(Re)Visualizing Rater Agreement: Beyond Single-Parameter Measures

David Eubanks, *Furman University*

Structured Abstract

- **Technique Identification**: A new graphical technique is presented for visualizing and assessing inter-rater agreement in discrete ordinal or categorical data, such as rubric ratings. To that aim, a chance-corrected Kappa with two new features is derived. First, it is based on interpreting ratings for each subject as vectors to visualize the data. This is done by creating two-dimensional vectors from a subject-rating summary table, sorting the vectors by their slopes, and plotting them in that order to create a trajectory that displays all the data in context. Second, it presents a graph and accompanying statistics (Kappa, *p*-value) for each pair of ratings in an organized display so that all useful comparisons of the data are visually displayed and statistically assessed. This information is presented on a logical grid, usually called *facets*. Kappa is calculated in the usual way, by referencing the actual results with an average of random rating assignments. This average becomes a reference line on each graph as a visual cue, as well. The statistical basis for the Kappa and significance testing are derived, and the test assumptions are specified.

- **Value Contribution:** The most commonly used statistics for inter-rater agreement, such as the Cohen Kappa or Inter-Class Correlation, give only a single parameter estimate of reliability from which to make judgments about ratings data. The technique presented here constructs graphs of all the data that allow visual inspection of the ratings versus a reference curve that represents chance-matching. The detailed reports on inter-rater agreement can show how to fine-tune ratings systems, such as understanding which parts of an ordinal scale are working best. This solves a practical problem for researchers who rely on rating-type classification by revealing which overall aspects of the rating system need

DOI: 10.37514/JWA-J.2017.1.1.10

to be improved and adds to the list of tools available for assessing rating reliability. In creating this approach to analysis of rater data, human usability is emphasized. Specifically, the use of geometry is designed to facilitate interpretability rather than being a mathematical derivation from first principles.

- **Technique Application:** Two applications are given, both involving social meaning-making. The first uses data from wine-judging to illustrate how the method can illuminate expertise in that domain. The results reproduce published findings that were based on a classical statistical method. A second sample application uses data from a university assessment of student writing in which ratings on a developmental scale are assigned by course instructors to their students. The rating program is an example of social meaning-making that can be used to generate larger data sets than are typical for classroom-based assessment programs. The analysis shows the strengths and weaknesses of the rating system in terms of reliability and demonstrates how that knowledge leads to improvements in assessment.

- **Directions for Further Research:** An argument is made for a public library of inter-rater data for empirical use by researchers. The social aspects of rating are discussed, and there is an illustration of the potential to derive new measures of inter-rater agreement from the meaning-making program that produces the data.

Keywords: assessment, inter-rater agreement, measurement, reliability, writing analytics

1.0 Technique Identification

Statistical measures in education, as elsewhere, are deterministic recipes for data summary and presentation that are predicated on explicit assumptions. Measures serve us well when they lead to insights that were hidden in the raw data. The main purpose of the method described in this section is to facilitate these insights by using intuitive graphs to represent rater data. Some readers may want to go directly to sections 2.0 and 3.0 to see examples and benefits first.

A new variation of existing inter-rater statistics is described below. It has two new features that make it easier to understand ratings data. The first is that the

raw data are turned into graphs so that we can more easily see patterns. The technique depends on turning the ratings on each subject into a vector with direction and distance. These are plotted one after another to create an arc that graphs all the ratings. Each of these arcs has a length that can be compared to complete rater agreement (arc length of one) and to the arc produced by hypothetical random raters. The proportion between randomness and perfect matching comprises the Kappa statistic.

Only basic knowledge of probability is needed to construct the new Kappa, specifically, the properties of binomial random variables: the formulas for their distributions, means, and variances, and the property of binomial distributions that for large enough sample sizes, they converge to the Normal distribution.

The R code to produce the graphs and statistics, as well as sample data, are available freely at Eubanks (2016a).

1.1 Inter-rater Agreement

The effective use of data to reach statistical conclusions depends on the quality of the measurements. In statistics, this quality is often determined by a *true value* plus *error*. A fundamental source of error is in classification—if the first observation of an object tells us it is a cat, but the second observation says that it is a dog, one of them (at least) is in error. The repeatability of measurement is referred to as the *reliability* of the measurement, and there are several existing statistical approaches to assessing reliability. Deciding which statistic is most appropriate depends on the type of data one has and the assumptions about it one is willing to make. See Haertel (2006) for an overview of reliability in educational measurement.

The focus of this paper is the assessment of reliability for measurements that classify observations into discrete categories. The categories may be nominal, as when a doctor classifies symptoms as a cold or the flu, or they may be ordinal, as in poor-to-excellent ratings scales in educational rubrics used to rate student work. Student writing samples are often judged by multiple raters using an ordinal rubric scale in order to assess rater agreement. If raters show a lack of sufficient agreement, this indicates a problem in process, training, definitions, standardization of review samples, or some other aspect of the complex task of classifying observations.

The method described below is forward-looking in the sense that it is most appropriate for large data sets with many raters and subjects. It takes advantage of the cheap computation and fast graphics of modern computers to produce intuitive graphs and multiple statistical summaries. These add interpretive value to the

inter-rater characteristics latent in the data. The example given in section 3.2 complements the theory by employing a method for rapidly generating large amounts of data on student learning. The unifying motivation is that educational assessment is a social activity that involves human judgment and interaction at all levels. As such, meaning-making is not an induction of a statistical Platonic ideal but a group exercise that will always be subject to later revision. This follows the agenda in Moss (2004) to "(a) expand the range of assessment practices considered sound and, more importantly, (b) illuminate taken-for-granted theories and practices of psychometrics for critical review" (p. 245).

We will consider cases where the number of scale outcomes is small (e.g., one to five on a typical Likert-type scale) and ratings are assigned on a sufficient quantity of subjects. In order to assess rater agreement, only cases with at least two raters are useful, but the raters need not be all the same, and the number of ratings per subject can vary. Rater agreement is compared to the frequency of agreement one would expect from random distribution of ratings. This "chance adjusted" agreement is intuitive: we would like to be able to detect if our raters are merely flipping coins to make judgments. In our case, random means that we choose random ratings for each subject in a way that preserves the overall original rating distribution (e.g., the fraction of 1s, fraction of 2s, and so on, present in the entire data set) and preserves the number of ratings per subject.

As an example, suppose two raters are scoring student portfolios as pass or fail according to a common standard, and we want to assess the reliability of the scoring methodology. The chance-correction idea is that if the scorers cannot agree more often than coin flips, the methodology is very poor. Once we have the scores in hand, they will have some overall frequency of passes and fails (e.g. 20% failing and 80% passing). Within the ratings, it is possible that the raters agree very often, but it is also possible that they agree infrequently on individual cases. It is possible that they *never* agree on the failing ratings in this case. The actual agreement rate, averaged over all cases, is compared to the random agreement we would expect from flipping a coin that is weighted to come up 20% tails (fail) and 80% heads (pass).

There is a long history of such chance-corrected statistics, perhaps starting with Galton (Galton, 1892), as noted by Smeaton (1985). The most well-known measures of inter-rater agreement may be Cohen's Kappa (Cohen, 1960) and a more general statistic by Fleiss (1971). There is considerable literature on the general topic, spanning statistics, education, psychology, medicine, and other fields. Readers are pointed to Agresti (2013) for the general topic of analysis of categorical data and to Gwet (2014) for a recent survey of inter-rater reliability measures. Fleiss (2003, p. 598-626) gives a statistical overview of the main

methods (including the relationship between the Fleiss Kappa and the inter-class correlation coefficient) and references more variations.

This paper develops a geometric approach to understanding rater agreement, using reports that are rich with data and relationships from which to make judgments. The goal is to provide practitioners a more useful way of analyzing inter-rater agreement as an exploratory tool. An example of the kind of question one might naturally want to ask about a set of ratings data is "do raters agree more about extreme cases (very good or very bad) than they do about intermediate ones?" This subsetting of the data is referred to as *conditional* rater agreement, a term from probability theory wherein the probability of event *A* is subject to the condition that event *B* has occurred (*A given B)* is $\Pr(A|B) \coloneqq \Pr(AB)/\Pr(B)$. In our case, the restriction *B* will be to consider only two rating responses at one time, e.g., the ones and twos. This is a way to ask, "Can the raters reliably tell the difference between a subject at level 1 and the same subject at level 2?"

Others have formulated conditional rater agreement statistics. Fleiss (1971) gives a single-column statistic, and Roberts (2008) gives a formula for pairs of outcomes. In Gwet (2014), one can find a chapter on the topic of conditional rater agreement. The approach here is different from those mentioned above, however. In chapter three of Gwet (2014), there is also a description of Kappa with the data conceived as vectors, but the vector lengths are not used. Only their squares are used as a calculation of rater agreement, like the Fleiss Kappa.

1.2 Visualizing Random Ratings

Cohen's idea was that it is not good enough to simply assess how often two raters agree on a categorical outcome for a particular subject: we should take into account how often we might see the same agreement "by chance." Exactly what that means has been the subject of debate (Powers, 2012), so we will first consider that question. In the following, we will proceed more along the lines of the Fleiss Kappa than the original Cohen Kappa. See Gwet (2014), chapter two, for a nice development of the Fleiss Kappa.

