Kappa and Rater Accuracy: Paradigms and Parameters

Educational and Psychological
Measurement

2017, Vol. 77(6) 1019–1047

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416663277

journals.sagepub.com/home/epm

Anthony J. Conger¹

Abstract

Drawing parallels to classical test theory, this article clarifies the difference between rater accuracy and reliability and demonstrates how category marginal frequencies affect rater agreement and Cohen's kappa (κ). Category assignment paradigms are developed: comparing raters to a standard (index) versus comparing two raters to one another (concordance), using both nonstochastic and stochastic category membership. Using a probability model to express category assignments in terms of rater accuracy and random error, it is shown that observed agreement (P_o) depends only on rater accuracy and number of categories; however, expected agreement (P_e) and κ depend additionally on category frequencies. Moreover, category frequencies affect P_e and κ solely through the variance of the category proportions, regardless of the specific frequencies underlying the variance. Paradoxically, some judgment paradigms involving stochastic categories are shown to yield higher κ values than their nonstochastic counterparts. Using the stated probability model, assignments to categories were generated for 552 combinations of paradigms, rater and category parameters, category frequencies, and number of stimuli. Observed means and standard errors for P_o , P_e , and κ were fully consistent with theory expectations. Guidelines for interpretation of rater accuracy and reliability are offered, along with a discussion of alternatives to the basic model.

Keywords

kappa, rater accuracy, interrater reliability, nominal scales

Introduction

The kappa coefficient (κ), introduced by Jacob Cohen over 50 years ago (Cohen, 1960), has been the major statistic for measuring interrater agreement for categorical

¹Purdue University, West Lafayette, IN, USA

Corresponding Author:

Anthony J. Conger, psyCATS, 1230 Main St., Crete, IL 60417, USA.

Email: info@psyCATS.net

data. Cohen's κ has been subject to criticisms and misunderstandings regarding the definition of chance agreement, variation associated with marginal frequencies or base rates, and relationship of κ to amount of agreement and to reliability. These criticisms have led to suggested replacements (e.g., Andres & Marzo, 2004; Brennan & Prediger, 1981).

A frequent question that is posed is, "Is κ the best available measure of agreement?" Kappa is defined as a measure of agreement beyond chance compared to the maximum possible beyond chance agreement. Watkins and Pacheco (2000) discuss, but do not agree with, a preference of some behavioral researchers for percent agreement as the appropriate measure of interobserver agreement because of its simpler calculation and interpretation. Watkins and Pacheco argue that percent agreement is flawed due to its not accounting for chance and its noncomparability across studies. However, Andres and Marzo (2004) argue that κ is difficult to interpret relative to agreement and offer their Δ coefficient in its place.

A related question is, "Does κ measure reliability?" Bakeman, Quero, McArthur, and Robinson (1977) state that κ would represent reliability only if an observer were compared to a known accurate standard, and otherwise state that in order to regard κ as reliability it is necessary to make essentially untestable classic parallel test theory assumptions. Actually, the comparison of a rater to a known standard corresponds to the classical test theory reliability *index* and not to the test theory reliability *coefficient*. Other analyses of κ have helped clarify this issue. For example, Light (1971) carefully delineated association among categories from agreement between raters and also discussed situations where items have a specified "correct" category membership. Kraemer (1982) helped clarify the difference between comparing a rating to truth and comparing ratings to one another, a distinction that at times is ignored. For example, Andres and Marzo (2004) stated that κ can be used with both, seemingly without consideration of the consequences. The stipulation by Bakeman et al (1977) describing rater accuracy as agreement with a known standard also confounds nonchance and spurious accuracy; however, Andres and Marzo (2004) clearly distinguish between nonchance accuracy and chance accuracy.

A frequent criticism of κ focuses on the "error of marginal dependency" (Von Eye & Von Eye, 2005, 2008) or marginal asymmetry (Andres & Marzo, 2004). One feature of rater marginals is the degree to which a rater's observed category proportions differ from homogeneity; however, two other types of differences between raters' marginals exist. First, two raters may have marginals that are unequal but proportional to each other. Alternatively, the marginals may be disproportionate. Marginal asymmetry of all types is seen as causing both spuriously high agreement (e.g., Green, 1981) and low κ values (e.g., Gwet, 2008). The link between agreement and κ involves the concept of chance agreement, as discussed by Brennan and Prediger (1981), Green (1981), and others. High agreement between raters is viewed at times as being indicative of agreed expert opinion (Brennan and Prediger, 1981) and at times as spurious agreement due to marginal asymmetry. Gwet (2008) considered the difference between rater marginals to be the problem underlying low κ

values. Brennan and Prediger (1981) offered a modified statistic, κ_{m} , in which chance agreement between raters ignored the rater marginals and was set to $1/n$. Xu and Lorber (2014) evaluated the effect of marginal skewness for 2×2 matrices on κ and five alternatives, concluding that Holley and Guilford's G (Holley & Guilford, 1964) be used instead of κ because G was the least sensitive to base rates. In contrast, Uebersax (1987), focusing on decision making, stated that κ should be sensitive to category marginal base rates. In general, there is disagreement about the appropriate sensitivity of κ to marginal asymmetry, and the focus on the rater marginals misses the determinants of the marginals: underlying category frequencies and rater accuracy and bias.

Uebersax (1987) and Green (1981) argued that κ can be interpreted properly only in the context of definite decision-making models but these models are generally unspecified. However, in the absence of knowledge of those specifics, varied assumptions are made about rater assignments or agreement. Green (1981) and Andres and Marzo (2004) proposed that raters are certain about their assignments for a proportion of objects within a category and guess for the remaining objects. The guessing may be equi-probable across all categories (Andres & Marzo, 2004; Green, 1981) or may match the base rate of category probabilities (Green, 1981; Uebersax, 1987). Cousineau and Laurencelle (2015, 2016) assumed a model in which a rater has a fixed probability of being correct with otherwise random guessing for all categories. Some (e.g., Bakeman et al., 1997) assume identically fallible observers, without specifically identifying the form of that fallibility. Others make no assumptions about the rater and focus on cells of the agreement matrix (e.g., Von Eye & Von Eye, 2005) or on rater marginals (e.g., Gwet, 2008; Hadzi-Pavlovic, 2010).

Related to the above issues, classic test theory (CTT) and its derivatives provide a variety of statistics and considerations for evaluating and interpreting associations among individual difference variables for populations or subpopulations. It is well known that reliability is influenced by the underlying score distributions (among other things), and care is required for their proper interpretation. For some applied and theoretic purposes, however, the standard error of measurement provides a more useful measure than does reliability (e.g., for estimating the "true" score of an individual knowing only the observed score). In CTT, under reasonable conditions, the standard error of measurement is relatively invariant throughout the distribution of observed scores; however, reliability and validity correlations are not independent of score distributions. Furthermore, CTT distinguishes between reliability *indices* (comparing an observed score to a true score) and reliability *coefficients* (comparing observed scores to one another). In an analogous way, comparing one rater to a standard is comparable to a classic reliability *index*, while comparing raters to one another is comparable to a reliability *coefficient*.

In like fashion, the assessment of raters' assignments to categories should be looked at relative to the intended use of the data. One use is simply to determine the category assigned to an object or person. For example, when assigning a psychological diagnosis, interest is in the accuracy of the diagnosis for each patient. In this case,

a statistic like the standard error of measurement would be the most useful measure of accuracy. A different use is the assignment of categories for the purpose of measuring individual differences and relating them to other measures: for example, assigning psychological disorders to individuals with the intent of associating them with other individual difference variables. In this case, a statistic like the reliability coefficient can be used to provide expectations for limits on the magnitude of the association. The former use, that is, focus on an individual's score, is independent of the distribution of the categories; however, in the latter case, reliability and association measures would be dependent on the distribution of individual differences, that is, category frequencies in the case of κ .

The purpose of this article is to clarify the difference between accuracy and reliability, clarify the meaning of kappa relative to category paradigms, and show how observed agreement, expected agreement, and kappa depend on category frequencies and category paradigms. This is accomplished by applying a simple rater assignment model (nonrandom probability of correct assignment with, otherwise, equi-probable guessing to all categories). Correspondingly, simulations varying category frequencies, rater accuracy, and number of stimuli demonstrate the sampling characteristics of rater agreement and κ and the convergence of generated data on theoretic expectations.

In this article, the investigation of kappa and rater accuracy begins by specifying properties of category systems. (This step is like defining a true score in CTT.) The second step is specification of a paradigm for the assignment of objects to the categories by raters. (This step is like defining the observed score and its relation to a true score.) This article develops five different theoretic paradigms for categorical data collection corresponding to reliability indices and reliability coefficients, with and without stochastic category membership. The third step is the evaluation of the assignments to categories by one or more raters. (This last step is like finding an intertest or interrater correlation.) With this systematic approach, it is possible to specify assumptions underlying paradigms of category membership, category assignment, and rater agreement and to investigate the consequences of deviations from these assumptions.

Theoretic Development

Category Membership

Specifying a category system is similar to defining true scores in CTT. At a minimum, nominal or categorical scaling requires mutually exclusive and together exhaustive categories. However, category membership, instead of being viewed as a certainty, could also be viewed as an expectation: much as a "Platonic" true score can be replaced by an "expected value" universe score as in generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Thus, at times, category membership might be stochastic rather than nonstochastic. Stochastic membership may arise because objects manifest in different phases (e.g., objects may have modal but

variable phases), or objects themselves may have only a derivative truth (e.g., using an expert's fallible ratings to specify category membership) or have insufficient information on unique or differentiating attributes or categories may share attributes. While categories traditionally require clear boundaries, more complex systems may possess fuzzy boundaries (see Dou et al., 2007).

Category System Properties. A system with N objects distributed across n categories can be described by the probability¹ that an object correctly belongs in a category; the distribution of errors of membership for an object across alternative categories; the distribution of accuracy across objects within the same category; and the overall distribution of accuracy across categories.

Let λ_{ik} represent the probability that object i is in category k , whether it belongs there or not. The λ_{ik} sum to unity for the n categories: $\sum_k \lambda_{ik} = 1$. Let π_{ik} denote the *nonchance* probability that object i in category k actually belongs to category k , and let ε_{ik} denote *chance* membership in category k , whether k is correct or not. Category membership probability, P_{ik} , is nonchance plus chance membership: $P_{ik} = \pi_{ik} + \varepsilon_{ik}$. In a stochastic category system, some ε_{ik} are different from zero. In a nonstochastic system, $\varepsilon_{ik} = 0$ and $P_{ik} = \pi_{ik} = 1$.

