Agreement and Kappa-Type Indices

Author(s): Jeroen De Mast

Source: *The American Statistician*, May, 2007, Vol. 61, No. 2 (May, 2007), pp. 148-153

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/27643866

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/27643866?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Taylor & Francis, Ltd. and *American Statistical Association* are collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician*

This content downloaded from
181.58.0.99 on Sat, 06 Dec 2025 04:20:12 UTC
All use subject to https://about.jstor.org/terms

Agreement and Kappa-Type Indices

Jeroen DE MAST

Kappa-type indices use the concept of agreement to express the reproducibility of nominal measurements. This article grounds kappa-type indices in statistical modeling, making explicit the underlying premises and assumptions. We critically review whether the interpretation of the kappa index as a chance-corrected probability of agreement can be substantiated. Further, we show that the so-called paradoxical behavior of the kappa index is explained from the fact that it is a measure of predictive association, rather than a pure measure of reproducibility. We discuss a number of alternative forms, critically examining whether they can be translated in tangible real-life interpretations.

KEY WORDS: Categorical data; Gauge capability analysis; Measurement system analysis; Nominal data; Reliability; Reproducibility.

1. INTRODUCTION

An important aspect of measurement systems is their reproducibility. Reproducibility is an aspect of the reliability (behavioral sciences) or precision (engineering sciences) of a measurement system. This article deals with reproducibility of measurements on a nominal scale. Nominal scales consist of unordered classes (Allen and Yen 1979), and nominal measurements classify objects into these classes. An example of nominal measurement in industry is the classification of production faults into a predefined system of categories (such as "machine related," "operator related," "material related," etc.). Other examples include the classification of complaints into complaint types, and the diagnosis of patients into disorder types. We will refer to the faults, complaints, and patients being classified as the measured objects. By *the measurement system* we refer to the quality inspector, helpdesk employee, or physician performing the classification, and in addition to the procedure, instructions and aids that he or she uses to do so.

Poor reproducibility could point to ineffective instructions, insufficient expertise of the appraisers, or a general lack of understanding of the property that the system is supposed to measure. During a reproducibility study a number of objects is measured (in the case of nominal measurements: classified) multiple times, possibly by multiple appraisers. Statistical experts could estimate a parameterized model for the joint distribution of the repeated measurements, and thus obtain a detailed understanding of the stochastic properties of the measurement system under study. But for the nonexpert it is more convenient to express the reproducibility of a measurement system in terms of a single or a combination of a few indices. This article deals with such indices for expressing the reproducibility of nominal measurements.

Kappa-type indices are popular for expressing reproducibility of nominal measurements in the psychometrical, biostatistical, and engineering sciences [see Kraemer (1988) for a discussion of the differences between the engineering context and the biological and behavioral contexts]. They are based on the concept of agreement. Standard accounts of kappa-type indices are Cohen (1960), Fleiss (1971), Conger (1980), and Davies and Fleiss (1982). Kappa-type indices are surrounded by quite some controversy. In this article we try to understand the behavior and value of various kappa indices. We approach the issue by making the lines of reasoning in papers such as the above-mentioned ones more explicit. Our expositions and discussions are based on more rigorous mathematical modeling. Further, where the above-mentioned papers define kappa-type indices solely in terms of sample statistics, we think it is more appropriate to define a solid experimental model and provide an exposition in terms of population parameters. Landis and Koch (1977), Kraemer (1979), and Tanner and Young (1985) provided accounts of kappa-type indices grounded in statistical modeling. None of these articles questioned and compared, however, the basic form of kappa-type indices (and in particular the so-called *chance correction*), which is what this article aims to do. The reader looking for a useful and recent source of references on the subject is referred to Kraemer, Periyakoil, and Noda (2002).

The next section defines our experimental model. The subsequent sections define and discuss various indices based on agreement. Our aim is to explicate the reasoning behind them, and to study how they can be interpreted in real-life terminology. In Section 7 (p. 152) we apply the various indices to an example, highlighting their different meanings.

2. EXPERIMENTAL DESIGN AND DATA MODEL

To assess the reproducibility of a nominal measurement procedure one collects data in the following manner. We have n objects, which are measured once by each of m appraisers on an unordered scale $\{1, 2, \ldots, a\}$. The data are denoted Y_{ij}, with $i = 1, \ldots, n$ indexing objects, and $j = 1, \ldots, m$ indexing appraisers.

