Double Entropy Inter-Rater Agreement Indices

Applied Psychological Measurement

2016, Vol. 40(1) 37–55

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621615592718

apm.sagepub.com

Andriy Olenko¹ and Vitaliy Tsyganok²

Abstract

The proper application of the most frequently used inter-rater agreement indices can be problematic for the case of a single target, for example, a psychotherapy patient, a student's thesis, a grant proposal, and the lifestyle in a country. The majority of indices that can handle this case assess either the deviation of ranks from some central/average value or the pattern of ranks' distribution. Contrary to other approaches, this article defines disagreement rating results using the unpredictability/complexity of scores. The article discusses alternative entropy methods for measuring inter-rater agreement or consensus in survey responses for the case of a single target. A new inter-rater agreement index is proposed. Comparisons between this index and the known inter-rater agreement measures show some limitations of the most frequently used indices. Various important methodological issues such as disagreement assumptions, average sensitivity, adjustments to deal with outliers, and missing or incorrectly recorded data are discussed. Examples of applications to actual data are presented.

Keywords

psychological statistics, agreement index, similarity measure, average sensitivity

Introduction

Assessments of inter-rater agreement appear in numerous applications in laboratory and field studies, various types of data analysis problems, organizational and applied psychology, and so on. The publications (LeBreton & Senter, 2008; Lin, Hedayat, Sinha, & Yang, 2002; Shoukri, 2004; von Eye & Mun, 2005; Wagner, Rau, & Lindemann, 2010) give excellent surveys of the field. Some most recent results and bibliography can be found in Gwet (2012); LeBreton, James, and Lindell (2005); Lin, Hedayat, and Tangm (2013); Lohse-Bossenz, Kunina-Habenicht, and Kunter (2014); and Shah (2011).

Inter-rater agreement indices were used and studied extensively in a number of contexts. For example, the article by Smith-Crowe, Burke, Kouchaki, and Signal (2013) reports that almost half of the articles published in various applied psychology journals used inter-rater agreement indices. Recently, such indices have been frequently applied to justify aggregating lower level data in multi-level and hierarchical models. Some statistical characteristics of inter-rater

¹La Trobe University, Melbourne, Australia

²Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine

Corresponding Author:

Andriy Olenko, Department of Mathematics and Statistics, La Trobe University, Melbourne, Victoria 3086, Australia.

Email: a.olenko@latrobe.edu.au

agreement indices were obtained and discussed for specific scenarios (e.g., Brown & Hauenstein, 2005; Burke & Dunlap, 2002; Fleiss, 1971; James, Demaree, & Wolf, 1984; Klemens, 2012).

Despite recent progress and results for the case of multiple targets, there has been remarkably little study on agreement estimates which can be also applied to a single target (see Baca-García et al., 2001; Cicchetti, Fontana, & Showalter, 2009; Cicchetti, Showalter, & Rosenheck, 1997; Lindell & Brandt, 1997; Lindell, Brandt, & Whitney, 1999). The most of the known indices do not allow for the estimation of chance agreement on a single target. However, there has been a significant interest in single target problems in applications where a group of raters is formed only for a single evaluation.

First, various online ratings that are based on situational random responses require agreement estimates. For example, the Faculty of 1000 (<http://f1000.com>) evaluates the quality of scientific articles based on the opinion of scientific experts. Now the Faculty of 1000 has more than 5,000 evaluators spread across 44 subject-specific faculties. A few other examples are as follows: Big online movie databases—for instance, Rotten Tomatoes, IMDb, and Netflix—collect and provide information about user rankings of movies; Amazon's users rank books and music; and Google asks customers to rate businesses on a 30-point scale. The number of responses in these ratings varies from few dozens to hundreds of thousands.

The second example situation is when users select raters of interest to perform an evaluation. For instance, patients can be examined by independent psychiatric practitioners. Knowing about a degree of agreement might help in taking decisions in the most difficult or unprecedented clinical cases.

Inter-rater agreement indices can also be used to examine consensus in survey responses. They can measure whether several respondents, who evaluate the same target, produce similar scores. Especially interesting cases are multi-level and hierarchical models, where each final score is a result of various aggregations of responses, see Example 1 in "Illustrative Examples" section. In such situations, generating methods of final scores can have rather complex structure to investigate them directly.

Another potentially interesting area is when human experts are replaced by some automatic labeling processes, for example, data mining classifiers, which are selected from a large number of potential methods. The reader can find more details about classical and new applications of agreement indices in Shah (2011), Smith-Crowe et al. (2013), and von Eye and Mun (2005).

Agreement indices are also widely applied in constructed-response assessments, especially when different raters can use partially different criteria. In such cases, data about scores of constructed-response items may have a hierarchical structure (see DeCarlo, Kim, & Johnson, 2011). Another important essay grading application is measuring the increase of consensus using scores of raters before and after an initial training.

Contrary to other approaches, this article defines disagreement rating results for the case of a single target using the unpredictability/complexity of scores. The focus in this article is on a new entropy-based agreement index and its properties. The most of the known agreement indices either assess the deviation of scores from some central (average) value or the pattern of scores' distribution. To overcome this limitation, both the spread and the variability of distributions have been assessed. In terms of the underlying rationale for this approach, the unpredictability/uncertainty of scores has been measured by applying the Shannon entropy twice. First, the degree of uncertainty for the frequencies of selected scores has been accessed; then, the degree of uncertainty in their concentration has been computed. Each time the Shannon entropy is computed and then normalized to the range $[0, 1]$. The double entropy index will be introduced using weighted averages of these normalized entropies.