For stochastic category systems, homogeneity of the distribution of errors is defined by $\varepsilon_{ik} = \varepsilon_i$, for every object i . Specifically, $\varepsilon_i = (1 - \pi_{ik})/n$. The membership probabilities for object i from category k thus become: $\lambda_{ik} = P_{ik} = \pi_{ik} + \varepsilon_i$ for the correct category k and $\lambda_{ik'} = \varepsilon_i$ for all incorrect categories k' , $k' \neq k$. Homogeneity of accurate membership for objects *within* a category is given by $\pi_{ik} = \pi_k$ and $\varepsilon_{ik} = \varepsilon_k$ for all i in k . Homogeneity of accurate membership *across* categories is defined by the above and a further assumption that $\pi_{ik} = \pi_s$ and $\varepsilon_{ik} = \varepsilon_s$, constants for all objects in all categories.

Category Membership Distribution: Category Heterogeneity

While categorical scales, at their most basic level, do not possess quantitative characteristics and statistical description is limited to the proportion of occurrence for each category, the proportions² of occurrence do possess interval-level scale properties. Proportions of occurrence can be represented as $(\mu_1, \mu_2, \dots, \mu_k, \dots, \mu_n)$. Regardless of which categories have which specific proportions, the mean proportion is a constant $(1/n)$. However, the variance of the proportions can be used to describe different distributions of category proportions: herein referred to as *category heterogeneity* (σ_M^2). The minimum category heterogeneity is 0, occurring when category proportions are all equal to $1/n$. The maximum category heterogeneity, $\max(\sigma_M^2) = (n-1)/n^2$, occurs when all of the objects are in one category. The possible maximum heterogeneity can be used to "standardize" the category heterogeneity across models with different numbers of categories (n) by using the ratio of the category heterogeneity relative to the maximum category heterogeneity:

$$\sigma^2_{\text{ratio}} = \sigma_M^2 / \max(\sigma_M^2) = n^2 \sigma_M^2 / (n - 1)$$

As will be shown below, category heterogeneity (σ_M^2) and the variance ratio (σ^2_{ratio}) are sufficient for the analytic comparison of differences in category proportions for agreement indices and κ . For example, for a 4-category system, the following three sets of category proportions, $\{.580, .140, .140, .140\}$, $\{.537, .231, .231, .001\}$, and $\{.441, .440, .060, .059\}$, look very different; however, they have essentially identical category heterogeneity ($\sigma_M^2 = .0363$) and quantitatively have the same effect on observed agreement, expected agreement, and κ .

Category Assignment by Raters

Models for raters' assignments of objects to categories are specified along the same lines used for category membership. While the most basic model assumes homogeneity of nonchance correct probabilities of assignment across categories and independence and homogeneity of errors of assignment within categories, these assumptions may not be tenable for all raters and categories. These issues will be discussed later.

Using a category system as defined above, properties of a rater's assignments are developed as follows: P_{ijk} represents the probability that object i is assigned by rater j to category k , whether it belongs there or not. For the n categories: $\sum_k P_{ijk} = 1$. For rater j , π_{ijk} denotes the *nonchance* probability that object i is assigned to its correct category k and ε_{ijk} denotes *chance* assignment to any category. Correct category assignment probability, P_{ijk} , is the sum of nonchance and chance assignment: $P_{ijk} = \pi_{ijk} + \varepsilon_{ijk}$.

In this article, it will be assumed that a rater's nonchance assignment probability is constant ($\pi_{ijk} = \pi_j$) and errors of assignment are homogeneous across objects and categories ($\varepsilon_{ijk} = \varepsilon_j$). Assignment probabilities for object i in category k thus become: $P_{ijk} = P_j = \pi_j + \varepsilon_j$ for the correct category k and ε_j for all incorrect categories $k', k' \neq k$. This system in many ways resembles the basic model of classic true score theory: an observed score X_{ik} for object i has an underlying true score T_i . The error E associated in measuring X_{ik} is random, it is independent of T_i , and its variance is constant across objects i and true scores T_i . For categories, the equivalent statements are that object i (truly) belongs in category T_k and is assigned to it by the rater with probability $P_j = \pi_j + \varepsilon_j$. Errors of assignment, ε_j , are homogeneous and independent of object or category. Randomness of errors is provided by assuming that erroneous assignments are distributed independently and uniformly across all categories. As with CTT, such ideal conditions may not be realized, but useful conceptual relationships result. Specific alternative models for assignments and errors can be specified and are discussed below.

Category Assignment Paradigms

This article investigates two different types of rater agreement paradigms: agreement indexing and agreement concordance. In the indexing paradigm, assignments made by a rater are compared to extant category memberships. Indexing is like the

reliability index in CTT, that is, the correlation between observed scores and true scores. In the concordance paradigm, assignments are made independently by two raters and are compared to one another. Concordance parallels the classic interform reliability coefficient, that is, the correlation between two sets of observed scores. For both indexing and concordance, category membership could be nonstochastic or stochastic; moreover, stochastic concordance has two further options: a dependent paradigm in which both raters view the same manifest phase of an object versus an independent paradigm in which each rater views independent manifest phases. In total, this article looks at five different paradigms: nonstochastic indexing, nonstochastic concordance, stochastic indexing, dependent stochastic concordance, and independent stochastic concordance.

The five paradigms can be illustrated by the following example: Imagine a study in which raters are being trained to classify a child's peer interactions as prosocial, neutral, aggressive, or passive. Using strict protocols, multiple video recordings are made to illustrate each type of behavior (known standards). Each trainee rates 30 video recordings. An individual trainee's ratings could be compared to the known standards (nonstochastic indexing) or two trainees could be compared to one another (nonstochastic concordance). After reaching criteria on the training task, the trainees are asked to classify behaviors as prosocial, neutral, aggressive, or passive based on actual classroom interactions. With "real" behavior, the interactions and behaviors vary across time and situation, leading to a stochastic paradigm. If a "master" rater evaluates the children over a period of time, the "master" ratings could be used as the "standard," albeit stochastic, by which children are classified. An individual trainee, observing children for a single time period, could be compared to this stochastic standard (stochastic indexing). Moreover, a pair of trainees could be compared to one another in either of two ways: they could both observe a child's classroom interactions at the same time period (dependent stochastic concordance) or they could observe interactions at different time periods (independent stochastic concordance).

A variety of characteristics are used to describe and compare agreement statistics for these five paradigms. For each paradigm (1) parameters describing category membership and category assignment by raters are specified; (2) a parameter describing category frequencies or heterogeneity is developed; (3) using the preceding parameters, a matrix of observed agreements and disagreements is derived; (4) formulae are developed for overall probability of agreement (P_o), expected probability of agreement (P_e), and the κ coefficient. Figure 1 provides a summary of the five paradigms and the statistics that are developed for them, and Figure 2 illustrates similarities and differences among the paradigms using a common set of parameter values.

Index Paradigms

Nonstochastic Indexing (NSI): Agreement Between One Rater and True Category Membership

Category Membership. The NSI paradigm assumes that each object manifests in only one category. The situation in which one rater's category assignment is compared to

category marginals	rater j versus category (index)		rater j versus rater j' (concordance)		
	non-stochastic	stochastic	non-stochastic	dependent stochastic	independent stochastic*
	true category	manifest category	true category	manifest category	manifest category
frequency μ_k	μ_k	$\mu_{s_k} = \pi_s \mu_k + \varepsilon_s$	μ_k	$\mu_{s_k} = \pi_s \mu_k + \varepsilon_s$	$\mu_{s_k} = \pi_s \mu_k + \varepsilon_s$
variance σ_M^2	σ_M^2	$\sigma_{M_s}^2 = \pi_s^2 \sigma_M^2$	σ_M^2	$\sigma_{M_s}^2 = \pi_s^2 \sigma_M^2$	$\sigma_{M_s}^2 = \pi_s^2 \sigma_M^2$
rater j marginals					
frequency m_{kj}	$\pi_j \mu_k + \varepsilon_j$	$\pi_j \mu_{s_k} + \varepsilon_j$	$\pi_j \mu_k + \varepsilon_j$	$\pi_j \mu_{s_k} + \varepsilon_j$	$\pi_j^* \mu_k + \varepsilon_j^*$
variance $\sigma_{m_j}^2$	$\pi_j^2 \sigma_M^2$	$\pi_j^2 \sigma_{M_s}^2$	$\pi_j^2 \sigma_M^2$	$\pi_j^2 \sigma_{M_s}^2$	$\pi_j^{*2} \sigma_M^2$
agreement matrix					
agree a_{kk}	$p_j \mu_k$	$p_j \mu_{s_k}$	$(p_j p_{j'} - \varepsilon_j \varepsilon_{j'}) \mu_k + \varepsilon_j \varepsilon_{j'}$	$(p_j p_{j'} - \varepsilon_j \varepsilon_{j'}) \mu_{s_k} + \varepsilon_j \varepsilon_{j'}$	$(p_j^* p_{j'}^* - \varepsilon_j^* \varepsilon_{j'}^*) \mu_k + \varepsilon_j^* \varepsilon_{j'}^*$
disagree $a_{kk'}$	$\varepsilon_j \mu_{k'}$	$\varepsilon_j \mu_{s_{k'}}$	$\pi_j \varepsilon_{j'} \mu_k + \varepsilon_j \pi_{j'} \mu_{k'} + \varepsilon_j \varepsilon_{j'}$	$\pi_j \varepsilon_{j'} \mu_{s_k} + \varepsilon_j \pi_{j'} \mu_{s_{k'}} + \varepsilon_j \varepsilon_{j'}$	$\pi_j^* \varepsilon_{j'}^* \mu_k + \varepsilon_j^* \pi_{j'}^* \mu_{k'} + \varepsilon_j^* \varepsilon_{j'}^*$
agreement statistics					
observed P_o	$\frac{(n-1)}{n} \pi_j + \frac{1}{n}$	$\frac{(n-1)}{n} \pi_j + \frac{1}{n}$	$\frac{(n-1)}{n} \pi_j \pi_{j'} + \frac{1}{n}$	$\frac{(n-1)}{n} \pi_j \pi_{j'} + \frac{1}{n}$	$\frac{(n-1)}{n} \pi_j^* \pi_{j'}^* + \frac{1}{n}$
expected P_e	$\frac{(n-1)}{n} \pi_j \sigma_{ratio}^2 + \frac{1}{n}$	$\frac{(n-1)}{n} \pi_j \sigma_{ratio}^2 + \frac{1}{n}$	$\frac{(n-1)}{n} \pi_j \pi_{j'} \sigma_{ratio}^2 + \frac{1}{n}$	$\frac{(n-1)}{n} \pi_j \pi_{j'} \sigma_{ratio}^2 + \frac{1}{n}$	$\frac{(n-1)}{n} \pi_j^* \pi_{j'}^* \sigma_{ratio}^2 + \frac{1}{n}$
kappa κ	$\frac{\pi_j [1 - \sigma_{ratio}^2]}{1 - \pi_j \sigma_{ratio}^2}$	$\frac{\pi_j [1 - \sigma_{ratio}^2]}{1 - \pi_j \sigma_{ratio}^2}$	$\frac{\pi_j \pi_{j'} [1 - \sigma_{ratio}^2]}{1 - \pi_j \pi_{j'} \sigma_{ratio}^2}$	$\frac{\pi_j \pi_{j'} [1 - \sigma_{ratio}^2]}{1 - \pi_j \pi_{j'} \sigma_{ratio}^2}$	$\frac{\pi_j^* \pi_{j'}^* [1 - \sigma_{ratio}^2]}{1 - \pi_j^* \pi_{j'}^* \sigma_{ratio}^2}$
where $\pi_j^* = \pi_j \pi_s$, $\varepsilon_j^* = \frac{[1 - \pi_j \pi_s]}{n}$, $p_j^* = \pi_j^* + \varepsilon_j^*$, and $\sigma_{ratio}^2 = \frac{\sigma_M^2}{\text{Maximum}(\sigma_M^2)}$					

