ASSESSING FIVE COMMON MEASURES OF INTEROBSERVER RELIABILITY: PROPOSING NEW REFINED MEASURES

BY

SHAWN M. CHMIL

A Thesis

Submitted to the Faculty of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Statistics

University of Manitoba

Winnipeg, Manitoba

0-612-32911-9

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

ASSESSING FIVE COMMON MEASURES OF INTEROBSERVER RELIABILITY: PROPOSING NEW REFINED MEASURES

BY

SHAWN M. CHMIL

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree

of

MASTER OF SCIENCE

ABSTRACT

It is frequently desired to determine the extent of agreement between two raters when the data are measured on an ordinal scale. Five common measures of interobserver reliability are the overall proportion of agreement, Cohen's kappa, weighted kappa, the disagreement rate and the concordance between raters.

A number of studies have assessed interobserver reliability including ones which have reservations about the measures of reliability and others which recognize several paradoxes. It is known that chance-corrected measures of agreement are prone to exhibit paradoxical and counter-intuitive results. Also, if measures are to be adjusted for chance agreement, then the guessing mechanism needs to be specified properly and precisely, as the current assumption that all observations are guessed is simply impractical.

The inadequacies of these measures are discussed and, in light of their deficiencies, new measures are proposed. The assumption that some but not all observations are guessed is used to develop three new measures of interobserver reliability, namely, partial-chance proportion, partial-chance kappa and the expected-chance proportion.

Simulations are used to compare the finite sample performance of these measures. In the simulations. the concordance between raters produced the best results. closely followed by partial-chance proportion. expected-chance proportion and partial-chance kappa. in terms of bias. efficiency and the empirical distributions of critical ratios.

Recommended measures of interobserver reliability are the concordance between raters. partial-chance proportion. expected-chance proportion and partial-chance kappa. Although the concordance between raters is highly advised, its usage should be cautioned as it is based on assumptions that are impractical in clinical practice.

ACKNOWLEDGMENTS

I owe very deep gratitude to my advisor, Dr. Bruce Johnston, for his expert counsel and patient encouragement. For his many helpful and constructive comments and suggestions, I am sincerely thankful.

I am also thankful to committee members Dr. Kenneth Mount and Dr. Shashi Seshia for their thought-provoking questions and fruitful discussions.

Finally, I am greatly indebted to my mother and father, who gave me life, and for their continued love and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Studies in medicine, epidemiology, sociology, psychology and psychiatry are conducted in which two or more raters independently examine a group of subjects to determine whether attributes are present or absent in each of the subjects. The subjects may be human or animal subjects, written materials, X-rays, etc. The usual goal of such studies is to evaluate how well the raters identify the attributes for any variable under investigation. Since the raters will invariably make some incorrect assessments, quantification of rater performance becomes an important statistical concern. When the set of attributes possessed by each subject is known without error, this becomes an issue of validity rather than agreement. However, in some situations an absolute standard is not known, and hence measurement of the extent of agreement among the different raters, *interobserver agreement*, is of primary interest.

The reliability of a classification procedure refers generally to the degree of reproducibility attained in repeated use of the procedure. Ideally, reproducibility would be measured by repeated evaluation of the same subjects by the same raters on different occasions. However, since a rater-subject pair may not be usually be used more than once, most studies designed to measure the reliability of an instrument employ a large group of subjects, who are considered representative of a population of interest. The subjects are then evaluated by a small group of raters, and the agreement displayed by the raters in classifying the subjects is used as a measure of reliability of the classification instrument. If agreement among the raters is high, then there is a possibility that the ratings do in fact reflect the dimension they are purported to reflect. If their agreement is low, on the other hand, then the usefulness of the variable rated is severely limited. It is futile to ask what is associated with the variable in question when one cannot trust those ratings.

In psychological investigations it frequently happens that two or more raters interview the same sample of subjects for the purpose of allocating them to various categories (Cohen 1960). For example, the raters may be clinical psychologists, the categories schizophrenic, neurotic or brain-damaged, and the subjects psychological test protocols; or the raters may be social psychologists, the categories various types of leadership, and the subjects small groups, etc. In such situations, one would desire assurance that the diagnosis given a patient is valid, i.e., actually serves the purpose intended. In the absence of ultimate criteria for validating psychological diagnosis, the question arises as to the degree of agreement between the raters.

The management of the comatose child is determined by the assessment of the level of consciousness and brainstem function (Gordon *et al.* 1983). A wide range of conditions may be associated with coma or impaired consciousness. Apart from acute brain damage, there are metabolic disorders and therefore, it is vital to be able to assess and to record changing states of altered consciousness reliably. Unfortunately, the level of consciousness cannot be directly measured and its estimation requires the interpretation of several clinical signs. The grouping of such signs has brought forth different types of scales, and some have included brain stem signs (Born *et al.* 1987). The three most commonly used coma scales in unconscious children are the Adelaide Scale, a pediatric modification of the Glasgow Coma Scale (Simpson and Reilly 1982) (Table A-1), the Jacobi Scale (Table A-2), and the 0-IV Scale (Table A-3). Agreement among different observers is an indispensable condition for the validation of an evaluation scale of consciousness disorders. Assessment of interobserver agreement is desirable since important judgmental decisions are made on the basis of the clinical information, and if the data are to be used in clinical research.

Necrotizing enterocolitis (NEC) is the most common acquired gastrointestinal emergency in the neonatal intensive care unit and is suspected when gastrointestinal signs and symptoms predominate (Kliegman and Fanaroff 1984). It occurs mainly in premature neonates, predominantly in the first two weeks of life, the incidence between 1 and 5 percent of admissions to the neonatal intensive care unit. Overall, NEC has a mortality of 20 to 40 percent. Its pathogenesis is still incompletely understood and there is no clinical sign or laboratory test that confirms the diagnosis. The interpretation of the abdominal radiograph is the most important factor in making a definitive diagnosis of NEC, with management strategies generally guided by NEC staging based on clinical and radiographic features (Bell *et al.* 1978). Although correct interpretation of abdominal radiographs is the single most important factor in diagnosing NEC, a wide range of interobserver variability in their interpretation has been suggested (Mata and Rosengart

1980; Markus *et al.* 1989). Clearly, observer variation is an important consideration in the interpretation of abdominal radiographs as the signs and diagnoses for which agreement is poor cannot be considered reliable.

A number of studies have assessed interobserver reliability (Shinar *et al.* 1987; Solari *et al.* 1989) including ones which have reservations about the measures of reliability (Kupper and Hafner 1989; Posner *et al.* 1990; Yager, Johnston and Seshia 1990) and others which recognize several paradoxes (Feinstein and Cicchetti 1990; Byrt, Bishop and Carlin 1993).

Numerous measures of interrater agreement have been used to quantify the degree of concordance among two raters, but it should be clear that there must be more to the measurement of interrater agreement than the arbitrary selection of an index of agreement. Five common measures of interobserver reliability are the overall proportion of agreement, Cohen's (1960) kappa, weighted kappa (Cohen 1968), the disagreement rate (Teasdale, Knill-Jones and Van der Sande 1978) and the concordance between raters (Kupper and Hafner 1989). This thesis will evaluate the five common measures of interobserver reliability and propose new refined measures. These measures will be described, contrasted and their properties illustrated in order to aid users with interpretation and selection. The reliability of these measures can be studied through simulation techniques.

2

LITERATURE REVIEW

Occasionally, the $k \times k$ table of joint categorical assignment frequencies of two raters (where each rater has made assignments to the same k-level nominal scale) has been treated as a contingency table. This having been done, many investigators have computed χ^2 over the table for use as a test of the hypothesis of chance agreement, and some have gone on to compute the contingency coefficient C as a measure of agreement (McNemar 1962). The defect of χ^2 in this context, and therefore of C, is that it indexes association and not necessarily agreement, which is the special kind of association of interest in reliability.

The simplest and most frequently used index to measure interobserver agreement has been the overall proportion of agreement, i.e., the ratio of the number of cases in which the raters agreed to the number of cases. This index suffers in that it includes agreements which can be accounted for by chance.

Different opinions have been stated on the need to incorporate chance-expected agreement into the assessment of interrater reliability. The presumptive reason for the *chance correction* is that the measuring instruments are often human observers, rather than inanimate technologic procedures, and that the subjective responses of the raters may sometimes agree by chance. Cohen's (1960) kappa adjusts for chance-expected agreement. and can be interpreted as the proportion of agreement after chance agreement is removed from consideration. This apparent virtue of the kappa coefficient has made it increasingly popular in studies of interobserver reliability, but many investigators are not aware of an important disadvantage, kappa is affected by prevalence, leading to two paradoxes in the kappa coefficient (Feinstein and Cicchetti 1990). The use of these particular measures in practice can be misleading as difficulties with their use and interpretation have been cited.

The development of weighted kappa (Cohen 1968) is motivated by studies where the relative seriousness of each possible disagreement could be quantified. It can be interpreted as the proportion of weighted agreement corrected for chance. The weights assigned are an integral part of how agreement is defined and therefore how it is measured with weighted kappa. Cohen's (1960) kappa makes no such distinction, implicitly treating all disagreement cells equally. Weighted kappa has been advocated as one of the preferred methods for the analysis of agreement data.

Properties of kappa and weighted kappa, in particular, approximations of their standard errors, have been given by Cohen (1960, 1968) and Everitt (1968). However, they are in error, having been derived from the contradictory assumptions of fixed marginal totals and binomial variation of cell frequencies. The errors seem to be in the direction of overestimation, so that their use results in conservative significance tests and confidence intervals. Everitt (1968) derived the exact variances of kappa and weighted kappa when the parameters are zero assuming a generalized hypergeometric distribution. Valid formulas for the approximate large-sample variances are given by Fleiss. Everitt and Cohen (1969), which do not require such assumptions.

The disagreement rate incorporates the magnitude of disagreement, and was first proposed in an article by Teasdale. Knill-Jones and Van der Sande (1978). It basically takes account of differences between raters, although it does not adjust for agreement expected by chance as in kappa. It has been previously suggested (Yager. Johnston and Seshia 1990) that the disagreement rate and kappa statistics may provide different yet complementary information about interobserver agreement, where the former provides a better measure of the degree of disagreement and the latter corrects for chance-expected agreement.

For assessing the extent of interrater agreement for multiple (nominal) response data, Kupper and Hafner (1989) derived a two-rater concordance statistic, the concordance between raters. This statistic is comparable to kappa in that there is an adjustment for chance-expected agreement, however, the assumption in the guessing mechanism differs between the two statistics. The adjustment for chance-expected agreement in the concordance between raters is based on the assumption that the observers are guessing by giving every one of the categories an equal chance of being observed, whereas the adjustment in kappa uses the marginal totals in the familiar approach to contingency tables. Furthermore, the adjustment made in either coefficient is based on the assumption that all subjects are guessed.

