Measurement System Analysis for Categorical Measurements: Agreement and Kappa-Type Indices

JEROEN DE MAST

Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands

WESSEL N. VAN WIERINGEN

Department of Mathematics, Vrije Universiteit, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands

The standard method for the assessment of the precision of measurement gauges—the gauge R&R method—is not applicable for measurement systems that measure on a nominal scale. For nominal measurements, the agreement method can be used, which expresses precision in the form of an index named κ. This paper aims to provide a consistent framework for the kappa method, based on statistical modeling. The defined probability models are used to study the statistical properties of proposed estimators. Various alternative forms of κ are discussed, as well as their relationship to the so-called paradoxes of kappa.

Key Words: Attribute Data; Gauge Capability; Kappa Coefficient; Nominal Data; Precision; Reproducibility.

An important aspect of measurement system analysis is the assessment of a measurement system's precision. The precision of a measurement system is its consistency across multiple measurements per object. The standard method to assess the precision of measurement systems that measure on a metric scale is the gauge R&R method (Burdick et al. (2003)). This papers deals with the precision of measurement systems that measure on a categorical scale. Categorical measurements can be nominal or ordinal (Allen and Yen (1979)). Both ordinal and nominal data are nonnumeric, and basic mathematical operations such as subtraction and addition are not defined. For ordinal scales (such as 'good', 'acceptable', 'bad') there is a defined order among the classes that it consists of. Nominal scales, on the contrary, consist of unordered classes (types A, B, C, for example).

This paper focuses on nominal measurements. A common example of nominal data in industry are classifications of production faults into defect types. Recording for each defect what sort of fault it is (using a classification system of prespecified types, such as 'machine related', 'operator related', 'material related', etc.), one collects information regarding which types of faults occur most frequently. Similarly, many companies classify complaints in complaint types, thus recording how frequently the various sorts of complaints occur. Classifying complaints into complaint types (and production faults into defect types) is measurement on a nominal scale. The 'measurement system' here is the helpdesk agents registering the complaints (or the operators filing the faults) plus the procedure and instructions that they use to do so. Note that also pass/fail inspection could be viewed as nominal measurement, and the methods discussed in this paper could be used to study their consistency.

To assess the precision of nominal measurements

Dr. de Mast is a Senior Consultant at the Institute for Business and Industrial Statistics (IBIS UvA) and Associate Professor at the University of Amsterdam. He is a member of the ASQ. His email address is jdemast@science.uva.nl.

Dr. van Wieringen is a Postdoctoral Scientist at the Vrije Universiteit of Amsterdam. His email address is wvanwie@few.vu.nl.

(that is, the consistency with which, for instance, defects or complaints are classified), one cannot resort to the standard gauge R&R method. With its interpretation of measurement precision based on the concept of standard deviation, it is not applicable to nominal scales. De Mast and Van Wieringen (2004) study the problem of measurement system analysis for ordinal measurements and Van Wieringen (2003) for binary measurements. This paper focuses on nominal measurements.

A widely applied and often discussed method to assess the precision of nominal measurements is the kappa method, or method of agreement. It assesses a measurement system's precision in terms of an index named κ, which was proposed originally by Cohen (1960). Originating from the fields of medical statistics, psychometrics, and biostatistics, the method has recently gained in popularity also in the practice of quality engineering and industrial statistics. It is presented in the AIAG Measurement System Analysis Manual (AIAG (2002)), part of the ASQ's Body of Knowledge for Six Sigma Black Belts, and included in software packages such as Minitab.

Upon reviewing the literature on the subject, we found the accounts that are given in the psychometrical and biostatistical literature not to provide a satisfactory basis for integration of the method in the industrial statistics and quality engineering sciences. The majority of articles on the method (e.g., Cohen (1960), Fleiss (1971), Conger (1980), Davies and Fleiss (1982)) describe the κ index almost exclusively in terms of sample statistics. They typically do not provide statistical models for the data, nor do they define the κ index as a population parameter. Because inferences based on the κ index typically concern the population and not just the sample, it is important to understand the sample κ index as an estimator for a population κ index. Moreover, grounding the kappa method in sound statistical modeling would enable the development of more precise definitions of terminology, concepts, and background assumptions, whereas the current literature abounds in imprecise formulations and lines of reasoning that are sometimes rather obscure.

The literature on latent-class models (e.g., Agresti and Lang (1993)) does provide probabilistic models for the study of the precision of nominal measurement systems. These models were developed against the background of cross-tabulations, and they typically model cell counts, whereas the models we propose in this paper model individual outcomes of measurements (the Y_{ij} in the notation introduced later). The latter results in probabilistic modeling, which links up better with the type of models that are used in the standard gauge R&R method. Furthermore, papers such as Agresti and Lang (1993) and Schuster and Smith (2002) focus on estimation of all model parameters (using the EM algorithm for example) and they ground inferences about measurement precision in the estimated parameters. The κ index is, in these approaches, sidestepped or mentioned briefly at best. The approach expounded in this paper, to the contrary, focuses on the definition and estimation of the κ index directly, rather than estimation of the model parameters.

It is the purpose of this paper to integrate measurement-system analysis methods based on agreement and kappa-types indices in the industrial statistics and quality-engineering literature. To do so, we critically assess the accounts provided by the psychometrical and biostatistical literatures, and in particular the papers by Cohen (1960), Fleiss (1971), Conger (1980), and Davies and Fleiss (1982). We aim to provide a consistent framework, based on sound statistical modeling, with precise definitions of assumptions and concepts and clearly described lines of reasoning. Moreover, the defined probability models will be used to study statistical properties of the estimators. The study of the kappa method from the perspective of the model that this paper proposes is—to our knowledge—novel, although Boyles (2001) and Van Wieringen (2003) use a similar type of model to study consistency of binary measurement systems.

For means of illustration, we will use an artificial data set. Suppose we consider classifications of complaints into five complaint types, which are numbered $1, 2, \ldots, 5$. Note that, although numerals are used to label categories, this is a nominal scale, and for that reason, the order among these categories has no meaning. To study to what extent employees classify complaints consistently, we could select a sample of 5 complaints (that is, the transcripts of the telephone conversation in which the complaint is put forward). Note that, for the sake of simplicity, the sample size is smaller than would be the case in a real precision assessment. Each of these complaints is presented to and classified by each of six appraisers. The data could look like the ones in Table 1. Because rows represent the classifications of the same complaint, minor variation within rows indicates that the classifications are quite consistent, whereas substantial variation implies a lack of consistency. The

TABLE 1. Fictitious Example of an Agreement Study

	Appraiser					
Complaint	1	2	3	4	5	6
1	1	2	1	1	1	1
2	2	2	2	2	3	3
3	4	4	4	4	4	4
4	2	1	3	1	1	1
5	3	3	3	3	3	3

next sections present a methodology for quantifying this type of assessment.