Of interest for reproducibility studies is the joint distribution of the $\{Y_{ij}\}_{j=1,\ldots,m}$, and in particular the association structure between the repeated measurements. We choose to model the Y_{ij} using a latent class model (the main alternative being a log-linear model). Although log-linear models are powerful means to analyze association structures (see Tanner and Young 1985), the advantage of latent class models is that the cause of the association among repeated measurements—the objects' true values with respect to the measured property—is modeled explicitly.

Jeroen de Mast is Senior Consultant at IBIS UvA, and Associate Professor at the University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands (E-mail: *jdemast@science.uva.nl*).

 ©*American Statistical Association* *DOI: 10.1198/000313007X192392*

This content downloaded from
181.58.0.99 on Sat, 06 Dec 2025 04:20:12 UTC
All use subject to https://about.jstor.org/terms

Consequently, the variation in the measurements is explicitly attributed to a systematic part (variation among true values) and a random part (measurement variation), a practice which resembles the typical manner in which reproducibility studies for ratio and interval scale measurements are modeled.

Following considerations in Goodman and Kruskal (1954), the true values $X_i, i = 1, \ldots, n$ could be assumed categorical or continuous. For nominal measurements, it seems more appropriate to assume the same unordered categorical scale for both the measurements Y_{ij} and the true values X_i. The X_i are assumed stochastically independent, and have a discrete distribution with parameters

$$p(k) := P(X_i = k) \quad , k = 1, \ldots, a, \text{ with } \textstyle\sum_{k=1}^{a} p(k) = 1. \tag{1}$$

As for the distribution of the Y_{ij}, we assume that given an object's true value X_i the m measurements $Y_{i1}, Y_{i2}, \ldots, Y_{im}$ are stochastically independent (the assumption of *local independence*, which is standard in latent class models). Moreover, the distribution of the $Y_{i1}, Y_{i2}, \ldots, Y_{im}$ depends on the true value X_i, and we define

$$q(k|\ell) := P(Y_{ij} = k | X_i = \ell), \tag{2}$$

thus specifying the distribution of the measurement errors. The model parameters $p(k)$, $k = 1, 2, \ldots, a$, and $q(k|\ell)$, $k, \ell = 1, 2, \ldots, a$, determine the distribution of the Y_{ij} and we have

$$\begin{aligned} P(Y_{ij} = k) &= \sum_{\ell=1}^{a} p(\ell) q(k|\ell) \\ &=: q(k) \text{ (marginal distribution).} \end{aligned} \tag{3}$$

If the concept of *true value* is thought to be problematic, it may help to view it as a hypothetical value that would be assigned to the object by an authoritative measurement system (such as the standard meter); see ISO (1993). It is used to model the dependence structure in the data.

De Mast and Van Wieringen (in press) provide a similar model in the case that each appraiser measures each object $s \geq 1$ times. Another alternative, also discussed by De Mast and Van Wieringen (in press), is to assume different stochastic properties for each appraiser j, and replace Equation (2) with

$$q_j(k|\ell) := P(Y_{ij} = k | X_i = \ell). \tag{4}$$

Finally, the stochastic properties of the measurements may depend on other factors than the true values X_i alone. For this reason, one could consider models with parameters of the form $q(k|\ell, u) = P(Y_{ij} = k | X_i = \ell, T_i = u)$, with T a factor affecting the stochastic properties of the measurements.

This article focuses on the definition of the index that is used to express the result of reproducibility studies. For this reason, the model is kept simpler, assuming a single measurement per appraiser per object. The conclusions carry over in straightforward manner to more complex experimental set-ups and models.

