Journal of Official Statistics, Vol. 28, No. 3, 2012, pp. 395–412

Mutual Information as a Measure of Intercoder Agreement

Ben Klemens[1]

In a situation where two raters are classifying a series of observations, it is useful to have an index of agreement among the raters that takes into account both the simple rate of agreement and the complexity of the rating task. Information theory provides a measure of the quantity of information in a list of classifications which can be used to produce an appropriate index of agreement. A normalized weighted mutual information index improves upon the traditional intercoder agreement index in a number of ways, key being that there is no need to develop a model of error generation before use; comparison across experiments is easier; and that ratings are based on the distribution of agreement across categories, not just an overall agreement level.

Key words: Intercoder agreement; Cohen's kappa.

1. Introduction

This article uses a measure of shared entropy between two classifiers to express an index for agreement between them. This new measure, which will be notated as P_I, will be compared to the traditional measure κ and several related measures. The new measure does away with the standard of applying a series of transformations to the observed rate of agreement, and instead uses information-theoretic accounting rules to calculate the total information in the agreed-upon observations relative to the total information in the individual ratings. The simplest means of measuring agreement is to take the ratio of ratings in agreement to the total number of ratings; the complexity-adjusted measure P_I takes the ratio of information in agreement to total information.

Consider the case of unordered, discrete categories; the case of ordered categories will be discussed as an extension below. Once the distribution of category choices for the individual raters is set, κ depends only on the percent of items which are given the same classification by both raters, while P_I also depends on the characteristics of the categories in which the agreements occur. Categories where the agreement rate is significantly larger than the rate that would occur given independent raters will have a positive effect on P_I, while categories where the agreement rate is smaller than the rate given independence will

[1] United States Census Bureau, CSRM, 4600 Silver Hill Road, Suitland, Maryland, 20233, U.S.A.
Email: ben.klemens@census.gov
Acknowledgments: Thanks to Joanne Pascale and Patricia Goerman of the Census Bureau's Center on Survey Methodology, Eric Slud and Yves Thibaudeau of the Census Center on Statistical Research and Methodology, and a number of exceptionally helpful anonymous reviewers. The computer code used to generate the data analysis, simulations, and figures in this article (based on the Apophenia library of functions for statistical computing (Klemens 2008)) is available upon request. Also, code for calculating P_I is available upon request for C, Python, and R. This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

lower P_I. The final score is an amalgam of each category's score, and broadly asks whether there is more weight on categories where there is more than chance agreement, or on categories where there is less than chance agreement.

We must separate the unordered case from the ordered case because the model of agreement at random changes depending on the situation. The literature is thus filled with variants of κ for different situations, and it is difficult to argue that it is valid to compare the statistic across studies where the underlying model of agreement at random is different. Conversely, the ratio of (possibly weighted) information in agreement to total information can be calculated for every situation in this range, meaning that comparison across studies has greater validity.

Empirically, the traditional measure κ and the new measure P_I tend to behave similarly. At the extremes, cases where κ is zero or one are cases where P_I is also zero or one, and the simulations below find that their power to distinguish small changes in the rate of agreement are remarkably similar. Thus, the alternate statistic here largely retains the empirical characteristics and general intuition associated with κ, while providing several conceptual benefits: there are fewer conceptual issues hindering comparisons across studies, one does not need to develop a model of error generation for different situations, and P_I will be shown to usefully distinguish between situations with strong agreement in one category and situations with more even agreement across several categories, even when κ considers the two situations to be identical.

By way of introduction, this article will begin with a short motivating example. Section 2 gives a brief overview of entropy accounting and defines P_I. Section 3 provides some simple examples and lemmata about the basic properties of P_I. Section 4 presents a simple simulation showing that the powers of κ and P_I are similar. Section 5 is the empirical section, in which the measure is applied to a test of the US Census Bureau's American Community Survey; in all cases, P_I and κ behave in a similar manner.

A motivating example. Consider the case of two raters classifying a sequence of observations into categories. The two rows represent the sequence of categorizations of the same data set by two raters, so each column represents two classifications of the same data point by two independent raters.

Table 1. Two coders (top, bottom) classified a series of 12 observations

1	1	1	1	1	1	2	2	2	3	3	3
1	1	1	1	1	1	3	3	2	2	2	3

The coders readily agree on which observations fall into category 1, but are muddled with regard to 2 and 3.

The simplest index for intercoder agreement is the percentage of classifications that are in agreement across raters which will range from 0% in the case of no items classified the same by both raters, to 100% in the case of full agreement. This is notated as P_o, and in this case with twelve items and eight correct classifications is 66.67%. For many purposes, this index is all that one needs, but it is sometimes unsatisfactory because it does not reflect the complexity of the task. If there are only two categories, two people picking categories by each flipping a coin would arrive at 50% agreement; if there are a hundred categories and the task is difficult, 50% agreement may be excellent.

Let the observed probability of the first rater choosing category i and the second choosing category j be p_{ij}. Then the marginal probability $p_{i\cdot}$ indicates the probability that the first rater chooses category i and $p_{\cdot j}$ indicates the probability that the second rater chose category j. Thus, given C categories, $P_o \equiv \sum_{i=1}^{C} p_{ii}$.

In some fields, authors distinguish between the theoretically ideal distribution p and the current set of ratings which provide a single empirical estimate of the distribution $\hat{p}$. The custom in the interrater agreement literature is to accept the current observed empirical distribution $\hat{p}$ as identical to p, so this paper will make no distinction and use p throughout.