Consider scores assigned by two lazy raters of student papers who simply flip coins and then assign a 1 for "good" or 2 for "poor," depending on whether the coin is heads or tails. With two ratings for each paper (one per reviewer), each row of the summary table (representing a paper under review) must be populated with either (2 ,0), (1 ,1), or (0 ,2). A (2,0) entry, for example, means that both raters assigned a rating of "poor," whereas a (1,1) entry means the raters disagreed, and (0,2) means both assigned "good" for a particular paper. With

random coin tosses, we would expect the (1 ,1) rows to occur twice as often as each of the other two types, according to the binomial distribution. Therefore, if we saw rows that look like Table 1, we might suspect that they were random. (Imagine this pattern aggregated over many rows in various orders of appearance.)

Table 1

Ideal Coin Flips

	Ratings of "Good"	Ratings of "Poor"
Paper 1	0	2
Paper 2	1	1
Paper 3	1	1
Paper 4	2	0

A visualization of this random reference data is created by plotting each row as a two-dimensional vector on a plane. These are concatenated in the usual head-to-tail manner, as displayed in Figure 1. If there were three raters instead of two (each flipping coins to assign ratings), a representative table would be (0 ,3),(1 ,2) ,(1 ,2) ,(1 ,2) ,(2 ,1) ,(2 ,1) ,(2 ,1) ,(3 ,0), according to the binomial distribution.

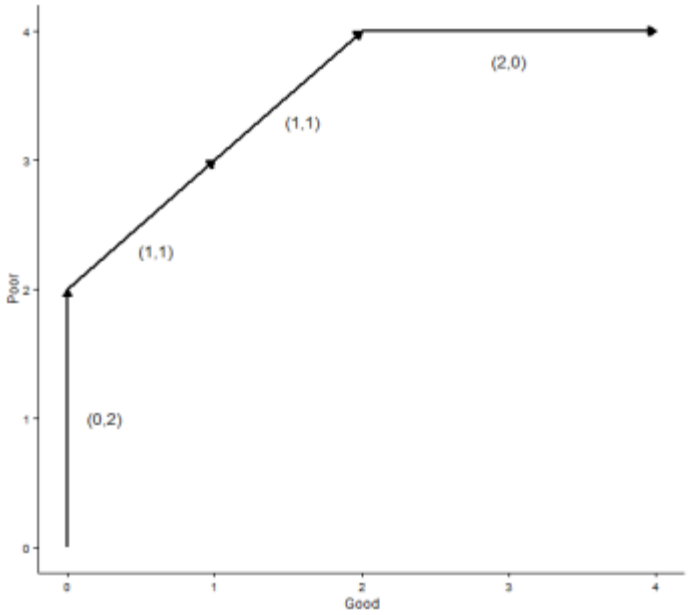

Figure 1. Random ratings with two raters.

Curves for increasing numbers of coin-flipping raters (n = 2, 4, 10, and 100) are shown in Figure 2, where the vectors have been scaled so that the vertical and horizonal dimensions sum to one (.5 plus .5 in this case). That transformation is done by dividing the data table by the total number of ratings and then plotting the vectors as before. In Figure 2, the curves flatten as *n* increases, becoming closer to the straight diagonal (dotted) line, which is the asymptotic case (an infinite number of coin-flippers). The length of such a vector path will be taken as a statistical measure of rater agreement and called λ. In Figure 2, the path lengths decrease as the number of random ratings *n* increases. Conceptually, it is easier to be fooled by two coin-flipping raters (longer path, meaning a higher chance agreement rate) than by 100 coin-flipping raters (shorter path, meaning a lower chance agreement rate).

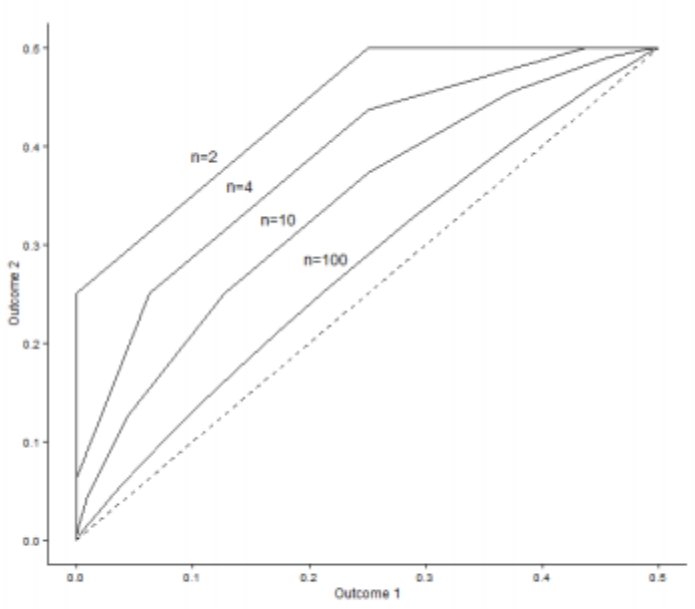

Figure 2. The effect of the number of random raters on λ.

By sorting the rows (which we can do since the order of cases carries no meaning for us), we can arrange them with steepest slopes first (from vertical to diagonal to horizontal), so that the vectors trace out a convex curve on or above the diagonal, as has been done with the ones in Figure 2. These curves have a very useful property. Although each of them shares the same proportion of each outcome, rater agreement increases with the length of the curve. Conceptually, a

straight diagonal line represents the worst possible agreement, with the two outcomes occurring at the same rate for every subject. As an example, imagine a set of good/poor ratings with two raters that looks like (1,1), (1,1) …. (1,1). In this case, the two raters disagreed in every case. The vectorization of (1,1) is a line that traces diagonally up to the right. They would combine to create a diagonal line like the one in Figure 2.

By contrast, if the raters only produce (0,2) and (2,0) ratings (complete agreement on each case), the vectors are all either vertical or horizontal. The resulting curve makes an inverted 'L' by tracing a line straight up and then horizontally to the right and represents the maximum possible match rate as well as the longest curve (with length one).

1.3 Calculating Geometric Agreement

Some preliminary notation is convenient: S is the number of subjects being rated (or classified), k is the number of categories in the classification, and the number of raters who assigned outcome j to subject i is n_{ij}, which is shown as a table with subjects as rows and outcomes as columns. An asterisk is used to denote a marginal sum in that table. For example, n_{*1} is the sum over all rows of column one. In calculating the vector length statistics, we will consider only two categories at a time, using only those subjects with at least two ratings within those categories and using N for the total number of eligible ratings in that set.

For many practical applications it is not reasonable to insist on a constant number of raters per subject, and that restriction is not necessary. Nor will rater characteristics be taken into consideration (e.g., not all raters have to rate all subjects).

Our understanding of what a random data set would look like is the basis for determining significant non-random matching. To do this, we follow Fleiss (1971) in finding the column frequencies in the table. Consider the ratings in column one and column two of the table: n_{ij} for $j \in \{1,2\}$. Considering only these two columns, the fraction of ratings in column one is $p_1 := b_{*1}/(b_{*1} + b_{*2})$, and the column two fraction is $p_2 = 1 - p_1$. The null hypothesis (random raters) is that the ratings are just binomial assignments across the two categories with the given probabilities (p_1, p_2). Using the binomial distribution for the number of raters in each row b_{i*}, $i = 1, \ldots, R$, we can construct a path like the ones in Figure 2 and calculate an expected length $\mu_{\lambda 0}$. Data that show more agreement than this expected random path is evidence for reliability of the measurements. By comparing the length λ to the mean random length $\mu_{\lambda 0}$ and using the standard

deviation of the random lengths σ_{λ_0}, we can perform a hypothesis test and generate a *p*-value (assuming enough ratings to use the Normal approximation).

Of course, most ratings scales have more than two outcome categories, which motivates the second idea. Instead of assessing the rating scale as a whole, we can study it in detail by computing conditional rater agreement *between each pair of outcome categories*. For example, with a three-point scale (e.g., poor, average, and good), there are three different comparisons: poor/average, average/good, and poor/good.

This disaggregation into multiple reports allows an assessment of the reliability of ratings *within* the scale. It could be that raters agree well on poor/good ratings but not on average/good ratings, because the former is an easier judgment. In fact, we should *expect* this pattern to be a characteristic of ordinal scales.

Given a pair of outcome categories, we measure the path length λ that represents actual agreement rates and then compute Kappa following Cohen's original (1960) formulation (recalling that μ_{λ_0} is the average path length for random ratings),

$$\kappa_\lambda := \frac{\lambda - \mu_{\lambda_0}}{1 - \mu_{\lambda_0}}.$$

The numerator is the difference between actual and average random path lengths, and the denominator is the difference between perfect matching ($\lambda = 1$) and the average random rate. The fraction therefore represents how much of the possible non-random agreement is actually demonstrated as a number between zero (actual ratings are the same as average random ones) and one (perfect actual rater agreement).

The Kappa statistic is provided in combination with the graph of the vector path and statistical confidence (a *p*-value). Together these can inform practical decisions about the quality of agreement among raters.