A number of assumptions underlying the use of standard statistical tests of reliability may not be valid when applied to many rating scales, as convincingly argued by Hall (1974): (1) scores are distributed normally. The distribution of scores obtained will depend on the degree of handicap in the rated sample, which will often contain many grossly abnormal individuals; (2) agreement is meaningful. For agreement to be a useful measure it should take account of both partial agreement and the total score distribution; (3) chance agreement is negligible. There is a certain level of agreement between two raters on an item that could be attained by chance alone. If one category of an item is consistently rated more frequently than others then the overall probability of agreement by chance will be higher; (4) total scores are meaningful. While total scores are normally more stable than item scores, they may lead to false results in calculating interrater reliability. Total score reliabilities may therefore give spurious values, so that the reliability of the individual item scores making up the total score should be examined; and (5) mean scores of both raters are similar. Correlation methods fail to take account of differences between the means, so that apparently good reliabilities can be obtained with significant differences between sets of scores, as well, the addition of a constant in order to correct one set of means will not correct the skewed form of score distribution that may be associated with such differences.

Hall (1974) states, further, that the test of choice for calculating reliability with rating scales should: (1) be distribution free; (2) allow credit for partial rater agreement; (3) correct for rater agreement due to chance alone; (4) make use of individual items in the rating scale; and (5) correct for differences in rater mean scores. One method which appears to meet these criteria satisfactorily is weighted kappa, introduced by Cohen (1968), however, Graham and Jackson (1993) identified serious problems with the use of weighted kappa, suggesting that weighted kappa behaves more like a measure of association than an index of agreement.

3

STATEMENT OF THE PROBLEM

Chance-corrected measures of agreement are prone to exhibit paradoxical and counter-intuitive results when used as measures of reliability. It will be demonstrated that these problems arise with both kappa and weighted kappa, and that the correction for *guessing* (knowledge-based decision making under uncertainty) needs to be carefully considered because these statistics change dramatically with this assessment.

The adjustment made for agreement expected by chance alone requires the guessing mechanism to be specified precisely. There is an implicit assumption in kappa, weighted kappa, and the concordance between raters that some rater's scores are based on perfect knowledge. Other rater's scores on a subject, based on less than perfect knowledge, are then guessed at with some of the guesses being correct (raters agree) and some incorrect (raters disagree). The adjustments in these statistics then are computed based on the assumption that all subjects are guessed, rather than assuming that only a subset of them are guessed, with the statistics differing in the assumptions for the probability of falling in the various categories.

It is desired to evaluate the five common measures of interobserver reliability, to determine the conditions under which the statistics can be readily used, and the limitations. Based on the failings of these current statistics, new measures of interobserver reliability are proposed in this thesis. These newly refined measures are induced by the assumptions that only some observations are subject to classification by chance, and secondly, the information of the number of disagreements arising can be used to determine a likely value for the number of subjects guessed.

Computer simulations are used to compare the finite sample performance of the measures of interobserver reliability. In particular, simulation is used to achieve two objectives: first, to evaluate the accuracy of the statistics and their variances; and second, to compare the empirical distributions of provided critical ratios under the hypothesis of no association between the two raters' examinations with the theoretical normal distribution.

4

METHODOLOGY

4.1 Examination of the Subjects

Consider a study in which two raters, say rater A and rater B, independently examine each of n subjects. Assume that each subject is examined by the raters within a short period of time (minimize the possibility of clinical change in the interval) and that the raters did not see each other examine the subjects (minimize any bias involved during the course of an examination). Following examination of the rth subject, $r = 1, 2, \ldots, n$, each rater must decide which one attribute, from $k \geq 2$ mutually exclusive and exhaustive (categorical) attributes for any variable under investigation, best describes the rth subject.

4.2 Symbols and Notation

Let n_{ij} denote the number of subjects assigned to category i by rater A and to category j by rater B; let

$$n_{i} = \sum_{j=1}^{k} n_{ij} \qquad \text{and} \qquad n_{j} = \sum_{i=1}^{k} n_{ij}$$

denote the total number of subjects assigned to category i by rater A and to category j by rater B, respectively. The resulting frequencies can be arranged in a $k \times k$ contingency table with cell frequencies n_{ij} and $\sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} = n$:

	Rater B				
Rater A	1	2	...	k	Total
1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
.					
.					
.					
k	n_{k1}	n_{k2}	...	n_{kk}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	n

The n subjects will be regarded as a sample of size n from some target population according to some characteristic of interest. Under the hypothesis of rater independence and, conditional on the marginal totals, the distribution of a single n_{ij} is hypergeometric (Everitt 1968)

$$\frac{\binom{n_{i.}}{n_{ij}}\binom{n-n_{i.}}{n_{.j}-n_{ij}}}{\binom{n}{n_{.j}}}. \qquad (1)$$

Therefore, the null expected value and null variance of n_{ii} based on (1) are, respectively,

$$E_0(n_{ii}) = \frac{n_{i.}\, n_{.i}}{n} \qquad \text{and} \qquad Var_0(n_{ii}) = \frac{n_{i.}(n-n_{i.})\,n_{.i}(n-n_{.i})}{n^2(n-1)}.$$

Also the sum of any number of the n_{ij} is a hypergeometric variable (Everitt 1968), and using this fact the null covariance of any two of the n_{ij} can be derived. The null covariance for any two elements in the same row, in the same column, and for any diagonally opposed elements has been shown by Everitt (1968) to be, respectively,

$$Cov_0(n_{ij}, n_{it}) = \frac{-n_{i.}\, n_{.j}\, n_{.t}(n-n_{i.})}{n^2(n-1)}, \qquad j \neq t,$$

$$Cov_0(n_{ij}, n_{sj}) = \frac{-n_{i\cdot}\, n_{s\cdot}\, n_{\cdot j}(n - n_{\cdot j})}{n^2(n-1)}, \qquad i \neq s,$$

$$Cov_0(n_{ij}, n_{st}) = \frac{n_{i\cdot}\, n_{s\cdot}\, n_{\cdot j}\, n_{\cdot t}}{n^2(n-1)}, \qquad i \neq s,\ j \neq t.$$

Let

$$T_o = \sum_{i=1}^{k} n_{ii} \qquad \text{and} \qquad T_{ow} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\, n_{ij}$$

denote the total number of agreements observed and the weighted total number of agreements observed, respectively. The w_{ij} are a set of weights indicating the level or amount of agreement. These weights are arbitrary and chosen by the experimenter.

Using the values derived for the null variances and null covariances of the n_{ij}, then it follows that

$$Var_0(T_o) = \frac{1}{n^2(n-1)}\Big\{\sum_{i=1}^{k} n_{i\cdot}(n - n_{i\cdot})\, n_{\cdot i}(n - n_{\cdot i}) \;+\; 2\sum_{i=1}^{k}\sum_{\substack{j=1\\ i<j}}^{k} n_{i\cdot}\, n_{j\cdot}\, n_{\cdot i}\, n_{\cdot j}\Big\},$$

$$Var_0(T_{ow}) = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}^2\, \frac{n_{i\cdot}(n - n_{i\cdot})\, n_{\cdot j}(n - n_{\cdot j})}{n^2(n-1)} \;+\; 2S\, w_{ij}\, w_{st}\, Cov_0(n_{ij}, n_{st}).$$

where S denotes the appropriate summation over the whole table.

The argument for finding the expected values is the same as that used in the familiar approach to contingency tables (Bhattacharyya and Johnson 1977). Then,

$$T_c = \sum_{i=1}^{k} \frac{n_{i\cdot}\, n_{\cdot i}}{n} \qquad \text{and} \qquad T_{cw} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\, \frac{n_{i\cdot}\, n_{\cdot j}}{n},$$

the total number of agreements expected by chance alone and the weighted total number of agreements expected by chance alone. respectively.

4.3 Five Common Measures of Interobserver Reliability

In a sample of n subjects. a number of agreements will arise based completely on perfect knowledge, i.e.. guessing will have played no role in these c agreements. The ratio of these two numbers gives the true proportion

$$P_t = \frac{c}{n}, \qquad 0 \le c \le n.$$

where c is generally unknown.

Overall Proportion of Agreement

The simplest agreement index is based on the proportion of subjects classified into the same category by the raters. It is given by

$$\hat{P}_a = \frac{T_o}{n},$$

and is known as the overall proportion of agreement. This statistic suffers in that it doesn't take into account agreements arising from guessing, nor does it reflect the magnitude of disagreements which could be close.

Kappa

Correcting $\hat{P}_a$ for agreement attributable to chance yields Cohen's (1960) kappa coefficient, defined by

$$\hat{\kappa} = \frac{T_o - T_c}{n - T_c}.$$

Paradoxical Results Produced by Kappa

Although $\hat{\kappa}$ is the most popular summary measure of agreement between two raters on a nominal scale, Feinstein and Cicchetti (1990) identified two paradoxes associated with its interpretation. These paradoxes arise because of the decision to impose a correction for chance agreement, making the assumption that the expected values of agreement should depend on the marginal totals. However, these marginal totals depend on the prevalence of the target trait and on the validity of the raters under study. The dependence of $\hat{\kappa}$ on prevalence can be explored, whereas, the investigation of the agreement of the two raters within the purview of validity will not be pursued further as it is beyond the scope of this thesis.

The first paradox of $\hat{\kappa}$ is that a high value of $\hat{P}_a$ can be drastically lowered by a substantial imbalance in the marginal totals either vertically or horizontally. The second paradox is $\hat{\kappa}$ will be higher if the imbalance in the corresponding marginal totals is asymmetrical rather than symmetrical.

Weighted Kappa

When the concept of full credit for complete agreement and varying amounts of partial credit for different off-diagonal $(i \neq j)$ cells seems natural in a given context, agreement is scaled so as to yield a ratio scale of positive agreement weights, w_{ij}, ranging down from some convenient maximum value assigned to the diagonal $(i = j)$ cells representing complete agreement. Several authors have suggested alternative ways of determining weights. Fleiss and Cohen (1973) suggested the squared error weights