Scott (1955) and Cohen (1960) adjusted the base value of P_o using the odds of agreement given random categorizations. Under Cohen's setup, the expected rate of agreement, P_e, assumes that each rater has a distinct discrete distribution of categories produced by tabulating the rate at which the first rater used each of the C categories, and similarly for the second rater. Then the odds of chance agreement between the two raters is

$$P_e \equiv \sum_{i=1}^{C} p_{i\cdot} p_{\cdot i}.$$

The chance-adjusted rate of agreement, Cohen's κ, is defined as

$$\kappa \equiv \frac{P_o - P_e}{1 - P_e}.$$

Applying this to the example here, $P_e = 0.375$ and $\kappa \approx 0.467$.

Scott's π assumes that both raters are drawing from the same discrete distribution, and the best estimator of the bins in that distribution is the sum of selections by both raters, $\tilde{p}_i \equiv (p_{i\cdot} + p_{\cdot i})/2$. Then, having generated one distribution for both raters,

$$P_e^{\text{Scott}} \equiv \sum_i \tilde{p}_i^2.$$

Replacing P_e in the definition for κ with P_e^{Scott} gives the definition of π.

With two categories that were used equally, $P_e = P_e^{\text{Scott}} = 0.5$; for 100 categories, each used with equal likelihood, $P_e = P_e^{\text{Scott}} = 0.01$, demonstrating that the adjustment typically does give a better rating to a value of P_o when there are many options than the same P_o when there are few.

Cohen (1960, p. 40) describes κ as "simply the proportion of chance-expected disagreements which do not occur." The literature typically refers to κ as a *chance-adjusted* or *chance-corrected* measure. However, the measure of chance-expected disagreement is not based on an explicit and plausible model of how chance errors occur.

Both Scott's and Cohen's chance-adjustments model rating at chance by imagining two random number generators spitting out two sequences of categories, without taking inputs of any sort (including the data to be categorized). That is, the model underlying P_e consists of no more than comparing independent draws from discrete distributions.

There exist occasions where it is worth considering the possibility of rating at random; for example, Kravitz et al. (2010) failed to reject the null hypothesis that pairs of peer reviews for a journal agreed on accept/reject/revise categories at rates no greater than

random. However, to test the hypothesis that $P_o = P_e$, one can use P_o without adjustment, rather than testing that $\pi = 0$ or $\kappa = 0$.

In the introductory example, the raters have no problem agreeing on type 1, but have difficulty with types 2 and 3, so the model of errors entirely at random seems implausible, and there may be a more accurate model that reflects raters who can accurately spot type 1 but commit errors with types 2 and 3.

Other situations afford other error processes that each require a different model: one rater might consider values under 0.5 to be class 1 and values over 0.5 to be class 2 while the other rater sets a cutoff at 0.6, or the first rater may tend to misclassify category 3 as category 4 while the other rater tends to misclassify 3s as 5s, or raters may overuse the *some other type* category but make no further errors. For any of these situations, using P_e or P_e^{Scott} as a chance-adjustment is using an already rejected model of the error process.

Uebersax (1992) discusses the lack of a plausible underlying error model at length; on the web, Uebersax (2010) offers a bibliography of a dozen and a half articles discussing issues and problems with κ, primarily regarding its underlying error model, its misinterpretation as a measure of true agreement, and its unexpected sensitivity to features like the number of categories. Cook (2005) also offers some critique of the statistic.

Even when assuming that errors occur entirely at random, subtle issues arise in determining the rate of chance-expected disagreement. We already saw an example of this in the difference between Scott's and Cohen's versions of P_e. Andrés and Marzo (2004) use an error model where raters are generally accurate but sometimes make errors entirely at random. Fleiss (1971) and Craig (1981) offer multi-rater generalizations of Scott's error models. When dealing with continuous variables, Barry and Mielke (1988) – offer an alternative error model. Krippendorff (2004) generalizes many of the above statistics by allowing an arbitrary metric for discrete, ordinal, continuous, data and another method of calculating the observed and expected rate of disagreement given more than two raters. The intuitive concept of a chance-adjustment may make sense, but there is no one-size-fits-all way of formalizing what rating at chance means.

If the calculation of P_e changes across studies, there is no meaningful way to compare cross-study κ statistics outside of using the statistics a rough heuristic. Conversely, the ratio of information in agreement to total individual information is calculable for any pair of distributions, and is therefore more easily compared across studies.

An information-theoretic measure. Information theory is built upon an accounting of information. The measure of the information held by distribution D is its *entropy*, and is notated $H(D)$. As will be demonstrated by some simple examples below, it is not unreasonable to take the entropy as a formal measure of the intuitive idea of the complexity of a rating task. For example, the entropy rises as the number of equiprobable bins increases. Common information accounting identities break down total entropy into several elements, one of which is the mutual information shared jointly.

By analogy to P_o, which is the count of elements in agreement divided by the total count of elements, the new index proposed here, P_I, is the count of information in agreement divided by the total information.

For a given distribution of ratings, entropy is well-defined, so the information-based index sidesteps the issue of determining the error generation model underlying a series of ratings. In the example above, the entropy for the top row is 1.5, and is identical for the

bottom row. The information shared in agreement, defined below, is 0.569, so the shared information as a ratio of average individual information is 0.379.

Now consider a second example, where the top row of ratings is identical to those before, but the bottom row of ratings has changed:

Table 2. Classifications as in Table 1, but the coders are not in full agreement regarding category 1.

1	1	1	1	1	1	2	2	2	3	3	3
1	1	1	1	2	3	1	2	2	1	3	3

Overall, the agreement rate is no better than before: P_o is still 66.67%. But the raters are not in perfect agreement with category 1 anymore, and instead get two-thirds agreement for each of the three categories. That is, the rate of agreement across categories is more consistent in this example than in the first.