1.4 Random Arc Length Statistics

The length of a vector arc for a pair of outcome categories (J_1 and J_2) is derived from those two columns of the ratings data table and given by a sum over the R rows that have at least two ratings in the J_1 and J_2 categories:

$$\lambda_{J_1 J_2} := \frac{1}{N} \sum_{i=1}^{R} \sqrt{(b_{iJ_1})^2 + (b_{iJ_2})^2}.$$

The sum ranges over each eligible row of the table, each of which gets turned into a vector length. The length is determined by the Pythagorean Theorem, which is the formula inside the sum. These lengths are averaged by dividing by the total number of ratings. This ensures that the maximum λ is *1*. That maximum can only occur if all the subjects have perfect rater agreement, a fact that follows from convexity but is also easy to visualize from the graph.

To clarify what is meant by "eligible rows," when a pair of outcomes is chosen to compute the arc length λ, the rows of data for those two outcomes may contain instances of (0, 0), (0, 1), or (1, 0). This means that for the subject with ratings in that row, a maximum of one rating fell into the two rating types under consideration. Since there is not enough data in such rows to make a judgment about rater agreement, they are omitted, and *N* becomes the total number of ratings in the remaining *R* rows.

The vector length λ of the usable *R* rows is compared to the mean length μ_{λ_0} obtained under the assumption of randomness. The null hypothesis assumes that the column proportions (the number of eligible ratings in the column divided by the total) are used in binomial sampling and that the number of samples is the number of ratings in a given row. For convenience, we will call these row counts n_i. The column proportions are calculated by adding up the column and dividing by the total, so that $p := b_{*J_1}/N$ is the proportion for outcome J_1, and $1 - p$ is the complementary proportion for J_2.

For each row $i = 1, \ldots, R$ the binomial distribution (representing our hypothetical random raters) can be used to find the average length of a binomial vector (x,y) where $x+y = n_i$ and the probability of x is p. Using the formula for the binomial distribution to find the probability of a ratings mix and then multiplying by the vector length of that set gives average row lengths

$$l_i := \sum_{j=0}^{n_i} \binom{n_i}{j} p^j (1-p)^{n_i - j} \sqrt{i^2 + (n_i - j)^2},$$

These are summed and scaled by dividing by *N* to obtain the average arc length under the null hypothesis,

$$\mu_{\lambda_0} = \frac{1}{N} \sum_{i=1}^{R} l_i.$$

The variance of the random arc lengths is calculated using the variance formula for the binomial distribution,

$$\sigma^2_{\lambda_0} = \frac{1}{N^2} \sum_{i=1}^{R} \sum_{j=0}^{n_i} \binom{n_i}{j} p^j (1-p)^{n_i - j} \left(\sqrt{i^2 + (n_i - j)^2} - l_i \right)^2.$$

For large enough *N*, the central limit theorem allows us to approximate the distribution of random lengths by a Normal distribution. Using the mean and variance above, *p*-values can be generated in the usual way to test the hypothesis that the observed ratings have a mean greater than μ_{λ_0} with a given confidence.

Because the random variables that comprise the random baseline (null hypothesis) case are binomials, software can simulate data to compare to the calculated values. This was done as a check to the code that produced the diagrams. The histogram in Figure 3 shows 10,000 simulated values for a sample data set, and the solid line traces the Normal distribution with μ_{λ_0} and σ_{λ_0} calculated using the formulas given above. The code and sample data to reproduce similar results are found in Eubanks (2016a).

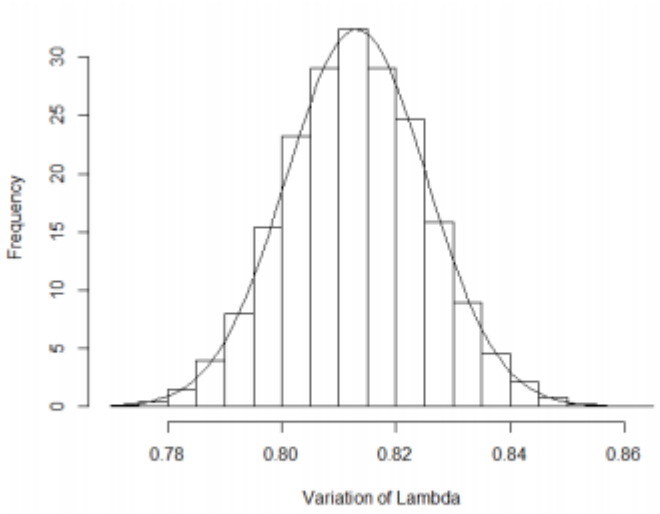

Figure 3. Comparison of simulated (bins) to calculated (curve) λ.

1.5 The Meaning of λ

As discussed above, the maximum length of λ is one (the inverted L shape), corresponding to perfect rater matching. For statistical derivations of Cohen's or Fleiss' Kappa, see Fleiss et al. (1969). With the Fleiss Kappa, one can have a match rate of zero, whereas the minimum λ is the length of the diagonal $\sqrt{p_1^2 + p_2^2}$, where (p_1, p_2) are the column proportions of the two outcomes under consideration. When squared, this quantity is the base match rate considered random for Fleiss. However, Fleiss compares this baseline probability to the actual combinatorial match rates within each row (ratings on a single subject)

using a sum over $\binom{n}{2}\Big/\binom{k}{2}$ for *n* raters, *k* of whom assigned the same outcome. There are some undesirable small-*n* effects of this. With two raters, a row (1, 1) counts as zero matches, whereas four rater responses of (2, 2) have a match rate of (1 + 1) /6 = 1/3. If the column proportions are (.5, .5), both of these cases represent the worst possible match rate. For the λ length calculation, (1,1) rows accumulate outcome 1 and outcome 2 at the same rate, which are drawn along a diagonal from (0, 0) toward (.5, .5). Two rows of (1, 1) ratings would be counted the same as one case of (2, 2). Thus, the definition of λ distance is more self-consistent than combinatorial matches. Rather than a worst case of zero matches, as with Fleiss, the λ distance has a worst case match rate when the accumulation of the outcomes exactly matches the column rates (p_1, p_2). In Figure 2, as the number of raters increases, the binomial distribution approaches the worst case rate (the diagonal line), which creates more efficient Kappas the larger *N* becomes.

Visually, the meaning of the λ statistic is intuitive as the length of a path, but we can relate it to match probabilities. For the sake of simplicity, assume there are a constant number of raters *n* per subject for each of the *R* subjects, so that the total number of ratings is $N = nR$. In this case, for each pair of outcomes we have

$$\lambda_{J_1J_2} := \frac{1}{R}\sum_{i=1}^{R}\sqrt{(b_{iJ_1}/n)^2 + (b_{iJ_2}/n)^2},$$

which is comparable to the same thing without the radical:

$$M_{J_1J_2} := \frac{1}{R}\sum_{i=1}^{R}(b_{iJ_1}/n)^2 + (b_{iJ_2}/n)^2 .$$

Here, $M_{J_1J_2}$ calculates the asymptotic match probability using the Fleiss formula as the number of raters increases to infinity. Using Jensen's Inequality (Jensen, 1906), we see that

$$M_{J_1J_2}/R \le \lambda^2_{J_1J_2} \le M_{J_1J_2},$$

so that the two measures are, in a certain sense, equivalent.

A direct comparison between the Fleiss and λ -based Kappas was made using a software simulation that created 100 sample data sets of 100 rows each, for probability distributions that ranged from (.50,.50) to (.99, .01) in increments of .01 (totaling about a million simulated subjects being rated). The proportion

comprising the left number of the pair was plotted against the difference in Kappa values. The differences between the two Kappas was small across the range of these proportions.

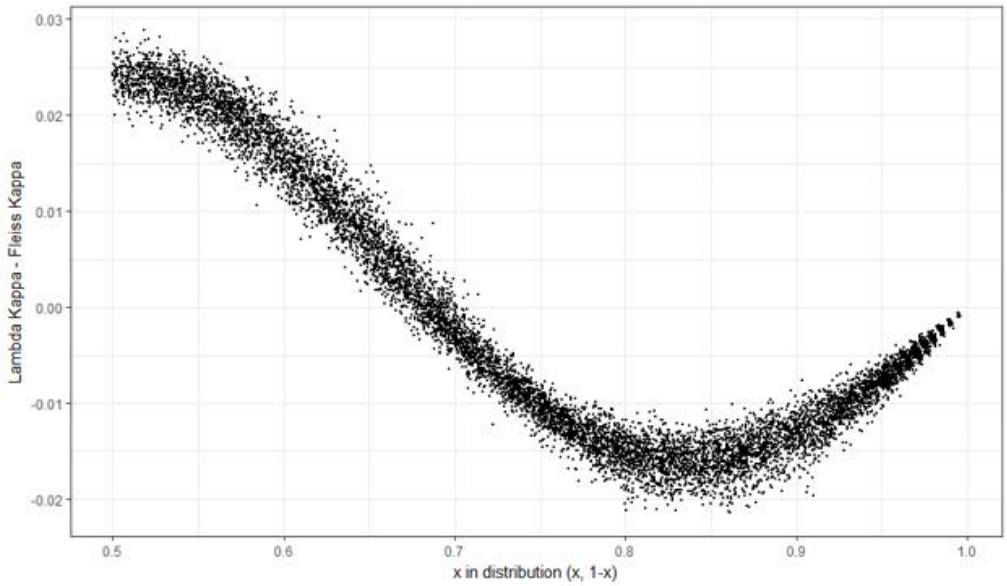

Figure 4. Comparison of λ-based Kappa and Fleiss Kappa in simulated data.