Statistical Modeling

In order to assess a measurement system's precision, one collects data from an experiment. Experiments for measurement-system analysis have a typical design: each of a collection of objects $i = 1, \ldots, n$ is measured repeatedly ($j = 1, \ldots, m$). These repetitions can be done by the same appraiser (in which case they are replications, and the assessment gives the intrarater consistency of the measurement system), or once by different appraisers (which gives interrater consistency), or repeatedly by different appraisers (allowing assessment of both intra- and interrater consistency). Returning to this matter in a later section, we assume an experimental design here in which each object is measured once by different appraisers. In the case of measurements on a nominal scale, each object is assigned a value from a finite and unordered set $\{1, 2, \ldots, a\}$. The measurement value that is assigned to object i by appraiser j is denoted Y_{ij}. If one conceives of nominal measurement as the classification of objects into categories, this value is the category to which the ith object is assigned by the jth appraiser.

We propose the following model for the data Y_{ij}. We have n objects, whose true values $X_1, \ldots, X_n$ we assume to be in the unordered set $\{1, 2, \ldots, a\}$. The X_i are assumed stochastically independent and have a discrete distribution with parameters

$$p(k) := P(X_i = k), \tag{1}$$

where

$$k = 1, \ldots, a, \text{ with } \sum_{k=1}^{a} p(k) = 1.$$

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. Moreover, the distribution of the Y_{i1}, Y_{i2}, …, Y_{im} depends on the true value X_i, and we define

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

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

$$P(Y_{ij} = k) = \sum_{\ell=1}^{a} p(\ell) q(k \mid \ell) =: q(k), \tag{3}$$

where $q(k)$ is the marginal distribution. If the concept of *true value* is thought to be problematic, it may help that this can be defined without getting stuck in ontological discussions as follows. The true value of an object (regarding the property under study) is the mean value that would be assigned to the object's property by an authoritative measurement system (such as the standard meter); see ISO (1993). The true values are a latent variable that induces a dependence structure in the data Y_{ij}. The true value underlying a categorical scale need not be categorical. However, if the true values would be continuous, this would likely induce an order in the categories, and one would expect an ordinal scale instead of a nominal one. Hence, the assumption that the underlying property is categorical.

If the repetitions $j = 1, \ldots, m$ are replications (that is, repeated ratings done by the same appraiser), the model defined by Equations (1) and (2) is the only option, but if repetitions correspond to measurements done by different appraisers, another option is to replace Equation (2) with

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

which gives a marginal distribution with probabilities

$$P(Y_{ij} = k) = \sum_{\ell=1}^{a} p(\ell) q_j(k \mid \ell) =: q_j(k). \tag{5}$$

We will refer to this model as the *heterogeneous appraisers model.*

Agreement, P_{a} and $\widehat{P}_{\mathrm{a}}$

Because standard deviation and correlation do not apply to nominal data, one has to express precision in terms of alternative concepts. In this paper, we will express precision of nominal measurements in terms of a probability of agreement. Two measurements of an object agree if they are identical. If a complaint is classified by two persons, there is agreement if both times the complaint is given the same classification. $P_{\text{agreement}}$ (or short: P_{a}) is the probability that two arbitrary measurements of an arbitrary object agree. Under the model specified by Equations (1) and (2), we have, for an object with true value $X_i = \ell$,

$$\begin{aligned} P_a(\ell) &:= P(Y_{ij_1} = Y_{ij_2} \mid X_i = \ell) \\ &= \sum_{k=1}^{a} q^2(k \mid \ell); \end{aligned}$$

and for an arbitrary object,

$$P_{\text{a}} := P(Y_{ij_1} = Y_{ij_2}) = \sum_{\ell=1}^{a}\sum_{k=1}^{a} p(\ell) q^2(k \mid \ell). \quad (6)$$

If one made the assumption that measurement consistency (and thus agreement) is homogeneous across objects, one would impose that $P_{\text{a}}(\ell) = P_{\text{a}}$ for $\ell = 1$, …, a, but usually this assumption is not made.

The definition of P_{a} is not given in this form in the literature. Instead, it (as well as the κ index) is typically defined as a sample statistic. We demonstrate that this statistic can be interpreted as an estimator of P_{a}. A statistic introduced by Fleiss (1971) is

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

(Fleiss's formula (3)), where $N_{ik} = \{\#j : Y_{ij} = k\}$. This statistic is an unbiased estimator of P_{a} because

$$\begin{aligned} \mathrm{E}\hat{P}_{\text{a}} &= \frac{1}{nm(m-1)} \sum_{i=1}^{n}\sum_{k=1}^{a} \mathrm{E}_{P_{X_i}} \mathrm{E}(N_{ik}(N_{ik}-1) \mid X_i) \\ &= \frac{m(m-1)}{nm(m-1)} \sum_{i=1}^{n}\sum_{k=1}^{a} \mathrm{E}_{P_{X_i}} q^2(k \mid X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n}\sum_{k=1}^{a}\sum_{\ell=1}^{a} p(\ell) q^2(k \mid \ell) = P_{\text{a}}. \end{aligned} \quad (7)$$

For the standard error of $\hat{P}_{\text{a}}$, we have

$$\begin{aligned} \mathrm{Var}(\hat{P}_{\text{a}}) = {} & \frac{1}{(nm(m-1))^2} \\ & \times \mathrm{EVar}\left(\sum_{i=1}^{n}\sum_{k=1}^{a} N_{ik}(N_{ik}-1) \middle| X_i \right) \\ & + \frac{1}{(nm(m-1))^2} \\ & \times \mathrm{VarE}\left(\sum_{i=1}^{n}\sum_{k=1}^{a} N_{ik}(N_{ik}-1) \middle| X_i \right) \\ = {} & A. + B. \end{aligned} \quad (8)$$

For the second term, we have (defining $M_k = \{\#i : X_i = k\}$)

$$\begin{aligned} B. &= \frac{1}{(nm(m-1))^2} \mathrm{Var} \sum_{i=1}^{n}\sum_{k=1}^{a} m(m-1) q^2(k \mid X_i) \\ &= \frac{1}{n^2} \mathrm{Var} \sum_{\ell=1}^{a}\sum_{k=1}^{a} M_\ell q^2(k \mid \ell) \\ &= \frac{1}{n} \sum_{\ell=1}^{a}\sum_{k=1}^{a} p(\ell)(1-p(\ell)) q^4(k \mid \ell). \end{aligned} \quad (9)$$

The first term equals (see Appendix A)

$$\begin{aligned} A. = {} & \frac{1}{nm(m-1)} \\ & \times \sum_{\ell=1}^{a}\sum_{k=1}^{a} p(\ell)(2q^2(k \mid \ell) + 4(m-2)q^3(k \mid \ell) \\ & \qquad + (6-4m)q^4(k \mid \ell)) \\ & + \frac{2}{nm(m-1)} \\ & \times \sum_{\ell=1}^{a}\sum_{k=1}^{a}\sum_{h=k+1}^{a} p(\ell)(6-4m) q^2(k \mid \ell) q^2(h \mid \ell). \end{aligned} \quad (10)$$

From the data in Table 1, the probability of agreement is estimated as $\hat{P}_{\text{a}} = 0.707$. This means the following: Given an arbitrary complaint, there is a 70.7% chance that two arbitrary appraisers give it the same classification.