Having observed this data pattern, the careful researcher might wish to throw out the error process for the first example (where type one was never rated in error) and develop a new model of the error process. In this case, it may be difficult to compare the two examples above, because both the ratings and the error generation model changed. Comparing π and κ across examples would have to be done under the understanding that these statistics are simply convenient heuristics which may have no serious real-world interpretation. Conversely, the complexity-adjusted measure comparing information in agreement to total agreement can be compared across configurations as easily as P_o could be.

Or, the researcher may believe that agreement at random is the correct model in both cases (so the run of ones in agreement from the first example was just good luck). In that case, it would be valid to use π or κ for both cases. Because neither P_o nor P_e changed from the first example to the second, both π and κ would give the same score to both of these examples. That is, these statistics are insensitive to the consistency of the raters (or lack thereof) across categories.

The counts for top and bottom rater did not change, so their individual entropies are identical. But as will be discussed extensively below, the more even distribution of shared information bears more information: it is 0.61 as opposed to the first example's value of 0.569. Colloquially, we would say that the second more consistent set of ratings is more informative—and it frequently is. For the typical study, all categories have roughly equal importance, and if they do not, we will see that it is easy to impose a weighting scheme so that weighted cases have equal importance, so one would not want an especially bad error rate on all but one or two categories.

The statistic presented here acknowledges that there are reasons to account for the complexity of a situation—it really is easier to put data into two bins than a hundred—but sidesteps the question of how errors are generated by measuring the complexity of the rating task directly. That is, the measure here replaces the *chance-adjusted index* with a *complexity-adjusted index*. It thus replaces the problem of modeling the error generation process, where the correct model must be chosen on a case by case basis and the easiest models are implausible for most cases, with the somewhat settled question of how we measure the complexity of a set of distributions.

Even if the theoretical basis of κ makes it inappropriate for many of the situations in which it is used, its prevalence in the literature indicates that it is a good heuristic that captures researchers' intuitions about a situation-adjusted agreement rate. Ideally, we

would have an index which has a more broadly applicable theoretical basis, but which captures similar characteristics of the data. The P_I statistic will prove to demonstrate this empirical similarity to κ.

2. Entropy and its Decomposition

Our index of intercoder agreement should be adjusted to fit the complexity of the situation, so it is natural to base the index on entropy, a common measure of a configuration's complexity. The entropy is equivalent to the minimum average number of binary digits (i.e., bits) that would be needed to transmit a message given a sufficiently clever encoding. Similarly, if two sets of categorizations share a great deal of mutual information (as defined below) and we have one coder's stream of ratings on hand, we would need fewer bits of information to determine the second stream. Readers unfamiliar with measuring information may wish to refer to a textbook exposition such as Pierce (1980) or MacKay (2003).

Formally, let there be a set of C bins, each with probability p_i for i in 1 to C—that is, a discrete distribution D. Then the entropy is defined as:

$$H(D) \equiv -\sum_{i=1}^{C} p_i \log_2(p_i). \tag{1}$$

Because $p_i \leq 1$ for all i, $\log_2(p_i)$ will be nonpositive, and the above expression will be nonnegative. For a single bin with probability one, $\log_2(1) = 0$, and $H(D) = 0$: because all observations will be in the same bin, seeing a new observation imparts no new information. For two equiprobable bins, $H(D) = 2(-\frac{1}{2}\log_2(\frac{1}{2})) = 1$, meaning that a single observation imparts one bit of new information. For ten equiprobable bins, $H(D) = 10\,(-0.1\log_2(0.1)) \approx 3.32$. Thus, as more options arise, entropy generally rises.

The limit of $x \log_2(x)$ as $x \downarrow 0$ is zero, so for the purposes of the entropy calculation, it is custom to let $0 \cdot \log_2(0) \equiv 0$ (Gallager 1968, footnote p. 18). This means that if bin i in a discrete distribution has $p_i = 0$, then it is irrelevant for the entropy calculation.

The entropy definition is applied to a conditional distribution of the first rater's choices (X) given the distribution of the second rater's choices (Y) as follows:

$$H(X|Y) = -\sum_{i=1}^{C}\sum_{j=1}^{C} p_{ij}\log_2\left(\frac{p_{ij}}{p_{\cdot j}}\right).$$

The *mutual information* is defined as a sum over the categories similar to the entropy, and can be written as the difference of unconditional and conditional entropies:

$$\begin{aligned} I(X,Y) &\equiv \sum_{i=1}^{C}\sum_{j=1}^{C} p_{ij}\log_2\left(\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}\right), \\ &= \sum_{i=1}^{C}\sum_{j=1}^{C}\left[p_{ij}\log_2\left(\frac{p_{ij}}{p_{\cdot j}}\right) - p_{ij}\log_2(p_{i\cdot})\right] \\ &= -H(X|Y) + H(X), \end{aligned}$$

so

$$H(X) = I(X,Y) + H(X|Y).$$

That is, the entropy of one configuration is the sum of the information shared with the other configuration and information that is residual given the other configuration. Symmetrically, $H(Y) = I(X, Y) + H(Y|X)$.