The graph in Figure 4 shows that the λ-based Kappa in the simulation is within .03 greater than the Fleiss Kappa when the distribution is uniform (.5, .5), and less than .02 less near the distribution (.85, .15). The simulation was run multiple times, yielding the same results, and the code is available from the author.

Although the vector-path λs were chosen primarily for their graphical usefulness, there is a way to interpret them with respect to match rates. Consider a subject that has been already rated by several raters on a poor/average/good scale, and for concreteness, imagine that the proportions are (.20, .30, .50). If two new independent raters rate the subject, what is the probability that they will match? It is reasonable to assume that their ratings will approximate the same distribution (i.e., 20% chance of poor rating, etc.). If so, the chance of both new raters choosing "poor" is (.20)(.20) = .04 (the multiplication is justified by independence). The probability of matching on any of the three ratings is .04 + .09 + .25 = .38. This is a measure of how easy it is to reach agreement for that one subject, and a measure over all subjects could be found by averaging them. The vector length is obtained by taking the square root of the sum of squares to get $\sqrt{.38} = .62$. This length can be interpreted by constructing a simpler probability

distribution over two outcomes (don't match, match) = (.38, .62). Two independent selections taken from this distribution match if and only if both values are "match." This has the same match rate as the original (.20, .30, .50) distribution, and this simplification to two outcomes can be accomplished no matter how many outcomes the original has. Intuitively, the simplification behaves like a logical AND operation on the original distribution and yields a parameter (.62 in this case) that can be visualized as a vector length but also represents a match probability.

1.6 Assumptions

The derivations above that create λ and hence Kappa make certain assumptions that are used in justifying probabilistic calculations. They are as follows:

1. As with the Fleiss Kappa, ratings are assumed to be independent. In practice, this is probably violated when a rater sits down to evaluate a stack of papers, because of the accumulating context. Also, different raters may have different rating styles, which also violates the independence assumption. This can be investigated empirically using item response theory if the raters are identified within the data set.
2. As with the Fleiss Kappa, the proportions of ratings for a given subject are assumed to accurately represent the ease or difficulty of rating that subject, measured by the matching rate. This is more likely to be true when the number of raters per subject is large.
3. As with the Fleiss Kappa, the null hypothesis assumes that all the raters are randomly assigning ratings according to the overall frequencies of ratings. This seems reasonable because it is actually possible. For example, raters could flip coins for pass/fail ratings, and this is a logical worst case for rater agreement.
4. In calculating the total λ, the lengths of the row vectors (each corresponding to a set of ratings for a single subject) are weighted by the number ratings in them, hence giving more weight to cases that have more raters. This assumption could be changed to weight each subject (i.e., each row of the data) equally. Computationally, this makes a difference only if the number of raters per subject varies widely and there is a relationship between the number of raters per subject and the associated match rates.

The caveats listed above are less likely to be problematic with larger data sets than with smaller ones.

2.0 Value Contribution

In educational measurement, "the concern of reliability is to quantify the precision of test scores and other measurements" (Haertel, 2006, p.65). General theories of reliability include applications from classical test theory (Haertel, 2006, p.67) and generalizability theory (p.87), but as Haertel notes (p. 99), these methods are more appropriate for continuous rather than discrete types of measurements. The classification of observations into ordinal or nominal categories has its own body of literature on reliability, and Haertel refers readers to Agresti (2013). Another source on that subject is Fleiss, Levin, & Paik (2003).

2.1 Categories of Reliability Measures

Stemler (2004) undertook a categorization of inter-rater reliability measures as estimates of either consensus, consistency, or measurement. Consensus estimates measure exact matches among raters, such as when nominal categories are used, and Stemler locates the Kappa statistics here. Consistency estimates relax this stringency and measure the tendency for raters to rate *in the same direction*. For example, if rater A consistently rates one category higher than rater B for the same subjects, this is a sign of consistency but not consensus. The Pearson correlation coefficient and Spearman's rank coefficient are given as examples of consistency measures. Stemler notes (p. 5) that these measures share with Kappa the problem of being "highly sensitive to the distribution of the observed data."

Stemler's measurement estimates include principal components analysis, generalizability theory, and a Rasch facets approach (item response theory). A recent example of a measurement approach is found in Wang (2017), where "scoring tasks are regarded as a test-like activity for raters, and accuracy ratings are obtained to evaluate their scoring proficiency" (p. 37). The analysis employed a many-facets Rasch model. The paper demonstrates a thoughtful use of a complex analysis that produces detailed reports on individual raters. This method can be an advantage to understanding the operation of a rating process, but it comes at a cost. Rasch-type analyses are complex and can have assumptions that the Kappas do not have, for example, the assumption that a unidimensional construct is being rated (Lane & Stone, 2006, p. 417). Nevertheless, item-response approaches like Rasch facets are useful for highly-detailed reporting on individual rater and subject statistics within a self-consistent mathematical model.

We can make order from this menu of possibilities by placing the rater agreement question into the context of Moss's (2004) argument for hermeneutics,

which Moss describes as "a holistic and integrative approach to interpretation of human phenomena that seeks to understand the whole in light of its parts, repeatedly testing interpretations against the available evidence until each of the parts can be accounted for in a coherent interpretation of the whole" (p. 245). As such, principal components analysis is useful as a high-level overview, and Rasch-type reports comprise dense details that are either interesting on their own or as a route to a holistic understanding. The pairwise facet graphs derived in the current work are somewhere in between, yielding a perspective with details that can be quickly used to identify patterns of interest. It is perhaps uniquely situated among the menu of options as an easy tool for exploratory analysis: the data requirements are minimal (ratings for anonymous subjects by anonymous raters), the algorithm is fast (e.g., a report on a half million ratings takes about a minute on a desktop computer), and the reports are easily understood.

While Moss wrote about reliability in educational measurement generally, the argument is particularly applicable to using human raters to assess student writing: a social construction that impacts learners. If a review committee decides that a student's portfolio is substandard and does not fulfill university requirements, this decision has consequences for the student. Similarly, conclusions about the quality of a writing program that lead to changes have consequences. Rather than beginning with an idealized "true score," such as is assumed to exist in item response theory, it is possible to use social construction as a foundation for analysis. This relates to Messick's (1986) proposal to consider the consequences of measurement as part of validity.

As an example of this, consider the rules of thumb that have been passed down as to what constitutes *good agreement.* Some statisticians pose arbitrarily-created thresholds like .75 as a minimum Kappa to qualify as good agreement (Landis and Koch, 1977). In fact, what constitutes usual practice and what is good practice emerges from data and its use in research communities and cannot be determined *a priori*. How much rater agreement should Amazon.com expect in its product ratings before it accepts them as useful to its business model? How much rater agreement is enough to do aggregate analysis of a writing program? How much is enough to ethically use ratings to determine the status of individual students? When should someone ask for a second or third opinion on a medical diagnosis? These are statistical questions, but they are equally social ones.

The geometric facets method, as with any statistical algorithm, is neutral with respect to its social implications. However, its design produces the best results with large data sets, making it a good tool to understand the relative strengths and weaknesses of an agreement-making social system emergent from ratings data. The vectorization of ratings is not motivated by a statistical problem.

It is done in order to create human-readable graphical interpretations of the relationships in ratings data.

2.2 Limitations of Kappa

Perhaps the most bothersome issue for users of existing Kappa or correlation coefficients is that a single measure of agreement is not very helpful. The reduction of a whole set of data into a single scalar value is a severe compression of the original data. As an analogy, it is more useful to have a graph of a random variable's distribution than just the point estimate of the mean. In describing rater agreement statistics, Agresti (2013, p. 432-436) warns that

> Controversy surrounds the utility of kappa and weighted kappa, partly because their values depend strongly on the marginal distributions. The same diagnostic rating process can yield quite different values, depending on the proportions of cases of the various types […]. In summarizing a contingency table by a single number, the reduction in information can be severe. An alternative is to find kappa separately for each outcome category […] (p. 435)

Traditional Kappa single-parameter measures of inter-rater agreement don't just lack nuance, but as Agresti notes, they can be misleading. This is because the distribution of ratings (for example, ratings skewed to one side of the scale) can affect the Kappa statistic, making it difficult to compare results from one study to another. Stemler (2004) reached the same conclusion, as noted earlier. Agresti suggests a conditional approach: compute a Kappa for each category. The paired comparisons that are the subject of this paper take that idea one step further by computing a Kappa for each intersection of two categories. This partially addresses the other problem that Agresti mentions: the effect of the underlying distribution on the Kappa statistic.

The distribution problem is worth considering in more detail. A Kappa for ratings (or a correlation coefficient, for that matter) on a five-point scale with rating distribution (.10, .10, .10, .30, .40) is not directly comparable to data with the distribution (.20, .20, .20, .20, .20). The advantage of comparing only two outcome types at a time is that the distribution only has one degree of freedom. This makes the problem more tractable, but does not remove it. For example, distribution of (.01, .99) is so heavily skewed toward the second type that most randomly assigned ratings will match by chance, creating an average random vector length $\mu_{\lambda 0}$ close to 1. This imbalance leaves little room to *exceed* the random rate and produce a meaningful Kappa. By contrast, a distribution of (.5, .5) is optimal, leaving the maximum possibility for actual ratings to demonstrate

their non-randomness. For more background on the subtleties of measuring reliability in the context of writing assessment, see Elliot, Plata, & Zelhart (1990, p 88).