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}, \qquad i, j = 1, \ldots, k,$$

and Cicchetti and Allison (1971) suggested the absolute error weights

$$w_{ij} = 1 - \frac{|i-j|}{k-1}, \qquad i, j = 1, \ldots, k.$$

Weighted kappa (Cohen 1968) has been used as an agreement index for ordinal data and is defined by

$$\hat{\kappa}_w = \frac{T_{ow} - T_{cw}}{n - T_{cw}}.$$

Using the approach in Section 4.2, Everitt (1968) derived the exact null variances of $\hat{\kappa}$ and $\hat{\kappa}_w$ to be

$$Var_0(\hat{\kappa}) = \frac{1}{n^2(n-1)(n-T_c)^2}\left\{\sum_{i=1}^{k} n_{i.}(n-n_{i.})n_{.i}(n-n_{.i}) + 2\sum_{\substack{i=1}}^{k}\sum_{\substack{j=1 \\ i<j}}^{k} n_{i.}\, n_{j.}\, n_{.i}\, n_{.j}\right\} \quad (2)$$

and

$$Var_0(\hat{\kappa}_w) = \frac{1}{(n-T_{cw})^2}\left\{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}^2 \frac{n_{i.}(n-n_{i.})n_{.j}(n-n_{.j})}{n^2(n-1)} + 2S\, w_{ij}\, w_{st}\, Cov_0(n_{ij}, n_{st})\right\}, \quad (3)$$

respectively, where S denotes the appropriate summation over the whole table. Fleiss, Cohen and Everitt (1969) found the large-sample variances of $\hat{\kappa}$ and $\hat{\kappa}_w$ to be estimable by

$$\hat{Var}(\hat{\kappa}) = \frac{1}{n(1-p_c)^4}\Big\{\sum_{i=1}^{k} p_{ii}[(1-p_c)-(p_{\cdot i}+p_{i\cdot})(1-p_o)]^2 + (1-p_o)^2\sum_{i=1}^{k}\sum_{\substack{j=1\\ i\neq j}}^{k} p_{ij}(p_{\cdot i}+p_{j\cdot})^2 - (p_o p_c - 2p_c + p_o)^2\Big\}$$

and

$$\hat{Var}(\hat{\kappa}_w) = \frac{1}{n(1-p_c)^4}\Big\{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}[w_{ij}(1-p_c)-(\overline{w}_{i\cdot}+\overline{w}_{\cdot j})(1-p_o)]^2 - (p_o p_c - 2p_c + p_o)^2\Big\},$$

respectively, and under $H_0 : \kappa = 0$, the estimated variances of $\hat{\kappa}$ and $\hat{\kappa}_w$ are, respectively,

$$\hat{Var}_0(\hat{\kappa}) = \frac{1}{n(1-p_c)^2}\Big\{\sum_{i=1}^{k} p_{i\cdot}\, p_{\cdot i}[1-(p_{\cdot i}+p_{i\cdot})]^2 + \sum_{i=1}^{k}\sum_{\substack{j=1\\ i\neq j}}^{k} p_{i\cdot}\, p_{\cdot j}(p_{\cdot i}+p_{j\cdot})^2 - p_c^2\Big\} \tag{4}$$

and

$$\hat{Var}_0(\hat{\kappa}_w) = \frac{1}{n(1-p_c)^2}\Big\{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{i\cdot}\, p_{\cdot j}[w_{ij}-(\overline{w}_{i\cdot}+\overline{w}_{\cdot j})]^2 - p_c^2\Big\}, \tag{5}$$

where

$$p_{ij} = \frac{n_{ij}}{n},\ p_{i\cdot} = \frac{n_{i\cdot}}{n},\ p_{\cdot j} = \frac{n_{\cdot j}}{n},\ p_o = \frac{T_o}{n},\ p_c = \frac{T_c}{n},\ \overline{w}_{i\cdot} = \sum_{j=1}^{k} w_{ij}\, p_{\cdot j} \text{ and } \overline{w}_{\cdot j} = \sum_{i=1}^{k} w_{ij}\, p_{i\cdot}.$$

An approximate significance test of κ, i.e., an approximate test of $H_0 : \kappa = 0$ versus $H_1 : \kappa > 0$, is accomplished by referring the critical ratio

$$Z(\hat{\kappa}) = \frac{\hat{\kappa}}{\sqrt{V_0(\hat{\kappa})}}$$

to the standard normal distribution, where $V_0(\hat{\kappa})$ can be replaced by either (2) or (4). A significance test of κ_w is defined analogously. $\hat{\kappa}$ and $\hat{\kappa}_w$, along with their respective variances are undefined when there is perfect agreement between the two raters, i.e., when all observations fall in a main diagonal cell.

The value of $\hat{\kappa}_w$ and its variance can be greatly influenced by the choice of weighting system. An obvious consequence of this is that the weights, however determined, must be set prior to the collection of the data. In the event that investigators use different weighting systems, comparison of $\hat{\kappa}_w$'s from different studies would prove difficult.

Under the squared error weighting system, Graham and Jackson (1993) suggested that $\hat{\kappa}_w$ should be regarded as a measure of association rather than an index of agreement. Amongst tables with the same marginal distributions, $\hat{\kappa}_w$ is dependent only on the overall correlation between row and column classifications and is not directly dependent on the propensity for exact agreement (data concentrated on the diagonal). Hence $\hat{\kappa}_w$ can appear insensitive to differences in $\hat{P}_a$ and large values of $\hat{\kappa}_w$ can be observed even when $\hat{P}_a$ is low.

The Disagreement Rate

Disagreements occur when two observers report different findings after examining the same subject. The frequency with which observers are in disagreement is a measure of the lack of reproducibility of the particular observation under test. First proposed in an article by Teasdale, Knill-Jones and Van der Sande (1978), the disagreement rate takes into account incorrect responses (disagreements). If on a subject, the first rater observes category i and the second rater records category j, then the disagreement score for that observation is $|i-j|$. If we use frequency counts n_{ij}, then

$$num = \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} |i-j|. \tag{6}$$

An adjustment to this absolute difference is made because differing scores lead to differing ranges. To standardize the change in scale of the absolute differences the term

$$den = 2 \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} [\max\{d_{ij} - 1, k - d_{ij}\}], \qquad \text{where } d_{ij} = (i+j)/2. \tag{7}$$

is calculated. The disagreement rate is then the ratio of (6) and (7), or

$$\hat{D} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} |i-j|}{2 \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} [\max\{d_{ij} - 1, k - d_{ij}\}]}.$$

$\hat{D}$ has a range of 0 to 0.5, with a lower value usually associated with a relatively larger kappa value. A relatively low kappa value despite a relatively low $\hat{D}$ is a reflection of the different properties between these two measures. The low $\hat{D}$ suggests that the disagreement was relatively small, whereas, a low kappa value does not take this into account but implies that chance played a major role.

The Concordance Between Raters

When determination regarding the presence or absence of exactly one nominal attribute is sufficient to describe each subject, the supposed measure of interrater agreement is the kappa statistic. In Cohen's (1960) kappa, the adjustment for chance-expected agreement depends on the marginal totals, however, this reliance on the marginal totals is not necessary.

Kupper and Hafner (1989) have developed a method for assessing the extent of interrater agreement when each unit is to be characterized by a (possibly empty) subset of $k \geq 2$ distinct nominal attributes. Except when $k = 2$ and the two attributes refer to the presence or absence of a single nominal trait, the k distinct nominal attributes should be defined so that the selection of any one attribute does not preclude the possible selection of any other.

Let $\mathcal{A}_i$ denote the subset of attributes for the ith subject chosen by rater A, and let Card$(\mathcal{A}_i) = A_i$, $0 \leq A_i \leq k$, be the random variable denoting the number of elements in the set $\mathcal{A}_i$. The set $\mathcal{B}_i$ and its cardinality B_i, $0 \leq B_i \leq k$, are defined analogously for rater B. Based on these definitions, it is informative to depict the data for the ith subject in the following table:

	Rater B		
Rater A	$\mathcal{B}_i$	$\overline{\mathcal{B}_i}$	Total
$\mathcal{A}_i$	Card$(\mathcal{A}_i \cap \mathcal{B}_i)$ $= X_i$	Card$(\mathcal{A}_i \cap \overline{\mathcal{B}_i})$ $= A_i - X_i$	A_i
$\overline{\mathcal{A}_i}$	Card$(\overline{\mathcal{A}_i} \cap \mathcal{B}_i)$ $= B_i - X_i$	Card$(\overline{\mathcal{A}_i} \cap \overline{\mathcal{B}_i})$ $= k - A_i - B_i + X_i$	$k - A_i$
Total	B_i	$k - B_i$	k

From the above table, the random variable X_i can be seen to be the number of attributes for the ith subject chosen by both raters, where $\max(0, A_i + B_i - k) \le X_i \le \min(A_i, B_i)$. Kupper and Hafner (1989) consider the agreement proportion

$$\hat{\pi}_i = \frac{X_i}{\max(A_i, B_i)},$$

and define the overall concordance between raters A and B to be the average of the $\hat{\pi}_i$'s,

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_i .$$

If both raters choose their subsets of attributes for the ith subject completely at random, then, conditional on $A_i = a_i$ and $B_i = b_i$, the null distribution of X_i is central hypergeometric (Johnson and Kotz 1969)

$$P_0(X_i = x_i) = \frac{\binom{a_i}{x_i}\binom{k-a_i}{b_i - x_i}}{\binom{k}{b_i}}, \qquad i = 1, \ldots, n,$$

where $\max(0, a_i + b_i - k) \le x_i \le \min(a_i, b_i)$. This conditional model for chance agreement on the ith subject derives from an underlying unconditional model which assumes that $A_i \sim \text{binomial}(k, \theta_A)$ and $B_i \sim \text{binomial}(k, \theta_B)$, and that A_i and B_i are independent. Under the assumption of random attribute selection, conditioning on the marginal frequencies in the previous table eliminates the nuisance parameters θ_A and θ_B from consideration. This approach to correction for chance agreement is based on the underlying assumption of rater-specific *a priori* equal probabilities of attribute selection (i.e., for the ith subject, rater A has the same probability θ_A of choosing each attribute).

Then, since

$$E_0(\hat{\pi}_i) = \frac{\min(a_i, b_i)}{k} \qquad \text{and} \qquad E_0(\hat{\pi}) = \frac{1}{nk}\sum_{i=1}^{n} \min(a_i, b_i) = \pi_0,$$

the concordance between raters (Kupper and Hafner 1989) is given by

$$\hat{C}_{AB} = \frac{\hat{\pi} - \pi_0}{1 - \pi_0}. \tag{8}$$

When each rater selects only one attribute to describe the ith subject, $a_i = b_i = 1$ for all i, and Equation (8) simplifies to

$$\hat{C}_{AB} = \frac{k\hat{P}_a - 1}{k - 1}. \tag{9}$$

Considering the special case when $a_i = b_i = 1$ for all i, the estimated large-sample variance of $\hat{C}_{AB}$ given by Kupper and Hafner (1989) is

$$\widehat{Var}(\hat{C}_{AB}) = \left(\frac{k}{k-1}\right)^2 \left[\frac{\hat{P}_a(1 - \hat{P}_a)}{n}\right],$$

and a significance test of C_{AB}, i.e., a test of $H_0 : C_{AB} = 0$ versus $H_1 : C_{AB} > 0$, is accomplished by referring the critical ratio

$$Z(\hat{C}_{AB}) = \frac{\hat{C}_{AB}}{\sqrt{Var_0(\hat{C}_{AB})}}$$

to the standard normal distribution, where the null variance of $\hat{C}_{AB}$ is

$$Var_0(\hat{C}_{AB}) = \frac{1}{n(k-1)}.$$

When the observed marginal proportions for the attributes selected by raters A and B are exactly the same, then (9) is always at least as large in value as the kappa statistic. The two measures are equal when each rater chooses each attribute an equal number of times, i.e., for data in the format of a $k \times k$ frequency table, the marginal proportions are all $1/k$. In general, however, Equation (9) can be smaller in value than kappa. The concordance between raters does not account for the magnitude of disagreements.

4.4 Newly Refined Measures of Interobserver Reliability

Three new measures of interobserver reliability described as alternatives are now proposed. These new measures represent the unique contribution of this thesis to the furthering development and utilization of measures of interobserver reliability.

An alternative approach is to assume that in a sample of n subjects, a certain number of observations will be guessed. For a certain number of subjects, c, there is no guessing involved and the rater's examinations result in agreement for each subject. It is envisaged that the information about the number of disagreements arising can be used to determine a likely value for the number of observations guessed. Three different strategies of differentiating between those observations that are guessed and those that are not guessed (based completely on perfect knowledge) are developed, leading to new measures of interobserver reliability. These three new measures are: (1) partial-chance proportion; (2) partial-chance kappa; and (3) expected-chance proportion.