3. PROBABILITY OF AGREEMENT

Reproducibility of nominal measurements will be expressed in terms of a probability of agreement. Two measurements of an object agree if they are identical (the object is classified in the same category both times). $P_{\text{agreement}}$ (or short: P_a) is the probability that two arbitrary measurements of an arbitrary object agree. Under the model specified by Equations (1)–(2), we have for an object with true value $X_i = \ell$:

$$\begin{aligned} P_a(\ell) &:= P(Y_{i\,j_1} = Y_{i\,j_2} | X_i = \ell) \\ &= \sum_{k=1}^{a} q^2(k|\ell), \end{aligned}$$

and for an arbitrary object:

$$P_a := P(Y_{i\,j_1} = Y_{i\,j_2}) = \sum_{\ell=1}^{a} \sum_{k=1}^{a} p(\ell)\, q^2(k|\ell). \tag{5}$$

Fleiss (1971) introduced the sample statistic

$$\hat{P}_a = \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{k=1}^{a} N_{ik}(N_{ik} - 1)$$

(Fleiss's formula (3)), where $N_{ik} = \{\#j : Y_{ij} = k\}$. De Mast and Van Wieringen (in press) show that $\hat{P}_a$ is an unbiased estimator of P_a.

P_a is quite a tangible expression of reproducibility; it can be translated quite easily in real-life implications. The main objection against its use, is that it is difficult to compare values across different scales. The chance of agreement is larger on a two-point scale than on a five-point scale by chance alone, and as a consequence, $P_a = 0.5$ on a two-point scale means something else than the same value on a five-point scale.

4. KAPPA-TYPE INDICES

Thinking about which values of P_a represent "good" measurement systems, and which represent "bad" ones, one should realize that a positive value of P_a does not automatically mean that the measurements are well reproducible. Even if appraisers would assign values to objects randomly, there would be some agreement. To correct for this phenomenon, Cohen (1960), Fleiss (1971), Conger (1980) and numerous others have introduced κ (kappa) type indices as a recentered and rescaled version of P_a. The traditional formula is

$$\kappa = \frac{P_{\text{observed}} - P_{\text{expected}}}{1 - P_{\text{expected}}}.$$

Here, P_{observed} and P_{expected} both denote probabilities of agreement. P_{observed} is the probability of agreement for the measurement system under study, while P_{expected} is the probability of agreement for a "chance" measurement system (i.e., a completely uninformative measurement system that assigns measurement values to objects randomly). The use of the words *observed* and *expected* is questionable here, and we shall instead use the more appropriate terminology

$$\kappa = \frac{P_{\text{agreement}} - P_{\text{agreement|chance}}}{1 - P_{\text{agreement|chance}}}, \tag{6}$$

This content downloaded from
181.58.0.99 on Sat, 06 Dec 2025 04:20:12 UTC
All use subject to https://about.jstor.org/terms

(resembling the terminology used by Lipsitz, Laird, and Brennan 1994).

Whereas the relevant range of P_a is $[P_{\text{agreement|chance}}, 1]$, the relevant range of κ-type indices is [0, 1], where 1 corresponds to the agreement that a perfect measurement system would attain, and 0 corresponds to the agreement that random measurements would attain. The probability of agreement of such random measurements will be denoted $P_{\text{agreement|chance}}$ (or short: $P_{a|c}$). To do this rescaling, we have to define how we conceive of a chance measurement system (i.e., we have to specify what we mean if we hypothesize about appraisers assigning values "randomly"). Different notions of a chance measurement system are advocated in the literature, leading to quite some controversy.

A popular elaboration is the one by Fleiss (1971). In this and similar articles, the line of reasoning is not made very explicit, and as a consequence it is difficult to give a precise exposition. The following is our reconstruction of what we believe is implicitly defined. Chance measurements are defined as measurements done at random (i.e., independent of the object being measured) and with a probability distribution equal to the marginal distribution of the measurement system under study when applied to the population under study. Denoting chance measurements by Z_{ij}, this amounts to

$$Z_{ij} \text{ are iid and } P(Z_{ij} = k) = q(k) \text{ for all } i, j, \text{ and } k. \quad (7)$$

Under these premises, the probability of agreement of chance measurements equals $P_{a|c}^{\text{Fleiss}} = P(Z_{ij_1} = Z_{ij_2}) = \sum_{k=1}^{a} q^2(k)$. Fleiss's (1971) estimator (his formula (5))

$$\hat{P}_{a|c}^{\text{Fleiss}} = \sum_{k=1}^{a} \frac{N_k^2}{(mn)^2},$$

(with $N_k = \{\#(i, j) : Y_{ij} = k\}$) has a small bias (see De Mast and Van Wieringen in press). We have

$$\kappa^{\text{Fleiss}} = \frac{P_a - P_{a|c}^{\text{Fleiss}}}{1 - P_{a|c}^{\text{Fleiss}}}$$

[actually, Fleiss (1971) gave only a definition in terms of sample statistics].