It is useful to break down the mutual information further. Define *weighted mutual information* as

$$I_w(X, Y) \equiv \sum_{i=1}^{C} \sum_{j=1}^{C} w_{ij} p_{ij} \log_2 \left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} \right), \tag{2}$$

where w_{ij} is a weighting between zero and one for the case where the first rater chose category i and the second chose category j. Because we are developing an index of matching, it makes sense to define

$$w_{ij} \equiv \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}.$$

Let the weighted mutual information using this weighting be $IA(X,Y)$—the information in agreement. It also makes sense to define the complement, information in disagreement, $ID(X, Y)$, using

$$w_{ij} \equiv \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases},$$

though it may be easier to define the information in disagreement as the residual mutual information after information in agreement is subtracted:

$$ID(X, Y) = I(X, Y) - IA(X, Y). \tag{3}$$

Then

$$H(X) = IA(X, Y) + ID(X, Y) + H(X|Y)$$

and

$$H(Y) = IA(X, Y) + ID(X, Y) + H(Y|X) \tag{4}$$

so

$$H(X) + H(Y) = 2IA(X, Y) + 2ID(X, Y) + H(X|Y) + H(Y|X)$$

One could informally rewrite the last equation as

$$(H(X) + H(Y))/2 = IA(X, Y) + \text{everything else}.$$

Our interest is in how much of the quantity on the left-hand side is accounted for by IA and how much is accounted for by everything else. Thus, the remainder of the article shall consider this statistics as a candidate for measuring intercoder agreement:

$$P_I(X, Y) \equiv \frac{IA(X, Y)}{(H(X) + H(Y))/2}.$$

A worked example. Consider this table of frequencies, in which two raters placed elements into categories 1, 2, and 3. The main table lists joint frequencies, which total to the individual rater frequencies to the right of and below the lines.

Table 3. A table of joint and marginal rating possibilities.

	1	2	3	
1	**0.2**	0	0	0.2
2	0.05	**0.06**	0.19	0.3
3	0.15	0.14	**0.21**	0.5
	0.4	0.2	0.4	1

The diagonal elements are in bold to indicate that they are the only elements of the joint distribution used for the calculation. The first term of IA is $0.2 \log_2(0.2/(0.2 \cdot 0.4)) \approx 0.264$. The second is $0.06 \log_2(0.06/(0.3 \cdot 0.2)) = 0.06 \log_2(1) = 0$; in this case 0.06 is the product of the individual frequencies $0.3 \cdot 0.2$, exactly what one would expect given independent raters. The third term is $0.21 \log_2(0.21/(0.5 \cdot 0.4)) \approx 0.015$, for a total $IA \approx 0.279$. The entropies are H (row) ≈ 1.485 and H (column) ≈ 1.52, so $P_I \approx 0.185$.

In this case, the measures give similar results: with $P_o = 0.2 + 0.06 + 0.21 = 0.47$ and $P_e = 0.2 \cdot 0.4 + 0.3 \cdot 0.2 + 0.5 \cdot 0.4 = 0.34$, we have $\kappa \approx 0.197$. For π, we first write down the mean marginal distribution (0.3, 0.25, 0.45), which gives $P_e^{\text{Scott}} = 0.355$ and $\pi \approx .178$.

3. Characteristics of P_I

This section will explore the characteristics of P_I, and where appropriate compare it to κ. Several artificial small-N examples will be considered. This section will also consider some variants, such as modifications for a weighted agreement measure, surveys with more than two raters, and follow-up questions.

Full agreement and full disagreement. Figure 1 gives an example of two raters in full agreement. At the top of the figure are two sequences of categorizations from the two raters, and in this example they match in all cases. At the lower left, the main part of the table is the joint probability distribution over the categories produced by the pair of raters. The distributions for the individual raters are summed to the right of and below the lines. In this case, the joint distribution has nonzero values only along the diagonal. The diagonal

1 1 1 1 2 2 2 2 3 3
1 1 1 1 2 2 2 2 3 3

	1	2	3	
1	**0.40**	–	–	0.40
2	–	**0.40**	–	0.40
3	–	–	**0.20**	0.20
	0.40	0.40	0.20	1

Agreement (P_o)	**100%**	
P_e	0.360	
κ	**1.000**	(0.000)
First rater entropy	1.522	
Second rater entropy	1.522	
Information in agreement	1.522	
P_I	**1.000**	(0.000)

Fig. 1. The case of full agreement.

elements and marginal totals are all that is necessary to compute the various statistics generated by the pair of configurations. Standard errors for κ and P_I were calculated using bootstrap, and are given in parentheses (Klar et al. 2002). For the case of full agreement here, there is zero variation in the indices.

As in Figure 1, if two raters agree on all observations, and $P_e < 1$, then $\kappa = P_I = 1$. There is effectively no complexity adjustment in this case: full agreement given two options and full agreement given a thousand options both get an index of $\kappa = P_I = 1$.

Two ways to be zero. There are two manners in which the list of frequencies for joint agreement fail to provide any additional information.

Figure 2 shows the first case, where both raters entirely disagree. At random, there would be about 11% agreement, so zero percent agreement is actually worse than random, and κ is negative. The index $P_I = 0$ tells us that having the diagonal of the joint distribution does not improve our knowledge of the individual raters' information at all, which is indeed the case if the diagonal is filled with zero entries. One can show that if two raters disagree on all observations, and $0 < P_e < 1$, then $\kappa < 0$ and $P_I = 0$.

The second way in which P_I can be zero is if the odds of agreement match the rate at which independent raters would agree, $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$. If this is the case for p_{ii}, for all i, then $P_o = P_e$, and so $\kappa = 0$.

The same holds for P_I when the agreement rate is equivalent to the independent rate. In the case of independent raters, information about one rater gives us no information about the other: $H(X) = H(X|Y)$, so $I(X,Y)$ is zero.

Consider the example of Figure 3, where one of the raters chose the same value every time. This is an extreme example of independent raters. There is agreement when the non-constant rater picks the constant rater's favorite category, but the agreement measures register the agreement as pure chance.