Partial-Chance Proportion

Define a concordance type statistic in its general form to be

$$Y = \frac{n - n_g}{n}, \tag{10}$$

where n_g is the number of guessed observations in a sample of n subjects. In a sample of n subjects, assume that there are G guessed and $n - G$ nonguessed observations with the chance of agreeing in the guess of p_a. With some reasonable value for p_a we can estimate G and use this to adjust the number of correct responses similarly to before. The information about the number of disagreements, X, can be used to give the most likely or maximum value for G. The distribution of X is

$$f_X(x; g, p) = \binom{g}{x}(1 - p_a)^x p_a^{g-x}, \qquad 0 \le x \le n,\ x \le g \le n,\ 0 \le p_a \le 1. \tag{11}$$

Now, if we observe x disagreements from a sample of n subjects, we are then primarily interested in determining the most likely or maximum number of guessed observations, $\tilde{g}$. Consequently,

$$f_X(x; \tilde{g} - 1, p_a) \le f_X(x; \tilde{g}, p_a) \ge f_X(x; \tilde{g} + 1, p_a), \qquad \text{for } 0 \le p \le 1, \tag{12}$$

which leads to the inequalities

$$\frac{x}{1 - p_a} - 1 \le \tilde{g} \le \frac{x}{1 - p_a}, \tag{13}$$

where the difference in the bounds on the inequality is one. Now the optimal value of G is the minimal value of $\tilde{g}$ from (13) and the boundary inequality $g \le n$ from (11). It may not be unique. There are two solutions for $\tilde{g}$ when $1/(1 - p_a)$ is an integer.

Since $x = n - T_o$, and if $p_a = 1/k$ (equal weighting), then $\tilde{g} = k(n - T_o)/(k - 1)$. Replacing n_g by $\tilde{g}$ in Equation (10) leads to the partial-chance proportion

$$\hat{P}_{pc} = \frac{kT_o - n}{n(k - 1)}. \tag{14}$$

Partial-Chance Kappa

With perfect agreement, the number of concordant instances would be n. Similar to the strategy used in $\hat{P}_{pc}$, since there are g guessed observations, of which a result in agreement, the number of observations that are nonguessed is simply $n-g$, whereas the maximum number of observations excluding those that are guessed correctly is $n-a$. The general form of partial-chance kappa is defined to be

$$\hat{\kappa}_{pc} = \frac{n-g}{n-a}. \qquad (15)$$

Again, since $x = n - T_o$, and if $p_a = 1/k$, then $\tilde{g} = k(n-T_o)/(k-1)$. Also, since $a = g - x$, Equation (15) reduces to

$$\hat{\kappa}_{pc} = \frac{kT_o - n}{n(k-2)+T_o}. \qquad (16)$$

Expected-Chance Proportion

Consider a random experiment consisting of repeated independent Bernoulli trials where p_d is the probability of disagreement between two raters at each individual trial (subject). Suppose that a number of subjects are independently examined by two raters and, assume for a moment, that the exact number of subjects rated is unknown. The information about the number of disagreements, X, resulting from these examinations can be used to determine the number of observations that were guessed.

Let the random variable V represent the number of trials (subjects) that were guessed, resulting in x disagreements. If x disagreements are observed and, for some constant b_x, the distribution of V is

$$f_V(v; x, p) = b_x \binom{v}{x} p_d^{\,x} (1-p_d)^{v-x}, \qquad v = x, x+1, \ldots, \; 0 \leq p_d \leq 1.$$

It can be shown that

$$E(V) = \frac{x+1}{p_d} - 1.$$

Therefore, an estimate of the number of guessed observations is

$$\hat{n}_v = \frac{x+1}{p_d} - 1,$$

and, if $p_d = 1 - 1/k$, then it follows that

$$\hat{n}_v = \frac{kx+1}{k-1}.$$

The assumption made during the experimental set-up that the number of subjects independently examined by the two raters is unknown shall now be relaxed. In order to compare the various measures of interobserver reliability via computer simulation, each measure is computed based on a fixed number of subjects, n, and hence $x = n - T_o$.

Replacing n_g by $\hat{n}_v$ in Equation (10) leads to the expected-chance proportion

$$\hat{P}_{ec} = \frac{kT_o - n - 1}{n(k-1)},$$

which can be rewritten as

$$\hat{P}_{ec} = \hat{C}_{AB} - \frac{1}{n(k-1)},$$

where $\hat{C}_{AB}$ refers to Equation (9). Using the approach of Everitt (1968) in Section 4.2, the exact null variance of $\hat{P}_{ec}$ is

$$Var_0(\hat{P}_{ec}) = \frac{k^2}{n^4(n-1)(k-1)^2}\{\sum_{i=1}^{k} n_{i\cdot}(n-n_{i\cdot})n_{\cdot i}(n-n_{\cdot i}) + 2\sum_{i=1}^{k}\sum_{\substack{j=1\\ i<j}}^{k} n_{i\cdot}\, n_{j\cdot}\, n_{\cdot i}\, n_{\cdot j}\},$$

and a significance test of P_{ec}, i.e., a test of $H_0 : P_{ec} = 0$ versus $H_1 : P_{ec} > 0$, is accomplished by referring the critical ratio

$$Z(\hat{P}_{ec}) = \frac{\hat{P}_{ec}}{\sqrt{Var_0(\hat{P}_{ec})}}$$

to the standard normal distribution.

The Concordance Between Raters: Revisited

One may arrive at the concordance between raters another way under seemingly different conditions. Consider a population in which each subject is examined by two raters according to some characteristic of interest, with some mean value of disagreements, μ_x. Assuming that the population of subjects, N_0, is comprised of C_0 perfect knowledge-based (nonguessed) agreements, and $N_0 - C_0$ guessed observations, and the chance of being in disagreement in the guess of p_d, then

$$(N_0 - C_0)p_d = \mu_x.$$

Furthermore, from a sample of n subjects, let the random variable X represent the number of disagreements between two raters. Assuming that there are c_0 (unknown) perfect knowledge-based agreements, and the chance of being in disagreement in the guess of $1 - 1/k$, then the number of observations that are guessed $(n - c_0)$ may be estimated. Since,

$$E(X) = (n - c_0)(1 - 1/k),$$

and if x disagreements are observed, where $x = n - T_o$, then it follows that,

$$(n - c_0) = \frac{n - T_o}{1 - 1/k} = \hat{n}_g,$$

i.e., $\hat{n}_g$ is an estimate of the number of guessed observations. Replacing n_g by $\hat{n}_g$ in Equation (10), it can be shown that

$$\frac{n - \hat{n}_g}{n} = \hat{C}_{AB},$$

where $\hat{C}_{AB}$ refers to Equation (9).

4.5 Simulation Method

All simulations are done on a 166 MHz PC using FORTRAN 77 with use of a random number generator adapted from a FORTRAN version of the Long Period random number generator (Press *et al.* 1992). This routine is based on the simple combination method of L'Ecuyer (1988) which efficiently combines two multiplicative linear congruential generators so as to obtain a generator whose period $(\approx 2.3 \times 10^{18})$ is the least common multiple of the individual periods. A shuffle is also implemented in this routine to remove low-order serial correlations, the shuffling algorithm is due to Bays and Durham as described in Knuth (1981).

The simulation technique employed to study the reliability of these measures consists of four parameters: (1) the number of subjects, n, sampled; (2) the number of perfect knowledge-based (nonguessed) agreements, c, in a sample of n subjects; (3) the number of categories of classification, k, on a given scale of measurement; and (4) an indicator as to the prevalence of the target trait. Prevalence of the observed entity $(d = 2)$ is demonstrated by c subjects observing the first category of the classification scale, otherwise, c will be evenly distributed amongst each of the categories along the main diagonal $(d = 1)$. The underlying marginal probabilities $(\phi_i, \phi_j; i, j = 1, \ldots, k)$ used to generate the guessed observations in a set of tables are uniform marginals $(\phi_i = \phi_j = 1/k$, for all i and $j)$.

Simulations were performed to compare the finite sample performance of the measures of interobserver reliability. The sample sizes n examined were 15 and 50, and the number of perfect knowledge-based agreements c was varied over the interval $[0, n]$. The number of categories of classification k was chosen to be 3 and 5. Under each combination of the parameters considered, 10^5 replicates of $k \times k$ tables were generated at random by a program written in FORTRAN 77. For each table, various sample measures were calculated: (a) for each measure of interobserver reliability, values of interobserver reliability; and (b) for provided measures of interobserver reliability, exact null variances, null and nonnull large-sample variance estimates, and critical ratios.

A certain number of the generated tables were discarded whenever any one of the sample measures was found undefined. Based on the nondegenerate samples, outcome measures for the simulations include the mean and empirical variance of each measure of interobserver reliability, which relates to bias and efficiency, respectively. For provided measures, the means of the large-sample variance estimates are compared to the empirical variances. Also, for provided measures, the sampling distributions of the critical ratios provided for $\hat{\kappa}$, $\hat{\kappa}_w$, $\hat{C}_{AB}$ and $\hat{P}_{ec}$ are studied by simulation under the hypothesis that the assignments by the two raters are independent. The empirical distributions of these critical ratios are compared with the theoretical normal distribution in terms of the mean, variance, skewness and kurtosis, for which the theoretical values are respectively, 0, 1, 0 and 0, and in terms of one-tailed areas.

The results that follow are based on the conditional distribution of the estimates, conditional on nondegenerate findings.

5

RESULTS

Kappa values range from -1.00 to +1.00; minus values reflect less than chance agreement, positive values suggest greater than chance agreement and a value of 0 indicates chance agreement (Cohen 1960). Landis and Koch (1977) provided the following labels to the corresponding ranges of kappa: less than 0 indicates poor agreement; 0 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; 0.81 to 1.00, almost perfect agreement. These divisions are clearly arbitrary, but now generally accepted, guidelines for interpreting kappa statistics in clinical studies.

5.1 Degenerate Samples

Table B-1 shows the number of degenerate samples at different combinations of n, c, k and d. When $d = 1$ (nonprevalent case), a degenerate result is unlikely to occur. It is suggested that degenerate samples are likely to occur when $d = 2$ (prevalent case) and, more likely to occur as c tends to n.

5.2 Bias

The mean values of the measures of interobserver reliability obtained for each combination of n, c, k and d are depicted in Figures 1-8. The goal of these figures is to give an idea of how each measure of interobserver reliability varies with c, compared to the true proportion P_t.

In the nonprevalent case (Figures 1-4), $\hat{P}_a$ is overly optimistic since agreement may occur by chance. Another measure which is positively biased is $\hat{\kappa}_{pc}$, although not as grossly inflated as $\hat{P}_a$. Partial-chance proportion appears to overestimate the true proportion slightly, while $\hat{P}_{sc}$ underestimates the true proportion slightly, but the bias appears to be of little practical importance. Looking at $\hat{\kappa}_w$, both weighting systems tend to give similar results, however, the absolute error weights tend to produce better results than the squared error weights.