The premise that chance measurements are independent of the objects' true values is uncontroversial, but the premise that they have the distribution defined by the marginal distribution $q(k)$ seems hard to defend. It is difficult to see why the distribution of chance measurements would be related to the $q(k)$ (and therefore to the $p(\ell)$, the distribution of the object's true values), especially given the first premise, that chance measurements are independent of the objects' true values. If chance measurements are conceptualized as being uninformative about the measured property, it is implausible that their distribution is in some way related to the $q(k)$. Unfortunately, the suggestive rather than precise descriptions of the premises made in many accounts obscure this discrepancy. Note that any other distributional assumption for the chance measurements is likely to be arbitrary as well, perhaps with one exception (as explained in the next section).

Because of this unconvincing and seemingly unsubstantiated premise that chance measurements are distributed as in (7), we think that the motivation for κ^{Fleiss} as a probability of agreement corrected for agreement by chance is weak. A better motivation for κ^{Fleiss} as an index to express reproducibility is to understand it as a measure of association. To be more precise, κ^{Fleiss} is a measure of predictive association based on Gini's measure of dispersion. The general form of measures for predictive association is (Hershberger and Fisher 2005)

$$r_{Y,X} = 1 - \frac{\Delta_{Y|X}}{\Delta_Y}, \quad (8)$$

with Δ a measure of dispersion, Δ_Y the dispersion of Y, and $\Delta_{Y|X}$ the conditional dispersion of Y given X (a strongly related class of measures are the proportional reduction in error, or PRE, measures of association, which are built around reduction in probability of misclassification instead of reduction in dispersion). For a categorical variable Y with probability distribution $(p_1, p_2, \ldots, p_a)$, the Gini dispersion is $\Delta_Y = 1 - \sum_{k=1}^{a} p_k^2$ (Gilula and Haberman 1995). Substituting in (8) that $\Delta_Y = 1 - \sum_{k=1}^{a} q^2(k)$ and that $\Delta_{Y|X} = 1 - \sum_{\ell=1}^{a} p(\ell) \sum_{k=1}^{a} q^2(k|\ell)$ gives κ^{Fleiss}. Taking instead for Δ_Y and $\Delta_{Y|X}$ the entropy and conditional entropy, one finds Theil's coefficient (Haberman 1988), which is thus a direct cousin of κ^{Fleiss}.

It is quite common to express measurement precision in terms of a measure of association (compare the intraclass correlation index for ratio and interval scale measurements). One should be aware, however, of an essential difference between a measure of association and a pure measure of reproducibility. A measure of association expresses reproducibility in relationship to the variation in the population of measured objects (sometimes called *part-to-part spread*, or *prevalence*). Consequently, measures of association are not useful for expressing a measurement system's reproducibility independent of a population of measured objects. Pure measures of reproducibility—such as measurement spread in Gauge R&R studies (Burdick, Borror, and Montgomery 2003), and P_a as defined above—express a measurement system's reproducibility independent of the population of objects being measured.

Consider the following example, with a nominal scale of $a = 2$ categories. A measurement system's statistical properties are given by $q(1|1) = 0.95$; $q(2|1) = 0.05$; $q(1|2) = 0.05$; and $q(2|2) = 0.95$ (the probability of agreement $P_a = 0.91$ is quite large). If one were to study this measurement system on a population of objects with distribution $p(1) = 0.50$ and $p(2) = 0.50$, one would find $P_{a|c}^{\text{Fleiss}} = 0.50$ and $\kappa^{\text{Fleiss}} = 0.81$. However, if one studied the same measurement system on a population of objects with distribution $p(1) = 0.95$ and $p(2) = 0.05$, one would find $P_{a|c}^{\text{Fleiss}} = 0.83$ and $\kappa^{\text{Fleiss}} = 0.45$. This dependence of κ^{Fleiss} on prevalence was noted by, for instance, Thompson and Walter (1988).