The new index is aimed to overcome some limitations of other well-known similarity/dissimilarity measures. For example,

- Many of classical measures are not normalized and, therefore, their results are not directly comparable.
- Indices similar to χ^2 use goodness-of-fit principles and depend only on the frequencies of selected scores but not on scores' actual values.
- Measures, that refer to a uniform distribution as the most dissimilar rating, often cannot distinguish between distributions with equidistant equal frequencies.
- Many similarity measures do not change gradually for discrete cases and can produce quite different values for similar rating results.

The article establishes properties of the new index and compares it with various known indices. Two modifications of the index designed to be insensitive to outliers and missing data are introduced and discussed. The article also gives unified representations of the known indices, which might be useful for other research problems. Potential practical applications are illustrated by examples. The study shows that the new indices perform overall well comparing with other well-known indices.

The rest of the article is organized as follows: The “Framework and Notation” section introduces the main notations and assumptions. A novel measure of inter-rater agreement is defined and discussed in the “Double Entropy Index” section. The “Comparison of Inter-Rater Agreement Measures” section gives unified representations of some well-known agreement/disagreement measures. Then it provides an insight into empirical behavior of the proposed index and its advantages over the related metrics. The “Adjusted Double Entropy Indices” section proposes two adjustments to deal with outliers and missing or incorrectly recorded data. Examples of applications to real data are presented in the “Illustrative Examples” section. Finally, conclusions and discussions are given in the “Directions for Future Research” section.

Framework and Notation

In this section, underlying assumptions, some basic notations, and a convenient way to represent rating results have been introduced.

All information about rating results can be presented by the four characteristics: the number of categories in the response scale, the number of raters, the selected score values, and the number of raters assigned each selected score. Below, the authors formalize this information for the following computations.

Let $n \geq 2$ be the number of possible scores (the number of categories into which assignments are made) and $I = \{i_1, \dots, i_k\}$ be a set of score values selected by raters. Without loss of generality, assume that $i_1 < i_2 < \dots < i_k$.

Let $m \geq 2$ be the total number of raters and r_i be the number of those who gave the score i . Denote $R = (r_1, \dots, r_n)$. If the score i was not given, then $r_i = 0$. Thus, $m = \sum_{i \in I} r_i$.

To visualize these notations, one can use the bar plot as shown in Figure 1. The horizontal axis of the plot shows the scores, and the vertical axis gives the frequencies r_i .

Agreement indices can be derived from various statistical and epistemological assumptions and yield differing results. The problem with the most of inter-rater agreement indices is that they either assess the deviation of the scores in I from some central value, for example, the mean or the median, or the deviation of the score distribution R from a specific pattern, for example, a uniform distribution. A new inter-rater agreement index will be defined by taking into account both the deviations.

Figure 1. Case of $n = 10$ and $R = (4, 5, 1, 1, 4, 0, 1, 2)$.

First, it is important to formally specify the most disagreement rating result. In this article, the approach which defines disagreement rating results has been adopted using the unpredictability of scores. Thus, it is impossible to conclude that some raters have more agreement about a target and prefer a particular score to the other scores. Note that some popular indices use a different approach. Namely, they define the most disagreement rating if a half of raters select the lowest score and the other half select the highest one. This article considers this scenario only as a partial disagreement because both the marginal scores are high, but diametrical (directly opposite preferences are usually highly correlated and can be treated as a partial agreement). In addition, in this situation, there is a high level of agreement inside of the each diametrical group. For example, all raters consider a target as un-ordinary but rank it very low or high because of their training or “like” or “dislike” preferences.

For given k , m , and n , the most disagreement rating result is when

- k elements of I are equidistantly located in the set $\{1, \dots, n\}$ with $i_1 = 1$ and $i_k = n$ (or the distances between consecutive selected scores vary at most by 1 if $k < n$ and equidistant placement is impossible, that is, $k - 1$ is not a divisor of $n - 1$);
- the values $\{r_{i_1}, \dots, r_{i_k}\}$ are the same (vary at most by 1 if k is not a divisor of m).

It means that the same number of raters selects each of k scores, and these k scores are swept in equidistant steps from 1 to n .

Figure 2 shows some examples of the most disagreement rating results for $n = 10$, $m = 20$, $k = 10$, 4, and 2, respectively.

Second, the situation of complete disagreement of m raters is referred to as an occurrence of the above two conditions and $k = \min(m, n)$. For example, for $n = 10$ and $m = 20$, it is shown in the first plot of Figure 2.

Note that the above definition describes the probability mass function of the uniform distribution on the set $\{1, \dots, n\}$ if n is a divisor of m . However, it differs from a uniform distribution for those cases where m raters cannot be equally split into n groups.

Double Entropy Index

In this section, the entropy measure to introduce a new index of inter-rater agreement/disagreement has been used. The motivation for index's definition and some marginal cases have been discussed. The sensitivity of the index has been estimated.

Figure 2. Examples of most disagreement rating results.

The new index describes the complexity of the scores. A natural candidate for that would be the entropy, a classical measure of complexity. However, it is not enough to use the entropy once as the problem involves two levels of complexity/uncertainty. The first one is pertinent to relative positions of the selected scores across the response scale. An additional level appears because some selected scores are more and some are less preferable by experts.