In the nonprevalent case, it is clear that $\hat{\kappa}$ and $\hat{C}_{AB}$ give very similar results throughout the range of values of c, and that they produce the best results. It is also suggested that the results obtained with the use of $\hat{D}$ were generally in accord with those using the complimentary measures of interobserver reliability.

In order to explore how $\hat{\kappa}$ and $\hat{\kappa}_w$ are affected by prevalence, simulations were performed for each combination of n, c and k, when $d = 2$ (Figures 5-8). The remaining measures of interobserver reliability are not affected by prevalence and are included merely for comparison, producing results similar to those when $d = 1$. As expected, high values of $\hat{P}_a$ were associated with low values of $\hat{\kappa}$ and $\hat{\kappa}_w$ when the raters place a preponderance of observations in one category. The choice of weighting system greatly influences the value of $\hat{\kappa}_w$ as illustrated in Figures 5-8. For higher values of $\hat{P}_a$, $\hat{\kappa}_w$ under the squared error weighting system is paradoxically altered to a lesser extent than under the absolute error weighting system. For larger values of c, both $\hat{\kappa}$ and $\hat{\kappa}_w$ grossly underestimate the true proportion, with this problem not repaired by larger sample sizes n.

It is clear from Figures 5-8 that $\hat{C}_{AB}$, $\hat{P}_{pc}$ and $\hat{P}_{cc}$ outperform the remaining measures of interobserver reliability when $d = 2$.

Note the somewhat irregular increase in value of interobserver reliability for $\hat{\kappa}$ and $\hat{\kappa}_w$ when c arrives at the upper boundary of its revised parameter space $(d = 2)$. This increase is due to the safeguards required by the simulation program, where the degenerate samples generated are discarded (see Table B-1).

5.3 Efficiency - Empirical Variances

The empirical variances of the measures of interobserver reliability obtained for each combination of the parameters are depicted in Figures 9-16. The goal of these figures is to give an idea of how the efficiency varies with c. For the sake of clarity in Figures 9-16, the empirical variance of $\hat{P}_{cc}$ is equivalent to that of $\hat{C}_{AB}$ as can be seen from the expression of $\hat{P}_{cc}$ in Section 4.4.

For the most part, it appears that the empirical variances tend to decrease when the number of categories of classification k increased from a value of 3 to 5. The situation is evident regardless of the indication as to the prevalence of the observed entity as illustrated in Figures 9-16, with the exception occurring for $d = 2$, when c arrives at the boundary of is revised parameter space.

When $d = 1$ (Figures 9-12), each measure attains its greatest efficiency when c is large and its smallest efficiency when c is small, with minor exceptions to $\hat{P}_{pc}$ and $\hat{\kappa}_{pc}$ where their smallest efficiency is realized when c is moderately low in value. It is clear from Figures 9-12 that $\hat{P}_a$ has the smallest empirical variance generally throughout the range of values of c (amongst agreement measures). For the mid- to upper range of values of c, competing measures include $\hat{\kappa}_{pc}$ which performs quite well, followed by $\hat{\kappa}$, $\hat{P}_{cc}$ ($\hat{C}_{AB}$) and $\hat{P}_{pc}$. For very small c, $\hat{P}_{pc}$ has moderately high efficiency relative to $\hat{P}_a$.

Looking at $\hat{\kappa}_w$ when $d = 1$, the empirical variance is greatly influenced by the choice of weighting system as illustrated in Figures 9-12. $\hat{\kappa}_w$ using absolute error weights is much more efficient than under the squared error weights. Nonetheless, the empirical variance is relatively large and, consequently, $\hat{\kappa}_w$ is unstable under either weighting system.

Figures 13-16 show how the empirical variances of $\hat{\kappa}$ and $\hat{\kappa}_w$ are affected by prevalence ($d = 2$), while the remaining measures produce similar results to those when $d = 1$.

Note that $\hat{\kappa}$ and $\hat{\kappa}_w$ attain their highest efficiency when c is moderately low in value and their lowest efficiency when c approaches the boundary of its revised parameter space.

Again, the empirical variance of $\hat{\kappa}_w$ is greatly influenced by the choice of weighting system as shown in Figures 13-16, with the choice of weighting system having a similar effect on efficiency as when $d = 1$. It is clear that the empirical variances of $\hat{\kappa}$ and $\hat{\kappa}_w$ are relatively large and, consequently, $\hat{\kappa}$ and $\hat{\kappa}_w$ are unstable.

Note the somewhat irregular decrease in value of the empirical variances of $\hat{\kappa}$ and $\hat{\kappa}_w$ when c arrives at the upper boundary of its revised parameter space ($d = 2$). The explanation for this decrease is drawn from the same argument provided earlier in Section 5.2 for the irregular increase in value of interobserver reliability for $\hat{\kappa}$ and $\hat{\kappa}_w$ (see Table B-1).

5.4 Efficiency - Large-Sample Variance Estimates

Figures 17-24 give the mean values of the large-sample variance estimates and the empirical variances of $\hat{\kappa}$, $\hat{\kappa}_w$ and $\hat{C}_{AB}$ obtained for each combination of the parameters. The goal of these figures is to give an idea of how the large-sample variance estimates compare to the empirical variances for the provided measures of interobserver reliability over the entire range of values of c.

What is clear in the nonprevalent case (Figures 17-20) is that for c close to the boundaries of its parameter space, the large-sample variance estimates of $\hat{\kappa}$, $\hat{\kappa}_w$ and $\hat{C}_{AB}$ do not differ greatly from the empirical variances. However, as c tends to move away from its boundaries, the estimated large-sample variances depart somewhat markedly from the empirical variances. This departure is consistent for $\hat{\kappa}$, $\hat{\kappa}_w$ and $\hat{C}_{AB}$, with the large-sample estimates overestimating the variances.

When $d = 2$ (prevalent case), Figures 21-24 seem to suggest that when c is small to moderate in value, the estimated large-sample variances of $\hat{\kappa}$ and $\hat{\kappa}_w$ do not differ greatly from the empirical variances. However, as c tends to approach the upper boundary of its revised parameter space, the large-sample variance estimates of $\hat{\kappa}$ and $\hat{\kappa}_w$ underestimate the variances. Since $\hat{C}_{AB}$ is not affected by prevalence, the results are similar to those when $d = 1$.

5.5 Empirical Distributions of Critical Ratios

Table B-2 gives the empirical central moments of the null distributions of critical ratios provided for $\hat{\kappa}$, $\hat{\kappa}_w$, $\hat{C}_{AB}$ and $\hat{P}_{cc}$, obtained for each combination of n, k and d, when $c = 0$. For the most part, the observed central moments are reasonably close to their expected values as illustrated in Table B-2.

For $\hat{\kappa}$ and $\hat{\kappa}_w$, the exact variance approach of Everitt (1968) produces slightly better results than the large-sample variance approach of Fleiss, Cohen and Everitt (1969) in terms of the means and variances. regardless of the indication as to the prevalence. The Z of $\hat{C}_{AB}$ also produces acceptable results with the means and variances close to their theoretical values of 0 and 1. respectively. However. the Z of $\hat{P}_{ec}$ produces mean values differing slightly from 0. approximately in the range of -0.1 to -0.2.

Looking at $\hat{\kappa}_w$ when $d = 1$ (nonprevalent case). the squared error weights of Fleiss and Cohen (1973) tend to approximate the expected values of the mean. variance and skewness slightly better than the absolute error weights of Cicchetti and Allison (1971). However. the Z of $\hat{\kappa}_w$ based upon the absolute error weights produces kurtosis values closer to the theoretical value of 0 than when the squared error weights are applied. When $d = 2$ (prevalent case), $\hat{\kappa}_w$ under the absolute error weights tend to produce slightly better results than under the squared error weights in terms of the mean. variance and kurtosis. However. the critical ratio $Z(\hat{\kappa}_w)$ under squared error weights tend to approximate the expected value of skewness slightly better than under the absolute error weighting system.

Table B-3 gives the empirical tail areas of the null distributions of critical ratios provided for $\hat{\kappa}$. $\hat{\kappa}_w$, $\hat{C}_{AB}$ and $\hat{P}_{ec}$, obtained for each combination of n, k and d, when $c = 0$. The results tended to closely parallel those based upon the central moments in Table B-2. Therefore. these results indicate that the null variances of $\hat{\kappa}$. $\hat{\kappa}_w$, $\hat{C}_{AB}$ and $\hat{P}_{ec}$ are valid for assessing levels of statistical significance.

6

DISCUSSION

6.1 Conclusions and Recommendations

It is generally appreciated that there is no perfect measure for summarizing any mass of data. When a $k \times k$ table is to be represented by only one measure, information is lost. The unwarranted presumption of sufficiently high agreement may lead to the use of a feasible but unreliable study procedure or technique, with attendant risk of drawing erroneous conclusions from the study results.

In conclusion, five common measures of interobserver reliability have been assessed, resulting in three newly refined measures being proposed. The simulation experiment confirms that these new measures prove to be very useful as they eliminate the problems encountered with the common measures. Due to the broad range of possible data configurations and underlying probability distributions generating the data, it is difficult to draw definitive conclusions from the simulations, and only general suggestions should be made.

A researcher who assumes that some of the results could have arisen due to guessing and then wants to adjust for this needs to clearly specify the guessing mechanism. Possible guessing circumstances and the associated measures are as follows:

(a) If we assume that all observations are guessed and further assume that the sample proportions are those used in the guessing, we would recommend $\hat{\kappa}$ or $\hat{\kappa}_w$. However, it might not be wise to use $\hat{\kappa}$ or $\hat{\kappa}_w$ in the circumstance when an observer places a preponderance of observations in one category, as these measures are known to produce two types of paradoxes. Furthermore, $\hat{\kappa}$ and $\hat{\kappa}_w$ are unstable as their variances are relatively large.

(b) If it is assumed that all observations are guessed and further assumed that the guessing is done by giving each of the categories of classification an equal chance of being observed, we would advise $\hat{C}_{AB}$ in practice. This measure is particularly applicable to studies in which the rates do not have *a priori* knowledge of the prevalence of the scores in the population. $\hat{C}_{AB}$ accurately estimates the true proportion and is stable for equal and unequal marginals. Also, the variance of $\hat{C}_{AB}$ is acceptably approximated.

(c) Assume that only some observations are subject to classification by chance. Then, if we assume that the raters are able to state and use some proportion for guessing, this leads to $\hat{P}_{pc}$, $\hat{\kappa}_{pc}$ and $\hat{P}_{ec}$. Each cf these measures indicate some presence of bias, however, not substantial enough in that it may be inappropriate to quote an index of agreement. In comparison to $\hat{C}_{AB}$, these measures estimate the true proportion moderately well and are relatively stable. Since $\hat{C}_{AB}$ is based upon the assumption that all observations are guessed, which is not a practical assumption often in a clinical setting, $\hat{P}_{pc}$, $\hat{\kappa}_{pc}$ and $\hat{P}_{ec}$ may appear as more attractive.

(d) The alternative is to assume that there is no guessing, in which case we would recommend either $\hat{P}_a$ or $\hat{D}$. The overall proportion of agreement is positively biased to a large extent and should only be recommend as a preliminary measure of interobserver reliability. The disagreement rate provides a measure of disagreement and should be recommended as a complimentary measure.