The observation that κ^{Fleiss} is a measure of association and not a pure measure of reproducibility explains behavior that is often described as paradoxical (Feinstein and Cicchetti 1990). The relationship between P_a and κ^{Fleiss} is strongly nonlinear, and as a result, small changes in P_a can result in dramatic changes in κ^{Fleiss}. For example, assuming an objects population with distribution $p(1) = 0.95$, $p(2) = 0.05$, a measurement system with properties $q(1|1) = 1.0$; $q(2|1) = 0.0$; $q(1|2) = 0.0$; and $q(2|2) = 1.0$ gives $\kappa^{\text{Fleiss}} = 1.0$, but a measurement system

This content downloaded from
181.58.0.99 on Sat, 06 Dec 2025 04:20:12 UTC
All use subject to https://about.jstor.org/terms

with properties $q(1|1) = 0.95$; $q(2|1) = 0.05$; $q(1|2) = 0.05$; and $q(2|2) = 0.95$ gives $\kappa^{\text{Fleiss}} = 0.45$. The strong sensitivity of κ^{Fleiss} for small changes in the $q(k|\ell)$ has as a consequence that the standard error of the estimator $\hat{\kappa}^{\text{Fleiss}}$ may be so large as to make it practically useless. Suppose, for example, that we measure $n = 100$ objects $m = 2$ times, and that the resulting classifications are

$$\begin{pmatrix} \#i : Y_{i1} = Y_{i2} = 1 & \#i : Y_{i1} = 1, Y_{i2} = 2 \\ \#i : Y_{i1} = 2, Y_{i2} = 1 & \#i : Y_{i1} = Y_{i2} = 2 \end{pmatrix} = \begin{pmatrix} 99 & 0 \\ 0 & 1 \end{pmatrix}. \quad (9)$$

The given data would result in $\hat{\kappa}^{\text{Fleiss}} = 1.0$, while

$$\begin{pmatrix} 98 & 1 \\ 0 & 1 \end{pmatrix}. \quad (10)$$

would give $\hat{\kappa}^{\text{Fleiss}} = 0.66$. This behavior is seen by many as leading to results that are difficult to interpret.

The following warning is also based on the fact that κ^{Fleiss} behaves as a measure of association. When expressing the results of a reproducibility study in terms of κ^{Fleiss}, it is important that the objects are a representative sample from the population of objects. In practice this means that it is vital that they are sampled randomly, and that care should be taken to avoid any selection bias. Trying to select objects such that an equal share of each category is in the sample (a prescription sometimes found in practice) can result in dramatic under- or overestimation.

The kappa index defined by Conger (1980) and Davies and Fleiss (1982) is based on essentially the same line of reasoning, but assumes model (4) (for details see De Mast and Van Wieringen in press).

5. KAPPA INDEX BASED ON UNIFORM CHANCE MEASUREMENTS

For chance measurements, the premise of randomness (i.e., values are assigned independently of the true values of the measured objects) is uncontroversial, but other distributional assumptions are hard to defend. However, of all choices, a uniform distribution can be given some justification. Any other distribution indicates that the measurements are not completely uninformative. Fleiss's chance measurements, for example, are informative of the $p(\ell)$. The uniform distribution can be defended as representing the maximally non-informative measurement system given a certain scale $\{1, \ldots, a\}$. Chance measurements are conceived as

$$Z_{ij} \text{ are iid and } P(Z_{ij} = k) = 1/a \text{ for all } i, j, \text{ and } k. \quad (11)$$

We have $P_{a|c}^{\text{Unif}} = P(Z_{ij_1} = Z_{ij_2}) = 1/a$. The corresponding kappa statistic

$$\kappa^{\text{Unif}} = \frac{P_a - 1/a}{1 - 1/a}$$

was proposed by Bennett, Alpert, and Goldstein (1954) and advocated by Brennan and Prediger (1981) and others. The value $1/a$ is a lower bound for P_a for measurement procedures whose statistical properties follow Equations (1) and (2) (De Mast and Van Wieringen in press), confirming that chance measurements, so defined, represent a maximally noninformative measurement system.