As a rule of thumb, the more skewed the distribution, the more difficult it is to demonstrate a significant Kappa, both in effect size and *p*-value. When comparing only two dimensions (rating or categorization types) at a time, we can at least aspire to agree on conventions and accumulate empirical exemplars as a guide. This is made simpler since two-dimensional comparisons can be compared symmetrically: a (.30,.70) distribution can be compared to a (.70, .30) distribution. The graphs of λ produced by the geometric method visually reveal when a distribution is skewed by the shape of the graph, so there is a built-in context when reading the statistics. An example is found in Figure 7, which is discussed in Section 3.3.

2.3 Advantages and Disadvantages of λ

For the practitioner, there are several advantages of the new method over existing Kappas. One is the visual presentation, which presents *all the data* in an organized manner for the reader (examples with explanations are found in the next section). One can literally see where agreement is and is not, due to the geometry. Second, the disaggregation of results into pairs of ratings creates all possible comparisons. This is somewhat possible with the conditional version of the Fleiss Kappa, but it seems to be little used in practice, and it has less resolution and is therefore harder to interpret. By contrast, the paired disaggregation makes it easy to see if an ordinal scale has rater agreement patterns that look ordinal (easier judgments should have higher rater agreement). Finally, the paired analyses reduces the distribution skew problem by isolating it to those pairs where data is sparse for one rating or another.

One disadvantage of the geometric method described in this paper is that it requires sufficient numbers of ratings to create the graph facets and statistics. The number of ratings needed depends on the scale used and the distribution of ratings. For example, with a pass/fail (two-point) scale that has evenly-distributed ratings, the method may be useful for as few as N = 20 ratings, but this would not be true for a five-point scale, where the lack of density in each of the conditional plots would likely be too small to be useful. A recent actual use of the method analyzed 23 student papers on a three-point rubric with three raters and a total of 56 ratings. This yielded usable results, giving Kappa = .74 ($p = .02$) with N = 23 when comparing the highest and lowest categories of the scale.

An additional consideration is that the path length λ and Kappa calculation make no use of rater identities, even when available. This makes the new method

most appropriate for large samples where we want to know about the characteristics of the rating system as a whole. In order to understand the behavior of individual raters, an item-response-type method would be more appropriate.

3.0 Technique Application

The preceding sections advocate inter-rater analysis as a way to understand social meaning-making, and the two examples in this section illustrate that principle. The first example will give an overview of interpreting the results, and the second example will be more detailed and focused on writing assessment.

3.1 Judging Wine Quality

Can wine drinkers distinguish quality of wine in a blind taste test? In an article entitled "An Examination of Judge Reliability at a major U.S. Wine Competition," Robert Hodgson (2008) gives an answer from four years of rating data. From the abstract: "Each panel of four expert judges received a flight of 30 wines imbedded with triplicate samples poured from the same bottle. […] Judges tend to be more consistent in what they don't like than what they do."

The author of that study was kind enough to share a data set of wine ratings. This sample comprised $R = 184$ wines that had been rated each by four judges using a four-point ordinal scale similar to Olympics medals, which are encoded here as 1 = no medal, 2 = bronze, 3 = silver, and 4 = gold, in ascending order of perceived quality. For example, the first row of data is (3,3,3,3), meaning that all four judges rated the wine as silver medal quality, the next-to-best classification.

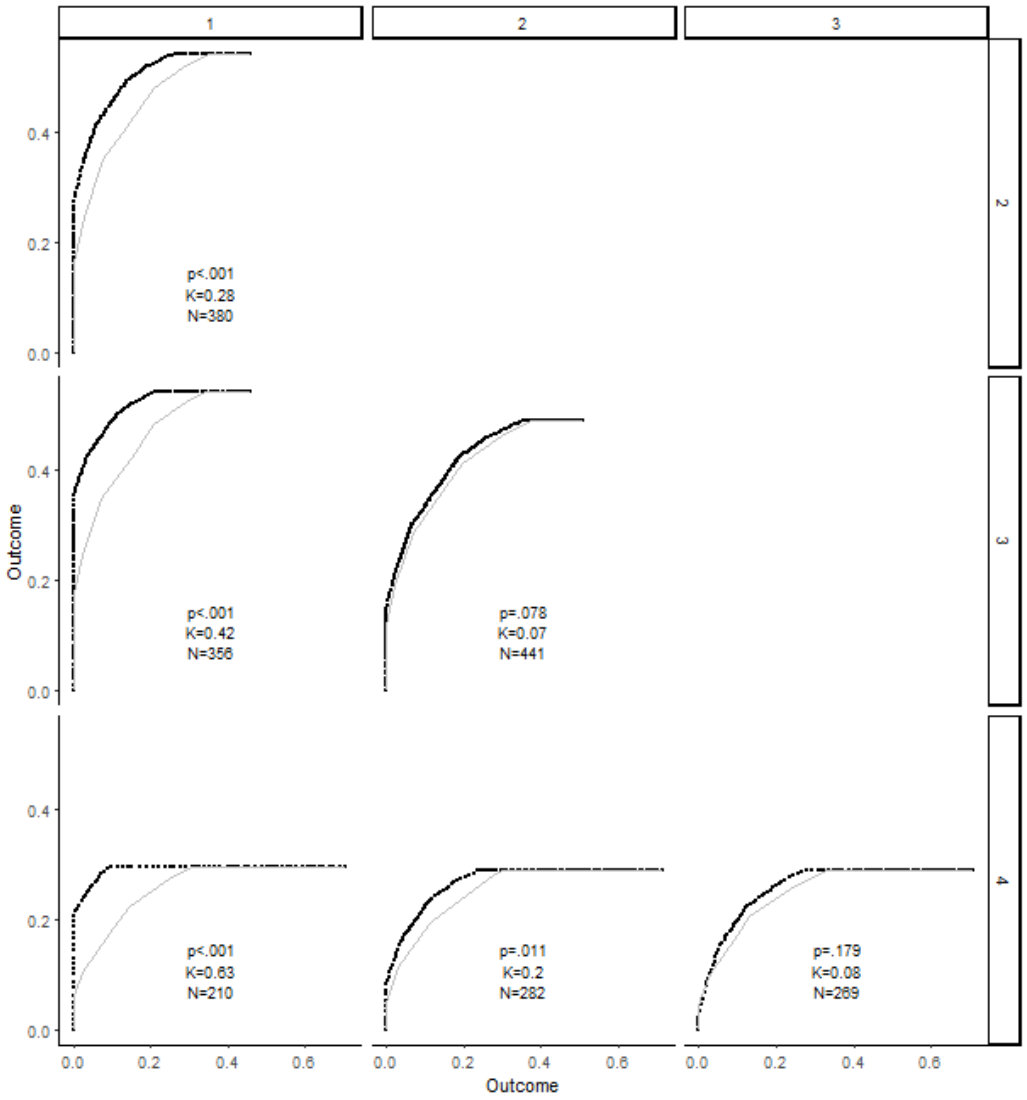

Figure 5. Wine judging agreement.

The graph facets and accompanying statistics in Figure 5 contain different perspectives on the wine ratings. A useful first scan of the graphs is to become aware of any skew present in the distributions. Notice the scales along the two axes—these are consistently applied for each graph for ease of comparison. The middle facet, which compares ratings of 2 (bronze) to 3 (silver) has a balanced distribution, with 2s and 3s each appearing about half the time. By contrast, the 1 (no medal) versus 4 (gold) is skewed, with approximately a (.7, .3) distribution. Recall that such imbalance makes it more difficult to generate large and significant Kappas, an effect that is visible on the graph. The dark dots each

represent one subject case (a wine sample), and the thin gray line shows the average by-chance results that were calculated using the formulas given earlier. Notice how the dots and the line overlap substantially more on the top (horizontal) of the graph than on the bottom (vertical). Two random selections from a (.7, .3) distribution will match at rates of (.49, .09), compared to a balanced (.5, .5) distribution, which matches at (.25, .25). In this case, however, the divergence between random and actual results is large and significant despite the moderate imbalance.

Next, inspect the Kappas on the main diagonal. These are intuitively the most difficult distinctions to make (e.g., bronze versus silver, silver versus gold). The agreement in distinguishing a no medal wine and a bronze medal gives K = .28, whereas the other two Kappas on the main diagonal are .07 and .08, and neither has a small *p*-value. This reiterates Hodgson's analysis of variance finding that judges agree more about what they don't like than what they do like.

Finally, notice that the Kappas increase as we inspect the sub-diagonal and then the bottom left comparison. Intuitively, these represent *easier* judgments (e.g., bronze versus gold). The easiest judgment (no medal versus gold) has the best agreement, with K = .63 and $p<.001$.

The wine-tasting results are interpretable as social meaning-making. For example, they suggest that choosing *a great wine* is a matter of aesthetic taste, which individuals are entitled to define for themselves. The results also suggest that it is possible to buy a bottle of wine for dinner guests that they may uniformly dislike.