6.2 Future Research

Often in practice, ethical and practical considerations limit the number of raters who can assess a patient within a short time of each other. As pointed out by Koran (1975a), studies of clinical reliability should focus on agreement between two physicians or perhaps three, as this more closely reflects clinical practice. In this thesis, focus was on the common case of two raters.

When given a diagnosis carrying out serious cost and risk consequences, a patient often seeks a second (or third or fourth) diagnostic opinion. Even the most careful and expert diagnostician using the best of diagnostic methods can make a mistake. There is potential to extend the work to the multi-rater case, allowing agreement among the multiple raters to be measured.

Sample sizes in clinical studies are often limited because of ethical considerations along with the practical difficulty in getting the same set of raters to examine patients within a short time of each other. Cicchetti and Fleiss (1977) and Cicchetti (1981) derived an empirically based formula for determining the approximate sample sizes required for the valid application of the kappa statistics, which is approximately $n \geq 2k^2$. This finding is of some comfort to investigators in contrast to the implied conservative estimate of $n \geq 200$, irrespective of the number of categories of classification k, due to Fleiss, Cohen and Everitt (1969).

However, studies have been conducted in which the sample sizes were small relative to Cicchetti's approximation. For instance, Teasdale, Knill-Jones and Van der Sande (1978) had 16 patients in their study employing a scale with 5 categories of classification, whereas, Yager, Johnston and Seshia (1990) had 15 patients in their study with 7 categories of classification in one scale. These sample size values have been similar in several studies of interobserver reliability (Koran 1975b). The sample sizes n examined in the simulations were 15 and 50, and the number of categories of classification k were chosen to be 3 and 5. These values were chosen arbitrarily, within a countless number of parameter combinations that could have been considered.

The underlying marginal probabilities used to generate the guessed observations in a set of tables were uniform marginals, where each rater has the same probability $1/k$ of choosing each category on the classification scale. Also, a deliberate attempt was made to demonstrate the situation where the raters place a preponderance of observations in one classification category, resulting in symmetrical unbalanced marginals.

Often, however, these marginals depend on the background of the two raters, aside from the prevalence of the target trait. There is potential to explore variations of the underlying probabilities used to generate a set of tables.

APPENDIX A

COMA SCALES

Table A-1: Adelaide Scale.[a]

Criterion	Score
Eyes open	
Spontaneously	4
To speech	3
To pain	2
None	1
Best verbal response	
Orientated	5
Words	4
Vocal sounds	3
Cries	2
None	1
Best motor response	
Obeys commands	5
Localise pain	4
Flexion to pain	3
Extension to pain	2
None	1

[a] A pediatric modification of the Glasgow Coma Scale used in the Adelaide Children's Hospital, South Australia, since 1977 takes neurological immaturity into account (Simpson and Reilly 1982).

Table A-2: Jacobi Scale.[a]

Criterion	Score
Best verbal response	
Orientated	5
Confused	4
Inappropriate	3
Incomprehensible	2
None	1
Best motor response	
Obeying	5
Localizing	4
Flexing	3
Extending	2
None	1
Eyes open	
Spontaneous	4
To speech	3
To pain	2
None	1
Ocular vestibular response	
Normal	5
Tonic-conjugate	4
Minimal-dysconjugate	3
No eye movements	2
Non-reacting pupil	1

[a] Gordon *et al.* (1983).

Table A-3: 0-IV Scale.[a]

Criterion	Score
Arouses spontaneously and to stimuli	0
Stuporous; spontaneous arousal rare; roused readily but briefly by stimuli	I
Spontaneous arousal absent; avoidance motor response to stimuli	II
Motor response to intense painful stimuli only	III
No response	IV

[a] Huttenlocher (1972); Seshia, S. S., Seshia, M. M. K., and Sachdeva (1977).

APPENDIX B

SIMULATION RESULTS

Table B-1: The number of degenerate samples out of 10^5 replicates in the simulation experiment.

	$d = 1$		$d = 2$	
c	$k = 3$	$k = 5$	$k = 3$	$k = 5$
(i) $n = 15$				
0	3	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	1	0
5	0	0	3	0
6	0	0	9	0
7	0	0	32	0
8	0	0	99	4
9	0	0	253	11
10	0	0	884	70
11	0	0	2499	308
12	0	0	7283	1632
13	0	0	21041	7842
14	0	0	55563	36084
15	0	0	100000	100000
(ii) $n = 50$				
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0

Table B-1: *(concluded)*.

	$d = 1$		$d = 2$	
c	$k = 3$	$k = 5$	$k = 3$	$k = 5$
(ii) $n = 50$				
10	0	0	0	0
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
16	0	0	0	0
17	0	0	0	0
18	0	0	0	0
19	0	0	0	0
20	0	0	0	0
21	0	0	0	0
22	0	0	0	0
23	0	0	0	0
24	0	0	0	0
25	0	0	0	0
26	0	0	0	0
27	0	0	0	0
28	0	0	0	0
29	0	0	0	0
30	0	0	0	0
31	0	0	0	0
32	0	0	0	0
33	0	0	0	0
34	0	0	0	0
35	0	0	0	0
36	0	0	0	0
37	0	0	0	0
38	0	0	1	0
39	0	0	1	0
40	0	0	5	0
41	0	0	7	1
42	0	0	31	0
43	0	0	95	4
44	0	0	250	16
45	0	0	825	68
46	0	0	2391	310
47	0	0	7151	1569
48	0	0	21183	7758
49	0	0	55367	36004
50	0	0	100000	100000

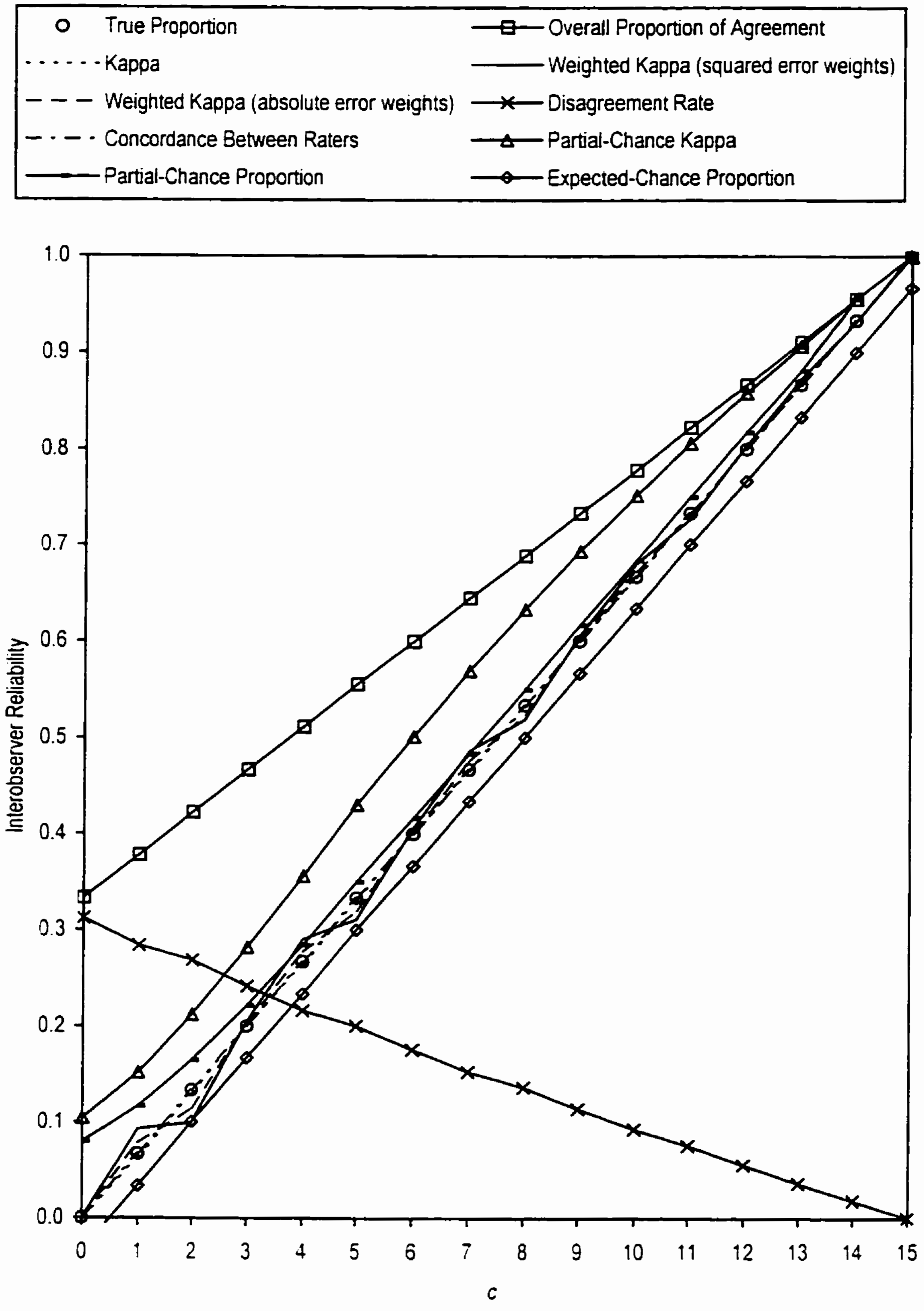


Figure 1: Mean values of the measures of interobserver reliability versus c when n =15, k =3, and d =1. Simulation results are based on the nondegenerate samples.