Under the heterogeneous appraisers model (Equations (1) and (4)), one would define P_{a}^{HA} as the probability that two arbitrarily selected but different appraisers J_1 and J_2 agree (with $P(J_1 = j_1) = 1/m$ and $P(J_2 = j_2 \mid J_1 = j_1) = 1/(m-1)$ for $j_2 \neq j_1$ and 0 for $j_2 = j_1$). Thus, we define

$$\begin{aligned} P_{\text{a}}^{\text{HA}} &:= P(Y_{iJ_1} = Y_{iJ_2}) \\ &= \frac{2}{m(m-1)} \sum_{j_1=1}^{m}\sum_{j_2=j_1+1}^{m} P(Y_{ij_1} = Y_{ij_2}) \\ &= \frac{2}{m(m-1)} \\ &\quad \times \sum_{j_1=1}^{m}\sum_{j_2=j_1+1}^{m}\sum_{\ell=1}^{a}\sum_{k=1}^{a} p(\ell) q_{j_1}(k \mid \ell) q_{j_2}(k \mid \ell). \end{aligned} \quad (11)$$

Both Conger (1980) and Davies and Fleiss (1982) seem to assume this model. They propose an estimator that is identical to our $\hat{P}_{\text{a}}$ (Davies and Fleiss's formula (2)). A calculation similar to Equation (7) shows that, under the model defined by Equation (4), $\mathrm{E}\hat{P}_\mathrm{a} = P_\mathrm{a}^\mathrm{HA}$.

Kappa-Type Indices and Chance Agreement

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 measurement system has good precision. Even if appraisers would assign values to objects randomly, there would be some agreement. By chance alone, one would expect better agreement on a two-category scale than on a five-category one, and this makes it difficult to interpret values of P_a independent of the used scale. To deal with this problem, Cohen (1960), Fleiss (1971), Conger (1980), and numerous others have introduced κ-type indices as a rescaled version of P_a. The traditional formula is

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

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 (that is, 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 in this paper the more appropriate terminology,

$$\kappa = \frac{P_\mathrm{agreement} - P_\mathrm{agreement|chance}}{1 - P_\mathrm{agreement|chance}} \tag{12}$$

(resembling the terminology used by Lipsitz et al. (1994)).

Whereas the relevant range of P_a is

$$[P_\mathrm{agreement|chance}, 1],$$

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 $P_\mathrm{agreement|chance}$ (or short: $P_\mathrm{a|c}$). To do this rescaling, we have to define how we conceive of a chance measurement system (that is, 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 different rescaling and thus different κ indices. We present the most current alternatives here and discuss their benefits and drawbacks. Notions of 'random measurements' will be defined in the form of chance models, which describe the behavior of hypothetical 'chance' measurement systems and which allow us to define $P_\mathrm{a|c}$.

Uniform Chance Measurements, $P_\mathrm{a|c}^\mathrm{Unif}$ and κ^Unif

A definition going back to Bennett et al. (1954), and which is advocated by (among others) Brennan and Prediger (1981), is to define a chance measurement system as one that assigns values to objects completely at random (meaning: independent of the object's true value) and with a uniform distribution,

$$P(Z_{ij} = k) = 1/a \tag{13}$$

for all k and Z_{ij} mutually stochastically independent, where the Z_{ij} denote observations that we would get from such a measurement system. Based on this conception of a chance measurement system, we define

$$\begin{aligned} P_\mathrm{a|c}^\mathrm{Unif} &:= P(Z_{ij_1} = Z_{ij_2}) = \sum_{\ell=1}^{a} p(\ell) \sum_{k=1}^{a} 1/a^2 \\ &= 1/a, \end{aligned}$$

and substitution in Equation (12) gives

$$\kappa^\mathrm{Unif} = \frac{P_\mathrm{a} - 1/a}{1 - 1/a}. \tag{13}$$

The version in sample statistics is $\hat{\kappa}^\mathrm{Unif} = (\hat{P}_\mathrm{a} - P_\mathrm{a|c}^\mathrm{Unif})/(1 - P_\mathrm{a|c}^\mathrm{Unif}) = (a\hat{P}_\mathrm{a} - 1)/(a - 1)$, which is unbiased and has a standard error that can be easily calculated from Equation (8). For the example in Table 1, where $a = 5$, the probability of agreement that random classifications would obtain is $P_\mathrm{a|c}^\mathrm{Unif} = 1/a = 20.0\%$. Thus, the measurement procedure's agreement ($P_\mathrm{a} = 70.7\%$, as computed earlier) is substantially larger than the agreement of random classifications on a five-point scale. The sample kappa index is $\hat{\kappa}^\mathrm{Unif} = (0.707 - 0.200)/0.800 = 0.633$.

The value $1/a$ is a lower bound for P_a for measurement systems for which statistical properties follow Equations (1) and (2), as can be seen as follows. We solve a minimization problem in the $a(a-1)$ parameters $q(k \mid \ell)$, $k = 1, \ldots, a-1$; $\ell = 1, \ldots, a$ (given $q(1 \mid \ell), \ldots, q(a-1 \mid \ell)$, $q(a \mid \ell)$ is fixed, so it is not a parameter in the minimization problem). For each

$1 \leq k_0 \leq a-1$ and $1 \leq \ell_0 \leq a$, we calculate the partial derivative of P_{a} with respect to $q(k_0 \mid \ell_0)$ as

$$\frac{dP_{\mathrm{a}}}{dq(k_0 \mid \ell_0)} = 2p(\ell_0)\left(q(k_0 \mid \ell_0) - \left(1 - \sum_{k=1}^{a-1} q(k \mid \ell_0)\right)\right).$$

Equating these to zero, we get, for each ℓ_0, the $a-1$ equations $q(1 \mid \ell_0) = q(2 \mid \ell_0) = \ldots = q(a-1 \mid \ell_0) = 1 - \sum_{k=1}^{a-1} q(k \mid \ell_0)$. The unique solution is $q(k_0 \mid \ell_0) = 1/a$ for all k_0 and ℓ_0. Thus, of all measurement systems that can be modeled as in Equations (1) and (2), the one with the lowest P_{a} is given by $P(Y_{ij} = k) = 1/a$ for all k, which happens to be equivalent to the chance measurement system defined in Equation (13), and the corresponding probability of agreement is $P_{\mathrm{a}} = 1/a$, which is, of course, identical to $P_{\mathrm{a|c}}^{\mathrm{Unif}}$ as defined above.