Figure 1. Formulae for agreement matrix and related statistics for different kappa paradigms.

a truth may seem artificial, just as specifying a real “True” score may seem artificial. However, actual membership of objects in categories could be established by construction or by assessment. This model has theoretic appeal as the most basic paradigm and provides a method for evaluating rater training. Assume the true category frequencies have category heterogeneity of σ_M^2 . For example, category proportions of $\{.70, .20, .10\}$ have a heterogeneity of .069 (see Figure 2).

Assignment Parameters. Category assignment by rater j involves the probability of correct assignment (p_j), nonchance correct assignment (π_j), and the error of assignment (ε_j). These are all co-determinant for an n -category system. That is, rater assignment in this initial paradigm is a single parameter model with $p_j = (n - 1)\pi_j/n + 1/n$ and $\varepsilon_j = (1 - \pi_j)/n$. However, interpretation of relationships, equations, and derivations is facilitated by the differential use of these three dependent characteristics. Assuming a rater with $\pi_j = .85$, then $\varepsilon_j = .05$ and $p_j = .90$ (Figure 2).

Rater Category Marginals. A second set of statistics involves the observed rater marginal for each category and the variability of the rater marginals. Since for each category the probability of correct assignment is $p_j = \pi_j + \varepsilon_j$, then for category k with marginal proportion μ_k , category k will be assigned $(\pi_j + \varepsilon_j)\mu_k$ times. Category k will also be assigned (incorrectly) with probability ε_j for each of the remaining $n - 1$ categories with total proportion of $\sum_k \varepsilon_j \mu_{k'}$, for $(k' \neq k)$. Combining all assignments,

Parameters		Category Membership		Category Assignment													
Parameter		Truth	Stochastic	Rater 1	Rater 2	Rater* 1 Rater* 2											
Non chance π		1.00	.91	.85	.79	.77											
Chance ϵ		.00	.03	.05	.07	.08											
Total $P = \pi + \epsilon$		1.00	.94	.90	.86	.85											
Category Var.		.069**	.057	.050	.043	.041											
* Denotes reparameterized rater values with $\pi_j^* = \pi_j \pi_s$, $\epsilon_j^* = (1 - \pi_j^* \pi_s^*)/n$, and $P_j^* = \pi_j^* + \epsilon_j^*$																	
** True category frequencies are .70, .20, & .10 for categories a,b, & c respectively (see Non-stochastic index paradigm).																	
Agreement Matrices		Rater vs. Membership (Index)		Rater vs. Rater (Concordance)													
rater 1	Non-stochastic		Stochastic		Non-stochastic		Dependent		Independent								
	truth		manifest		rater 2		rater 2		rater 2								
	a	b	a	b	a	b	a	b	a	b							
	c	freq	c	freq	c	freq	c	freq	c	freq							
a	.63	.01	.01	.65	.60	.01	.01	.62	.52	.05	.05	.62	.49	.07	.06	.62	
b	.04	.18	.01	.22	.03	.19	.01	.23	.04	.16	.02	.22	.04	.17	.02	.23	
c	.04	.01	.09	.14	.03	.01	.11	.15	.04	.02	.08	.14	.04	.02	.10	.15	
freq	.70	.20	.10		.67	.21	.12		.62	.23	.15		.60	.24	.17		
Note: Category agreement values are in bold. Category frequencies for respective rows and columns are in italics.																	
Agreement Statistics																	
Observed Po	.900			.900		.781		.781		.704							
Expected Pe	.509			.479		.472		.448		.448							
Kappa κ	.796			.808		.585		.603		.462							

Figure 2. Numerical illustration of parameters, agreement matrices, and statistics for different paradigms.

the probability that rater j assigns objects to category k is $(\pi_j + \varepsilon_j)\mu_k + \varepsilon_j \sum_k \mu_k$. Thus, the observed marginal for rater j for category k is a linear function of the true marginal proportion: $m_{jk} = \pi_j \mu_k + \varepsilon_j$. To the degree that π_j is less than 1, rater marginals will be *regressed* toward the mean proportion ($1/n$). Furthermore, since ε_j is a constant across categories, rater j 's category heterogeneity σ_{mj}^2 equals $\pi_j^2 \sigma_M^2$. In the illustration, Rater 1 marginals are equal to $.85\mu_k + .05$ (i.e., $\{.645, .220, .135\}$) and have a variance of $(.85)^2(.069) = .050$ (Figure 2).

Observed Agreements and Disagreements Between Rater Assignments and True Categories. Agreements and disagreements between the rater and true category can be put into a square nonsymmetrical matrix, with agreement cells (a_{kk}) on the diagonal and disagreement cells ($a_{k'k}$) off the diagonal. In this basic paradigm, these cells are simply a function of the true category proportions, μ_k , and the rater probabilities P_j and ε_j . For assignment k' and true category k : $\Pr(a_{kk}) = P_j \mu_k$ for ($k' = k$) and $\Pr(a_{k'k}) = \varepsilon_j \mu_k$ for ($k' \neq k$).

Summing the correct assignments over all categories, observed agreement between rater j and the true category is the following: $Po_{ji} = \sum_k a_{kk} = \sum_k P_j \mu_k = P_j$ (since $\sum_k \mu_k = 1$). Thus, Po_{ji} equals the probability of a correct assignment P_j , including both nonchance (π_j) and chance (ε_j) correct assignments, and Po_{ji} is unaffected by the distribution of category membership (σ_M^2 or σ_{mj}^2). Focusing on π_j , the observed agreement (Figures 1 and 2) can be expressed as

$$Po_{ji} = \{(n-1)/n\}\pi_j + 1/n$$

Expected Agreement Between Rater Assignments and True Categories. Assuming a traditional chance agreement model,³ expected agreement (Pe) is the sum of cross products of the category marginal proportions for the rater and the truth: $Pe = \sum_k m_{jk} \mu_k = \sum_k (\pi_j \mu_k + \varepsilon_j) \mu_k = \pi_j \sum_k \mu_k^2 + \varepsilon_j \sum_k \mu_k$. By expressing ε_j in terms of π_j and expressing μ_k^2 in terms of the corresponding variance and mean of marginals, this expression reduces to

$$Pe_{ji} = n\pi_j \sigma_M^2 + 1/n \quad \text{or} \quad Pe_{ji} = \{(n-1)/n\}\pi_j \sigma_{\text{ratio}}^2 + 1/n$$

Both forms for Pe_{ji} indicate that expected agreement is a function of the category heterogeneity variance weighted by the rater's probability of nonchance correct assignment and of the number of categories. Increased category heterogeneity increases the expected agreement and this effect, in turn, is enhanced by higher levels of a rater's nonchance assignment. Ignoring $\pi_j = 0$, expected agreement equals the number of categories ($1/n$) only when category heterogeneity is 0, that is, category frequencies are each equal to $1/n$. $Pe_{ji} = n\pi_j \sigma_M^2 + 1/n$ suggests that the expected agreement will increase, as well, with the number of categories; however, the alternative form indicates this effect is somewhat moderated by the limits imposed by the number of categories on the variance of category marginals.

Kappa Coefficient: κ_{jt} . Cohen's κ compares the increment of observed agreement over chance agreement to the maximum possible increment. Expressed in terms of proportions: $\kappa = (Po - Pe)/(1 - Pe)$. Based on the derived equations for Po_{jt} and Pe_{jt} , κ_{jt} can be expressed as: $\kappa_{jt} = \pi_j(1 - \sigma_{ratio}^2)/(1 - \pi_j\sigma_{ratio}^2)$.

This ratio of two linear forms of π_j shows that the agreement between rater assignments and true category memberships is affected by only two values: nonchance correct assignment probability, π_j ; and relative category heterogeneity, σ_{ratio}^2 . For fixed π_j , the maximum possible value for k occurs when $\sigma_{ratio}^2 = 0$. Since the impact of σ_{ratio}^2 is greater on the numerator than on the denominator, κ_{jt} will otherwise be reduced over its possible maximum to the degree that category heterogeneity increases: $\kappa_{jt} = .796$ in the example (see Figure 2).

As discussed above, if $\sigma_{ratio}^2 = 0$, expected agreement, Pe_{jt} , is simply $1/n$. From this it follows that if category heterogeneity is 0 then $\kappa_{jt} = \kappa_n$, the statistic recommended by Brennan and Prediger (1981). However, category heterogeneity is a characteristic of the true marginals and cannot simply be ignored (as is implicit for κ_n). Under the condition of equal category proportions, it is also true that κ_{jt} will equal π_j , the degree of nonchance correct assignment.

Stochastic Indexing (SI): Agreement Between One Rater and Stochastic Category Membership

Category Membership. Consider a category system that does not meet the null error of membership assumption (i.e., $\varepsilon_s = 0$). The existence of error associated with object membership can be specified as $\varepsilon_s > 0$. While each object truly belongs to one and only one category, its temporary appearance or information about it may make it look like it belongs to a different category. If the $\varepsilon_s > 0$, observed category assignments would involve error from both the category system and the rater.