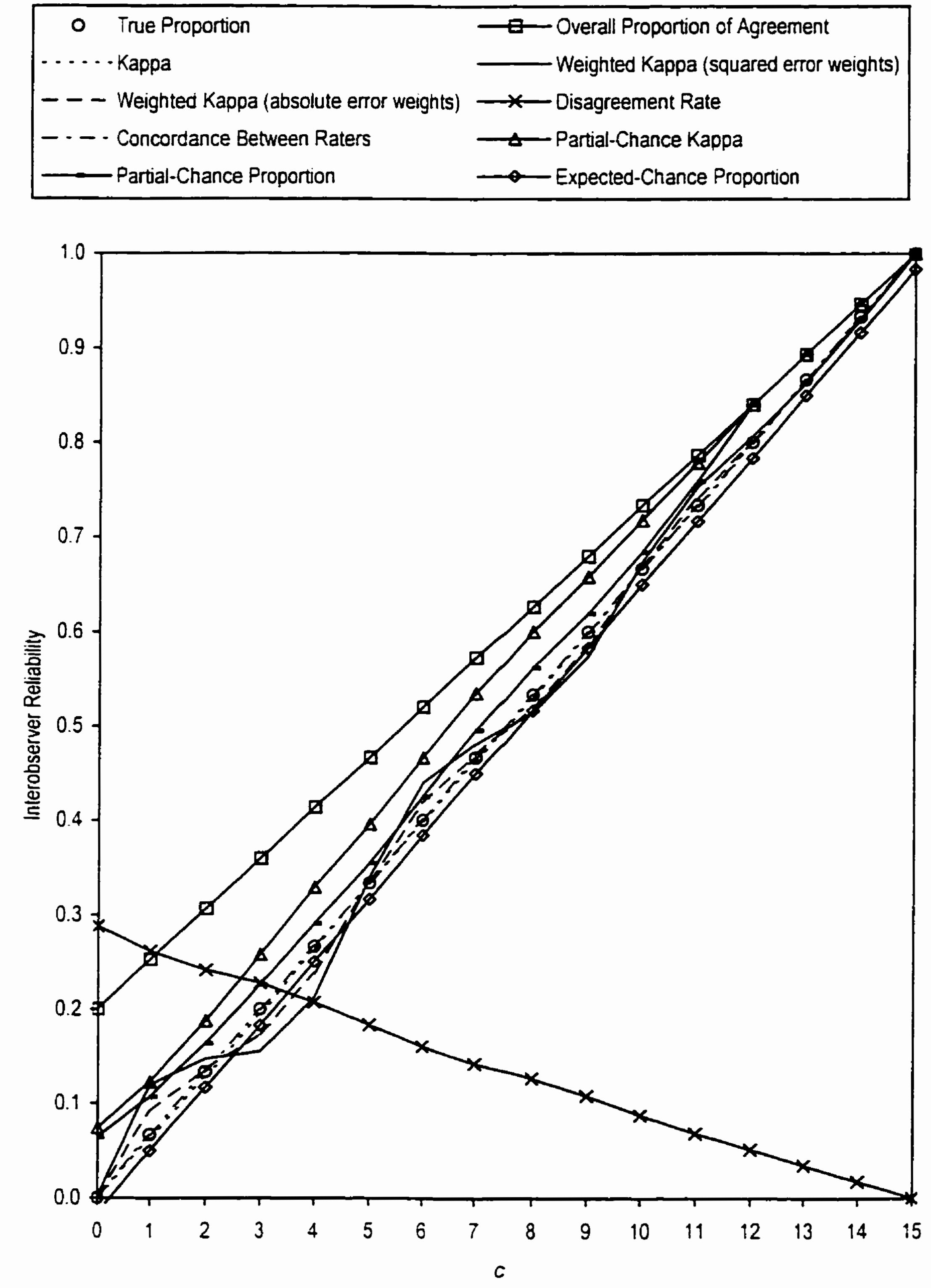


Figure 2: Mean values of the measures of interobserver reliability versus c when n =15, k =5, and d =1. Simulation results are based on the nondegenerate samples.

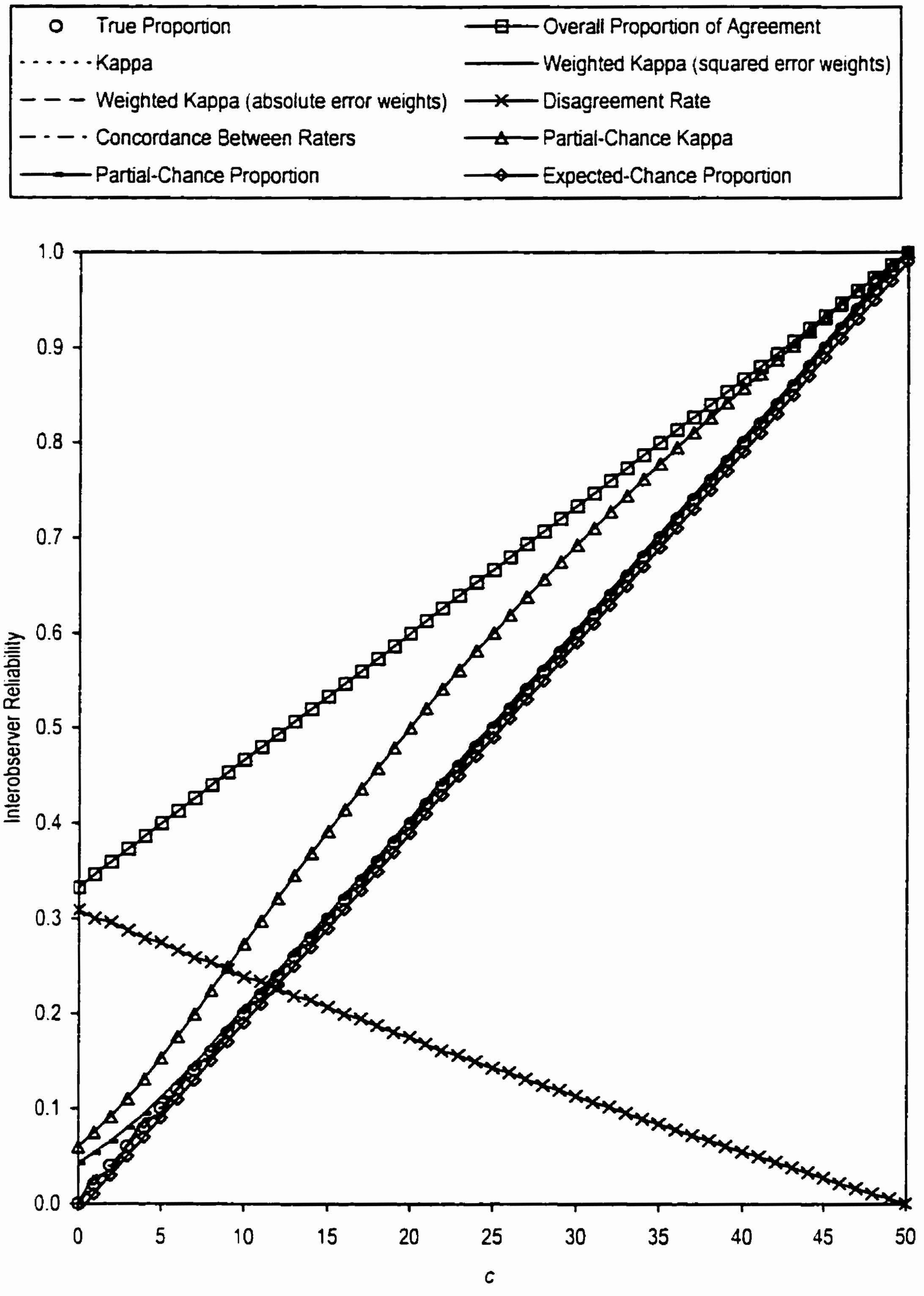


Figure 3: Mean values of the measures of interobserver reliability versus c when $n=50$, $k=3$, and $d=1$. Simulation results are based on the nondegenerate samples.

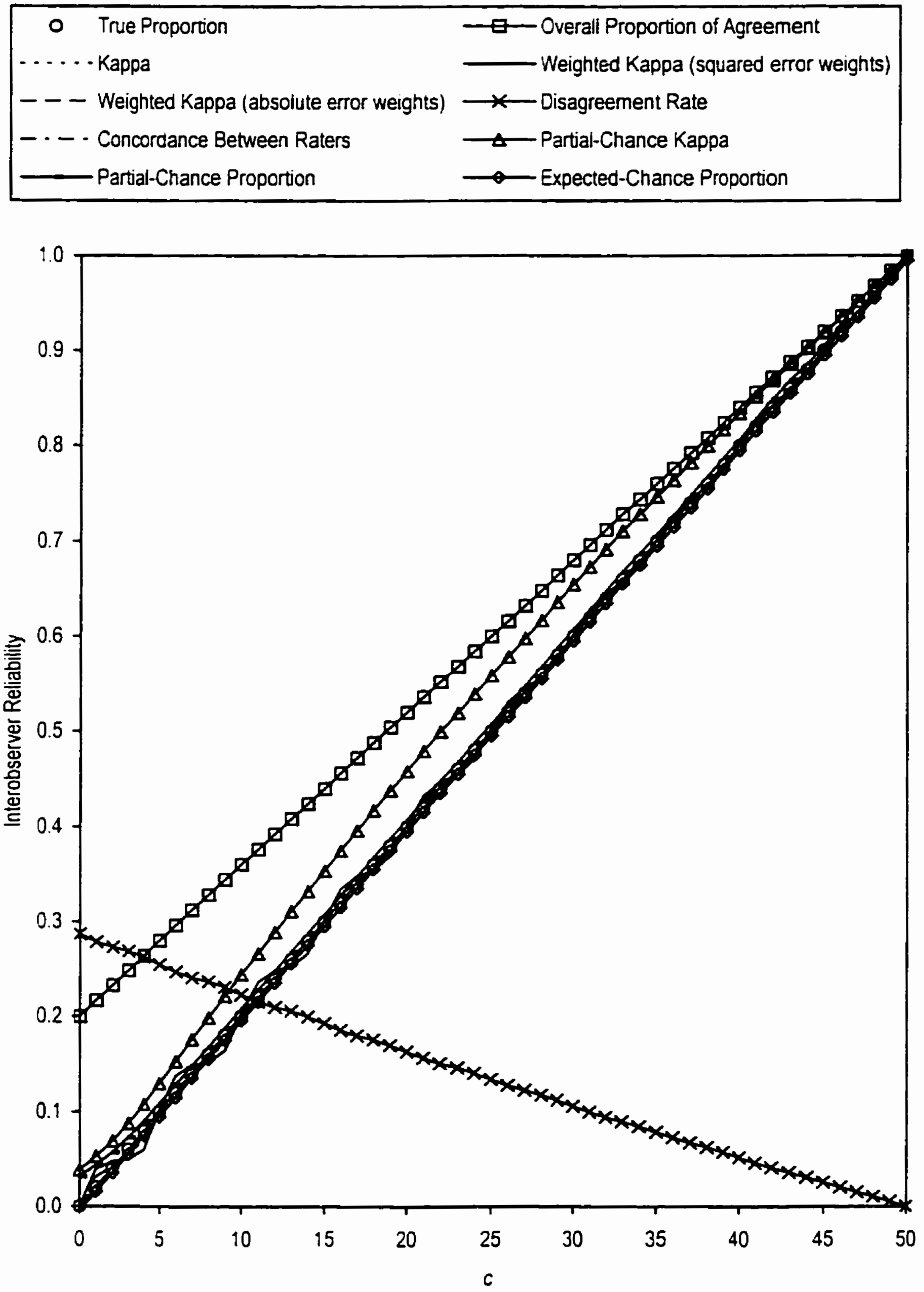


Figure 4: Mean values of the measures of interobserver reliability versus c when n=50, k=5, and d=1. Simulation results are based on the nondegenerate samples.