Several objections against Equation (13) are raised in the literature. Scott (1955) states 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." Granting that this objection sounds convincing at first, closer scrutiny shows that it is hard to understand precisely what the argument is. In our notation, Scott claims that the definition of a chance measurement system as in Equation (13) would be based on the assumption that $p(\ell) = 1/a$ for $\ell = 1, \ldots, a$ (or else, that it is based on the assumption that the $q(k \mid \ell)$ or $q(k)$ are all equal to $1/a$). It should be realized that a chance measurement system is a hypothetical concept: appraisers are not really assigning values at random, we are just hypothesizing what it would look like if they did, to use this as a reference. In the κ index, we compare two measurement systems: a real system with parameters $p(\ell)$ and $q(k \mid \ell)$ (neither of which is assumed to be distributed uniformly, nor are the $q(k)$) and a hypothetical chance system that is conceived of as having a uniform distribution. In other words, the uniform distribution is not used to model the real system under study, but as a notion of what we understand by 'random'. One can disagree that 'random' should mean 'uniformly distributed', but one cannot claim that the model in Equation (13) implies that either the $p(\ell)$ or the $q(k \mid \ell)$ (nor the $q(k)$) are uniformly distributed, because that is neither implied nor assumed.

A second objection made by Scott (1955) is that adding more categories to the nominal scale results in a larger value of κ even if these extra categories are not used by the appraisers. The fact that P_{a} is downscaled more for small a than for large a is not objected to by Scott: "By chance alone, one would expect better agreement on a two-category than on a five-category scale", and the correction of this phenomenon is precisely the purpose of using κ instead of P_{a}. But does this create a spurious effect if some categories are never used? Let us make matters precise. Suppose we start with a scale having $a_0 = 2$ categories and that $P_{\mathrm{a}} = 0.5$. Equation (13) gives that $\kappa^{\mathrm{Unif}} = 0$, meaning that this measurement system has a precision comparable to the precision one would expect from a chance measurement system having two categories. Now we add three more categories ($a_1 = 5$), but we assume that these are never used by the appraisers (that is, $q(k \mid \ell) = 0$ for $k \geq 3$). The same probability of agreement $P_{\mathrm{a}} = 0.5$ now corresponds to $\kappa^{\mathrm{Unif}} = 0.38$. This result can be justified: $P_{\mathrm{a}} = 0.5$ on a five-category scale *is* more precise than $P_{\mathrm{a}} = 0.5$ on a two-category scale (because it distinguishes on a finer scale), but if this extra precision is caused by the fact that the measurement system returns only two categories, this measurement system has another problem: its accuracy (or validity) is poor.

The extreme version of this line of thought perhaps makes the counterpoint even more clear: consider a measurement system with the following statistical properties ($a = 5$):

$$\text{For all } \ell = 1, \ldots, 5: \quad q(1 \mid \ell) = 0.99;\ q(k \mid \ell) = 0.0025 \quad \text{for } k = 2, 3, 4, 5 \qquad (15)$$

(i.e., a measurement system that 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 $\kappa^{\mathrm{Unif}} = 0.96$. But on second thought, the precision of this measurement system actually is very good: measurement spread is practically nil, and the repeatability and consistency are almost 100%. The measurement system has another problem, namely its accuracy (or validity). The analogue for numerical measurement systems is the case that a system returns the value 3.1415 (say) independent of the object being measured. The measurement spread is zero, and hence the precision is perfect, but its accuracy is poor. The κ index defined in Equation (13) only measures precision (or repeatability, reliability, consistency), not confounding this aspect with accuracy or other aspects of the quality of measurement systems. For practical purposes, keeping precision and accuracy separated makes sense.

Assuming the heterogeneous appraisers model (Equations (1), (4)) we get

$$\kappa^{\text{Unif,HA}} = \frac{P_{\text{a}}^{\text{HA}} - 1/a}{1 - 1/a}.$$

Fleiss's Chance Model, $P_{\text{a|c}}^{\text{Fleiss}}$ and κ^{Fleiss}

The operational definition in Equation (13) of a chance measurement system is to some extent arbitrary, and other definitions are current. Fleiss (1971) implicitly defines the chance measurement system as one that follows the model

$$P(Z_{ij} = k) = r(k), \tag{16}$$

where the Z_{ij} are stochastically independent. (Note that this is our elaboration of Fleiss's briefly formulated ideas). Under the chance model in Equation (16), we get

$$P_{\text{a|c}}^{\text{Fleiss}} := P(Z_{ij_1} = Z_{ij_2}) = \sum_{k=1}^{a} r^2(k)$$

and

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

This definition of the chance measurement system is underdetermined: It specifies a system up to the parameters $r(k)$, and because the chance measurement system is a hypothetical entity, we cannot collect data Z_{ij} from it to estimate them. Although Fleiss (1971) does not state so explicitly, the elaboration of his approach suggests that the idea is to equate them to the marginal distribution of the measurement system under study, that is,

$$r(k) := q(k), \qquad k = 1, \ldots, a,$$

and the $q(k)$ can, of course, be estimated from the Y_{ij}. Thus, a chance measurement system is conceived of as one that classifies objects randomly (independent of the objects' true values) but with a distribution equal to the marginal distribution of the measurement system under study when it is applied to the population of objects under study. As an estimator, Fleiss proposes (his formula (5))

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

with $N_k = \{\#(i,j) : Y_{ij} = k\}$ and

$$\hat{\kappa}^{\text{Fleiss}} = \frac{\hat{P}_{\text{a}} - \hat{P}_{\text{a|c}}^{\text{Fleiss}}}{1 - \hat{P}_{\text{a|c}}^{\text{Fleiss}}}$$