To assess agreement among multiple scores for a subject, use

1. the distribution of the selected scores $i_1, \dots, i_k$ in the set $\{1, \dots, n\}$;
2. the distribution of the frequencies, r_i , $i \in I$, of the selected scores.

For each of the above distributions, the Shannon entropy will be computed and then normalized to the range $[0, 1]$. The double entropy index will be defined as an average of the normalized entropies. In the “Adjusted Double Entropy Indices” section, the adjusted index will be introduced using weighted averages of these entropies.

Step 1: To describe the spread of the scores $\{i_1, \dots, i_k\}$, use the distances

$$d_j = i_{j+1} - i_j, \quad j = 1, \dots, k-1, \quad (1)$$

between consecutive selected scores. An additional distance between i_k and i_1 has been introduced by

$$d_k = \begin{cases} (n - i_k) + (i_1 - 1) + \left\lfloor \frac{n-1}{k-1} \right\rfloor, & \text{if } k > 1, \\ n, & \text{if } k = 1, \end{cases} \quad (2)$$

where $\lfloor x \rfloor$ is the largest integer less than or equal to x .

Note that all distances are positive integers, and their sum is the same for each k regardless of actual values of $\{i_1, \dots, i_k\}$. This sum equals

$$d = \begin{cases} n - 1 + \lceil \frac{n-1}{k-1} \rceil, & \text{if } k > 1, \\ n, & \text{if } k = 1. \end{cases} \quad (3)$$

Using the set of distances $\{d_1, \dots, d_k\}$, one can normalize them to sum to 1 and introduce the corresponding fractions $P = \{p_1, \dots, p_k\}$ by the formula $p_j = d_j/d$, $j = 1, \dots, k$.

To characterize disagreement among the k selected scores, use the entropy

$$H(P) = - \sum_{j=1}^k p_j \ln p_j. \quad (4)$$

Step 2: To describe the distribution of the frequencies r_i , $i \in I$, normalize them and introduce the corresponding fractions $Q = \{q_i, i \in I\}$ by the formula

$$q_i = \frac{r_i}{\sum_{i \in I} r_i}, \quad i \in I. \quad (5)$$

If the score i was not given by raters, then $r_i = 0$, and we obtain $q_i = 0$, $i \notin I$.

To characterize the degree of unpredictability for Q , use the entropy

$$H(Q) = - \sum_{i \in I} q_i \ln q_i. \quad (6)$$

Note that all $q_i > 0$ if $i \in I$, because the corresponding $r_i \geq 1$.

Step 3: To get a normalized index taking values between 0 and 1, first, normalize each of the entropies $H(P)$ and $H(Q)$ to the range $[0, 1]$:

$$H^*(P) = \begin{cases} \frac{H(P) - \min_P H(P)}{\max_P H(P) - \min_P H(P)} = \frac{H(P) - B_{n,k}}{A_{n,k} - B_{n,k}}, & \text{if } k < n, \\ 1, & \text{if } k = n, \end{cases} \quad (7)$$

$$H^*(Q) = \frac{H(Q)}{\max_{k,Q} H(Q)} = \frac{H(Q)}{C_{n,m}}, \quad (8)$$

where $\max_P H(P) = A_{n,k}$, $\min_P H(P) = B_{n,k}$, and $\max_{k,Q} H(Q) = C_{n,m}$. The details of computations of the values $A_{n,k}$, $B_{n,k}$, and $C_{n,m}$ and implementations in R code can be found in the Supplementary Materials.

Second, the normalized index of disagreement $\kappa_0(P, Q)$ as the average of the normalized entropies has been defined:

$$\kappa_0(P, Q) = \frac{H^*(P) + H^*(Q)}{2}. \quad (9)$$

The normalized index $\kappa_0(P, Q)$ takes values in the interval $[0, 1]$. There exist score distributions for which $\kappa_0(P, Q)$ reaches the both endpoints 0 and 1. Namely, if all raters give the same score, which indicates perfect agreement, then $p_1 = q_1 = 1$ and $\kappa_0(P, Q) = 0$; if each score is selected by $\lfloor m/n \rfloor$ or $\lfloor m/n \rfloor + 1$ raters, then $\kappa_0(P, Q) = 1$.

Finally, the corresponding index of inter-rater agreement $\kappa(P, Q)$ is defined by

$$\kappa(P, Q) = 1 - \kappa_0(P, Q). \quad (10)$$

It follows from Equations 9 and 10 that $\kappa(P, Q)$ increases when $H^*(P)$ or $H^*(Q)$ decreases. Both $H^*(P)$ and $H^*(Q)$ make equal contributions to $\kappa(P, Q)$.

Similar to r_{wg} and a_{wg} , see the discussion and references in Brown and Hauenstein (2005), there are no sampling distributions that completely specify agreement or disagreement and can be used to test hypotheses about statistical significance of κ . Simulation studies analogous to the ‘‘Comparison of Inter-Rater Agreement Measures’’ section suggest values in the range 0.6 to 0.7 as a reasonable cutoff for κ , while values from 0 to 0.6 should be considered as unacceptable levels of agreement. The article by Cicchetti et al. (2009) presents a more detailed discussion of the rationale underlying the use of 0.7 as the expected level of chance agreement. If κ is frequently used in the same circumstances, then the expected level of chance agreement can be adjusted using results of an initial trial period. A new level can be easily implemented in the developed R code.