Several objections against κ^{Unif} are raised in the literature. Scott (1955) stated that "The index is based on the assumption that all categories ... have equal probability of use $[1/a]$ by both [appraisers]. This is an unwarranted assumption [in many real-life situations] .... The phenomena being coded are likely to be distributed unevenly." This criticism seems misguided, however. Scott criticized the logic behind κ^{Unif} for assuming that $p(1), \ldots, p(a) = 1/a$, which he finds—justly—an illegitimate assumption ("The phenomena being coded are likely to be distributed unevenly"). But nowhere in the definition of κ^{Unif} is it assumed or implied that the $p(\ell)$ are uniformly distributed (nor is this assumption made about the $q(k|\ell)$ or the $q(k)$). Instead, it is assumed that the distribution of the chance measurements has no relation with the $p(\ell)$ or the $q(k)$—which is in line with the premise that chance measurements are independent of the objects being measured.

A second issue comes to light if we study the next example. Consider a measurement system with the following statistical properties ($a = 5$):

$$\text{For all } \ell = 1, \ldots, 5\text{: } q(1|\ell) = 0.99;$$

$$q(k|\ell) = 0.0025 \text{ for } k = 2, 3, 4, 5, \quad (12)$$

(i.e., a measurement system which virtually always returns the value 1 independent of the object being measured). This measurement system is of course useless, and one could be puzzled to find that $\kappa^{\text{Unif}} = 0.96$. But on second thought, the reproducibility of this measurement system actually is very good. Measurement spread is practically nil, and results are almost 100% repeatable, and this is what the high value of κ^{Unif} reflects. Instead of a reproducibility problem, the measurement system has another problem, namely its poor accuracy (or validity). The analogue for numerical measurement systems is the case that a system returns the value 3.14 (say) independent of the object being measured. The measurement spread is zero, and hence its reproducibility is perfect, but its accuracy is poor. Contrary to the high value of κ^{Unif}, we have $\kappa^{\text{Fleiss}} = 0$. Again, we see that κ^{Fleiss} is not a pure measure of reproducibility, but confounds this property with other properties.

Note that κ^{Unif} is not a measure of predictive association, but a pure measure of reproducibility. Consequently, it does not suffer from what is described by many as interpretation problems or paradoxical behavior associated with κ^{Fleiss} (as described above). For example, for the data in (9) and (10), $\hat{\kappa}^{\text{Unif}} = 1.0$ and $\hat{\kappa}^{\text{Unif}} = 0.98$, respectively (compared to 1.0 and 0.66 for $\hat{\kappa}^{\text{Fleiss}}$).

6. NUMBER OF DISTINGUISHABLE CLASSES

Another index based on P_a is $\nu = aP_a$. It could be loosely interpreted as the number of classes a measurement system can discern with perfect precision (a precise interpretation is developed in the Appendix). If $P_a = 1/a$, then $\nu = 1$, indicating

This content downloaded from
181.58.0.99 on Sat, 06 Dec 2025 04:20:12 UTC
All use subject to https://about.jstor.org/terms

that the procedure can discern but one class (i.e., it gives no information). The maximal value of ν is a, indicating that the procedure can discern perfectly all a classes of the scale. Writing $\mathbf{P} = (p(1), \ldots, p(a))'$ and $\mathbf{Q} = [q(k|\ell)]$, consider this example.

$$\mathbf{P} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}.$$

In words: given a true value of 1 or 2, we will get a 1 or a 2 (with equal chances); given a true value of 3 or 4, we get a 3 or a 4. Here, $\nu = 4 \times 0.5 = 2$, indicating that the measurement system can perfectly distinguish between two classes. In this extreme case, these classes can be easily identified: combine 1 and 2 into a class, and 3 and 4. The given measurement system is as informative as

$$\mathbf{P} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

which also has $\nu = 2$. A more realistic example:

$$\mathbf{P} = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.3 \\ 0.3 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 0.47 & 0.47 & 0.03 & 0.03 \\ 0.47 & 0.47 & 0.03 & 0.03 \\ 0.05 & 0.05 & 0.45 & 0.45 \\ 0.05 & 0.05 & 0.45 & 0.45 \end{pmatrix},$$

has $\nu = 1.7$ ($P_a = 0.42$). This indicates that although the system measures on a four-point scale, it conveys less information than a system that perfectly distinguishes on a two-point scale. But the system is substantially more informative than chance measurements. In fact, the measurement system is precisely as informative as the following one, which measures on a two-point scale

$$\mathbf{P} = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 0.94 & 0.06 \\ 0.10 & 0.90 \end{pmatrix}.$$

Also in this case, $\nu = 1.7$ (but $P_a = 0.85$). The ν-index seems valuable especially for scales with a larger number of classes.