Is P_I positive? Zero is not the minimum for either statistic. Cohen (1960 p. 42) comments that "if [κ] is less than zero . . . , it is likely to be of no further practical interest." In cases of generally competent raters doing a reasonable rating task, we can expect both κ and P_I to be positive; in other cases, the instrument or methods are likely too flawed to be useful. But it is worth discussing why and how both indices are negative, even if researchers may never encounter such situations in real-world data.

	1	2	1	2	1	2	3	1	3	2
	2	1	3	1	2	3	2	2	1	3

	1	2	3	
1	–	0.30	0.10	0.40
2	0.20	–	0.20	0.40
3	0.10	0.10	–	0.20
	0.30	0.40	0.30	1

Agreement (P_o)	**0%**	
P_e	0.340	
κ	**–0.515**	(0.003)
First rater entropy	1.522	
Second rater entropy	1.571	
Information in agreement	0.000	
P_I	**0.000**	(0.000)

Fig. 2. An example of full disagreement.

1	1	1	1	1	1
1	1	2	2	2	3

	1	2	3	
1	**0.33**	0.50	0.17	1.00
	0.33	0.50	0.17	1

Agreement (P_o)	**33%**	
P_e	0.333	
κ	**0.000**	(0.000)
First rater entropy	0.000	
Second rater entropy	1.459	
Information in agreement	0.000	
P_I	**0.000**	(0.000)

Fig. 3. One rater chose the same ranking in all cases. This is a special case of independent raters.

Define

$$R_i \equiv \frac{p_{ii}}{p_{i\cdot}p_{\cdot i}}.$$

If $R_i = 1$, then the joint agreement frequency is equivalent to the case of independence, where $P_{ii} = p_{i\cdot}p_{\cdot i}$. If $R_i > 1$, then joint agreement occurs with greater frequency than would occur under independence, and if $R_i < 1$, then joint agreement occurs less often than it would given independence.

The total P_I is the sum of one element for each category, and researchers may get value from individual elements of the sum. As above, we can determine whether the term is positive, negative, or zero by comparing p_{ii} to $p_{i\cdot} \cdot p_{\cdot i}$, but one could do a full per-term analysis comparable to the per-term breakdown of κ by Fleiss (1971) or James (1983). If $R_i > 1$, then bin i adds a positive term to the numerator of P_I; if $R_i < 1$, bin i adds a negative term; if $R_i = 1$ then it adds nothing.

This is akin to κ, whose numerator is $P_o - P_e = \sum_i (p_{ii} - p_{i\cdot}p_{\cdot i})$, and so each bin changes the numerator depending on whether p_{ii} is greater than, less than, or equal to $p_{i\cdot}p_{\cdot i}$.

Figure 4 shows a case analogous to negative correlation. In this case, if we had the full contingency table, and saw that the first rater had chosen a, then we would guess that the second rater had chosen b. But if we were given only the diagonal entries— $(a, a) = 0.1$ and $(b, b) = 0.1$—then we would place even odds on the second rater choosing a or b. In a sense, the partial table gives the misleading impression that there is no correlation between elements when the correlation is actually negative. In this case, $P_I \approx -0.264$.

1	1	1	1	1	2	2	2	2	2
1	2	2	2	2	1	1	1	1	2

	1	2	
1	**0.10**	0.40	0.50
2	0.40	**0.10**	0.50
	0.50	0.50	1

Agreement (P_o)	**20%**	
P_e	0.500	
κ	**–0.600**	(0.008)
First rater entropy	1.000	
Second rater entropy	1.000	
Information in agreement	–0.264	
P_I	**–0.264**	(0.003)

Fig. 4. Analogue to negative correlation.

1 1 1 2 1 2 2 2 3 3
1 1 1 1 2 2 2 2 3 3

	2	2	3	
1	**0.30**	0.10	–	0.40
2	0.10	**0.30**	–	0.40
3	–	–	**0.20**	0.20
	0.40	0.40	0.20	1

Agreement (P_o)	**80%**	
P_e	0.360	
κ	**0.688**	(0.007)
First rater entropy	1.522	
Second rater entropy	1.522	
Information in agreement	1.009	
P_I	**0.663**	(0.006)

Fig. 5. More matches on the rare case.

More informative agreement. In Figure 6, there are only a few cases of *C*, but the raters agreed on which they were. In Figure 5, the raters could not agree on which items were the rare type *C*. By the measure of count of *A*s, *B*s, and *C*s, and the count of agreements, the cases are identical. That means that κ is identical in both cases. Note the diagonal elements: in the first case they are (0.4, 0.3, 0.1) and in the second (0.3, 0.3, 0.2), and the more even pattern of agreement rates leads to a larger value of P_I than the more spread-out diagonal. One can show that, all else equal, shifting weight from a category where R_i is larger than the mean R across categories to a category where R_i is smaller than the mean R will raise P_I. As in the examples above, such a shift can be done without affecting P_o or P_e, meaning that κ is invariant.

3.1. A Few Variants

Many of the variants of the interrater agreement problem, such as situations where raters could be approximately correct, or where some questions like follow-ups are strongly related to prior questions, or where there are several raters, have been explored in the literature on κ, and could be applied to P_I analogously. In all cases, the problem of determining what P_e means in the given variant is no longer relevant. It is still sensible to compare the ratio of weighted information in agreement to individual information across studies, but with the caveat that the weighting scheme should be meaningful across studies.

1 1 1 1 2 2 2 3 2 3
1 1 1 1 2 2 2 2 3 3

	1	2	3	
1	**0.40**	–	–	0.40
2	–	**0.30**	0.10	0.40
3	–	0.10	**0.10**	0.20
	0.40	0.40	0.20	1

Agreement (P_o)	**80%**	
P_e	0.360	
κ	**0.688**	(0.007)
First rater entropy	1.522	
Second rater entropy	1.522	
Information in agreement	0.933	
P_I	**0.613**	(0.006)

Fig. 6. More matches on the common case.