In stochastic paradigms, the *manifest* occurrence of objects appearing in a category is a function of the accurate membership probability, π_s , and the probability of category occurrence, μ_k . Since for each category, the probability of correct membership is $P_s = \pi_s + \varepsilon_s$, then for category k with marginal frequency μ_k , category k will be manifest $(\pi_s + \varepsilon_s)\mu_k$ times. However, category k will occur with probability ε_s for each $k' \neq k$ category for a total of $\sum_k \varepsilon_s \mu_{k'}$. Combining all occurrences, the probability that category k occurs is $\pi_s \mu_k + \varepsilon_s \mu_k + \sum_{k' \neq k} \varepsilon_s \mu_{k'}$. Since $\sum_{(all\ k')} \varepsilon_s \mu_{k'} = \varepsilon_s$, the manifest marginal is the linear function: $\mu_{sk} = \pi_s \mu_k + \varepsilon_s$. Since $\pi_s < 1$, the manifest marginals will be regressed toward the mean proportion ($1/n$) and their variance will be proportionately reduced by π_s^2 , that is, $\sigma_{Ms}^2 = \pi_s^2 \sigma_M^2$.

The parameter for stochastic paradigm examples in Figure 2 is .91 with an error parameter of .03 for each category. That is, 94% of the time an object will look like it belongs in its correct category and 3% of the time it will look like it belongs in each of the other two categories. The resulting manifest category proportions are shown in Figure 2, along with the stochastic category heterogeneity of .057 and variance ratio of .257, representing a nearly 20% reduction in category heterogeneity.

Assignment Parameters. As in NSI, assignment by rater j involves the single parameter model $P_j = \pi_j + \varepsilon_j$ with $\varepsilon_j = (1 - \pi_j)/n$. However, the probability that rater j places an object in the correct category (k) is $P_j P_s = (\pi_j + \varepsilon_j)(\pi_s + \varepsilon_s)$, a two-parameter model.

Rater Category Marginals. The rater marginal for each category and the variability of the rater marginals depend on π_j and π_s . For each manifest category, the probability of correct assignment is P_j . For category k with marginal frequency μ_{s_k} , category k will be assigned $P_j \mu_{s_k}$ times. However, category k will also be assigned with probability ε_j for each of the $k' \neq k$ remaining categories for a total of $\sum_{k'} \varepsilon_j \mu_{s_{k'}}$ times. Combining all assignments, the probability that rater j assigns objects to category k is $P_j \mu_{s_k} + \varepsilon_j \sum_{k'} \mu_{s_{k'}}$ for $k' \neq k$. This reduces to $\pi_j \mu_{s_k} + \varepsilon_j$, since $\sum_{k'} \mu_{s_{k'}} = 1$. Thus, the rater marginal is a linear function of the manifest marginal. However, because manifest marginals are also regressed, observed rater marginals will be even further regressed toward the mean proportion ($1/n$): $m_{jk} = (\pi_j \pi_s) \mu_k + (1 - \pi_j \pi_s)/n$. The variance of the category marginals for rater j is $\sigma_{mj}^2 = \pi_j^2 \sigma_{Ms}^2 = (\pi_j^2 \pi_s^2) \sigma_M^2$. The product of the rater and stochastic parameters, $\pi_j \pi_s$, functions like the single rater parameter, π_j , in the nonstochastic model. These differences are illustrated in Figure 2.

Observed Agreements and Disagreements Between Rater Assignments and Manifest Categories. In the stochastic paradigm, raters' assignments are based on the manifest categories. Agreements and disagreements with the manifest categories follow the same pattern as in NSI. Substituting μ_{s_k} for μ_k into the agreement and disagreement cells of the NSI paradigm (Figure 1) yields: $\Pr(a_{kk}) = P_j \mu_{s_k}$ and $\Pr(a_{kk'}) = \varepsilon_j \mu_{s_{k'}}$.

The probability of a correct assignment to manifest category k is P_j . Summing over all categories, the total proportion of correct assignments (Po_{js}) is $\sum_k P_j \mu_{s_k}$, which reduces to P_j (since $\sum_k \mu_{s_k} = 1$). Po_{js} can also be expressed as $\{(n - 1)/n\} \pi_j + 1/n$ showing equality with the observed agreement in the nonstochastic model (as shown in Figure 1 and illustrated in Figure 2). Thus, Po_{js} equals the probability of correct assignment to category k (i.e., P_j) and is equal to Po_{jt} in the NSI paradigm. Po_{js} is the sum of nonchance (π_j) and chance (ε_j) correct assignments and is unaffected by category variance σ_M^2 or the nonstochastic parameter, π_s .

Expected Agreement Between Rater Assignments and Manifest Categories. Expected agreement between the assignments and the manifest categories, denoted Pe_{js} , is obtained from the cross products of the category marginal proportions: $\sum_k (m_{jk} \mu_{s_k})$. Using the above equations for μ_{s_k} and m_{jk} : $Pe_{js} = \sum_k \{ \pi_j \pi_s \mu_k + (1 - \pi_j \pi_s)/n \} \{ \pi_j \mu_k + \varepsilon_j \}$ or simply, $Pe_{js} = n \pi_j \pi_s^2 \sigma_M^2 + 1/n$. Using the variance ratio σ_{ratio}^2 or stochastic marginal variance ratio $\sigma_{Sratio}^2 = \sigma_{Ms}^2 / \max(\sigma_M^2)$, Pe_{js} can be expressed as follows: $Pe_{js} = \{(n - 1)/n\} \pi_j \pi_s^2 \sigma_{ratio}^2 + 1/n$ or $\{(n - 1)/n\} \pi_j \sigma_{Sratio}^2 + 1/n$.

The above equations point to the fact that Pe_{js} for SI is less than Pe_{jt} for NSI (Figures 1 and 2). The sole reason is the regressed variance associated with the manifest category marginals compared to the true category marginals.

Kappa Coefficient: κ_{js} . The κ coefficient for SI can be found from the above equations for Po_{js} and Pe_{js} : $\kappa_{js} = \pi_j(1 - \pi_s^2\sigma_{ratio}^2)/(1 - \pi_j\pi_s^2\sigma_{ratio}^2)$ or $\pi_j(1 - \sigma_{Sratio}^2)/(1 - \pi_j\sigma_{Sratio}^2)$.

Agreement between the rater's assignments and the category memberships is affected by both π_j and π_s , the measures of nonchance correct agreement and correct category manifestation, and also, by σ_{ratio}^2 , the relative degree to which the category frequencies deviate from homogeneity. The impact of both π_s and σ_{ratio}^2 is greater on the numerator. It is also worth noting that κ_{js} will equal or exceed κ_{jt} . This is surprising, in that one would expect stochastic error to reduce agreement. One way to accept the effect is to note that while observed agreement is equal for both NSI and SI, the expected agreement is less for the stochastic paradigm because of greater regression of rater and manifest category marginals. On the other hand, observed agreement, while equal to NSI observed agreement, includes an additional component of spurious agreement: "correctly" assigning the object to its manifest but wrong true category. The increase in κ illustrated in Figure 2 is small (.808 vs. .796); however, the category membership stochastic parameter for the example is reasonably high (.91).

If category heterogeneity is zero, stochastic κ will equal π_j , the value of the maximum nonstochastic κ . Otherwise, κ will decrease as category heterogeneity increases.

Concordance Paradigms

Nonstochastic Concordance (NSC): Agreement Between Two Raters

Category Membership. NSC requires that objects unambiguously belong in one category although their specific category membership need not be known. Raters' independent assignments of objects to categories are used to establish rater agreement and κ . As in NSI, true category heterogeneity is σ_M^2 . For the illustration, category proportions remain {.70, .20, .10} (Figure 2).

Assignment Parameters. Assignments by each rater involve a single parameter: π_j for the first rater and $\pi_{j'}$ for the second. When raters are parallel, $\pi_j = \pi_{j'}$. For the illustration of nonparallel raters, the first rater has $\pi_j = .85$ and the second has $\pi_{j'} = .79$ (Figure 2).

Rater Category Marginals. Since rater assignments are independent, category marginals of one rater would not affect those of the other; however, the rater marginals would be unequal (i.e., asymmetric) if rater accuracies differed. The marginal proportion for category k for rater j remains $m_{jk} = \pi_j\mu_k + \varepsilon_j$ with variance $\sigma_{mj}^2 = \pi_j^2\sigma_M^2$ (Figure 2).

The ratio of the rater variances is of interest in the concordance paradigms and equals the ratio of the squared nonchance correct assignment parameters: $\sigma_{mj}^2/\sigma_{mj'}^2 = \pi_j^2/\pi_{j'}^2$. This relationship assumes that the variance of the marginals is nonzero. If σ_{mt}^2 is zero, then σ_{mj}^2 and $\sigma_{mj'}^2$ should both be zero. Of course, *observed* estimates

of σ_{mj}^2 and $\sigma_{mj'}^2$ depend on the sampling distributions of the marginal variances and could or could not be zero regardless of σ_M^2 . Since both sets of marginals are linear functions of the true category frequencies, in theory, the observed rater marginals would be exactly proportional to one another.

Observed Agreements and Disagreements Between Raters. Agreement on category k (i.e., a_{kk}), arises from both raters assigning category k whether or not it is correct. If category k is correct, the probability of both raters agreeing is $P_j P_{j'} \mu_k$. If a different category k' , $k' \neq k$, is correct, the probability that both raters assign category k is $\varepsilon_j \varepsilon_{j'} \mu_{k'}$. Total assignment to a_{kk} is thus $P_j P_{j'} \mu_k + \sum_{k' (k' \neq k)} \varepsilon_j \varepsilon_{j'} \mu_{k'}$. Since $\sum_{k'} \mu_{k'} = 1$, $\Pr(a_{kk}) = (P_j P_{j'} - \varepsilon_j \varepsilon_{j'}) \mu_k + \varepsilon_j \varepsilon_{j'}$. The disagreement cells ($a_{kk'}$, $k \neq k'$) involve rater j assigning k and rater j' assigning k' . Three basic events are possible: rater j assigns k when k is true (rater j' does not) or rater j' assigns k' when k' is true (rater j does not) or neither assigns the correct category (k''). Total assignment to $a_{kk'}$ is thus, $P_j \varepsilon_{j'} \mu_k + \varepsilon_j P_{j'} \mu_{k'} + \sum_{k''} \varepsilon_j \varepsilon_{j'} \mu_{k''}$. Since $P_j \varepsilon_{j'} \mu_k = \pi_j \varepsilon_{j'} \mu_k + \varepsilon_j \varepsilon_{j'} \mu_k$ and $\varepsilon_j P_{j'} \mu_{k'} = \varepsilon_j \pi_{j'} \mu_{k'} + \varepsilon_j \varepsilon_{j'} \mu_{k'}$, this reduces to: $\Pr(a_{kk'}) = \pi_j \varepsilon_{j'} \mu_k + \varepsilon_j \pi_{j'} \mu_{k'} + \varepsilon_j \varepsilon_{j'}$ (Figures 1 and 2).