(Fleiss's formula (7)). This is the κ index that Minitab (version 14) computes. From the data in Table 1, we compute $\hat{P}_{\text{a|c}}^{\text{Fleiss}} = 0.260$ and therefore $\hat{\kappa}^{\text{Fleiss}} = (0.707 - 0.260)/0.740 = 0.604$. Note that $\hat{P}_{\text{a|c}}^{\text{Fleiss}}$ is not an unbiased estimator:

$$\text{E}\hat{P}_{\text{a|c}}^{\text{Fleiss}} = \frac{n-1}{n} \sum_{k=1}^{a} q^2(k) + \frac{m-1}{mn} P_{\text{a}} + \frac{1}{mn} \tag{17}$$

(see Appendix B) and therefore,

$$\text{E}\hat{P}_{\text{a|c}}^{\text{Fleiss}} = \frac{n-1}{n} P_{\text{a|c}}^{\text{Fleiss}} + \frac{m-1}{mn} P_{\text{a}} + \frac{1}{mn}.$$

Conceiving of random measurements as in Equation (16) instead of Equation (13) is more natural for some, and it is a valid option. It simply means that one uses a different conception of a chance measurement system as a reference. Below we discuss important differences in interpretation between κ^{Unif} and κ^{Fleiss}, providing the reader with enough material to make up his own mind as to which version he prefers.

Both κ^{Unif} and κ^{Fleiss} depend on the distribution $(p(1), \ldots, p(a))$ of true values in the population. There are two distinct reasons for this dependence. The first reason is that the measurement system is not necessarily equally consistent for each object. For this reason, P_{a} was defined as a weighted average of the probabilities of agreement given an object's true value (Equation (6)). However, if the probability of agreement were homogeneous (i.e., $P_{\text{a}}(\ell) = P_{\text{a}}$ for all ℓ), then one would want a κ index to be independent of properties of the population of objects. This is the case for κ^{Unif}, but for κ^{Fleiss}, there is an additional dependence on $(p(1), \ldots, p(a))$. Consider the following example, with a nominal scale of $a = 2$ categories. A measurement system's statistical properties are given by $q(1 \mid 1) = 0.95$, $q(2 \mid 1) = 0.05$, $q(1 \mid 2) = 0.05$, and $q(2 \mid 2) = 0.95$ (the probability of agreement $P_{\text{a}} = 0.91$ 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 $P_{\text{a|c}}^{\text{Fleiss}} = \sum q^2(k) = 0.50$ 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 $P_{\text{a|c}}^{\text{Fleiss}} = 0.83$ and $\kappa^{\text{Fleiss}} = 0.45$ (note that $\kappa^{\text{Unif}} = 0.81$ in both cases). Thus, κ^{Fleiss} depends strongly on $p(1)$ and $p(2)$ (or *prevalence*, as it is called in epidemiology) and confounds measurement precision with properties of the population of objects. This means that the precision of a measurement system is expressed only in relation to a certain population of objects.

A second difference between κ^{Unif} and κ^{Fleiss} comes to light if we apply both procedures to a measurement system virtually always returning the value 1 independent of the object being measured (as defined in Equation (15)). We have shown that $\kappa^{\text{Unif}} = 0.96$, and the interpretation is that this system indeed has perfect precision, but its accuracy is very poor. But for this system, $\kappa^{\text{Fleiss}} = 0$, and we see that κ^{Fleiss} confounds precision and accuracy.

What is undeniably an undesirable property of κ^{Fleiss} and its estimator $\hat{\kappa}^{\text{Fleiss}}$ is that 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 a population of objects with distribution $p(1) = 0.95$, $p(2) = 0.05$ (on an $a = 2$ point scale), a measurement system with properties $q(1 \mid 1) = 1.0$, $q(2 \mid 1) = 0.0$, $q(1 \mid 2) = 0.0$, and $q(2 \mid 2) = 1.0$ gives $\kappa^{\text{Fleiss}} = 1.0$; but a measurement system with properties $q(1 \mid 1) = 0.95$, $q(2 \mid 1) = 0.05$, $q(1 \mid 2) = 0.05$, and $q(2 \mid 2) = 0.95$ gives $\kappa^{\text{Fleiss}} = 0.45$. The strong sensitivity of κ^{Fleiss} for small changes in the $q(k \mid \ell)$ 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

$$\begin{pmatrix} 99 & 0 \\ 0 & 1 \end{pmatrix}.$$

This table should be read as

$$\begin{pmatrix} \#i : Y_{i1} = Y_{i2} = 1 & \#i : Y_{i1} = 1, Y_{i2} = 2 \\ \#i : Y_{i1} = 2, Y_{i2} = 1 & \#i : Y_{i1} = Y_{i2} = 2 \end{pmatrix}.$$

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

$$\begin{pmatrix} 98 & 1 \\ 0 & 1 \end{pmatrix}$$

would give $\hat{\kappa}^{\text{Fleiss}} = 0.66$ (the corresponding values of $\hat{\kappa}^{\text{Unif}}$ are 1.0 and 0.98, respectively). This behavior is seen by many as leading to results that are difficult to interpret, and these problems have come to be known under the name 'paradoxes of the kappa' (so called by Feinstein and Cicchetti (1990), but the dependence of κ^{Fleiss} on prevalence was described earlier by Thompson and Walter (1988)).

Conger's Chance Model, $P_{\text{a}|\text{c}}^{\text{Conger}}$ and κ^{Conger}

A form of the κ index that is also commonly used was defined by Conger (1980) and Davies and Fleiss (1982). This index is the adaptation of κ^{Fleiss} to the heterogeneous appraisers model (Equations (1) and (4)), and thus uses P_{a}^{HA} in its numerator. The chance model is

$$P(Z_{ij} = k) = r_j(k) \qquad (Z_{ij} \text{ stochastically independent}). \tag{18}$$

For given appraisers j_1 and j_2, we have $P(Z_{ij_1} = Z_{ij_2}) = \sum_{k=1}^{a} r_{j_1}(k) r_{j_2}(k)$. For randomly selected j_1 and j_2, we have

$$P_{\text{a}|\text{c}}^{\text{Conger}} = \frac{2}{m(m-1)} \sum_{j_1=1}^{m-1} \sum_{j_2=j_1+1}^{m} \sum_{k=1}^{a} r_{j_1}(k) r_{j_2}(k)$$

(which is equivalent to the definition of Davies and Fleiss). As the model in Equation (16), also the model in Equation (18) is underdetermined. As before, the parameters of the chance measurement system are related to the marginal distribution of the measurement system under study: $r_j(k) := q_j(k)$ (see Equation (5)). We have

$$\kappa^{\text{Conger}} = \frac{P_{\text{a}}^{\text{HA}} - P_{\text{a}|\text{c}}^{\text{Conger}}}{1 - P_{\text{a}|\text{c}}^{\text{Conger}}}.$$