Ordinal scales. For ordinal scales, researchers may want to give partial credit for ratings that are close to each other but not a match. The weighted κ puts some weight on near agreement; the same partial-credit principle can be applied to $I\,A$. Instead of the simple binary weighting used throughout this paper, one could define $I\,A_{\text{ord}}$ using a weighting such as

$$w_{ij} \equiv \begin{cases} 1 & i = j \\ 1/2 & 0 < |i-j| \leq 1 \\ 1/4 & 1 < |i-j| \leq 2 \\ 0 & 2 < |i-j| \end{cases}.$$

The choice of weighting scheme is subjective and should encode subject-specific knowledge of how much value a near miss provides. In Van der Wulp and Van Stel (2009, table 3), there is a good example of an asymmetric 5 × 5 table of weights comparing different levels of triage as given by a rater during an emergency and by a post-emergency gold standard. Their table of weights was intended for calculation of a weighted κ, but could be used for a weighted P_I without modification.

So long as we define $I\,D$ using Equation 3, the accounting identity of Equation Set 4 continues to hold for $I\,A_{\text{ord}}$, and the basic facts about P_I below—it is one at full agreement, is zero when agreement is at chance or there is no agreement, and so on—will have natural extensions to the ordinally-oriented version, $I\,A_{\text{ord}}/(\frac{1}{2}(H(X)+H(Y)))$.

New alphabets. Given the sequence of letters *i*, *a*, *m*, *a*, *d*, *o*, *g*, there is benefit to reinterpreting it as a sequence of words: *i*, *am*, *a*, *dog*. The same may hold with a sequence of questions. Question #2 may be a follow-up to Question #1, so valid response sequences might include:

Table 4. Valid options for Question 2 are contingent on the response to Question 1.

	Q1	**Q2**
1.	1A	2A
2.	1A	2B
3.	1B	2A
4.	1B	2C
5.	1C	[valid skip]
6.	1D	[valid skip]

Thus, if the first rater chose 1A for the first question and the second chose 1C, then the first rater will have a rating for question 2 (either 2A or 2B) while the second rater will have a blank for question 2. One can recode the pair of responses into word-like sets, such as the six rows of the table.

The method of aggregating questions into sets is especially useful in conjunction with a weighting table, because some aggregate questions are similar to others, such as how #3 and #4 in the table above both have Q1 in common but disagree on Q2, so it may be reasonable to assign a weight of 0.5 to a pair of ratings (3,4) or (4,3). The full matrix of weights might look like:

Table 5. A weighting matrix appropriate for the questions in Table 4

	1	**2**	**3**	**4**	**5**	**6**
1	1	0.5	0	0	0	0
2	0.5	1	0	0	0	0
3	0	0	1	0.5	0	0
4	0	0	0.5	1	0	0
5	0	0	0	0	1	0
6	0	0	0	0	0	1

Such aggregations from letters to words are natural in information theory, and thus natural for the information-based agreement measures discussed in this article. The method could readily be applied to traditional agreement measures such as κ as well, using a reasonable method of constructing P_e.

Many raters. Fleiss (1971) accommodates several raters by taking the mean of the agreement rate across raters. Not surprisingly, there is some awkwardness about resolving what P_e means in the new context. Although typically described as a generalization of Cohen's κ, Fleiss's version of P_e is the natural multi-rater generalization of P_e^{scott}, and in the case of two raters reduces to Scott's π.

The concept of aggregating several pairwise components applies naturally to P_I. If we have three raters, X, Y, and Z, then we now have three sets of individual rating distributions, $p_i^X, p_i^Y,$ and $p_i^Z,$ $i \in 1, \ldots, C$, and three sets of pairwise ratings, with diagonal terms $p_{ii}^{XY}, p_{ii}^{XZ},$ and $p_{ii}^{YZ},$ $i \in 1, \ldots, C$. Writing P_I for raters X and Y as $2IA(X,Y)/(H(X)+H(Y))$, it is natural to add together the additional information in agreement terms and the individual entropy terms to get a ratio of total information in agreement to total information in individual ratings:

$$P_I(X,Y,Z) = \frac{2IA(X,Y) + 2IA(X,Z) + 2IA(Y,Z)}{(H(X)+H(Y)) + (H(X)+H(Z)) + (H(Y)+H(Z))}.$$

This retains the spirit of the two-rater measure, as one could write the sum

$$(H(X)+H(Y)) + (H(X)+H(Z)) + (H(Y)+H(Z))$$
$$= 2IA(X,Y) + 2IA(X,Z) + 2IA(Y,Z) + \text{everything else}.$$

As in the two-dimensional case, full agreement means that the 'everything else' portion is zero and all weight is in the $I\,A(\cdot,\cdot)$ terms; full disagreement means that the $I\,A(\cdot,\cdot)$ terms are zero.

4. Simulation and Sensitivity

I ran a series of simulations using two raters to gauge the power of κ and P_I to detect small differences in rater efficacy. To simplify the simulation, let the first rater be the 'gold standard' rater, who draws from ten bins with equal probabilities. The second is a rater who agrees with the first rater with probability p^c.

The goal of the simulation was to test the power of the statistics to distinguish between one set of data based on raters agreeing at a rate p_c and a second set based on raters agreeing at rate $p_c - \epsilon$.