Overall agreement, $P_{ojj'}$, between the assignments of the two raters does not consider whether the agreement is correct or incorrect. Summing agreement cells, the total of correct assignments is $\sum_k (a_{kk}) = \sum_k \{ (P_j P_{j'} - \varepsilon_j \varepsilon_{j'}) \mu_k + \varepsilon_j \varepsilon_{j'} \} = \sum_k P_j P_{j'} \mu_k - \sum_k \varepsilon_j \varepsilon_{j'} \mu_k + n \varepsilon_j \varepsilon_{j'}$. This reduces to: $P_{ojj'} = \{ (n-1)/n \} \pi_j \pi_{j'} + 1/n$, since $\sum_k \mu_k = 1$. Note that observed agreement is a function of the product $\pi_j \pi_{j'}$ plus spurious agreements and it is unaffected by the distribution of category membership (σ_M^2 or σ_{mj}^2).

Because the product $\pi_j \pi_{j'}$ is less than either π_j or $\pi_{j'}$, NSC observed agreement will be less than NSI observed agreement for either rater. In the example (Figure 2), observed agreement of .781 for two raters is a reduction from the single rater observed agreement index values of .900 and .860 (latter is not shown).

Expected Agreement Between Raters' Assignments. Assuming a traditional chance agreement model, $P_{ejj'}$ is obtained from the sum of cross products of the category marginal proportions: $P_{ejj'} = \sum_k (m_{jk} m_{j'k}) = \sum_k (\pi_j \mu_k + \varepsilon_j) (\pi_{j'} \mu_k + \varepsilon_{j'})$. By expressing ε_j in terms of π_j and expressing μ_k^2 in terms of the corresponding variance and mean of marginals, the overall expression is as follows: $P_{ejj'} = n \pi_j \pi_{j'} \sigma_M^2 + 1/n$ or $\{ (n-1)/n \} \pi_j \pi_{j'} \sigma_{ratio}^2 + 1/n$.

Except for replacing the single rater parameter π_j with the product of $\pi_j \pi_{j'}$, the forms for NSC are identical with those for NSI (Figure 1). As previously discussed, as category heterogeneity increases, expected or chance agreement increases. This effect is enhanced by higher nonchance rater accuracy. The NSC expected agreement for the two raters is .472 (Figure 2).

Kappa Coefficient: $\kappa_{jj'}$. Following its basic definition, using $P_{ojj'}$ and $P_{ejj'}$, $\kappa_{jj'}$ can be expressed as $\kappa_{jj'} = \pi_j \pi_{j'} [1 - \sigma_{ratio}^2] / [1 - \pi_j \pi_{j'} \sigma_{ratio}^2]$. In this form, it is clear that the agreement between the raters' assignments is affected by the product $\pi_j \pi_{j'}$ and

σ^2_{ratio} , the category heterogeneity. As in NSI, the impact of σ^2_{ratio} is greater on the numerator than it is on the denominator and $\kappa_{jj'}$ will be increasingly reduced as σ^2_{ratio} nears its maximum value. It is noteworthy that $\kappa_{jj'}$ is the same for all pairings of raters with the same product $\pi_j\pi_{j'}$. Because $\pi_j\pi_{j'}$ is the product of π_j and $\pi_{j'}$, concordance agreement κ will be smaller than the κ_{jt} index agreement for either rater. In the example (Figure 2), $\kappa = .585$.

In general, if the category heterogeneity is zero, κ_{jj} between two raters equals $\pi_j\pi_{j'}$ (.672 in the example). It is worth noting a parallel to CTT. If the raters are parallel ($\pi_j = \pi_{j'} = \pi$), the $\kappa_{jj'}$ coefficient for parallel raters is the square of the κ_{jt} index for a single rater. Likewise, the reliability coefficient $\rho_{xx'}$ between parallel observed scores is the square of the reliability index ρ_{xt} , the correlation between observed scores and true scores. That is, $\kappa_{jj'} = \kappa_{jt}^2$ just as $\rho_{xx'} = \rho_{xt}^2$. For nonparallel raters, the relationship is equally interesting: $\kappa_{jj'} = \kappa_{jt}\kappa_{j't}$.

Agreement Between Two Ratets for Stochastic Data

Assumptions for the stochastic categories are as for SI. Additionally, assumptions about the raters' assignments follow those for NSC. There are, however, two distinct conditions under which the rater assignments can be made. In the earlier example, the two trainees could be compared to one another in two ways: they could both observe a child's classroom interactions over the same time period (dependent stochastic concordance [DSC]) or they could observe interactions over different time periods (independent stochastic concordance [ISC]). In DSC, if there were a category error, it would be the same for both raters. In ISC, raters could be observing different manifest categories for the object.

Dependent Stochastic Concordance (DSC): Agreement Between Two Ratets on the Same Stochastic Category Manifestations

Category Membership. In DSC different raters see the same phases or manifestations for a given object. The expected manifest category marginal, μ_{sk} , and the category heterogeneity, σ^2_{Ms} , are the same for both raters. Stochastic category proportions and heterogeneity are as developed for the SI paradigm.

Assignment Parameters. As in NSC, assignments use a single parameter for each rater: π_j for the first rater and $\pi_{j'}$ for the second. Since raters' assignments are independent, marginals for each would not be affected by the other. Moreover, each rater would have a marginal similar to that in the SI paradigm. Furthermore, the variance of the category marginals for rater j will be $(\pi_j^2\pi_s^2)\sigma^2_M$ as in the one rater stochastic paradigm (illustrated in Figure 2).

As in NSC, the ratio of raters' marginal variances equals the ratio of their squared nonchance correct assignment parameters: $\sigma^2_{mj}/\sigma^2_{mj'} = (\pi_j^2\pi_s^2)\sigma^2_M/(\pi_{j'}^2\pi_s^2)\sigma^2_M = \pi_j^2/\pi_{j'}^2$. This relationship assumes that the true variance is nonzero. Since both sets

of marginals are linear functions of the true category frequencies, the raters' marginals would be proportional.

Observed Agreements and Disagreements Between Raters' Assignments. In DSC both raters see an object in the same phase, hence the probability of agreement (or disagreement) would follow the same pattern as in NSC (Table 1). Nonstochastic agreement is a function of the rater parameters and true category occurrence marginals (μ_k). In comparison, stochastic agreement would be a function of the rater parameters and manifest marginals (μs_k). Substituting μs_k into corresponding NSC equations (Figure 1) yields the following equations:

$$\Pr(a_{kk}) = (P_j P_{j'} - \varepsilon_j \varepsilon_{j'}) \mu s_k + \varepsilon_j \varepsilon_{j'} \quad \text{and} \quad \Pr(a_{kk'}) = \pi_j \varepsilon_{j'} \mu s_k + \varepsilon_j \pi_{j'} \mu s_{k'} + \varepsilon_j \varepsilon_{j'}$$

Overall observed agreement of the raters, $Po_{jj'}$, does not consider whether the agreement is correct or incorrect. Furthermore, agreement is not based on the manifestation being correct. Summing over all agreement cells, $Po_{jj'}$, is $\sum_k (a_{kk}) = \sum_k \{ (P_j P_{j'} - \varepsilon_j \varepsilon_{j'}) \mu s_k + \varepsilon_j \varepsilon_{j'} \}$. As in NSC, this reduces to the following: $Po_{jj'} = \{ (n-1)/n \} \pi_j \pi_{j'} + 1/n$, which is the same as for NSC (Figures 1 and 2). It is the probability of correct plus spurious agreements and is unaffected by σ^2_M or σ^2_{Ms} .

Expected Agreement Between Rater Assignments. Expected agreement, $Pe_{jj'}$, is obtained from the cross products of the raters' category marginal proportions: $\sum_k (m_{jk} m_{j'k})$. Using $m_{jk} = (\pi_j \pi_s) \mu_k + (1 - \pi_j \pi_s)/n$, $Pe_{jj'} = \sum_k \{ (\pi_j \pi_s) \mu_k + (1 - \pi_j \pi_s)/n \} \{ (\pi_{j'} \pi_s) \mu_k + (1 - \pi_{j'} \pi_s)/n \}$. Expanding the product and expressing μ_k^2 in terms of the corresponding variance ($\pi_s^2 \sigma^2_M$) and mean of marginals ($1/n$), the overall expression is given in terms of both variances of true and manifest marginals:

$$Pe_{jj'} = n \pi_j \pi_{j'} \pi_s^2 \sigma^2_M + 1/n = n \pi_j \pi_{j'} \sigma^2_{Ms} + 1/n$$

or

$$Pe_{jj'} = \{ (n-1)/n \} \pi_j \pi_{j'} \pi_s^2 \sigma^2_{ratio} + 1/n = \{ (n-1)/n \} \pi_j \pi_{j'} \sigma^2_{Sratio} + 1/n$$

These equations indicate that DSC expected agreement is less than $Pe_{jj'}$ for the corresponding NSC paradigm (Figures 1 and 2). The reason for this is that the variance associated with the manifest category marginals is regressed compared to the true category marginals.

Kappa Coefficient: $\kappa_{jj'}$. Based on the above, $\kappa_{jj'}$ for DSC can be expressed as

$$\kappa_{jj'} = \pi_j \pi_{j'} (1 - \pi_s^2 \sigma^2_{ratio}) / (1 - \pi_j \pi_{j'} \pi_s^2 \sigma^2_{ratio}) = \pi_j \pi_{j'} (1 - \sigma^2_{Sratio}) / (1 - \pi_j \pi_{j'} \sigma^2_{Sratio})$$

The DSC $\kappa_{jj'}$ depends on $\pi_j \pi_{j'}$ and σ^2_{Sratio} . The impact of category heterogeneity is greater on the numerator, hence, κ will be increasingly reduced as σ^2_{Sratio} nears its maximum value. However, DSC $\kappa_{jj'}$ exceeds the NSC $\kappa_{jj'}$ (Figures 1 and 2). As for

NSI versus SI, this arises from identical observed agreements, but a lower expected agreement for the stochastic paradigms due to regressed marginals.