3.2 Assessments of Undergraduate Writing

There is great interest in reliably assessing the writing ability of students. A common method employs a rubric that describes aspects of writing such as knowledge of conventions or organization and sets different levels of achievement that are appropriate to learners. Raters use the rubric to assign ratings accordingly to student written works, and the ratings are then aggregated for various purposes. For example, one might be interested in detecting average change over time in the qualities of student writing as measured by the rubric ratings. The degree to which raters agree determines how usable the ratings are; if the level of agreement does not exceed random number generation (the baseline for the Kappa statistics), the ratings are probably unusable. In practice, it is easy to see how some ratings might have more agreement for some rating types than others, because some judgments are easier than others—as we saw with wine judging. The low ratings may find more agreement than high ratings, for example, if the former is more rules-based (e.g., grammar and spelling) and the latter is more aesthetic. Or it might be the

case that the ratings have more agreement on both ends because exceptional cases are easier to recognize, and the middle ratings have less agreement because they require finer judgments. These sorts of questions cannot be answered using single-parameter statistics like existing Kappas or correlation coefficients.

The title of this paper is a nod to Brian Huot's (2002) *(Re) Articulating Writing Assessment*, a small book with a high density of ideas. From these, I will cherry-pick one quote that I underlined in my copy of the book: "We need to begin thinking of writing evaluation not so much as the ability to judge accurately a piece of writing or a particular writer, but as the ability to describe the promise and limitations of a writer working within a particular rhetorical and linguistic context" (p. 107). Who better to make these descriptions than the teaching faculty who sweat over student writing assignments and get to know their students over weeks of feedback and interaction? This should be sufficient exposure for professionals to reach conclusions about the writing abilities (in the sense of promise and limitation) of their students. It turns out to be a simple procedure to ask for and collect that information. See Eubanks (2008) for the original description of this idea.

Ratings of students as writers from a small (2,700 undergraduates) private liberal arts university were generated in this way, by surveying faculty at the end of each term to rate student writing ability. In contrast to the more usual method of rating individual pieces of writing, course instructors were asked if they had observed student writing during the 15-week term to a degree that would let them make a judgment and to use a common rubric to rate students as writers in the discipline. The scale used has five points, from *Basic Skills* (scored numerically as zero) to *Ready to Graduate* (a score of four), and the instructors were encouraged to place each student in a continuum between these to choose the best rating. *Basic Skills* in this case means that the student was judged as not yet producing college-level work, and *Ready to Graduate* represents the ideal skill level that a college graduate should have.

The method described has some weaknesses as an assessment device. In particular, it is well known that writing skill in one discipline does not automatically transfer to another (cf, Yancey, Robertson, & Taczak, 2014; National Research Council, 2012). So, in this case, we are assessing both in-discipline and out-of-discipline writing as if they were the same thing, placing the focus on general writing skills that do overlap from one discipline to another. We should expect lower levels of rater agreement than if all the ratings came from the same type of writing. Similar considerations apply to rating heterogenous student portfolios.

In defense of this general approach to assessment, liberal arts students are asked to write in various genres, and the expectations of some external

stakeholders include *written communication* in general terms. Moreover, if we want to understand longitudinal growth of writers, it will include multiple genres because undergraduates generally do not immediately specialize. They take a variety of general education requirements and may change majors.

An advantage of the assessment method described is that it is easy to generate large data sets. In this case, the ratings gathered from four terms in the academic years 2015-16 and 2016-17 yielded data on 3,209 unique students and covered 93% of the undergraduate student body for each of the two academic years. The inter-rater calculation requires at least two ratings for a student in a given term to assess rater agreement within a term. This resulted in a usable data set with 1,765 students, totaling 3,916 ratings over the two academic years. For purposes of inter-rater agreement, student ratings in different terms were considered as different *students*, because student writing abilities are expected to increase over time (the data support this). Therefore, ratings taken within one academic term were considered a snapshot sufficient to assess rater agreement. Because of the disaggregated comparisons in the conditional rater agreement statistics, each rating may be used in more than one comparison, and the total number of comparable ratings within the report was 9,454. The graphs and statistics below summarize these comparisons.

3.3 Reading the Conditional Agreement Graphs

For simplicity, first consider a pair of rating outcomes. Perhaps the most important distinction in student writers is the difference between beginning college students and those about to graduate. After four years of instruction, can faculty members tell the difference?

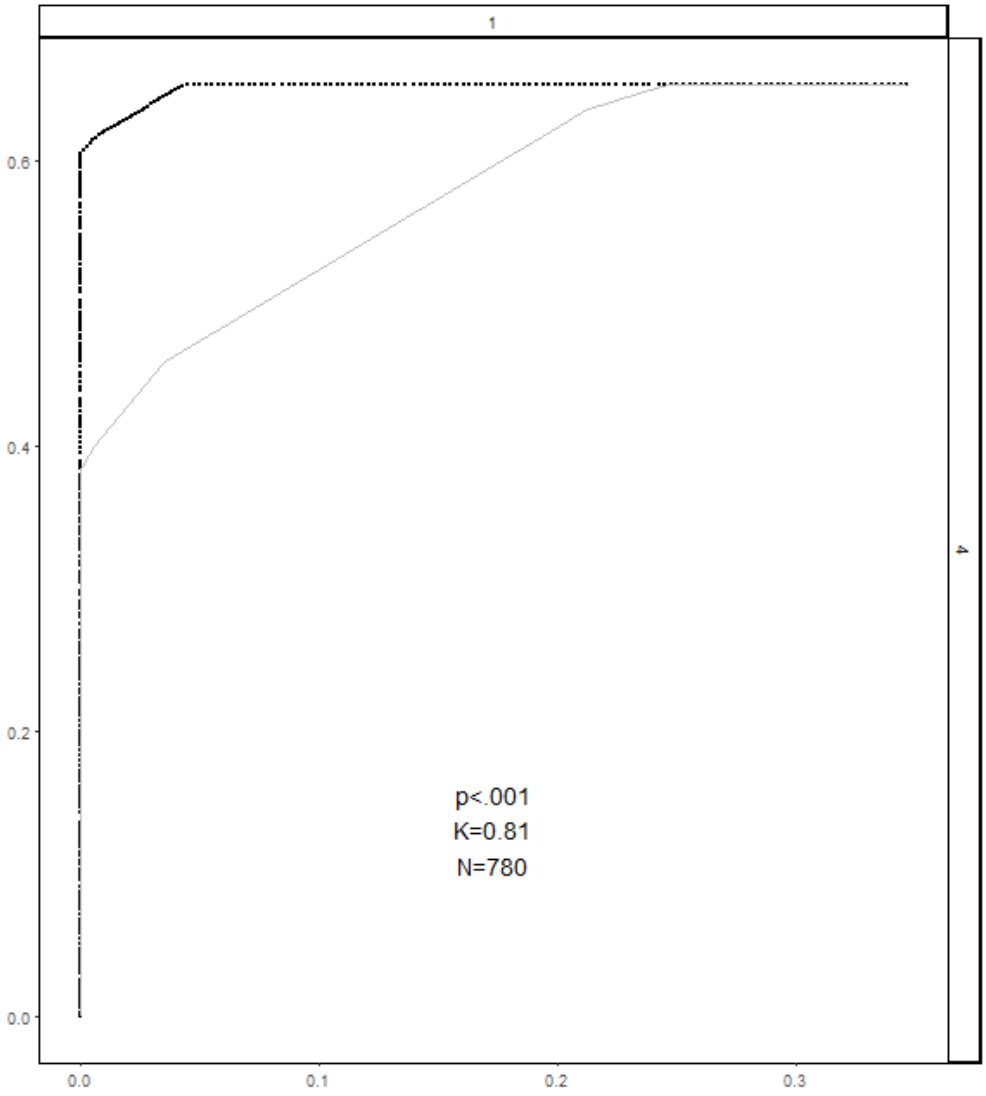

Figure 6. Comparison of 1- and 4-ratings of student writing.

The graph in Figure 6 was generated by first identifying those subjects with at least two ratings of a combination of *one* (the level of a beginning college student, one rating level above *Basic Skills*) or *four* (what we hope graduates achieve) within the data sample. The column frequencies are computed as (b_{*1}/N, b_{*4}/N), where N is the total number of ratings remaining in these two columns. The total number of ratings for a subject is used together with the column frequencies to compute the expected binomial vectors for the given distribution. These are sorted by slope, from vertical to horizontal, and combined to make the thin solid reference "chance" line that appears on the graph as a visual guide.

Above this reference line are dots that represent the vector addition of the actual data, again sorted by slope to trace out a convex shape. Dots are used so that the density of data is evident from the graph. Dots above the reference line comprise evidence of inter-rater agreement higher than random. The *p*-value gives the statistical significance of this, versus the null hypothesis that the ratings were randomly assigned from the overall distribution. That distribution can be read off the graph: the vertical axis, representing the four-ratings (see the key on the right side and top of the graph), accounts for more than 60% of the ratings in this subset, with the remaining ones being one-ratings. Random assignment of ratings to students from that (.60, .40) distribution results in the average curve traced out by the thin line. That is the baseline *chance level* of agreement for that facet. The further the dotted line is from the thin reference line, the higher Kappa will be, and the smaller the *p*-value will be.

The actual agreement in the ratings is represented by the arc of darker-appearing dots along the left and top of the graph. Each dot is one student (or more precisely, one student-term). The space between dots (the vector length) represents the ratings for that student, as explained in Section 1. Horizontal segments and vertical segments represent complete rater agreement that the student is a *one* or *four*, respectively. Note that the number of these exact agreement cases far exceeds the number expected by chance. The diagonal segment of the dotted line represents cases where the ratings were mixed: students who received both one and four ratings. These are relatively rare.