The simulation ran this procedure for 1,000 tests:

1. Generate first set, where both raters classify 500 items into ten bins. They agree with probability p^c.
2. Generate second set, where both raters classify 500 items into ten bins. They agree with probability $p^c - \epsilon$.
3. Calculate κ and standard error of κ (via bootstrap) for both runs.
4. Calculate P_I and standard error of P_I (via bootstrap) for both runs.
5. Run a standard t-test testing the null hypothesis $H_0 : (\kappa \text{ given } p_c) > (\kappa \text{ given } p_c - \epsilon)$. Mark whether the test rejected the null or failed to reject at the 95% confidence level.
6. Run a standard t-test testing the null hypothesis $H_0 : P_I \text{ given } p_c > P_I \text{ given } p_c - \epsilon$. Mark the result, as with the test on κ.

Representative results are printed in Table 6, including the proportion of runs where κ or P_I successfully rejected the null hypothesis and distinguished between the two samples at the 95% confidence level. The two statistics are remarkably close, and demonstrate power within 1% of each other in every case.

Table 6. The power of κ and P_I to resolve differences in p_c given different base values of p_c and difference ϵ

p_c	ϵ	κ power	P_I power
0.3	0.01	0.541	0.544
0.5	0.01	0.540	0.531
0.7	0.01	0.557	0.561
0.9	0.01	0.650	0.641
0.95	0.01	0.673	0.673
0.3	0.02	0.678	0.676
0.5	0.02	0.665	0.665
0.7	0.02	0.700	0.707
0.9	0.02	0.803	0.800
0.95	0.02	0.857	0.857
0.3	0.05	0.935	0.934
0.5	0.05	0.918	0.918
0.7	0.05	0.943	0.943
0.9	0.05	0.989	0.989
0.95	0.05	0.998	0.998

The examples above showed that as consistency of rating across categories improved, P_I changed while κ remained constant. Thus, if one needed a test of the consistency of raters across categories, there is no need to run simulations: κ is by construction insensitive to changes in consistency and is therefore unusable in such a situation. However, consistency and agreement rate are effectively unrelated, so the ability of P_I to discriminate between

more and less consistent patterns says nothing about its ability to better discriminate between situations with different agreement rates.

5. Empirical Results

The Census Bureau's Center on Survey Measurement ran a series of tests on the methods underlying the American Community Survey, including tests of new questions about internet usage and English/Spanish comparisons. The interviews were recorded, and several raters classified each interviewer's utterance in categories such as *asked as worded*, *major change*, *correctly coded response*, or *inaudible*, and classified respondent utterances in categories such as *directly answered*, *asked for clarification*, or *refused to answer*. It is these classifications that will be the data set under analysis below.

The test survey itself was conducted in mid-2011, and the questions include questions about individuals in the household, such as military service and parental place of birth/ancestry. Because the responses to the questions themselves are irrelevant to the issue of intercoder agreement, and the responses are protected by Title 13 of the U.S. Code, I do not detail the actual questions or responses.

Within the several categorization tasks associated with an interview, there were four 'check only one' parts, regarding

- the extent to which the interviewer's initial asking of the question adhered to the question script (six options, including exact reading, major change, skipped),
- the respondent's subsequent response (nine options, including codeable answer, qualified/uncertain answer, don't know, refused)
- the classification of the final answer (eight options, including codeable answer, qualified/uncertain answer, don't know, refused), and
- the evaluation of the interviewer's data entry (three options: matches respondent answer, does not match, other/unclear if matches).

I omitted rating tasks that allowed multiple responses or did not have a pair of raters with more than 100 ratings in common. The statistics were calculated where possible for each pair of raters and each of the above categories, producing 336 sets of statistics, as shown in Figure 7 and graphically in Figure 8.

For each subcategory, there were between 28 and 140 pairs of raters. For a given pair of raters, the average number of pairs of ratings ranges from 240 to almost 2,000. Different pairs of raters rated different combinations of the four subcategories of *all nonlanguage*, so the count in subcategories will not sum to 140. All pairs who classed nonlanguage

Section	Pairs	Average count	κ (std err)	P_I (std err)
All nonlanguage	140	396	0.58 (0.11)	0.52 (0.08)
Interviewer query	28	511	0.69 (0.05)	0.60 (0.07)
Response	56	512	0.57 (0.10)	0.51 (0.07)
Data entry	28	506	0.54 (0.10)	0.49 (0.06)
Final outcome	56	495	0.54 (0.10)	0.49 (0.06)
Language	140	345	0.93 (0.04)	0.91 (0.05)

Fig. 7. For each section, the count of rater pairs with more than 100 ratings, the average number of ratings for each pairing, and the mean and standard error of κ, and P_I.

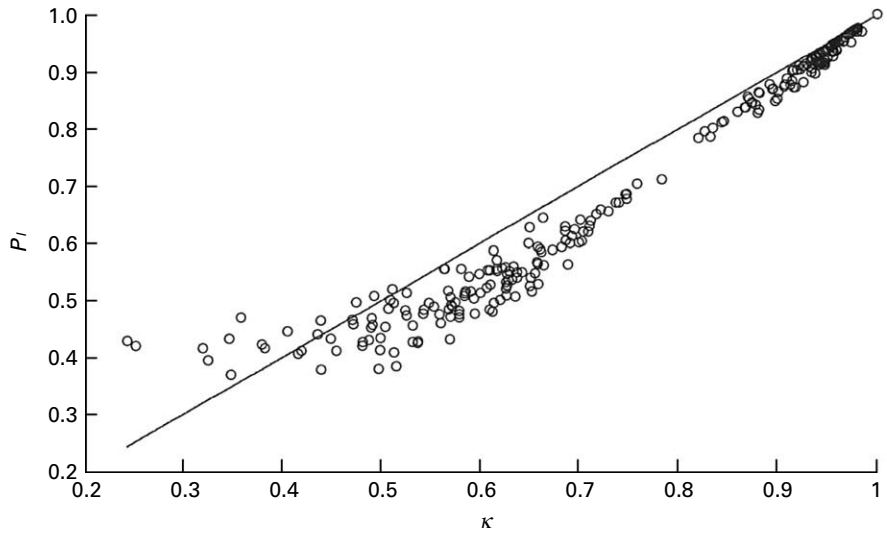

Fig. 8. The values of κ and P_I for the data coding described in Section 5.

categories also classed the language. The numeric value of P_I tends to be slightly smaller than κ. But for samples of this size, with competent raters, the two statistics tend to track each other closely. In this data set, the standard errors for P_I are smaller for most nonlanguage questions.