Independent Stochastic Concordance (ISC): Agreement Between Two Raters on Independent Stochastic Category Manifestations

Category Membership. In ISC different raters may see different phases or manifestations for a given object; however, the expected manifest category marginal, μ_{sk} , and the expected category heterogeneity, σ_{Ms}^2 , are the same for both raters.

Assignment Parameters. Category assignments follow the same assumptions as for DSC with identical equations for the rater marginals and the variability of the rater marginals. However, for ISC, because of independence of category manifestations, the product of the rater and stochastic parameters, $\pi_j\pi_s$, functions like the single rater parameter, π_j , in NSC. A reparameterization, in which $\pi_j^* = \pi_j\pi_s$, $\varepsilon_j^* = (1 - \pi_j^*)/n$, and $P_j^* = \pi_j^* + \varepsilon_j^*$ allows the agreement statistics to be easily derived and puts them in the same form as for the other paradigms (Figure 2 illustrates reparameterization). Thus, the rater marginal for category k can be written with reparameterization as follows: $\mu_{jk} = \pi_j^*\mu_k + \varepsilon_j^*$.

Observed Agreements and Disagreements Between Raters' Assignments. Derivation of agreement (a_{kk}) and disagreement ($a_{kk'}$) cells focuses on the rater marginals within each manifest category for each true category. In particular, for agreement to occur both raters independently assign true category k with probability $P_j^*P_{j'}^*\mu_k$ or both assign the same wrong category k when k' is true with probability $\varepsilon_j^*\varepsilon_{j'}^*\mu_{k'}$. Combined over k and all $k' \neq k$, this yields:

$$\Pr(a_{kk}) = P_j^*P_{j'}^*\mu_k + \sum_{k'(k' \neq k)} \varepsilon_j^*\varepsilon_{j'}^*\mu_{k'} = (P_j^*P_{j'}^* - \varepsilon_j^*\varepsilon_{j'}^*)\mu_k + \varepsilon_j^*\varepsilon_{j'}^*$$

Using similar derivations, the disagreement cells can be expressed as

$$\Pr(a_{kk'}) = \pi_j^*\varepsilon_{j'}^*\mu_k + \varepsilon_j^*\pi_{j'}^*\mu_{k'} + \varepsilon_j^*\varepsilon_{j'}^*$$

The forms of these expressions are the same as for NSC (Figure 1).

Overall agreement, $Po_{jj'}$, does not consider whether the agreement is correct or incorrect nor whether the manifest category is correct. Observed agreement simply requires the summation of the agreement cells: $Po_{jj'} = \{(n-1)/n\}\pi_j^*\pi_{j'}^* + 1/n$. Using reparameterized values, ISC observed agreement has the same form as NSC (Figure 1). It is the probability of correct plus spurious agreements and is unaffected by σ_M^2 or σ_{ms}^2 ; however, ISC observed agreement is less than for NSC or DSC (Figure 2).

Expected Agreement Between Rater Assignments. Expected agreement, $Pe_{jj'}$, is obtained from the cross products of the observed marginal proportions. Since the marginals are

the same as for DSC, expected agreement will be the same (Figure 1) and likewise is less than expected agreement for NSC. The reason for this is the regressed variance associated with the manifest category marginals.

Kappa Coefficient: $\kappa_{jj'}$. Based on $Po_{jj'}$ and $Pe_{jj'}$, $\kappa_{jj'} = \pi_j^* \pi_{j'}^* (1 - \sigma_{\text{ratio}}^2) / (1 - \pi_j^* \pi_{j'}^* \sigma_{\text{ratio}}^2)$, as shown in Figure 1. The ISC paradigm has the smallest $Po_{jj'}$ and $\kappa_{jj'}$ of the concordance paradigms. This is illustrated in Figure 2 with observed agreement (.781, .781, .704) and $\kappa_{jj'}$ (.585, .603, .462) for the NSC, DSC, and ISC paradigms, respectively.

Interpretation Guidelines

The aforementioned developments distinguish among the five different paradigms and show that Po , Pe , and κ share common forms and properties for all of the paradigms. Furthermore, the derivations show that Po is independent of category frequencies, depending only on rater accuracy. Pe and κ also depend on rater accuracy but are affected by category membership frequencies through category variance (expressed relative to maximum possible variance as σ_{ratio}^2). Moreover, Pe varies directly with σ_{ratio}^2 and κ varies inversely with σ_{ratio}^2 . For stochastic paradigms, manifest category variance (σ_s^2) is regressed toward zero resulting in higher κ values for SI and DSC compared to NSI and NSC, respectively.

Kappa has often been described as difficult to interpret and guidelines have been issued to assist with interpretation (Landis & Koch, 1977; Watkins & Pacheco, 2000). However, these guidelines do not distinguish between index and concordance paradigms, nor take into account category frequency distributions, nor differentiate between rater accuracy and rater agreement. Figure 3 suggests interpretive quality ranges for index κ (left ordinate axis) and concordance κ (right ordinate axis) for different category variances and rater accuracy parameters (π_j for index paradigms and $\pi_j \pi_{j'}$ for concordance paradigms). The quality range limits for concordance κ s are the squared limits of index κ s, reflecting the relationship between concordance and index values. Figure 3 also shows the reduction in the numerical value of κ and corresponding degradation of quality as category variance is increased.

Quality ranges for nonrandom rater accuracy, π_j or $\pi_j \pi_{j'}$, are the same as for the corresponding index or concordance κ with 0 variance (see Figure 3) and are independent of category heterogeneity. Because observed agreement, Po , involves the number of categories (n), its quality ranges are not given in Figure 3; rather, quality for Po can be attributed based on the underlying π_j (or $\pi_j \pi_{j'}$) value, which can be estimated from Po (i.e., from $(nP_o - 1)/(n - 1)$).

As category variance increases, rater accuracy and agreement will not change but the equality between accuracy and κ diminishes, and κ will be less than the corresponding rater accuracy values. A further difference in the meaning of κ comes into play regarding category variance: The higher the variance, the greater the probability of *chance* agreement. Thus, the numerical value of κ decreases as category variance

Figure 3. Quality ranges for index and concordance kappas and nonchance rater accuracy.

increases (see Figure 3); however, interpretation of the numerical value of κ should not change. For example, an index κ of .50 represents a “moderate” quality for the measurement of individual differences whether it comes from a homogeneous ($\sigma^2_{\text{ratio}} = 0$) condition with underlying rater accuracy of .50 or comes from a high heterogeneous category variance condition ($\sigma^2_{\text{ratio}} = .75$) with rater accuracy of .80.

Simulations

The relationships of basic agreement statistics to underlying category and rater parameters provide theoretic expected values for sample statistics but the expected values and standard errors of sample statistics need to be confirmed or developed. For example, the expected value and standard error of the observed sample agreement P_o is

anticipated to follow a binomial distribution. Observed values of κ and expected agreement are expected to be unbiased and to depend only on the rater and category parameters. The observed standard deviation of κ within conditions should be comparable to the standard error of κ provided by Fleiss, Cohen, and Everitt (1969). Thus, simulated category assignments that vary N , category variance, and rater parameters will be used to estimate basic agreement statistics and their characteristics.

The concordance paradigm equations (Figure 1) indicate that P_o , P_e , and κ depend on the product of the rater parameters and not on the individual parameters. While the expected values are not dependent on the individual parameters, other distributional characteristics might be so dependent. Thus, simulated category assignments will be used to estimate basic agreement statistics for different pairs of π_j and $\pi_{j'}$ (which have the same product, $\pi_j\pi_{j'}$). The basic statistics of the different pairs will be compared for the concordance paradigms.

Simulation of Category Assignments

Simulated category assignments were generated to confirm and explore the above issues. Data were simulated using a three-category system {a, b, c} with two-category frequency distributions {9:2:1} and {5:4:3} used proportionally with each of four total frequencies (12, 24, 48, and 96). These category proportions provided a clear contrast of category heterogeneity variances (.0880 and .0046) with maximum and minimum nonzero, nonequal frequencies for $N = 12$. While category frequencies (variance and N) and paradigms were strictly controlled, other conditions were governed by probabilities. Stochastic parameter values of .9 and .8 were used for the SI, DSC, and ISC models. The stochastic parameters randomly established the manifest category "provided" to the "raters." For DSC, one manifest category was selected for both raters; for ISC, manifest categories were independently selected for each rater. A hypothetical rater j would correctly place an object in its correct (or manifest category) with probability P_j or into each alternative category with probability $\varepsilon_j = (1 - \pi_j)/3$. To provide a range of rater nonchance ($\pi_j > 0$) values, data were generated using rater nonchance parameters of .6, .7, .8, and .9 for the index paradigms and (.6, .6), (.7, .7), (.8, .8), and (.9, .9) for the concordance paradigms. The above-mentioned combinations yielded 256 simulation models.

Further simulation conditions were used to evaluate predictions about the product of nonchance parameters in the concordance paradigms. Thus, simulations were generated for rater parameter pairings of (.9, .4) and (.72, .5) matching the product of the concordance parameter pairing of (.6, .6). All three of these pairings ($\pi_j\pi_{j'} = .36$) should yield identical results for P_o , P_e , and κ . For the basic concordance paradigms, this added another 80 simulation models.

Finally, simulations were conducted for three different "chance" conditions: rater chance ($\pi_j = 0$ for each rater), stochastic chance ($\pi_{j'} = 0$) for SI, DSC, and ISC (each category has a 1/3 probability of occurrence regardless of the true category), and an all-chance model combining rater chance ($\pi_j = 0$) and stochastic chance ($\pi_{j'} = 0$) for

SI, DSC, and ISC. These 216 simulation conditions bring the total number of simulation models to 552.

For each simulation model, category assignments were randomly generated for each "stimulus" as presented (truth or stochastic manifestation) according to the rater's nonchance parameter (.6, .7, etc.). For example, for ISC, $N = 12$, $\sigma^2_{\text{ratio}} = .3958$, stochastic $\pi_{j'} = .8$, with raters at $\pi_j = .9$ and $\pi_{j'} = .4$, each rater was independently provided a probabilistic manifestation of the true category. If category A was true, A had an 86.7% (80% + 6.7%) chance of manifesting, while B and C each had a 6.7% chance of manifesting. The first rater assigned the "correct" (A, B, or C as manifested) category with probability .933, or each "incorrect" (not manifest) category with probability .033, while the second rater, with an independently chosen manifestation, assigned the "correct" category with probability .60 and each "incorrect" category with probability .20. The simulated assignments of the two raters were aggregated over the 12 stimuli into a category-by-category agreement matrix for each replication trial.