A sensitivity of the index to changes in scores can be used to investigate its performance and reliability (it also characterizes its sensitivity to outliers). Namely, the following two-step procedure is used to compute an average sensitivity of the index:

- Calculate the maximum absolute deviation of κ for each admissible list of scores of m raters when one of scores has been changed.
- Calculate the mean of the maximum absolute deviations from Step 1 over all admissible scores of m raters.

Figure 3 demonstrates the average sensitivity of the index as a function of the number of raters m and the number of categories n . To obtain this figure, the index κ was computed for m Monte Carlo simulated scores and recalculated for all possible changes in one score. Then the maximum absolute differences of κ were averaged over all replications. Monte Carlo simulations with 10,000 replications for each group size $m = 2, \dots, 30$ and the numbers of categories 5, 10, and 15 were performed.

Figure 3 demonstrates that the average sensitivity decreases rather quickly as the number of raters m increases. On the average, for a large number of raters, the distribution of selected scores is closer to a uniform distribution, when there are fewer categories. Therefore, the right-most part of Figure 3 exhibits a slightly smaller average sensitivity if the number of categories is smaller.

Using these plots, one can compute an average level of sensitivity for a given number of raters, and vice versa. For example, based on the Monte Carlo simulation results, to get the average level of index’s sensitivity at most 0.05, the recommended number of raters is 15 or more. However, 15 raters might not be realistic, and very high sensitivity is not required in many applications. At the same time, for typical applications with 6 raters, Figure 3 indicates levels of sensitivity in the range 0.08 to 0.15 depending on a response scale. These levels are quite acceptable in practice.

Figure 3 demonstrates that more raters are needed to obtain a higher sensitiveness of the index. This is true not only for κ but also for most of inter-rater agreement indices. There are

Figure 3. Average sensitivity for 5-, 10-, and 15-level response scales.

various situations when a large number of raters is appropriate, for instance, in examining raters' consensus during a training period. DeCarlo et al. (2011) gave an example of a large language assessment, where 34 raters scored the first item and 20 raters scored the second item. More examples can be found in Kim (2009) and Xi and Mollaun (2009). However, in many performance assessments (e.g., essay exams), the large number of raters is not realistic. Thus, some adjustments of the index that are robust to outliers and missing data are required. They are introduced in the "Adjusted Double Entropy Indices" section.

Comparison of Inter-Rater Agreement Measures

In this section, the double entropy index has been compared with the most frequently used indices of inter-rater agreement. The most of these indices involve determining the extent to which either each rater's score differs from the mean or the median item rating, or the score distribution differs from the uniform distribution.

The more sound theoretical basis introduced in the "Double Entropy Index" section improves the validity of inter-rater comparisons. In this section, a set of simple examples has been given, backed by the authors' intuition, about what sorts of raters' results should correspond to higher or lower inter-rater agreements.

The authors start by unifying definitions of these indices using the notations in this article.

Sample Standard Deviation S

The sample standard deviation of ratings was proposed and studied as an agreement index by Schmidt and Hunter (1989). It was introduced to measure the dispersion from the mean rating. Its low values indicate that the ranks tend to be close to the mean.

In the authors' notations, it can be computed by the formula

$$S = \sqrt{\frac{\sum_{i \in I} r_i(i - M)^2}{m - 1}}, \quad (11)$$

where $M = \sum_{i \in I} i \cdot r_i / m$ is the mean rating. The standard error of the mean $SEM = S / \sqrt{m}$ was used to construct confidence intervals around M to assess significant levels of agreement among the raters.

Coefficient of Variation CV

The coefficient of variation is a normalized measure of dispersion obtained from the sample standard deviation S by $CV = S / M$.

Adjusted Average Deviation Index $AD_{M_{adj}}$

Burke, Finkelstein, and Dusig (1999) studied the average deviation around the mean or the median as an index of inter-rater agreement. It is based on the same principle as S but uses the absolute differences which usually produce more robust results.

Using the notations of this article, AD_{μ} for a single subject is computed as follows:

$$AD_{\mu} = \frac{\sum_{i \in I} r_i |i - \mu|}{m}, \quad (12)$$

where μ is the mean M or the median M_d rating.

Compared with the original AD_M index, the adjusted $AD_{M_{adj}}$ index introduced by Lohse-Bossenz et al. (2014) gives less biased estimates within smaller groups. It can be computed as follows:

$$AD_{M_{adj}} = \frac{2m - 1}{2m(m - 1)} \sum_{i \in I} r_i |i - M| = \frac{2m - 1}{2(m - 1)} AD_M. \quad (13)$$

As $AD_{M_{adj}}$ is not a normalized index, a basic rule for interpreting significant levels of agreement is whether the AD_{μ} is smaller than $n/6$. The further discussion of properties of this index and additional references can be found in Smith-Crowe et al. (2013).

r_{wg} and r_{wg}^* Indices

The index r_{wg} was introduced by James et al. (1984). The main rationale for it was to define no agreement as random responses over all categories.

Using the authors' notations, r_{wg} may be represented as $r_{wg} = 1 - S^2 / \sigma_{EU}^2$, where σ_{EU} is the expected variance. The value $\sigma_{EU}^2 = (n^2 - 1) / 12$ is mostly used, which corresponds to the discrete uniform distribution. If $S^2 > \sigma_{EU}^2$, then the truncation of S^2 to σ_{EU}^2 is applied which results in $r_{wg} = 0$.

The modified index $r_{wg}^* = 1 - S^2 / \sigma_{MV}^2$, where $\sigma_{MV}^2 = 0.5 \cdot (n^2 + 1) - 0.25(n + 1)^2$ is the maximum variance, was recommended and discussed by Lindell et al. (1999).