As an estimator for $P_{\text{a}|\text{c}}^{\text{Conger}}$, Davies and Fleiss (1982) propose (their formulas (3) and (4))

$$\hat{P}_{\text{a}|\text{c}}^{\text{Conger}} = \frac{2}{m(m-1)} \sum_{j1=1}^{m} \sum_{j2=j1+1}^{m} \sum_{k=1}^{a} \frac{L_{j1k}}{n} \frac{L_{j2k}}{n}$$

(for each j and k: $L_{jk} = \{\#i : Y_{ij} = k\}$). It is a biased estimator:

$$\begin{aligned} &\mathrm{E}\hat{P}_{\text{a}|\text{c}}^{\text{Conger}} \\ &\quad = \frac{n-1}{n} P_{\text{a}|\text{c}}^{\text{Conger}} \\ &\qquad + \frac{1}{n} \sum_{j_1=1}^{m} \sum_{j_2=j_1+1}^{m} \sum_{k=1}^{a} \sum_{\ell=1}^{a} p(\ell) q_{j_1}(k \mid \ell) q_{j_2}(k \mid \ell) \end{aligned}$$

(see Appendix C). Davies and Fleiss's κ index is defined as

$$\hat{\kappa}^{\text{Conger}} = \frac{\hat{P}_{\text{a}} - \hat{P}_{\text{a}|\text{c}}^{\text{Conger}}}{1 - \hat{P}_{\text{a}|\text{c}}^{\text{Conger}}}.$$

It has $\hat{P}_{\text{a}}$ in its numerator because, under the heterogeneous appraisers model (Equations (1) and (4)), $\mathrm{E}\hat{P}_{\text{a}} = P_{\text{a}}^{\text{HA}}$ (see below, Equation (11)).

From the data in Table 1, we get $\hat{P}^{\text{Conger}}_{\text{a|c}} = 0.251$ and therefore $\hat{\kappa}^{\text{Conger}} = (0.707 - 0.251)/0.749 = 0.609$. Note that, for $m = 2$, $\hat{\kappa}^{\text{Conger}}$ reduces to the original definition of the κ index by Cohen (1960). Besides being applicable only if the repetitions $j = 1, \ldots, m$ relate to the levels of a factor and are not just replications, the qualifications made about the behavior of κ^{Fleiss} apply to κ^{Conger} as well.

Intra- and Interrater Agreement

So far, we have studied the analysis of experiments in which each object is measured once by different appraisers, allowing the estimation of interrater consistency. Here we discuss briefly how to proceed in the situation that each object is measured $s \geq 2$ times by each of $m \geq 2$ appraisers. The data are denoted Y_{ijh}, $i = 1, \ldots, n$; $j = 1, \ldots, m$; $h = 1, \ldots, s$. For the true values X_i, the model in Equation (1) is retained. The data are now modeled as

$$q_j(k \mid \ell) := P(Y_{ijh} = k \mid X_i = \ell).$$

For each of the m appraisers, we can define a probability of agreement, namely,

$$\begin{aligned} P_{a,\text{intra}}(j) &:= P(Y_{ijh_1} = Y_{ijh_2}) \\ &= \sum_{\ell=1}^{a} \sum_{k=1}^{a} p(\ell) q_j^2(k \mid \ell). \end{aligned}$$

The probability that a randomly selected appraiser agrees with himself can be defined as follows. Let J be the randomly selected appraiser, and let $P(J = j) = 1/m$ for all j, then

$$\begin{aligned} P_{a,\text{intra}} &:= P(Y_{iJh_1} = Y_{iJh_2}) \\ &= \frac{1}{m} \sum_{j=1}^{m} \sum_{\ell=1}^{a} \sum_{k=1}^{a} p(\ell) q_j^2(k \mid \ell), \end{aligned}$$

which could be interpreted as intrarater agreement. It can be estimated (unbiasedly) by

$$\hat{P}_{a,\text{intra}} = \frac{1}{mns(s-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} \sum_{k=1}^{a} N_{ik}(j)(N_{ik}(j) - 1),$$

with $N_{ik}(j) = \{\#h : Y_{ijh} = k\}$.

Interrater agreement is the probability that two different appraisers agree. For two randomly selected but different appraisers J_1 and J_2, we have

$$\begin{aligned} P_{\text{a,inter}} &:= P(Y_{iJ_1h_1} = Y_{iJ_2h_2}) \\ &= \frac{2}{m(m-1)} \sum_{j_1=1}^{m} \sum_{j_2=j_1+1}^{m} P(Y_{ij_1h_1} = Y_{ij_2h_2}) \\ &= \frac{2}{m(m-1)} \\ &\quad \times \sum_{j_1=1}^{m} \sum_{j_2=j_1+1}^{m} \sum_{\ell=1}^{a} \sum_{k=1}^{a} p(\ell) q_{j_1}(k \mid \ell) q_{j_2}(k \mid \ell), \end{aligned}$$

which could be estimated (unbiasedly) by

$$\begin{aligned} \hat{P}_{\text{a,inter}} &= \frac{2}{m(m-1)ns^2} \\ &\quad \times \sum_{i=1}^{n} \sum_{j_1=1}^{m} \sum_{j_2=j_1+1}^{m} \sum_{k=1}^{a} N_{ik}(j_1) N_{ik}(j_2). \end{aligned}$$

Rescaling to correct for chance agreement, one could define κ-type indices along the lines discussed earlier. Furthermore, one could consider defining an overall κ index based on

$$P_{\text{a,overall}} = P(J_1 = J_2) P_{\text{a,intra}} + P(J_1 \neq J_2) P_{\text{a,inter}}.$$

Discussion and Conclusion

The current literature in quality engineering on methods for the assessment of precision of nominal measurements is limited. In the biostatistical, medical, and psychometrical sciences, to the contrary, the literature on this subject is extensive and can be used as a basis for the development of methods suitable for quality engineering. This paper critically reviews methods based on agreement and κ-type indices, and especially the model-based development in this paper is new.

Nominal scales are equipped with only the simplest arithmetical operations and relations, such as equivalence ('='). For this reason, it is hard to imagine that metrics for precision of nominal measurements can be based on concepts that are substantially more advanced than agreement. Whether and how P_{a} should be rescaled, however, is a more controversial matter.

Whether one finds the logic underlying κ^{Unif}, or κ^{Fleiss} and κ^{Conger}, convincing or not, an equally important angle is to look at the effects of either definition. The first important difference is that $P^{\text{Fleiss}}_{\text{a|c}}$ and $P^{\text{Conger}}_{\text{a|c}}$ are based on a property of the population of measured objects. This makes the rescaling that κ^{Fleiss} and κ^{Conger} apply specific for a population of objects. The adjustment applied in κ^{Unif} only depends on properties of the measurement system under study (namely its scale). Second, κ^{Unif} separates precision from issues related to a measurement system's accuracy, while these issues are confounded in κ^{Fleiss} and κ^{Conger}.