The Kappa statistic, shown as .81 on the graph, is the fraction of the available "room" above the reference line that is accounted for by the data—a kind of effect size. In this case, the actual data is *81%* of the way between the baseline chance line (Kappa = 0) and perfect agreement (Kappa = 1), which would appear as an inverted L, with only vertical and horizontal segments. The N = 780 given on the graph is the number of ratings.

The single graph in Figure 6 already shows the usefulness of being able to visually inspect rater agreement, but it is much more powerful when the whole scale is shown at once. To do this, we produce multiple graphs like the one in Figure 6, one graph *facet* for each possible pair of ratings. In this case the zero-through-four scale has five responses, and therefore 4 + 3 + 2 + 1 = 10 facets, one for each pair of response types. The graph with all facets shown is presented in Figure 7.

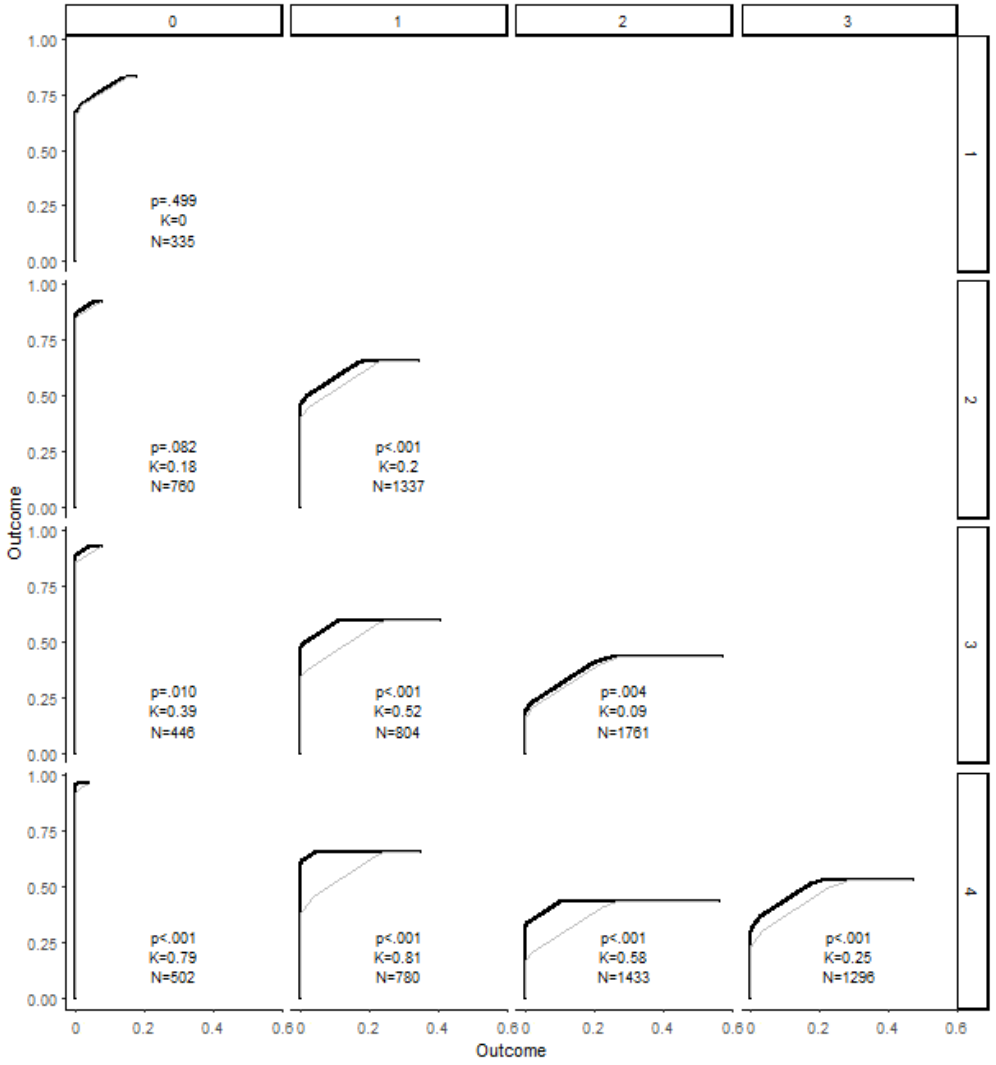

Figure 7. Inter-rater facets for faculty assessments of student writing ability, with 0 = lowest, to 4 = highest.

The comparison of each of the rubric outcome categories with the others is obtained by consulting the row and column indexes in the facets display. Within each plot, as before, the dots represent individual student-terms being rated and trace out the observed path. As before, the thinner line is the expected curve for random assignments with the given distribution of scores, numbers of raters (which varies by student), and number of student-terms. The calculated *p*-value is given on each plot, where a low enough *p*-value (depending on what alpha level is chosen) rejects the hypothesis that the assigned scores are binomial random variables. The *N* is the total number of ratings (not subjects).

Notice that the ratings are very unbalanced in the leftmost column, which compares *Basic Skills* ratings to each of the others. This is because there are relatively few *Basic Skills* ratings to begin with. The top left facet, which compares zero to one rating scores, is usable, but the rest in that column should be viewed with suspicion when making judgments from the statistics. In this way, the facets can separate out distributional skew that becomes problematic in a single parameter measure of reliability.

The diagonal of the facets contains outcomes that are adjacent. For example, the top left compares outcome 0 to outcome 1, and the bottom right is 3 compared to 4. Because the ratings are ordinals, we would expect that there would be less agreement between raters on outcomes that are on this main diagonal, because these should be the most difficult decisions for a rater to make. Deciding between a 3 and 4 rating should require a finer assessment than deciding between a 2 and 4, and this is even more applicable for the 1 and 4 comparison in Figure 6.

A useful way to navigate the facet graphs is to look at the main diagonal, the top left (0,1) comparison down and right to the (3,4) comparison. These comprise the most difficult ratings on an ordinal scale for the reasons just mentioned. Recalling that a small *p*-value is a measure of statistical confidence that the ratings are not random, the top left facet, which compares zero to one ratings (*Basic Skills* to entry-level college writing) provides no evidence that the ratings match more than random assignment would. The other Kappas on the main diagonal range between .09 and .25 with small *p*-values, and one can see concomitant divergence of the darker line (the dots are too close together to see individuals) and the thin reference line. This first look shows that faculty raters agree more about which students are writing at the level of a college graduate than they do about entry-level writing.

Proceeding inward, inspect the first sub-diagonal. These are the (0,2), (1,3), and (2,4) comparisons, which should be easier judgments than the ones on the main diagonal. We should expect to see larger values for Kappa on the sub-diagonal for that reason. And this is so, with Kappas of .18, .52, and .58, although the first of these is impugned by a high *p*-value. Moreover, the (0, 2) facet suffers from a very imbalanced data set: the fraction of zero ratings is only about 10%. As with any kind of statistics, fewer samples means less confidence. As educators, we are delighted that there are not more ratings of *Basic Skills*, but as researchers it is an annoyance.

The second sub-diagonal should have even higher Kappas, and this is true, even for the unbalanced (0,3) facet. The (1,4) facet is the one in Figure 6. Ignoring the leftmost zero column of the facets, the overall pattern is convincing evidence that the scale behaves like an ordinal scale should: the ends of the main

diagonal have higher agreement than the middle, and the Kappas increase as we move left and down.

3.2 Implications and Use of Findings

In the example just given, the analysis of writing ratings revealed a likely problem with the instructors' definition of *Basic Skills* writers. This finding persisted over two years (the second year of data replicated the first year in this respect), and so the Writing Program director led an effort to see if this lack of discernment was experienced by instructors in the Freshman Writing Program. First-year undergraduates enroll in a required writing seminar taught by an instructor from the faculty at large, probably not someone trained in composition or rhetoric. In meetings with these faculty members, there was consensus that they did not, in fact, agree on minimal standards for undergraduate writing and that there was a need to do so. The variation in grades for these seminars reinforced that opinion; after adjusting for the high school grades and standardized test scores of students, the variance in writing seminar grades seemed unacceptably high. This lack of consistency represents a lost opportunity and an unfairness to first-year students. Consequently, the university has collected a range of first-year writing samples to use for classification, with the goal of faculty creating a university standard for entry-level writers. A first review of these has taken place and has revealed at least two interesting observations. The first is that minimum expectations are high and that papers are not littered with easily-identified mistakes like grammar and spelling. The most common comment was a desire to see a creative synthesis of the paper's subject, and "not just a Wikipedia article." The second observation is that the rater from the writing support center thought that the faculty's minimum standard was too high.

The work just described is just the first step in addressing the consistency issue, and there is much left to do. The point of discussing it here is that this issue would not have been noticed without the detailed inter-rater statistics provided by the Kappa facets.

3.3 Validity

Inter-rater agreement only tells us about the consistency of these data; it does not tell us that the ratings actually measure writing ability for the data presented in Section 3.2. It could be that instructors just assign higher ratings to more senior students. This is a question of validity, and some initial work has been done on that for the writing ratings described above. A sketch of current results is provided here for context.

Students who arrive at the university with higher (standardized) high school grades receive higher writing ratings on average. These increase over time with no evidence of a Matthew Effect (Stanovich, 1986). Support for convergent and divergent validity comes from the relationship of students' scores on Advanced Placement (AP) tests taken in high school to the instructor ratings as many as five or six years later. Writing-intensive AP courses statistically lead to higher writing scores being assigned for each of six independent natural sample sets (three years of AP data times two years of writing assessments).