Simulated category assignments were compared to category manifestations (index paradigms) or to each other (concordance paradigms) and were aggregated over the respective number of stimuli (12, 24, 48, or 96) into category-by-category agreement matrices. Basic statistics were computed from each agreement matrix for 300 replications of each simulation condition. Altogether, 12,744,000 random category assignments by raters were generated.

Three different Excel programs were written for the simulations: a stochastic index program for SI and NSI ($\pi_s = 1$), a dependent stochastic concordance program for DSC and NSC ($\pi_s = 1$), and an independent stochastic concordance program for ISC. Each program was modified for either the .0880 or .0046 category proportions and allowed different stochastic and rater parameter values to be entered. The programs allowed N to be reduced from 96 to 48, 24, and 12. For each model, P_o , P_e , and κ were computed for each of 300 replications. These were aggregated over the replications to obtain observed means and standard deviations and the number of κ s equal to 1. Table 1 provides the expected and observed ranges of means of P_o , P_e , and κ for the diverse simulation models. Overall, expected P_o ranged from .33 to .94, expected P_e ranged from .33 to .57, and expected kappa ranged from .00 to .90.

Simulation Results

Observed Agreement (P_o). Observed agreement was obtained from the rater agreement matrices within each of the 552 models. The sample P_o follows a binomial distribution with expected value P_o and standard deviation (standard error) of $\sigma(P_o) = \sqrt{P_o(1 - P_o)/N}$. Under the assumptions used in this article, P_o depends only on the nonchance accuracy parameter π (Figure 1). Accordingly, P_o had a minimum expected value of .333 (rater $\pi = 0$, all paradigms) and a maximum of .933 (rater $\pi = .90$, NSI). The maximum standard error is at $P_o = .50$, decreasing quadratically as P_o decreases or increases. The resulting range of standard errors as N increases from 12

Table 1. Range of Observed^a and Expected Means for Basic and Chance^b Models.

Statistic	Index		Concordance		
	NSI	SI	NSC	DSC	ISC
Po					
Basic	.73 to .94 .73 to .93 ^c	.72 to .94 .73 to .93	.56 to .88 .57 to .87	.56 to .88 .57 to .87	.47 to .77 .49 to .77
Chance	.33 to .35 .33	.32 to .94 .33 to .93	.33 to .34 .33	.33 to .88 .33 to .87	.32 to .35 .33
Pe					
Basic	.34 to .57 .34 to .57	.34 to .53 .34 to .53	.34 to .55 .34 to .55	.34 to .52 .34 to .51	.34 to .51 .34 to .51
Chance	.33 to .34 .33	.33 to .38 .33	.33 to .34 .33	.33 to .38 .33	.33 to .34 .33
K					
Basic	.47 to .90 .47 to .90	.50 to .90 .50 to .90	.25 to .81 .25 to .81	.25 to .82 .28 to .81	.14 to .66 .15 to .65
Chance	-.01 to .01 .00	-.01 to .90 .00 to .90	-.01 to .01 .00	-.01 to .80 .00 to .81	-.02 to .02s .00
# of models					
Basic	32	64	48	96	96
Chance	8	56	8	72	72

^aEach range in the table is based on the means of the respective models in the indicated paradigm and chance condition. Each model's mean is based on 300 simulations. ^bChance indicates either rater chance ($\pi_j = 0$) or stochastic chance ($\pi_s = 0$) or both. ^cValues in italics represent the range of the expected means.

to .96, would be (.136 to .048) at the minimum Po, increasing to (.144 to .051) at Po = .50 and decreasing to (.072 to .025) for the maximum Po.

The average sample values of Po (among all 552 models) were less than .01 units from expected values (100% were within .02 units of expected values). The average difference between observed and expected means for basic and chance models for the five paradigms ranged from -.001 to .001 with corresponding root mean square differences (RMS) ranging from .004 to .006. The small RMS values indicated a close fit of means for each simulation condition within the paradigms. The observed standard deviation of Po was calculated within each of the 552 models and showed average differences between observed and expected values ranging from -.001 to .001 with RMS ranging from .003 to .005. The observed standard deviations appear unbiased and show an excellent fit to the expected values.

Expected Agreement (Pe). This article used Pe in the traditional sense: as a measure of how much raters would agree by chance based on their individual category marginals. Expected values of Pe (see Figure 1) depend on rater accuracy, category variance, and paradigm. Pe had a minimum expected value of .333 (all rater and stochastic chance models) and a maximum of .571 (rater $\pi = .90$, NSI, high variance).

Pe was calculated for each replication for each of the 552 models. Differences between observed and expected means for the 552 individual models ranged from $-.01$ to $.05$ with an overall mean difference or residual of $.002$ for nonchance models (RMS = $.005$) and $.006$ (RMS = $.011$) for chance models. The fit was poorest for SI and DSC, stochastic chance, $N = 12$ and $N = 24$ models: mean difference $.016$ (RMS = $.022$).

Kappa Coefficient (κ). κ was obtained from the observed Po and Pe values for each replication within each of the 552 models. Differences between observed and expected means ranged from $-.005$ to $.004$ units (98% were within $.002$ units of expected values). There was an overall mean difference of $.001$ for nonchance models (RMS = $.005$) and $-.003$ (RMS = $.009$) for chance models. The fit was poorest for SI and DSC, corresponding, perhaps, to the higher instability of Pe for these two models.

The estimated standard error of κ , $s(\kappa)$, provided by Fleiss et al. (1969) was calculated on the expected rater marginals for each of the 552 models. Observed standard deviations for κ were computed for the replications within each of the models and were compared to the expected values. The difference between the observed and the estimated values had a mean of $-.001$ and a root mean square residual of $.012$. The larger residuals were noted for NSI and SI and may arise from the fact that $s(\kappa)$ was not designed for comparing one rater to a fixed marginal. Larger residuals (not shown) were also associated with low N and high category heterogeneity models. Larger residuals also occurred for models with a high expected value for κ . This latter factor was likely due to range compression of the observed values of κ . The poorer fit under these conditions is similar to a finding by Gwet (2016). Range compression was specifically noted for low N and high κ index models: for the 6 NSI and SI models with $N = 12$ and rater $\pi = .9$, from 43% to 49% of the κ s (out of 300 replications) equaled 1. With $N = 24$, the corresponding models had 19% to 23% of the κ s equal to 1.

Rater Accuracy Product Versus Distinct Values. The concordance paradigm equations (Figure 1) indicate that Po, Pe, and κ values depend only on the product of the individual rater nonchance parameters and not the distinct individual values (interpair differences). This proposition was evaluated for each of 40 models by using three sets of rater parameters (.6, .6), (.72, .5), and (.9, .4) having the same product (.36). The 40 models are based on 4 levels of N by 2 levels of category heterogeneity by 5 models (NSC, DSC .9, DSC .8, and ISC .9, ISC .8). The different pairs of rater parameters produced virtually identical means for Po, Pe, and κ (within $.002$ of each other and of expected values). The largest differences among the three pairs always occurred for $N=12$ and were in either a DSC or ISC model.

The standard error of κ (Fleiss et al., 1969) depends on the marginal distribution for each rater. While models with rater pairs (.6, .6), (.72, .5), and (.9, .4) should yield identical mean values for Po, Pe, and κ , they have different marginals and, thus,

different standard errors for κ . In particular, with everything else held constant, $s(\kappa)$ *reduces* in magnitude as raters differ from each other. A combination of a very accurate rater with an inaccurate rater (e.g., .9, .4) provides a smaller standard error than if both raters are equivalent (e.g., .6, .6). The observed standard errors for P_o did not vary with interpair differences. The observed standard deviations of κ varied slightly with interpair differences and, accordingly, fit the expected standard errors.

Simulation Summary

In total, 12,744,000 assignments by raters to three categories were generated controlling for number of objects assigned, proportion of objects belonging to the categories, probability that objects manifest in categories, and the probability that raters would put them in their apparent categories. All of simulations used a homogeneous error of assignment model. Some assignments compared ratings to apparent category memberships (index paradigms) and some compared raters to one another (concordance paradigms). Models were used to test the hypothesis that for concordance paradigms P_o , P_e , and κ relied only on the product of rater accuracies and not on the individual rater accuracies.

Simulations indicated that observed P_o , P_e , and κ were unbiased estimators of underlying theoretic values. The standard errors of P_o and κ were also closely approximated by observed standard deviations. While estimates of P_e were unbiased, P_e did seem unstable for low N and low variance. Also, the simulations showed the product of rater accuracies was sufficient for determining values of P_o , P_e , and κ . Observed standard deviations of κ varied with interpair differences corresponding to expected values (Fleiss et al., 1969).

Discussion

The overall purpose of this article was to clarify some long-standing issues about rater agreement and kappa. One goal was to clarify the difference between the accuracy of category assignments and the reliability of category assignments. The analogous delineation with CTT is between the accuracy of measuring an attribute of an object and the reliability of individual differences. The accuracy of measuring an attribute is often assumed to be independent (within a range) of the object while the corresponding reliability of measurement is known to be a function of the accuracy of measurement and the distribution of individual differences. Observed agreement, P_o , is strictly a function of rater accuracy while κ functions exactly like reliability. Many issues surrounding agreement and kappa are related to this distinction.

In this article, the investigation of kappa and rater accuracy began by specifying properties of category systems. A simple system of nonstochastic membership in mutually exclusive categories was liberalized by introducing stochastic category membership: objects appear to belong to the correct category with probability π_s , otherwise, they appear to belong to any category randomly (including the correct

one) with probability ε_j . The frequency of membership in categories was captured by the concept of category heterogeneity or variance: a concept analogous to true score variance in classical test theory, except that higher category variance is similar to lower true score variance. This concept of category variance allowed accuracy and agreement statistics to be derived without reference to specific category frequencies or proportions.