The account so far has focused on explication and clarification of prevailing ideas and approaches in the literature, less emphasizing the evaluation of these ideas. We hope that this account enables the industrial-statistics and quality-engineering sciences to study these ideas effectively and develop their own approaches. In this final section, we discuss what we think will work in the context of quality technology.

The problem with many indices for precision is that, although the extreme values have a clear interpretation, the intermediate values are less clearly translated into real-life implications. This makes the question of how large or small an index for precision should be in order to indicate an acceptable measurement system hopelessly arbitrary. Turning to the κ index, it is especially P_{a} (more than κ itself) that is easily interpretable in real-life terms: the chance that repeated measurements are identical is quite a tangible expression of precision. Normalizing P_{a} into a κ index turns it into a more abstract number. For that reason, we would prefer to assess a measurement system's precision in terms of two numbers: P_{a}, because it is a tangible quantity, and $P_{\text{a}|\text{c}}^{\text{Unif}}$ (or, if that is one's preference, $P_{\text{a}|\text{c}}^{\text{Fleiss}}$ or $P_{\text{a}|\text{c}}^{\text{Conger}}$) to have a reference. Even if one combines these two in a single index, we would recommend reporting the intermediate results P_{a} and $P_{\text{a}|\text{c}}$ as well.

The kappa/agreement method originates in the methodology of diagnostic tests. Sensitivity and specificity are two other concepts that are frequently used to express the useability of these tests. They are defined as (see, for instance, Ingelfinger et al. (1983))

$$\begin{aligned}&\text{Sensitivity}\\ &\quad = P(\text{positive diagnosis} \mid \text{disorder is present})\\ &\text{Specificity}\\ &\quad = P(\text{negative diagnosis} \mid \text{disorder is not present}).\end{aligned}$$

Agreement and sensitivity/specificity do not address identical issues. Agreement is a measure exclusively for precision. Sensitivity and specificity combine precision and accuracy: sensitivity and specificity can be poor due to poor precision, poor accuracy, or a combination of both. More in general, we are in favor of evaluating a measurement's precision and accuracy separately. Therefore, we would ourselves not use indices that confound these two aspects. This is one of the reasons why we would prefer κ^{Unif} to κ^{Fleiss} or κ^{Conger}.

Appendix A

Let $(N_1, N_2, \ldots, N_{\text{a}})$ have a multinomial $(m;\, q_1, q_2, \ldots, q_{\text{a}})$ distribution. Then

$$\begin{aligned}&\text{Var}(N_k(N_k-1))\\ &\qquad = 2m^{(2)}q_k^2 + 4m^{(3)}q_k^3 + \left(m^{(4)} - m^{(2)}m^{(2)}\right)q_k^4,\end{aligned}$$

because the lower moments of the multinomial distribution are

$$\begin{aligned}\text{E}N_k &= mq_k\\ \text{E}N_k^2 &= mq_k + m^{(2)}q_k^2\\ \text{E}N_k^3 &= mq_k + 3m^{(2)}q_k^2 + m^{(3)}q_k^3\\ \text{E}N_k^4 &= mq_k + 7m^{(2)}q_k^2 + 6m^{(3)}q_k^3 + m^{(4)}q_k^4.\end{aligned}$$

Furthermore, we have

$$\begin{aligned}&\text{Cov}(N_k(N_k-1); N_h(N_h-1))\\ &\qquad = \text{E}N_k^{(2)}N_h^{(2)} - m(m-1)q_h^2\text{E}N_k^{(2)}\\ &\qquad\quad - m(m-1)q_k^2\text{E}N_h^{(2)} + m^2(m-1)^2q_k^2q_h^2\\ &\qquad = (m^{(4)} - m^{(2)}m^{(2)})q_k^2q_h^2,\end{aligned}$$

because (mixed factorial moments)

$$\text{E}(N_1^{(r_1)} \cdots N_{\text{a}}^{(r_{\text{a}})}) = m^{(\sum r_i)} q_1^{r_1} \cdots q_{\text{a}}^{r_{\text{a}}}.$$

Combing these results, we get

$$\begin{aligned}&\text{E}\sum_{i=1}^{n}\text{Var}\left(\sum_{k=1}^{a}N_{ik}(N_{ik}-1)\middle| X_i\right)\\ &\quad = \text{E}\sum_{i=1}^{n}\sum_{k=1}^{a}\text{Var}(N_{ik}(N_{ik}-1) \mid X_i)\\ &\qquad + 2\text{E}\sum_{i=1}^{n}\sum_{k=1}^{a}\sum_{h=k+1}^{a}\text{Cov}\big(N_{ik}(N_{ik}-1);\\ &\qquad\qquad N_{ih}(N_{ih}-1)\big|X_i\big)\\ &\quad = \sum_{i=1}^{n}\sum_{k=1}^{a}\text{E}\Big(2m^{(2)}q^2(k \mid X_i)\\ &\qquad\qquad + 4m^{(3)}q^3(k \mid X_i)\\ &\qquad\qquad + (m^{(4)} - m^{(2)}m^{(2)})q^4(k \mid X_i)\Big)\\ &\qquad + 2\sum_{i=1}^{n}\sum_{k=1}^{a}\sum_{h=k+1}^{a}\text{E}(m^{(4)} - m^{(2)}m^{(2)})\\ &\qquad\qquad \times q^2(k \mid X_i)q^2(h \mid X_i)\\ &\quad = nm^{(2)}\sum_{\ell=1}^{a}\sum_{k=1}^{a}p(\ell)\Big(2q^2(k \mid \ell)\\ &\qquad\qquad + 4(m-2)q^3(k \mid \ell)\end{aligned}$$

$$+ (6 - 4m)q^4(k \mid \ell)\Big)$$

$$+ 2nm^{(2)} \sum_{\ell=1}^{a} \sum_{k=1}^{a} \sum_{h=k+1}^{a} p(\ell)(6 - 4m)q^2(k \mid \ell) \times q^2(h \mid \ell).$$

