \frac{2500 + 2500}{10000} = .5$$

$$kappa = \frac{p - e(\kappa)}{1 - e(\kappa)} = \frac{.9 - .5}{1 - .5} = \frac{.4}{.5} = .8$$

$$P_1 = \frac{(A1 + B1)/2}{N} = \frac{(50 + 50)/2}{100} = \frac{50}{100} = .5$$

$$e(\gamma) = 2P_1(1 - P_1) = 2 * .5(1 - .5) = 1 * .5 = .5$$

$$AC1 = \frac{p - e(\gamma)}{1 - e(\gamma)} = \frac{.9 - .5}{1 - .5} = \frac{.4}{.5} = .8$$

1.7.2 Skewed distribution

$$p = \frac{A + D}{N} = \frac{90 + 0}{100} = .9$$

$$e(\kappa) = \frac{A1}{N} * \frac{B1}{N} + \frac{A2}{N} * \frac{B2}{N} = \frac{95 * 95}{100 * 100} + \frac{5 * 5}{100 * 100} = \frac{9025 + 25}{10000} = .905$$

$$kappa = \frac{p - e(\kappa)}{1 - e(\kappa)} = \frac{.9 - .905}{1 - .905} = \frac{-.005}{.095} = -.0526$$

$$P_1 = \frac{(A1 + B1)/2}{N} = \frac{(95 + 95)/2}{100} = \frac{95}{100} = .95$$

$$e(\gamma) = 2P_1(1 - P_1) = 2 * .95(1 - .95) = 1.9 * .05 = .095$$

$$AC1 = \frac{p - e(\gamma)}{1 - e(\gamma)} = \frac{.9 - .095}{1 - .095} = \frac{.805}{.905} = .8895$$

Table 1-7 Comparison of kappa and AC1 for balanced and skewed distributions shown in Table 1-5 showing that kappa gives a strange result for a skewed distribution

	Balanced Distribution	Skewed Distribution
kappa	0.8	-0.05
AC1	0.8	0.89

This worked example shows that kappa gives a result for a skewed distribution that fails the common sense test and confirms the work of various researchers (Feinstein & Cicchetti, 1990; Gwet, 2002a; Stegmann & Lucking, 2005; Blood & Spratt, 2007). It also supports Gwet's claim that AC1 is a "more robust chance-corrected statistic that consistently yields reliable results" (Gwet, 2002a).

References

Aberson, C. (2002). *Interpreting Null Results: Improving Presentation and Conclusions with Confidence Intervals*. **Journal of Articles in Support of the Null Hypothesis 1**(3): 36-42.

Blood, Emily & Spratt, Kevin F. (2007). *Disagreement on Agreement: Two Alternative Agreement Coefficients*. **SAS Global Forum 2007**.

Chan, Y. H. (2003). *Biostatistics 104: Correlational Analysis*. **Singapore Medical Journal 44**(12): 614-619.

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

Dancey, Christine P. & Reidy, John (2002). **Statistics Without Maths for Psychology: Using SPSS for Windows**. Essex, England, Prentice Hall.

Daniel, Wayne W. (1977). **Introductory Statistics with Applications**. Boston, Houghton Mifflin Company.

DiEugenio, Barbara & Glass, Michael (2004). *The kappa statistic: a second look.* **Computational Linguistics 30**(1).

Dumais, S. T. (1991). *Improving the retrieval of information from external sources*. **Behavioral Research Methods, Instruments & Computers 23**(2): 229-236.

Dumais, Susan. T. (2003). *Data-driven approaches to information access*. **Cognitive Science 27**: 491-524.

Feinstein, Alvan R. & Cicchetti, Domenic V. (1990). *High agreement but low kappa: I. The problems of two paradoxes*. **Journal of Clinical Epidemiology 43**(6): 543-549.

Foltz, Peter W. (1990). *Using latent semantic indexing for information filtering*. **Conference on Office Information Systems**, Cambridge, MA.

Foltz, Peter W., Gilliam, Sara & Kendall, Scott A. (2000). *Supporting content-based feedback in online writing evaluation with LSA*. **Interactive Learning Environments 8**(2): 111-129.

Foltz, Peter W., Kintsch, Walter & Landauer, Thomas K. (1998). *The Measurement of Textural Coherence with Latent Semantic Analysis*. **Discourse Process 25**(2&3): 285-307.