Index a_{wg}

This index was introduced by Brown and Hauenstein (2005) to address some limitations of r_{wg} and r_{wg}^* indices. It has a similar interpretation to r_{wg} . The index estimates agreement by the proportion of observed agreement between raters to the maximum possible disagreement assuming that the mean rating is given.

Using the authors' notations, a_{wg} is defined by

$$a_{wg} = 1 - \frac{2(m-1)S^2}{m((i_1 + i_k)M - M^2 - i_1 \cdot i_k)}. \quad (14)$$

The further discussion of a_{wg} can be found in Wagner et al. (2010).

Spectral Consistency Factor $\mathcal{K}$

The weighted spectrum of the set I was used to assess inter-rater agreement by Zgurovsky, Totsenko, and Tsyganok (2004). The index defined the no-agreement case as a distribution of ranks with the minimal sum of the variance and the deviation from a uniform distribution.

In the authors' notations, the index $\mathcal{K}$ can be expressed as

$$\mathcal{K} = \left(1 - \frac{AD_M + H(Q)}{G \sum_{i=1}^n |i - (n+1)/2| + \ln(n)} \right) \cdot (1 - z), \quad (15)$$

where $G = m/(n \ln(m) \ln(n))$ is a scaling factor,

$$z = \mathbf{I}(i_1 = 1) \cdot \mathbf{I}(i_k = n) \cdot \prod_{i=1}^{k-1} \mathbf{I}(r_i = r_{i+1}) \cdot \prod_{j=2}^{k-1} \mathbf{I}(i_{j+1} - i_j = i_2 - i_1), \quad (16)$$

and the indicator function $\mathbf{I}(\cdot)$ is defined by

$$\mathbf{I}(x) = \begin{cases} 1, & \text{if } x \text{ is true,} \\ 0, & \text{if } x \text{ is false.} \end{cases} \quad (17)$$

The multiplier $1 - z$ was introduced to truncate $\mathcal{K}$ to 0 when the equal numbers of raters select equidistant scores (the case of the uniform discrete distribution).

Unfortunately, the normalization above gives negative values of $\mathcal{K}$ for some ratings. Therefore, the authors propose to use the following correction:

$$\mathcal{K}_0 = \left(1 - \frac{AD_M + H(Q)}{(n-1)/2 + \ln(n)} \right) \cdot (1 - z), \quad (18)$$

which ensures that $\mathcal{K}_0 \in (0, 1)$.

Weighted Pairing Index MR

This index was introduced by Cicchetti et al. (1997) and can be computed as

$$MR = \frac{\sum_{i,l \in I} r_i r_l (1 - |i - l|/n) - \sum_{i \in I} r_i^2}{m(m-1)}. \quad (19)$$

The main rationale for the index is to determine whether the ratings of experts are too far from the group average.

Pearson's Statistic χ^2

χ^2 is used to test goodness of fit and establishes whether or not an observed rating distribution differs from a theoretical distribution. It can be computed by $\chi^2 = n/m \sum_{i=1}^n (r_i - m/n)^2$, where m/n is an expected rating asserting complete disagreement (a uniform distribution).

Table 1. Values of Indices for Example Patterns.

Case	S	CV	$AD_{M_{adj}}$	r_{wg}	r_{wg}^*	a_{wg}	$\mathcal{K}_0$	MR	χ^2	κ
1	2.95	0.54	2.57	0.00	0.57	0.19	0.00	0.65	0	0.00
2	2.91	0.49	2.52	0.00	0.58	0.20	0.3	0.66	1	0.01
3	3.44	0.63	3.08	0.00	0.42	-0.11	0.00	0.61	30	0.20
4	4.62	0.84	4.62	0.00	0.00	-1.00	0.00	0.53	80	0.35
5	2.56	0.47	2.57	0.20	0.68	0.38	0.53	0.74	80	0.46
6	0.97	0.19	0.75	0.89	0.95	0.91	0.69	0.89	39	0.71
7	0.51	0.21	0.51	0.97	0.99	0.96	0.82	0.95	80	0.85
8	0.31	0.05	0.18	0.99	1.00	0.99	0.93	0.98	144	0.93
9	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	180	1.00

For the number of raters $m = 20$, Figure 4 shows some example patterns for which the above indices were compared. The corresponding values of the indices are shown in Table 1. Similar results were also obtained for AD_M , AD_{M_d} , and $\mathcal{K}$ but are not shown in the article. The cases in Figure 4 are numbered from left to right and continuing on in the next row. It follows from the assumptions that the degree of agreement should increase from Case 1 to Case 9.

Note that the indices S , $AD_{M_{adj}}$, a_{wg} , MR, and χ^2 are not normalized to $[0,1]$. The indices S , CV, and $AD_{M_{adj}}$ are disagreement measures, that is, their values should decrease from Case 1 to Case 9.

The test examples provide important marginal cases that help to compare various indices. Respective contributions of $H^*(P)$ and $H^*(Q)$ to $\kappa(P, Q)$ can be easily seen from Equations 7 and 8. The normalized entropy $H^*(P)$ equals 1 for Cases 1 to 4 with equidistant selected scores; then, it decreases to 0 for the cases when scores form a group. At the same time, the normalized entropy $H^*(Q)$ gradually decreases from 1 to 0.