Appendix B

$$\begin{aligned}
&\mathrm{E}\hat{P}^{\text{Fleiss}}_{\text{a}|\text{c}} \\
&= \frac{1}{m^2 n^2} \sum_{k=1}^{a} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \mathrm{E} N_{i_1 k} N_{i_2 k} \\
&= \frac{1}{m^2 n^2} \\
&\quad \times \sum_{k=1}^{a} \sum_{i_1=1}^{n} \sum_{\substack{i_2=1 \\ i_2 \neq i_1}}^{n} \sum_{\ell_1=1}^{a} \sum_{\ell_2=1}^{a} p(\ell_1)p(\ell_2) \\
&\qquad \times \mathrm{E}(N_{i_1 k} N_{i_2 k} \mid \\
&\qquad\qquad X_{i_1} = \ell_1, X_{i_2} = \ell_2) \\
&\quad + \frac{1}{m^2 n^2} \sum_{k=1}^{a} \sum_{i=1}^{n} \sum_{\ell=1}^{a} p(\ell) \mathrm{E}(N_{ik}^2 \mid X_i = \ell) \\
&= \frac{n-1}{n} \sum_{k=1}^{a} \sum_{\ell_1=1}^{a} \sum_{\ell_2=1}^{a} p(\ell_1)p(\ell_2)q(k \mid \ell_1)q(k \mid \ell_2) \\
&\quad + \frac{1}{mn} \sum_{k=1}^{a} \sum_{\ell=1}^{a} p(\ell)\Big(q(k \mid \ell) + (m-1)q^2(k \mid \ell)\Big) \\
&= \frac{n-1}{n} \sum_{k=1}^{a} q^2(k) + \frac{m-1}{mn} P_{\text{a}} + \frac{1}{mn}.
\end{aligned}$$

Appendix C

Defining

$$N_{ijk} = \begin{cases} 1, & \text{if } Y_{ij} = k \\ 0, & \text{if } Y_{ij} \neq k \end{cases},$$

we have

$$\begin{aligned}
&\frac{m(m-1)}{2} \mathrm{E}\hat{P}^{\text{Conger}}_{\text{a}|\text{c}} \\
&= \frac{1}{n^2} \sum_{j_1=1}^{m} \sum_{j_2=j_1+1}^{m} \sum_{k=1}^{a} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \mathrm{E} N_{i_1 j_1 k} N_{i_2 j_2 k} \\
&= \frac{1}{n^2} \sum_{\substack{j_1, j_2, k, \\ i_1 \neq i_2}} \sum_{\ell_1=1}^{a} \sum_{\ell_2=1}^{a} p(\ell_1)p(\ell_2) \mathrm{E}(N_{i_1 j_1 k} N_{i_2 j_2 k} \mid \\
&\qquad\qquad X_{i_1} = \ell_1, X_{i_2} = \ell_2) \\
&\quad + \frac{1}{n^2} \sum_{j_1=1}^{m} \sum_{j_2=j_1+1}^{m} \sum_{k=1}^{a} \sum_{i=1}^{n} \sum_{\ell=1}^{a} p(\ell) \mathrm{E}(N_{i j_1 k} N_{i j_2 k} \mid \\
&\qquad\qquad X_i = \ell) \\
&= \frac{n-1}{n} \sum_{j_1} \sum_{j_2} \sum_{k} \sum_{\ell_1} \sum_{\ell_2} p(\ell_1)p(\ell_2)q_{j_1}(k \mid \ell_1) \\
&\qquad \times q_{j_2}(k \mid \ell_2) \\
&\quad + \frac{1}{n} \sum_{j_1} \sum_{j_2} \sum_{k} \sum_{\ell} p(\ell)q_{j_1}(k \mid \ell)q_{j_2}(k \mid \ell) \\
&= \frac{n-1}{n} \sum_{j_1} \sum_{j_2} \sum_{k} q_{j_1}(k)q_{j_2}(k) \\
&\quad + \frac{1}{n} \sum_{j_1} \sum_{j_2} \sum_{k} \sum_{\ell} p(\ell)q_{j_1}(k \mid \ell)q_{j_2}(k \mid \ell).
\end{aligned}$$

Acknowledgment

This work was supported by the Center for Medical Systems Biology (CMSB) established by the Netherlands Genomics Institute/Netherlands Organization for Scientific Research (NGI/NWO).

References

AGRESTI, A. and LANG, J. B. (1993). "Quasi-Symmetric Latent Class Models, with Application to Rater Agreement". *Biometrics* 49, pp. 131–139.

AIAG (2002). *Measurement System Analysis; Reference Manual*, 3rd ed. Automotive Industry Action Group, Detroit, MI.

ALLEN, M. J. and YEN, W. M. (1979). *Introduction to Measurement Theory*. Brooks/Cole, Monterey, CA.

BENNETT, E. M.; ALPERT, R.; and GOLDSTEIN, A. C. (1954). "Communications Through Limited Response Questioning". *Public Opinion Quarterly* 18(3), pp. 303–308.

BOYLES, R. A. (2001). "Gauge Capability for Pass–Fail Inspection". *Technometrics* 43(2), pp. 223–229.

BRENNAN, R. L. and PREDIGER, D. J. (1981). "Coefficient Kappa: Some Uses, Misuses, and Alternatives". *Educational and Psychological Measurement* 41, pp. 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(4), pp. 342–354.

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

CONGER, A. J. (1980). "Integration and Generalization of Kappas for Multiple Raters". *Psychological Bulletin* 88(2), pp. 322–328.

DAVIES, M. and FLEISS, J. L. (1982). "Measuring Agreement for Multinomial Data". *Biometrics* 38(4), pp. 1047–1051.

DE MAST, J. and VAN WIERINGEN, W. N. (2004). "Measurement System Analysis for Bounded Ordinal Data". *Quality and Reliability Engineering International* 20(5), pp. 383–395.

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

Fleiss, J. L. (1971). “Measuring Nominal Scale Agreement Among Many Raters”. *Pychological Bulletin* 76(5), pp. 378–382.

Ingelfinger, J. A.; Mosteller, F.; Thibodeau, L. A.; and Ware, J. H. (1983). *Biostatistics in Clinical Medicine.* Macmillan Publishing, New York, NY.

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

Lipsitz, S. R.; Laird, N. M.; and Brennan, T. A. (1994). “Simple Moment Estimates of the κ Coefficient and Its Variance”. *Applied Statistics* 43(2), pp. 309–323.

Schuster, C. and Smith, D. A. (2002). “Indexing Systematic Rater Agreement with a Latent-Class Model”. *Psychological Methods* 7(3), pp. 384–395.

Scott, W. A. (1955). “Reliability of Content Analysis: The Case of Nominal Scale Coding”. *Public Opinion Quarterly* 19(3), pp. 321–325.

Thompson, W. D. and Walter, S. D. (1988). “A Reappraisal of the Kappa Coefficient”. *Journal of Clinical Epidemiology* 41(10), pp. 949–958.

Van Wieringen, W. N. (2003). *Statistical Models for the Precision of Categorical Measurement Systems.* PhD-thesis, University of Amsterdam.