Freshmen surveys ask students to rate their abilities relative to their peers in a number of areas, including writing, academics, leadership, physical health, and creativity. Of these, academic and writing self-efficacy are the most predictive of the faculty-assigned writing scores. However, academic self-efficacy is the more important of the two, and a factor analysis of the self-ratings show that students associate writing with creativity, which is not predictive of instructor-assigned writing scores. This echoes Rezaei and Lovorn's (2010) finding that "The results showed that raters were significantly influenced by mechanical characteristics of students' writing rather than the content even when they used a rubric." Interestingly, this quantitative finding contradicts what faculty reviewers said about minimum standards (at the end of Section 3.2), so there may be a lack of common agreement about what creativity is.

This is not the end of the story, but a full validity analysis will not be possible until four years of ratings have been accumulated (e.g. to control for survivorship). At first look, however, it seems that this trust-the-teacher, crowd-sourced method of collecting data is useful in assessing student writing in the aggregate. Beyond that, it seems to have salutatory effects on instructors, who can meaningfully contribute to the assessment process without a large investment of time. One professor in the sciences volunteered that she uses the rating scale as a pedagogical tool, asking students to reflect on their own development at the beginning and end of the course and then comparing it to her own assessments. Her conclusion was that her students were "too hard on themselves."

4.0 Future Directions

Rater agreement is a key reliability measure for many types of data. In testing the mathematics and software used to generate the graphs, a variety of rating types were used. These include Amazon.com reviews of on-demand videos, book ratings, dating website ratings, bottled water quality ratings, airport cleanliness, and Olympic figure-skating scores. Educational data included several varieties of rubric ratings on different-sized scales, including portfolio scoring by

faculty reviewers, course grades, course evaluations, and student peer review scores.

Given the volume of ratings data that is potentially available, we can imagine a database of de-identified rating data with descriptions and statistics. This would be a useful empirical library for consultation by researchers, and such a library would create new possibilities for larger-scale projects and meta-analyses.

Extending the idea that ratings are social constructs first and statistical data second may usefully contribute to the dialog that extends from Moss (2004) to Mislevy (2004) and beyond (itself a hermeneutics practice). An illustrative example will serve as a marker for a larger conversation.

One of the 583,987 rows of Amazon.com test data contains ratings for the on-demand video *South of Nowhere, Season 1*, with 59 reviews. These range from one star (lowest) to five (highest) with a distribution of (.07, .05, .03, .14, .71) for that video. How we make sense of this depends on our purpose. If we want to do an analysis of on-demand video ratings in order to predict consumer behavior, we may want to average the ratings as scalars (4.4). In that case we might want to use standard deviation of ratings as a measure of self-consistency (reliability), since that is how confidence intervals are created. On the other hand, an interested consumer may glance at the ratings as a heuristic filter and then start reading comments for more information, e.g., "I was very pleased with my recent purchase of season one of south of nowhere on DVD. I had read some bad things about this set but I bought it anyway, I was a big fan of the show back then and I just wanted to be able to watch it again. – hearttofalion26." This browsing may be targeted at understanding possible deficiencies to minimize the risk of disappointment. That question operates on two levels—1) can I glean useful information in this instance, and 2) is this is a reliable method for choosing products in general? (For the whole data set, the largest Kappa among the ten facets was K = .27, comparing one-star and four-star ratings. Generally, rater agreement is low.)

Suppose we assume that there is a *true* rating of the video and further assume that the probability of a rating type (one to five stars) being the true value is given by the distribution. Then there is a 71% chance that *South of Nowhere, Season 1* is a five-star video, and a 7% chance that it is a one-star video *in reality*. Under this assumption the estimation of the number of correct classifications in the data set is identical to a match rate calculation. Looking at this the other way around, if we like the idea of counting rater matches as a reliability measure, then by implication we can embrace this epistemology about true scores—the statistics stay the same.

Or we might prefer democracy, where the rating category with the maximum number of ratings is the *true* assessment. For example, imagine a portfolio pass/fail review with three raters. This is guaranteed to produce the *true* score, because there will always be a majority. If we use a three-point scale instead, we would sometimes require a fourth rating to reach a majority. The rule for classifying cases has implications for how we calculate reliability, since any vote that is not a majority vote is wrong regardless of its other properties. With the practices defined, it is possible to construct a custom version of rater agreement that reflects this program. A chance-corrected Kappa in this case would compare the actual number of majority votes to a suitable random benchmark. Note that *majority votes* is different from *number of matches*, and this would lead to a new Kappa.

These examples are intended to illustrate the idea that statistical reliability can emerge organically from the purpose of the project, which may involve tradition, social consequence, bureaucratic feasibility, local politics, pedagogy, research findings, and so on. Describing these derivations and their empirical consequences would ensure a fascinating scholarly conversation with potential benefits to students. It would also clarify when cases are comparable and work toward a shared theoretical and empirical understanding of rater agreement as an integral part of classification. More thoughts on the subject are found in Eubanks (2016b), including a discussion about having *too much* reliability.

Other future work could include developing variations of graphical displays of ratings data. One experimental variation that rescales unbalanced data is available in the code provided in Eubanks (2016a). Another question concerns the shape of the graph in Figure 4. Its regularity suggests a simple relationship between the geometric and Fleiss Kappas that can be derived mathematically.

5.0 Resources

Code to produce the reports found in this paper, as well as sample data sets, can be found at Eubanks (2016a). The particular code and data used to make the figures is also available from the author via personal communication at david.eubanks@furman.edu.

Author Biography

David Eubanks holds a PhD in Mathematics from Southern Illinois University and currently serves as Assistant Vice President for Assessment and Institutional Effectiveness at Furman University. He has worked on the practical side of assessing student learning, including student writing, since the late nineties. His research focuses on writing assessment, predictive analysis, and survey research, including the development of methods and software tools. Some of these applications can be found online at http://github.com/stanislavzza.

Acknowledgments

This research was conducted under an award from the National Science Foundation (1544239) entitled Collaborative Research: The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses. Additionally, the project would not have been possible without the support of Furman University.

The computations and figures for this paper were produced in R (R Core Team, 2015). The R inter-rater package from Gamer, Lemon & Singh (2012) was very helpful.

I am deeply grateful to Norbert Elliot for his feedback and guidance in preparing this manuscript and to the anonymous reviewers, who led me to rethink several key points.

References

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics*, *44*(2), 539–548.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322.

Elliot, N., Plata, M., & Zelhart, P. F. (1990). *A program development handbook for the holistic assessment of writing.* Lanham, MD: University Press of America.

Eubanks, D. A. (2008). Assessing the general education elephant. *Assessment Update*, *20*(4), 4.

Eubanks, D. A. (2016a) Interrater facets. *Github Repository*. http://github.com/stanislavzza/Inter-Rater-Facets

Eubanks, D. A. (2016b) Rethinking interrater agreement. *Assessment Update*, *28*(4), 8-14. DOI: 10.1002/au.30065

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Fleiss, J. L., Cohen, J., & Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*(5), 323.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Galton, F. (1892). *Finger prints*. London: Macmillan and Company.

Gamer, M., Lemon, J., & Singh, I. F. P. (2012). Irr: Various coefficients of interrater reliability and agreement. *R package version 0.84*. Retrieved from https://CRAN.R-project.org/package=irr

Gwet, K. (2015). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement*, *76*(4), 609-637.

Gwet, K. L. (2014). *Handbook of inter-rater reliability:Tthe definitive guide to measuring the extent of agreement among raters.* Gaithersburg, MD: Advanced Analytics, LLC.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.

Hodgson, R. T. (2008). An examination of judge reliability at a major US wine competition. *Journal of Wine Economics*, *3*(02), 105-113.

Kelly Riley, D., & Whithaus, C. (2016). A theory of ethics for writing assessment. [Special issue]. *Journal of Writing Assessment*, *9*(1). Retrieved from http://journalofwritingassessment.org/article.php?article=99

Landis, J. R. and Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lane, S, & Stone, C. (2006), Performance Assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 386-431). Westport, CT: American Council on Education/Praeger.

Messick, S. (1986). The once and future issues of validity: Assessing the meaning and consequences of measurement. *ETS Research Report Series*, 1986(2).

Moss, P. A. (2004). The meaning and consequences of "reliability". *Journal of Educational and Behavioral Statistics, 29*(2), 245-249.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn.* Cambridge, UK: Cambridge University Press.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on Defining Deeper Learning and 21st Century Skills, J.W. Pellegrino & M.L. Hilton (Eds.). Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Powers, D. M. (2012). The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 345-355). Association for Computational Linguistics.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18-39.

Roberts, C. (2008). Modelling patterns of agreement for nominal scales. *Statistics in Medicine, 27*(6), 810–830.

Smeeton, N. C. (1985). Early history of the kappa statistic. *International Biometric Society, 41*(3).

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 22*, 360-407.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1-11.

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing

assessments using a mixed-methods approach. *Assessing Writing, 33*, 36-47.

Yancey, K. B., Robertson, L., & Taczak, K. (2014). *Writing across contexts: Transfer, composition, and sites of writing.* Logan, UT: Utah State University Press.