Gerald & Wheatley (1970). **Applied Numerical Analysis**. Addison-Wesley.

Graesser, Arthur C., Wiemer-Hastings, Peter, Wiemer-Hastings, Katja, Harter, Derek & The Tutoring Research Group (2000). *Using latent semantic analysis to evaluate the contributions of students in AutoTutor*. **Interactive Learning Environments. [Special Issue, J. Psotka, guest editor] 8**(2): 129-147.

Gwet, Kilem (2001). **Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement Between Two or Multiple Raters**. Gaithersburg, MD, STATAXIS Publishing Company.

Gwet, Kilem (2002a). *Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters*. **Statistical Methods for Inter-Rater Reliability Assessment 1**.

Gwet, Kilem (2002b). *Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity*. **Statistical Methods for Inter-Rater Reliability Assessment 2**.

Haley, Debra, Thomas, Pete, De Roeck, Anne & Petre, Marian (2007a). *Measuring Improvement in Latent Semantic Analysis-Based Marking Systems: Using a Computer to Mark Questions about HTML*. **Proceedings of the Ninth Australasian Computing Education Conference (ACE2007)**, Ballarat, Victoria, Australia, Australian Computer Society Inc.

Haley, Debra Trusso, Thomas, Pete, De Roeck, Anne & Petre, Marian (2005). *A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications*. **International Conference on Recent Advances in Natural Language Processing'05**, Borovets, Bulgaria.

Haley, Debra Trusso, Thomas, Pete, De Roeck, Anne & Petre, Marian (2007b). *Tuning an LSA-based Assessment System for Short Answers in the Domain of Computer Science: The Elusive Optimum Dimension*. **1st European Workshop on Latent Semantic Analysis in Technology Enhanced Learning**, Heerlen, The Netherlands.

Huo, Ming, Zhang, He & Jeffrey, Ross (2006). *An Exploratory Study of Process Enactment as Input to Software Process Improvement*. **WoSQ'06**, Shanghai, China.

Landauer, Thomas K. & Dumais, S. T. (1997). *A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge*. **Psychological Review 104**(2): 211-240.

Manning, C. D. & Schütze, H. (1999). **Foundations of Statistical Natural Language Processing**. Cambridge, Massachusetts, MIT Press.

Nakov, Preslav, Valchanova, Elena & Angelova, Galia (2003). *Towards Deeper Understanding of the LSA Performance*. **Proceedings of Recent Advances in Natural Language Processing '03**, Borovetz, Bulgaria.

Perez, Diana, Gliozzo, Alfio, Strapparava, Carlo, Alfonseca, Enrique, Rodriquez, Pilar & Magnini, Bernardo (2005). *Automatic Assessment of Students' free-text Answers underpinned by the combination of a Bleu-inspired algorithm and LSA*. **Proceedings of the 18th International FLAIRS Conference**, Clearwater Beach, Florida.

Purpura, Stephen & Hillard, Dustin (2006). *Automated Classification of Congressional Legislation*. **The 7th Annual International Conference of Digital Government Research '06**, San Diego, CA, ACM.

Rowntree, Derek (2004). **Statistics Without Tears: A Primer for Non-Mathematicians**. Boston, Pearson Education, Inc.

Sebastiani, Fabrizio (2002). *Machine Learning in Automated Text Categorization*. **ACM Computing Surveys 34**(1): 1-47.

Stegmann, Jens & Lucking, Andy (2005). *Assessing Reliability on Annotations (1): Theoretical Considerations*, University of Beilefeld.

Thompson, Bruce (2002). *What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes*. **Educational Researcher 31**(3): 25-32.

Wade-Stein, Dave & Kintsch, Eileen (2003). *Summary Street: Interactive computer support for writing.* Technical Report from the Institute for Cognitive Science. University of Colorado, USA.

Wiemer-Hastings, Peter (1999). *How Latent is Latent Semantic Analysis*. **Proceedings of the 16th International Joint Conference on Artificial Intelligence**, Stockholm, Sweden.

Wiemer-Hastings, Peter, Wiemer-Hastings, Katja & Graesser, Arthur C. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. **Artificial Intelligence in Education**. Lajoie, S. P. and Vivet, M. Amsterdam, IOS Press.

Yang, Hui, Callan, Jamie & Shulman, Stuart (2006). *Next Steps in Near-Duplicate Detection for eRulemaking*. **Proceedings of the Sixth National Conference on Digital Government Research**.