Rater assignments to categories followed simple assumptions: a rater correctly assigns objects to categories with probability π_j , otherwise objects are assigned to any category with probability ε_j . The probability ε_j is analogous to the standard error of measurement in classical test theory and it is independent of the individual category and of the category frequency distribution. In addition, different paradigms were developed to describe data collection procedures for raters making category assignment. Two paradigms describe the assignments of a single rater to nonstochastic (NSI) or stochastic (SI) category memberships. These index paradigms correspond to a reliability index, that is, correlating observed scores to true scores. They differ from three concordance paradigms in which raters are compared to one another. One paradigm considers assignments to nonstochastic category memberships (NSC). The others consider either both raters making assignments to identical stochastic memberships (DSC) or each rater making assignments to independent stochastic memberships (ISC).

Formulae were developed for each paradigm to describe observed agreement (Po), expected rater marginals and agreement by "chance" (Pe), and a κ coefficient. For all paradigms, derivations indicated that Po is a function of rater accuracy (or product of rater accuracies) and the number of categories. Po is not affected by marginal asymmetry (unequal marginals) and can be expected to be invariant over changes in category heterogeneity; however, Po is affected by data collection paradigms. Po is higher for index paradigms and is affected by stochastic category membership only for the ISC paradigm. The reliance of Po on number of categories is a function of splitting the error ($1 - \pi$) over n . While increasing n would seem to be a good idea, it would likely result in making the basic assumption of error homogeneity untenable.

Pe is based on chance agreement between a rater's assignments and category membership (Index paradigms) or between raters' assignments (Concordance paradigms). The formulae for Pe are similar in form for all paradigms: all include the rater(s) accuracy times the category variance (true or stochastic, depending on the paradigm) and the number of categories. The inclusion of category variance in Pe makes it clear that expected agreement is not invariant over changes in population heterogeneity and the more homogeneous the category frequencies, the more that Pe approaches $1/n$.

The equations derived for Cohen's κ under the assumptions of this article show that κ is dependent on three characteristics: data collection paradigm, rater(s) accuracy, and category variance. Other notable features follow from the equations: κ increases with rater accuracy but decreases with increased category heterogeneity; holding heterogeneity constant, index κ will always be larger, tending to be at least

the square root of concordance κ , stochastic κ will exceed the corresponding nonstochastic κ . The relationship of κ to category variance also has explicit meaning. High category variance occurs when most of the objects are in just one category, a condition analogous in CTT to having all true scores equal. With classical reliability measures, the less the true score variance, the lower the reliability coefficient. This does not mean that the measure is not good: it means that the measure is not good at differentiating among the available individuals. However, in CTT, given a constant standard error of measurement, reliability can be estimated for other true score distributions. The formulae for κ provide the same possibility. Given an estimate of rater accuracy parameters, the basic κ could be used to estimate κ for different category variances, including the homogeneous frequency variance of zero. Of course, the most appropriate estimate would take into account the category frequency distribution in which individual differences were to be investigated. Under the assumptions made in this article, κ is a function of category variance modified by rater accuracy; however, differences between rater marginals per se do not affect κ . Marginal asymmetry would affect κ only to the degree that it was based on differential rater biases.

Quality ranges were suggested to assist in the interpretation κ relative to rater accuracy and to category variance. Except for two extremes ("near perfect" and "poor"), the four ranges are of equal size for index κ ; however, the same "qualities" are then associated with the squared range limits for concordance κ , representing the relationship between index and concordance values.

The formulae derived for agreement statistics were evaluated by simulating rater assignments for different membership and assignment assumptions for the different paradigms. Averages and standard deviations of P_o , P_e , and κ were compared to their underlying theoretic expected values within and across paradigms, category heterogeneity conditions, and parameter values. All comparisons indicated that P_o and κ means and standard deviations fit expected values very well. P_e revealed a slight degree of instability but this had no obvious consequence on the expected value of the standard error for κ . Holding the rater accuracy product constant, differences between raters did not matter for P_o , P_e , and κ and varied appropriately for the Fleiss et al. (1969) standard error for κ . Standard errors reasonably varied inversely with sample sizes with the smallest sample size of $N = 12$ showing major ceiling effects for κ .

The assumptions made in this article are likely idealistic; however, they serve a heuristic function, leading to some very useful findings:

- Distinguishing between category membership and rater assignment allowed the former to serve as a condition for varying the latter.
- Allowing category membership to have a stochastic aspect liberated the comparison of a fallible rater to an infallible one to include the more realistic comparison of a fallible rater to a (possibly) less fallible one.
- Stochastic category membership provided for dependent versus independent concordance paradigms.

- Replacing specific category frequencies with a single category heterogeneity parameter allowed P_o , P_e , and κ to be expressed in common forms, showing the impact of category heterogeneity on P_e and κ .

The assumptions used in this article represent a starting point for considering alternative assumptions. As mentioned earlier, assumptions about category membership could differ from the homogeneous error model included in this article. For example, for each category, items could have a membership difficulty level. Some items could be easy for all raters and other might be so difficult that assignment would be at chance level. Similarly, as mentioned by Green (1981) and assumed by Andres and Marzo (2004), rater accuracy in a given category could be “certain” for a subset of items and totally random for the remainder. However, assignments and errors for different raters would be independent only if the “certain” subsets were independent. In fact, if the subsets were independent and included homogeneity of accuracy across categories, the results would coincide with those of this article. Although they were interested in evaluating measures of global agreement and deviations from a basic model, the basic model of Cousineau and Laurencelle (2015, 2016) was similar to the one used in this article. Although not focusing on a parametric model, Cousineau and Laurencelle (2015, 2016) used a systematic method of disagreement at the individual rater level and noted that under conditions of high category variance, systematic disagreement (bias) between raters did not give a downward bias to κ , rather it gave an upward bias to the observed rate of agreement. Different assumptions could also be made about rater assignments. For example, Green (1981) mentioned the possibility that errors of assignment might match the base rates of category frequencies. In such a case, it is very likely that κ would increase if rater error patterns match one another, particularly if the pattern matches the distribution of category frequencies (base rate confirming patterns). However, raters could share base rate compensating biases, in which lesser represented categories are assigned when there is uncertainty, and these shared biases would also increase κ . On the other hand, if rater biases are in opposition to one another or are randomly different from one another, κ is likely to be reduced. It seems that this latter type of marginal asymmetry should underlie concerns of differential rater bias. Theoretic derivations and simulations with these assumptions would help address the issue of how to use agreement matrix asymmetry to detect and measure bias. Derivations of formulae and simulations of category assignments seem to require a degree of parametric assumptions. It does not seem that category assignments and κ can be understood well by simulating values of P_o or cells of agreement matrices without fully understanding how those values represent underlying processes and parameters.

Not all issues of interest have been discussed in this article. In particular, estimation of the underlying category frequency and rater and stochastic parameters has not been addressed. Estimation of true category frequencies and category variance is clearly of interest, as would be the estimation of the stochastic parameter for category manifestation. Furthermore, estimates of the rater accuracy parameters would be very useful, particularly when one rater might be quite superior to another. These issues

have already been pursued; however, space limitations require that these issues be addressed in a separate article.

Authors' Note

Anthony J. Conger is now at psyCATS, Crete, Illinois, and is Professor Emeritus at Purdue University.

Acknowledgments

Special acknowledgement is given to Judith A. Dygdon, Roosevelt University, for her helpful suggestions in organizing and clarifying this article.

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.

Notes

1. Technically, object i either belongs in a category or does not, as described in the deterministic model. The term *probability* is used to capture the idea of stochastic membership and relative frequency of occurrence of objects across categories.
2. Proportion or probability is preferred to frequency because it allows for comparisons, without reference to N , the total number of objects being classified.
3. Cousineau and Laurencelle (2016) discuss alternative conceptualizations of chance agreement.

References

- Andres, A. M., & Marzo, P. F. (2004). Delta: A new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology*, 57, 1-19. doi: 10.1348/000711004849268
- Bakeman, R., Quero, V., McArthur, D., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2, 357-370. doi:10.1037/1082-989X.2.4.357
- Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687-699. doi: 10.1177/001316446002000104
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46. doi:10.1177/001316446002000104
- Cousineau, D., & Laurencelle, L. (2015). A ratio test of interrater agreement with high specificity. *Educational and Psychological Measurement*, 75, 979-1001. doi: 10.1177/0013164415574086

- Cousineau, D., & Laurencelle, L. (2016). An unbiased estimate of global interrater agreement. *Educational and Psychological Measurement, 76*, 1-22. doi:10.1177/0013164416654740
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- Dou, W. A., Ren, Y. A., Wu, Q. A., Ruan, S. C., Chen, Y. D., Bloyet, D. B., & Constans, W. E. (2007). Fuzzy kappa for the agreement measure of fuzzy classifications. *Neurocomputing, 70*, 726-734. doi:10.1016/j.neucom.2006.10.007
- Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327. doi:10.1037/h0028106
- Green, S. B. (1981). A comparison of three indexes of agreement between observers: Proportion of agreement, G-index, and kappa. *Educational and Psychological Measurement, 41*, 1069-1072. doi:10.1177/001316448104100415
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*, 29-48. doi:10.1348/000711006X126600
- Gwet, K. L. (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement, 76*, 609-637. doi:10.1177/0013164415596420
- Hadzi-Pavlovic, D. (2010). Sample size for kappa. *Acta Neuropsychiatrica, 22*, 199-201. doi:10.1111/j.1601-5215.2010.00479.x
- Holley, J. W., & Guilford, J. P. (1964). A note on the G index of agreement. *Educational and Psychological Measurement, 24*, 749-753. doi:10.1177/001316446402400402
- Kraemer, H. C. (1982). Estimating false alarms and missed events from interobserver agreement: Comment on Kaye. *Psychological Bulletin, 92*, 749-754. doi:10.1037/0033-2909.92.3.749
- Landis, J. R., & Koch, G. G. (1977). The measurement of interobserver agreement for categorical data. *Biometrics, 33*, 159-174. doi:10.2307/2529310
- Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76*, 365-377. doi:10.1037/h0031643
- Uebbersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin, 101*, 140-146. doi:10.1037/0033-2909.101.1.140
- Von Eye, A., & Von Eye, M. (2005). Can one use Cohen's kappa to examine disagreement? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*, 129-142. doi:10.1027/1614-2241.1.4.129
- Von Eye, A., & Von Eye, M. (2008). On the marginal dependency of Cohen's κ . *European Psychologist, 13*, 305-315. doi:10.1027/1016-9040.13.4.305
- Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education, 10*, 205-212. doi:10.1023/A:1012295615144
- Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *Journal of Consulting Psychology, 86*, 1219-1227. doi:10.1037/a0037489