Table 1 demonstrates that the index κ performs well on the test examples. Its values gradually increase with increasing agreement between raters' scores. In addition, it follows from its definition that κ has various properties required for inter-rater agreement indices. Namely, it is invariant under score shifts, that is, κ is the same for I and $I' = \{i'_1, \dots, i'_k\}$ if $i'_j - i_j = \text{const}$ for all $j = 1, \dots, k$. The index is monotonically decreasing when the distance $i_{j+1} - i_j$ increases. κ decreases if one of m raters changes his or her score to a new one in such a way that the new set I' consists of more elements than the original I , that is, $I' = I \cup \{j\}$, where $j \in 1, \dots, n$ and $j \notin I$.

All other indices have various limitations. In the discussion below, Figure 4 and the corresponding results in Table 1 have been referred to list some of these limitations:

- The indices r_{wg} and $\mathcal{K}_0$ cannot distinguish between distributions with equidistant equal frequencies, for example, Cases 1, 3, and 4. Thus, these indices might give rather similar results for two different cases: when all raters chose ranks randomly and when the raters formed several groups (schools) that shared quite different opinions.
- The indices S , CV, $AD_{M_{adj}}$, r_{wg} , r_{wg}^* , a_{wg} , $\mathcal{K}_0$, and MR specify complete disagreement for Case 4, that is, these indices indicate when two equal groups of raters have opposite opinions (a half of raters assign the score 1 and the other half the score n) instead of unpredictability of scores.
- The indices S , CV, $AD_{M_{adj}}$, r_{wg}^* , a_{wg} , and MR do not suggest complete disagreement for Case 1 of complete unpredictability (the uniform distribution on all scores). They will demonstrate various levels of agreement even if raters assign ranks at random.

Figure 4. Example patterns.

Table 2. Values of Indices for the Case of Outliers.

S	CV	$AD_{M_{adj}}$	r_{wg}	r_{wg}^*	a_{wg}	$\mathcal{K}_0$	MR	χ^2	κ	κ^*
0.51	0.21	0.51	0.97	0.99	0.96	0.82	0.95	76.26	0.85	0.85
1.75	0.62	0.87	0.63	0.85	0.56	0.75	0.88	71	0.49	0.81

- The index $\mathcal{K}_0$ is “discontinuous,” that is, substantially vary between similar patterns (Cases 1 and 2). It means $\mathcal{K}_0$ is not robust to outliers. It may produce misleading results when a single rank impacts the overall agreement.
- The indices r_{wg} , r_{wg}^* , $\mathcal{K}_0$, and a_{wg} do not change gradually having big gaps between their values. It may cause problems in selecting an appropriate critical region/level to test an agreement hypothesis.
- Pearson’s statistic χ^2 does not work well for non-uniform distributions which demonstrate high degrees of agreement (Case 6 vs. Cases 4 and 5). Also, it depends only on the frequencies of selected scores, but not on their actual values. Thus, quite different Cases 4, 5, and 7 have the same χ^2 value.

Adjusted Double Entropy Indices

It is well known that values of inter-rater agreement indices may significantly change if a new expert assigns a score, which is quite different from the existing scores. Particularly, it is a serious problem in applications with a small number of raters and a relatively large response scale.

In this section, two adjusted indices that can handle outliers and missing or incorrectly recorded data have been introduced.

In situations where the identification and removal of outliers are controversial issues, reporting both the unadjusted and adjusted values is methodologically sound.

Adjusted Index κ^*

First, situations when the developed double entropy index κ might be sensitive to outliers have been considered. Namely, values of κ can significantly change when a new rater gives a score too far from the other scores, for example, Figure 5 and the corresponding Table 2. Some other indices in the “Comparison of Inter-Rater Agreement Measures” section demonstrate similar sensitivity too. For κ , this sensitivity is caused by significant changes in the component $H^*(P)$ for small k .

To overcome this limitation, the adjusted index κ^* has been proposed. To define it, the normalized entropy $H^*(Q)$ as before and the adjusted $H^*(P)$ which is computed after censoring small r_i have been used. For instance, if all r_i less than $0.2M_r$ ($M_r = \sum_{i=1}^k r_i/k$) are truncated, then $I' = \{i_l \in I : r_{i_l} \geq 0.2M_r\}$ is used instead of I to compute $H^*(P)$.

For example, for the scores in Figure 5, the values of κ^* are given in Table 2. It demonstrates that the adjusted κ^* is less sensitive to marginal scores than the index κ . Note that κ and κ^* give same results for all cases in Table 1.

Adjusted Index κ_*

Equation 9 defines κ_0 as the weighted sum of $H^*(P)$ and $H^*(Q)$ with equal weights of 0.5. It can be generalized to the weighted sum $(1 - \alpha)H^*(P) + \alpha H^*(Q)$, where $\alpha \in (0, 1)$. For example, if $\alpha = k/n$, we obtain

Figure 5. Case of outliers.

Table 3. Values of Indices for the Case of Incorrect Records.

S	CV	$AD_{M_{adj}}$	r_{wg}	r_{wg}^*	a_{wg}	$\mathcal{K}_0$	MR	χ^2	κ	κ_*
2.11	0.28	1.77	0.46	0.78	0.46	0.47	0.76	14	0.59	0.34
2.45	0.34	2.09	0.27	0.7	0.33	0.4	0.72	10	0.06	0.11

$$\tilde{\kappa}_0(P,Q)=\frac{n-k}{n}H^*(P)+\frac{k}{n}H^*(Q). \tag{20}$$

The index $\tilde{\kappa}_0$ gives higher weights to $H^*(Q)$ and smaller weights to $H^*(P)$ for those k which are closer to n . Therefore, $\tilde{\kappa}_0$ is less sensitive to changes in the spread of a distribution and more sensitive to changes of distribution’s shape than κ_0 when $k \approx n$. Thus, its application is preferable for large values of k which indicate a significant divergence in raters’ assessments. The adjustment is useful if some of k scores were missing or incorrectly recorded.