For each pair of raters and each segment, there is a single point in the plot of Figure 8 with coordinates (κ, P_I). The $\kappa = P_I$ line is printed for reference, and we see that, for the bulk of the cases, the numeric value of P_I follows that of κ closely, although it is typically about 0.05 smaller.

6. Conclusion

This article proposes replacing the chance-adjusted rate of agreement with a complexity-adjusted rate of agreement.

The main problem with chance-adjustment is that it must be based on a solid model by which errors are generated, yet both Scott's π and Cohen's κ are based on a completely at random decision model which is implausible for most real-world situations. To correct for this, the literature offers variants of κ for rating entirely at random, rating given some fixed rate of accuracy, rating given ordered categories, rating given multiple raters whose discrete distributions need to be aggregated before κ is calculated, and so on. But such a diversity of forms means that researchers need to do up-front work to determine which member of the κ family is most appropriate for the given situation, and it is invalid to compare κ across studies. Conversely, the same P_I calculation retains its applicability for any pair of raters, because the entropy accounting does not rely on a specific error model. Just as it is meaningful to compare the count of ratings in agreement to the total count of ratings across a broad range of situations, it is meaningful and valid to compare the proportion of information in agreement to total individual information across a broad range of situations.

The change to a more sound theoretical basis improves the validity of cross-scenario comparisons, but does not throw out our intuition about what sort of rater behavior should be given higher or lower ratings. In the simulation and ACS example, we find a great deal of similarity in how both old and new indices gauge the adjusted level of agreement, so intuition about κ largely transfers to intuition about P_I. Thus, it is reasonable to use P_I as a plug-in replacement for κ, with the added benefit of not needing to develop an appropriate model for P_e for each given situation.

The one key distinction between P_I and κ is in measuring consistency across categories: were raters very good with one category but bad with all the others, or did raters do as well across all categories? The P_I statistic rates consistent ratings more highly than equivalent ratings with the same P_o and P_e but less consistency. This advocates leaning toward using P_I in situations where consistency is valuable, and κ in situations where the breakdown of agreement across categories is not relevant.

7. References

Andrés, A.M. and Marzo, P.F. (2004). Delta: A New Measure of Agreement Between Two Raters. British Journal of Mathematical and Statistical Psychology, 57, 1–19.

Barry, K.J. and Mielke, P.W., Jr (1988). A Generalization of Cohen's Kappa Agreement Measure to Interval Measurement and Multiple Raters. Educational and Psychological Measurement, 48, 921–933.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20, 37–46.

Cook, R.J. (2005). Kappa and Its Dependence on Marginal Rates. In Encyclopedia of Biostatistics, P. Armitage and T. Colton (eds). London: Wiley.

Craig, R.T. (1981). Generalization of Scott's Index of Intercoder Agreement. Public Opinion Quarterly, 45, 260–264.

Fleiss, J.L. (1971). Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, 76, 378–383.

Gallager, R.G. (1968). Information Theory and Reliable Communication. New York: John Wiley and Sons.

James, I.R. (1983). Analysis of Nonagreements Among Multiple Raters. Biometrics, 39, 651–657.

Klar, N., Lipsitz, S.R., Parzen, M., and Leong, T. (2002). An Exact Bootstrap Confidence Interval for κ in Small Samples. Journal of the Royal Statistical Society. Series D (The Statistician), 51, 467–478.

Klemens, B. (2008). Modeling with Data: Tools and Techniques for Statistical Computing. Princeton, NJ: Princeton University Press.

Kravitz, R.L., Franks, P., Feldman, M.D., Gerrity, M., Byrne, C., and Tierney, W.M. (2010). Editorial Peer Reviewers Recommendations at a General Medical Journal: Are They Reliable and Do Editors Care? PLoS ONE, 5(4), el0072.

Krippendorff, K. (2004). Content Analysis: An Introduction to its Methodology. Thousand Oaks, CA: Sage.

MacKay, D.J.C. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge: Cambridge University Press.

Pierce, J.R. (l980). An Introduction to Information Theory: Symbols, Signals, and Noise, (Second Revised Edition). New York: Dover Press.

Scott, W.A. (1955). Reliability of Content Analysis: The Case of Nominal Scale Coding. Public Opinion Quarterly, 19, 321–325.

Uebersax, J. (1992). Modeling Approaches for the Analysis of Observer Agreement. Investigative Radiology, 27, 738–743.

Uebersax, J. (2010). Kappa Coefficients. http://www.john-uebersax.com/stat/kappa.htm. [Online; accessed 25 July 2011].

Van der Wulp, I. and Van Stel, H.F. (2009). Adjusting Weighted Kappa for Severity of Mistriage Decreases Reported Reliability of Emergency Department Triage Systems: A Comparative Study. Journal of Clinical Epidemiology, 62, 1196–1200.

Received April 2012
Revised July 2012