7. DISCUSSION AND CONCLUSION

All indices discussed in this article can be defended and have their use, but their meanings are not alike. We wish to demonstrate these differences from an example, trying to capture each index's meaning in nontechnical terminology. We study a measurement system producing measurement values on a five-point nominal scale. The distribution of the true values $\mathbf{P}$ and the distribution of the measurement values given the true value $\mathbf{Q}$ are given by

$$\mathbf{P} = \begin{pmatrix} 0.12 \\ 0.03 \\ 0.50 \\ 0.30 \\ 0.05 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 0.80 & 0.12 & 0.03 & 0.02 & 0.03 \\ 0.12 & 0.80 & 0.03 & 0.02 & 0.03 \\ 0.02 & 0.02 & 0.90 & 0.03 & 0.03 \\ 0.03 & 0.00 & 0.00 & 0.95 & 0.02 \\ 0.00 & 0.00 & 0.20 & 0.10 & 0.70 \end{pmatrix}.$$

The raw probability of agreement equals $P_a = 0.80$, which means the following. Given a randomly selected object, there is a 80% chance that two arbitrary appraisers assign the same value to it. This result is quite tangible, and therefore easily interpreted. This is fairly larger than $P_{a|c}^{\text{Unif}} = 0.2$, which indicates that chance measurements (done purely at random and with a uniform distribution) have a probability of agreement of 20%.

Fleiss's index κ^{Fleiss} equals 0.71 (based on $P_{a|c}^{\text{Fleiss}} = 0.33$). This number expresses the degree of (predictive) association between repeated measurements of an object, but only provided that objects are sampled from a population with an identical distribution as $\mathbf{P}$. The result is well below the ideal value of 1.0, but it is hard to give it a more tangible interpretation. This is a general problem of abstract association indices: the extreme values have a clear interpretation, but it is difficult to substantiate that the values in between convey information beyond the establishment that reproducibility is somewhere in between perfect and nil. It makes the question of how large or how small κ^{Fleiss} should be in order to indicate an acceptable reproducibility hopelessly arbitrary.

The kappa index based on uniform chance measurements is $\kappa^{\text{Unif}} = 0.75$. The value expresses the probability of agreement in excess of the agreement that maximally noninformative measurements (done randomly and uniformly distributed) would obtain, and normalized to the unit interval. Also this result is more abstract.

Finally, $\nu = 4.01$, which means that the information that this measurement system provides is comparable to that provided by a system that perfectly distinguishes on a four-point scale.

APPENDIX: INTERPRETATION OF ν

A precise formulation of the interpretation of $\nu = aP_a$ is as follows. We refer to the measurement system under study as system M (with a_M classes). Let its reproducibility be characterized by ν_M, and let $A = \lfloor \nu_M \rfloor$ (the largest integer strictly smaller than ν_M) and $B = \nu_M - A$. We construct a hypothetical system N that has identical reproducibility as M, but clearly interpretable properties. System N measures on a $(A+1)$-point scale. It is applied in a population of objects with distribution $\mathbf{P}_N = (1/(A+1), 1/(A+1), \ldots, 1/(A+1))$. The stochastic properties of N are specified by the $(A+1) \times (A+1)$ matrix

$$\mathbf{Q}_N = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & q & 1-q \\ 0 & \cdots & 0 & 1-q & q \end{pmatrix},$$

where $q = (1+\sqrt{B})/2$. System N classifies objects with perfect reproducibility in A classes (the last class combines objects with true values A and $A+1$). Objects classified in this combined class could be further subclassified into two classes with probability of agreement equal to $(B+1)/2$. Another property of N is that

This content downloaded from
181.58.0.99 on Sat, 06 Dec 2025 04:20:12 UTC
All use subject to https://about.jstor.org/terms

$\nu_N = \nu_M$. Namely,

$$\begin{aligned} \nu_N &= (A+1)\sum_{\ell=1}^{A+1} p(\ell)\sum_{k=1}^{A+1} q^2(k|\ell) \\ &= A - 1 + 2\left(q^2 + (1-q)^2\right) \\ &= A + B = \nu_M. \end{aligned}$$

These results can be interpreted as follows. The reproducibility of M is comparable to that obtained by measurement system N (i.e., $\nu_M = \nu_N$), which can perfectly classify objects into A classes, the last of which it could further subdivide into two subclasses into which it can distinguish objects with probability of agreement equal to $(B+1)/2$.