The adjusted index κ_* is defined by $\kappa_*(P,Q)=1-\tilde{\kappa}_0(P,Q)$. For example, the second subplot of Figure 6 shows that one of scores 8 was mistakenly recorded as 2. The numerical values of the indices are given in Table 3. It is clear that the adjusted κ_* is less sensitive to incorrect records.

Figure 7 shows respective contributions of $H^*(P)$ and $H^*(Q)$ to the adjusted double entropy index $\kappa_*(P,Q)$. Monte Carlo simulations were used to generate random responses of six raters over a 10-level scale and to compute κ_* . For each random response, the figure shows a value of κ_* versus the corresponding values of $H^*(P)$ and $H^*(Q)$. Similar to κ , the adjusted double entropy index increases when $H^*(P)$ and $H^*(Q)$ decrease.

Illustrative Examples

The examples in this section present typical applications of the indices for a situational selection of raters from a larger population. The agreement of inter-rater judgments is investigated. For the sake of space, tables with the detailed values of all of the other agreement indices have not been included.

Figure 6. Case of incorrect records.

Figure 7. κ_* versus $H^*(P)$ and $H^*(Q)$.

Example 1: Assessment of Life Satisfaction

This example illustrates possible applications of the double entropy inter-rater agreement indices to examine consensus in survey responses. Life satisfaction is a complex index, and individual scores can vary substantially.

In the example, life satisfaction data from the Household, Income and Labour Dynamics in Australia survey and Statistics Canada's General Social Survey have been analyzed. The distributions of individuals' responses in the 2010 Australian and 2011 Canadian surveys presented in Ambrey and Fleming (2014), Figure 1, Bonikowska, Helliwell, Hou, and Schellenberg (2014), and Table 3, respectively, have been used. The 11-point response scale with 0 meaning *very dissatisfied* and with 10 meaning *very satisfied* was used. Figure 8 provides a graphical representation of the percentage distribution of scores.

Note that in both surveys, there are a number of categories with rather low percentage frequencies which bias index estimates. In this case, the adjusted index κ^* provides a better solution and effectively corrects for the low remote frequencies. Indeed, the adjusted κ^* equals 0.65 and 0.64 for the Australian and Canadian surveys, respectively, when unadjusted indices are 0.15 and 0.14. The analysis demonstrates moderate levels of agreement in individuals' responses. The levels of respondent's agreement are roughly the same for the both countries.

Figure 8. Distributions of life satisfaction scores.

Example 2: Prompt Difficulty

In this example, the double entropy index is applied to estimate degrees of agreement for constructed-response assessments.

Data from Table 5.14 presented in Lim (2009) have been analyzed. The University of Michigan offers an advanced-level English proficiency test which is similar to the IELTS and TOEFL. Examinees respond to prompts that are drawn from a larger pool of prompts. Among other problems, Lim studied raters' perceptions of prompts. For each prompt, 10 raters answered to the following question: "Compared with the average prompt in the pool of writing prompts, is this prompt easier, about average, or more difficult to get a high score on?" The following 5-point Likert-type scale was used: 1 = *clearly easier*, 2 = *somewhat easier*, 3 = *about average*, 4 = *somewhat more difficult*, and 5 = *clearly more difficult*.

Figure 9 provides a graphical representation of the percentage distribution of scores for two prompts.

As expected from the data, the value $\kappa = 0.04$ for Prompt 12 shows little agreement among raters. On the contrary, we get $\kappa = 0.71$ for Prompt 46. It demonstrates moderate agreement in individuals' responses. Notice that the two cases are easily distinguished by κ despite of a rather low level of sensitivity for $n = 5$ and $m = 10$ given by Figure 3. The example shows that the same group of raters can have very different levels of agreement for distinct single targets. The values of κ match well the observed distribution patterns of prompts.

Overall, the observed values of the indices in all three examples appear to be consistent with visual and other assessments of the degree of inter-rater agreement.

Directions for Future Research

This article develops novel agreement indices and shows some application of a new assessment procedure of the chance agreement, assuming no prior standard rating. It demonstrates that the new indices are superior and more closely capture consensus/disagreement in comparison with other inter-rater agreement indices for a single target.

Figure 9. Distributions of scores for two prompts.

Although the discussion on definitions and properties of the indices in relation to a single target data have been focused, the derived results can be extended and applied to other research questions. The future work includes extending of the proposed statistics to various testing scenarios. Some of them are listed below.

- It is typically assumed that a lack of agreement is generated by a uniform distribution. For the illustrations in this article, the authors followed this approach too. Replacing the entropy by the relative entropy allows using other distributions, such as those that would be caused by response biases (LeBreton & Senter, 2008).
- The article mainly deals with introducing and discussing properties and applications of the new indices for single target cases. It would be interesting to extend the results and statistical analysis to multiple target cases using various aggregation methods, for example, the unweighted group mean, the response data-based weighted mean, and the confidence-based weighted mean considered by Wagner et al. (2010).
- It would be interesting to generalize the results to the scenarios which take into account relative rater's competence in a group.
- The indices proposed in the article were developed for discrete quantitative data and can also be used for ordinal data. An important problem is the investigation of the impact of aggregating/rounding real-valued measurements to a finite set of equidistant scores on changes in inter-rater agreement indices (see, for example, Zgurovsky et al., 2004). Various applied problems may also require generalizations to a finite set of non-equidistant categories.
- It would be interesting to extend the approach based on additive functions in Equations 9 and 20 to other transformations. For example, multiplicative functions of $H^*(P)$ and $H^*(Q)$ can be studied. It would be also important to further clarify respective contributions of $H^*(P)$ and $H^*(Q)$ to novel indices (see, for example, Figure 7).

Acknowledgments

The authors are grateful for the referees' careful reading of the article and many detailed comments and suggestions, which helped to improve an earlier version of this article. The authors are also grateful to Drs. C. L. Ambrey, J. Dawes, and C. M. Fleming for providing their data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Additional examples and materials on testing hypotheses about inter-rater agreement are available in the subsection “Agreement Coefficient of Research Materials” on the website: <https://sites.google.com/site/olenkoandriy/>. It also contains the mathematical derivations, data, and R code used in this article.

References

- Ambrey, C. L., & Fleming, C. M. (2014). Life satisfaction in Australia: Evidence from ten years of the HILDA Survey. *Social Indicators Research, 115*, 691-714.
- Baca-García, E., Blanco, C., Sáiz-Ruiz, J., Rico, F., Diaz-Sastre, C., & Cicchetti, D. V. (2001). Assessment of reliability in the clinical evaluation of depressive symptoms among multiple investigators in a multicenter clinical trial. *Psychiatry Research, 102*, 163-173.
- Bonikowska, A., Helliwell, J. F., Hou, F., & Schellenberg, G. (2014). An assessment of life satisfaction responses on recent statistics Canada surveys. *Social Indicators Research, 118*, 617-643.
- Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the rwg indices. *Organizational Research Methods, 8*, 165-184.
- Burke, M. J., & Dunlap, W. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*, 159-172.
- Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2*, 49-68.
- Cicchetti, D. V., Fontana, A., & Showalter, D. (2009). Evaluating the reliability of multiple assessments of PTSD symptomatology: Multiple examiners, one patient. *Psychiatry Research, 166*, 269-280.
- Cicchetti, D. V., Showalter, D., & Rosenheck R. (1997). A new method for assessing interexaminer agreement when multiple ratings are made on a single subject: Applications to the assessment of neuropsychiatric symptomatology. *Psychiatry Research, 72*, 51-63.
- DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*, 333-356.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.
- Gwet, K. L. (2012). *Handbook of inter-rater reliability*. Gaithersburg, MD: Advanced Analytics.
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85-98.
- Kim, Y. H. (2009). An investigation into native and non-native teachers judgments of oral English performance: A mixed methods approach. *Language Testing, 26*, 187-217.
- Klemens, B. (2012). Mutual information as a measure of intercoder agreement. *Journal of Official Statistics, 28*, 395-412.
- LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding rWG, rWG, rWG(J), and rWG(J). *Organizational Research Methods, 8*, 128-138.
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815-852.
- Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment* (Doctoral dissertation). Retrieved from <http://hdl.handle.net/2027.42/64665>

- Lin, L., Hedayat, A. S., Sinha, B., & Yang, M. (2002). Statistical methods in assessing agreement. *Journal of the American Statistical Association*, 97, 257-270.
- Lin, L., Hedayat, A. S., & Tang, Y. (2013). A comparison model for measuring individual agreement. *Journal of Biopharmaceutical Statistics*, 23, 322-345.
- Lindell, M. K., & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement*, 21, 271-278.
- Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, 23, 127-135.
- Lohse-Bossenz, H., Kunina-Habenicht, O., & Kunter, M. (2014). Estimating within-group agreement in small groups: A proposed adjustment for the average deviation index. *European Journal for Work & Organizational Psychology*, 23, 456-468.
- Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology*, 74, 368-370.
- Shah, M. (2011). Generalized agreement statistics over fixed group of experts. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Lecture Notes in Computer Science: Vol. 6913. Machine learning and knowledge discovery in databases* (pp. 191-206). Berlin, Germany: Springer.
- Shoukri, M. M. (2004). *Measures of interobserver agreement*. Boca Raton, FL: Chapman & Hall.
- Smith-Crowe, K., Burke, M. J., Kouchaki, M., & Signal, S. M. (2013). Assessing interrater agreement via the average deviation index given a variety of theoretical and methodological problems. *Organizational Research Methods*, 16, 127-151.
- von Eye, A., & Mun, Y. E. (2005). *Rater agreement: Manifest variable methods*. Mahwah, NJ: Lawrence Erlbaum.
- Wagner, S. M., Rau, C., & Lindemann, E. (2010). Multiple informant methodology: A critical review and recommendations. *Sociological Methods & Research*, 38, 582-618.
- Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps?* (TOEFL iBT Research Report No. TOEFLiBT-11). Princeton, NJ: Educational Testing Service. Retrieved from <http://files.eric.ed.gov/fulltext/ED507804.pdf>
- Zgurovsky, M. Z., Totsenko, V. G., & Tsyganok, V. V. (2004). Group incomplete paired comparisons with account of expert competence. *Mathematical and Computer Modelling*, 39, 349-361.