[Received October 2006. Revised February 2007.]

REFERENCES

Allen, M. J., and Yen, W. M. (1979), *Introduction to Measurement Theory*, Monterey: Brooks/Cole.

Bennett, E. M., Alpert, R., and Goldstein, A. C. (1954), "Communications Through Limited Response Questioning," *Public Opinion Quarterly*, 18, 303–308.

Brennan, R. L., and Prediger, D. J. (1981), "Coefficient Kappa: Some Uses, Misuses, and Alternatives," *Educational and Psychological Measurement*, 41, 687–699.

Burdick, R. K., Borror, C. M., and Montgomery, D. C. (2003), "A Review of Methods for Measurement Systems Capability Analysis," *Journal of Quality Technology*, 35, 342–354.

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

Conger, A. J. (1980), "Integration and Generalization of Kappas for Multiple Raters," *Psychological Bulletin*, 88, 322–328.

Davies, M., and Fleiss, J. L. (1982), "Measuring Agreement for Multinomial Data," *Biometrics*, 38, 1047–1051.

De Mast, J., and Van Wieringen, W. N. (in press), "Measurement System Analysis for Categorical Data: Agreement and Kappa Type Indices ," *Journal of Quality Technology*.

Feinstein, A. R., and Cicchetti, D. V. (1990), "High Agreement but Low Kappa," *Journal of Clinical Epidemiology*, 43, 543–549; 553–558.

Fleiss, J. L. (1971), "Measuring Nominal Scale Agreement Among Many Raters," *Pychological Bulletin*, 76, 378–382.

Gilula, Z., and Haberman, S. J. (1995), "Dispersion of Categorical Variables and Penalty Functions: Derivation, Estimation, and Comparability," *Journal of the American Statistical Association*, 90, 1447–1452.

Goodman, L. A., and Kruskal, W. H. (1954), "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, 49, 732–764.

Haberman, S. J. (1988), "Association, Measures of," in *Encyclopedia of Statistical Sciences* (vol. 1), eds. S. Kotz and N. L. Johnson, Chichester: Wiley.

Hershberger, S. L., and Fisher, D. G. (2005), "Measures of Association," in *Encyclopedia of Statistics in Behavioral Science* (Vol. 3), eds. B. Everitt and D. Howell, Chichester: Wiley, pp. 1183–1192.

ISO (1993), *Guide to the Expression of Measurement Uncertainty* (1st ed.), Geneva: International Organization for Standardization.

Kraemer, H. C. (1979), "Ramifications of a Population Model for κ as a Coefficient of Reliability," *Psychometrika*, 44, 461–472.

——— (1988), "Assessment of 2×2 Associations: Generalization of Signal-Detection Methodology," *The American Statistician*, 42, 37–49.

Kraemer, H. C., Periyakoil, V. S., and Noda, A. (2002), "Tutorial in Biostatistics: Kappa Coefficients in Medical Research," *Statistics in Medicine*, 21, 2109–2129.

Landis, J. R., and Koch, G. G. (1977), "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, 33, 159–174.

Lipsitz, S. R., Laird, N. M., and Brennan, T. A. (1994), "Simple Moment Estimates of the k-Coefficient and its Variance," *Applied Statistics*, 43, 309–323.

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

Tanner, M. A., and Young, M. A. (1985), "Modeling Agreement Among Raters," *Journal of the American Statistical Association*, 80, 175–180.

Thompson, W. D., and Walter, S. D. (1988), "A Reappraisal of the Kappa Coefficient," *Journal of Clinical Epidemiology*, 41, 949–958.

This content downloaded from
181.58.0.99 on Sat, 06 Dec 2025 04:20:12 UTC
All use subject to https://about.jstor.org/terms