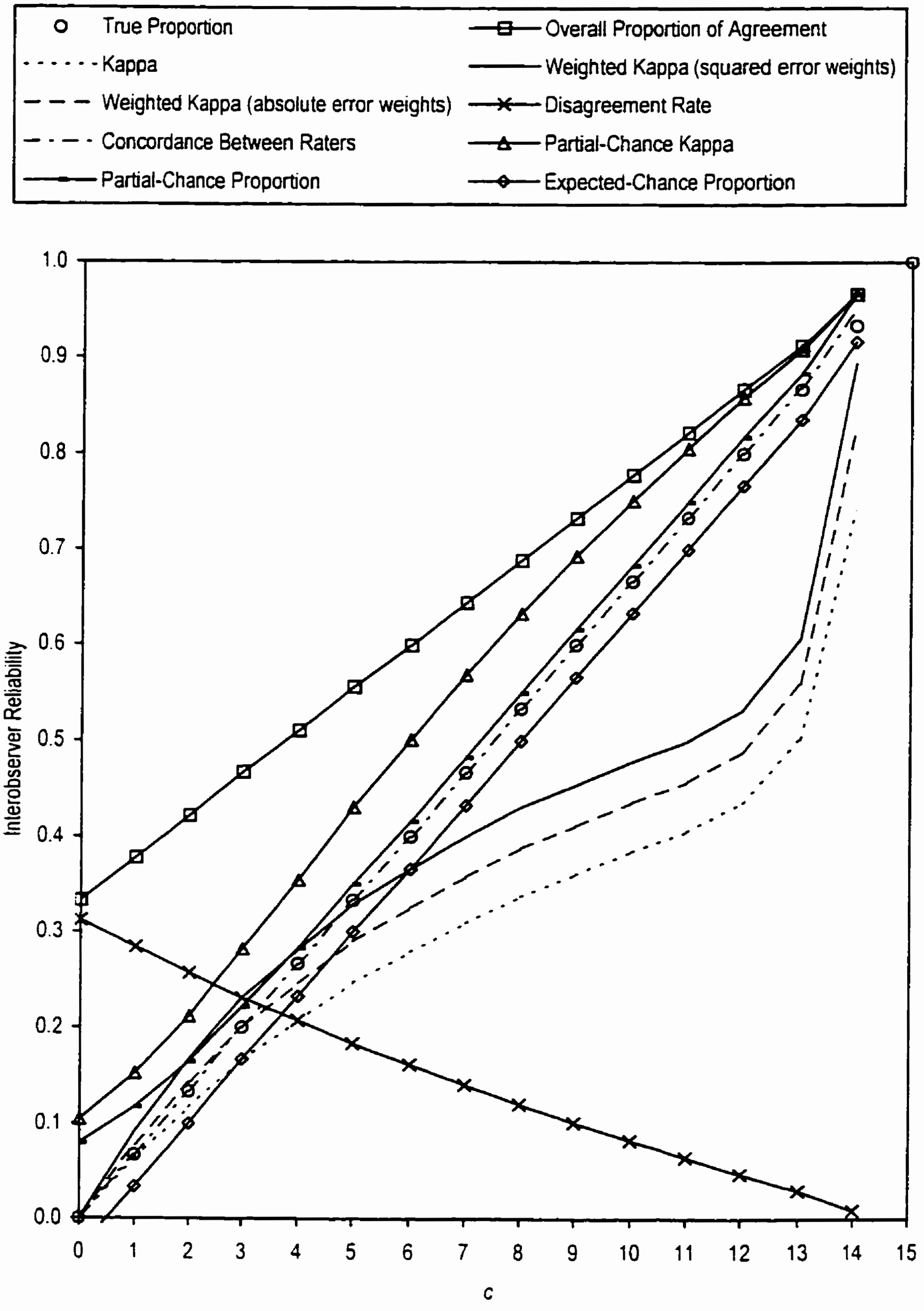


Figure 5: Mean values of the measures of interobserver reliability versus c when $n=15$, $k=3$, and $d=2$. Simulation results are based on the nondegenerate samples.

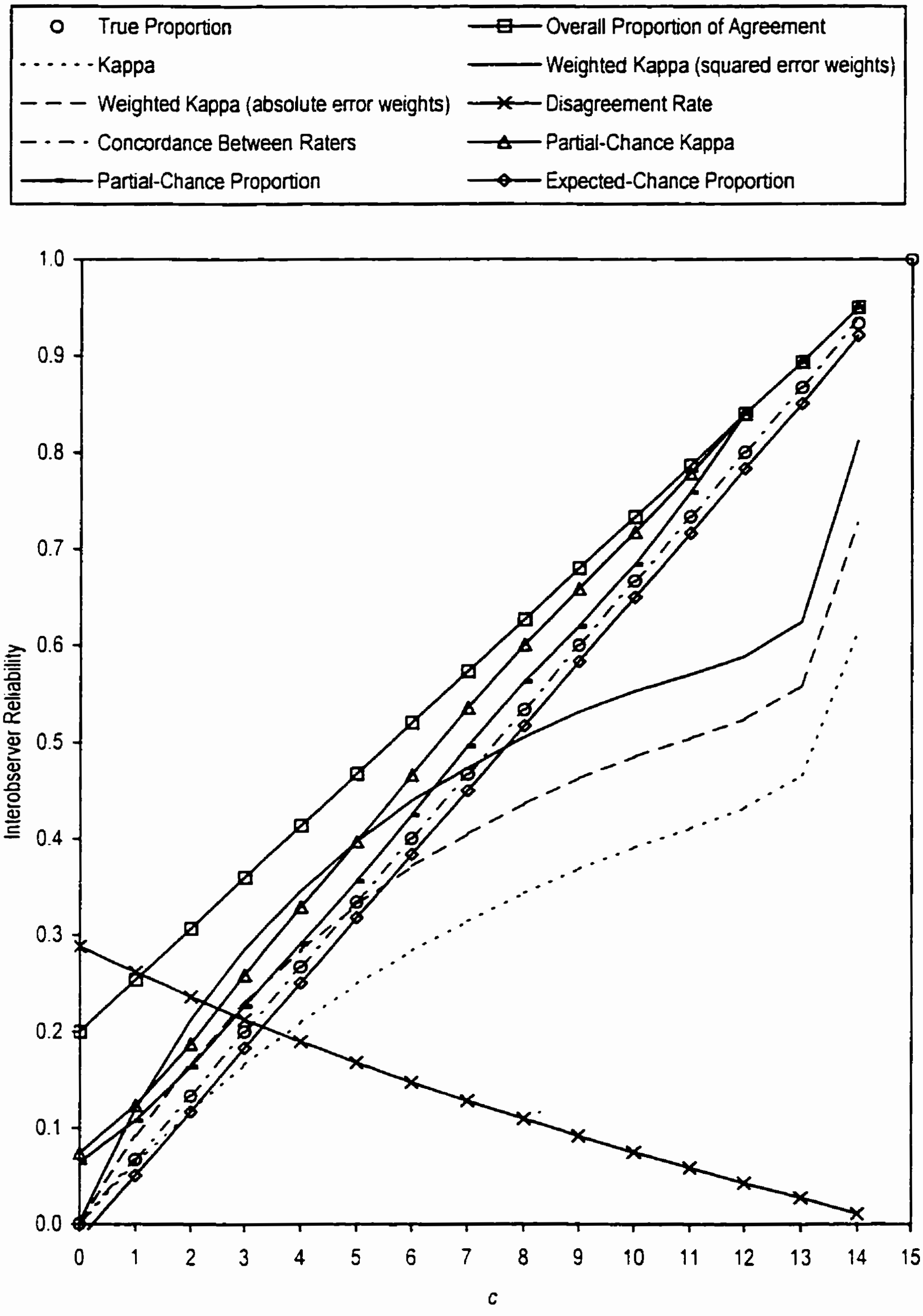


Figure 6: Mean values of the measures of interobserver reliability versus *c* when $n = 15$, $k = 5$, and $d = 2$. Simulation results are based on the nondegenerate samples.

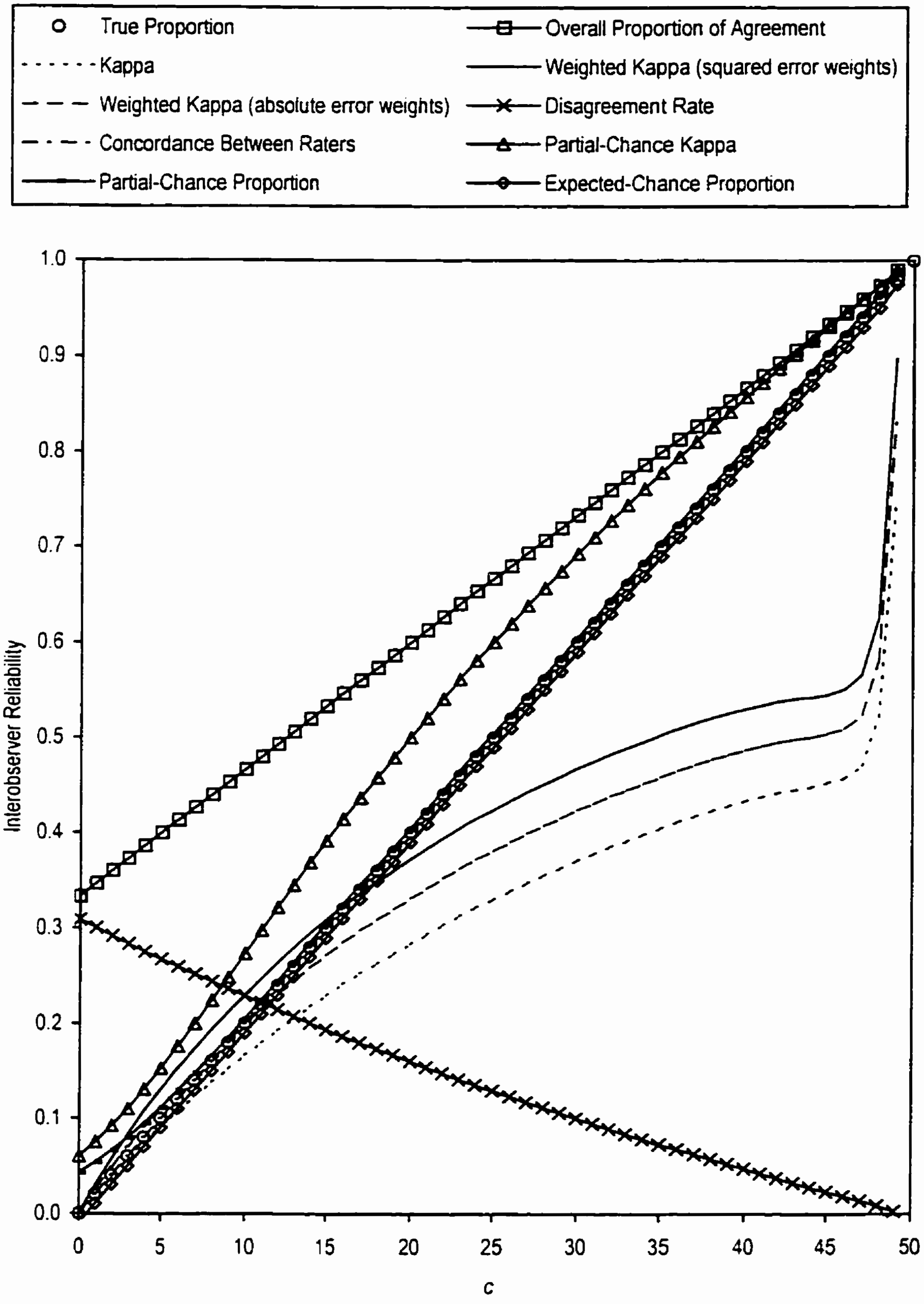


Figure 7: Mean values of the measures of interobserver reliability versus c when n =50, k =3, and d =2. Simulation results are based on the nondegenerate samples.

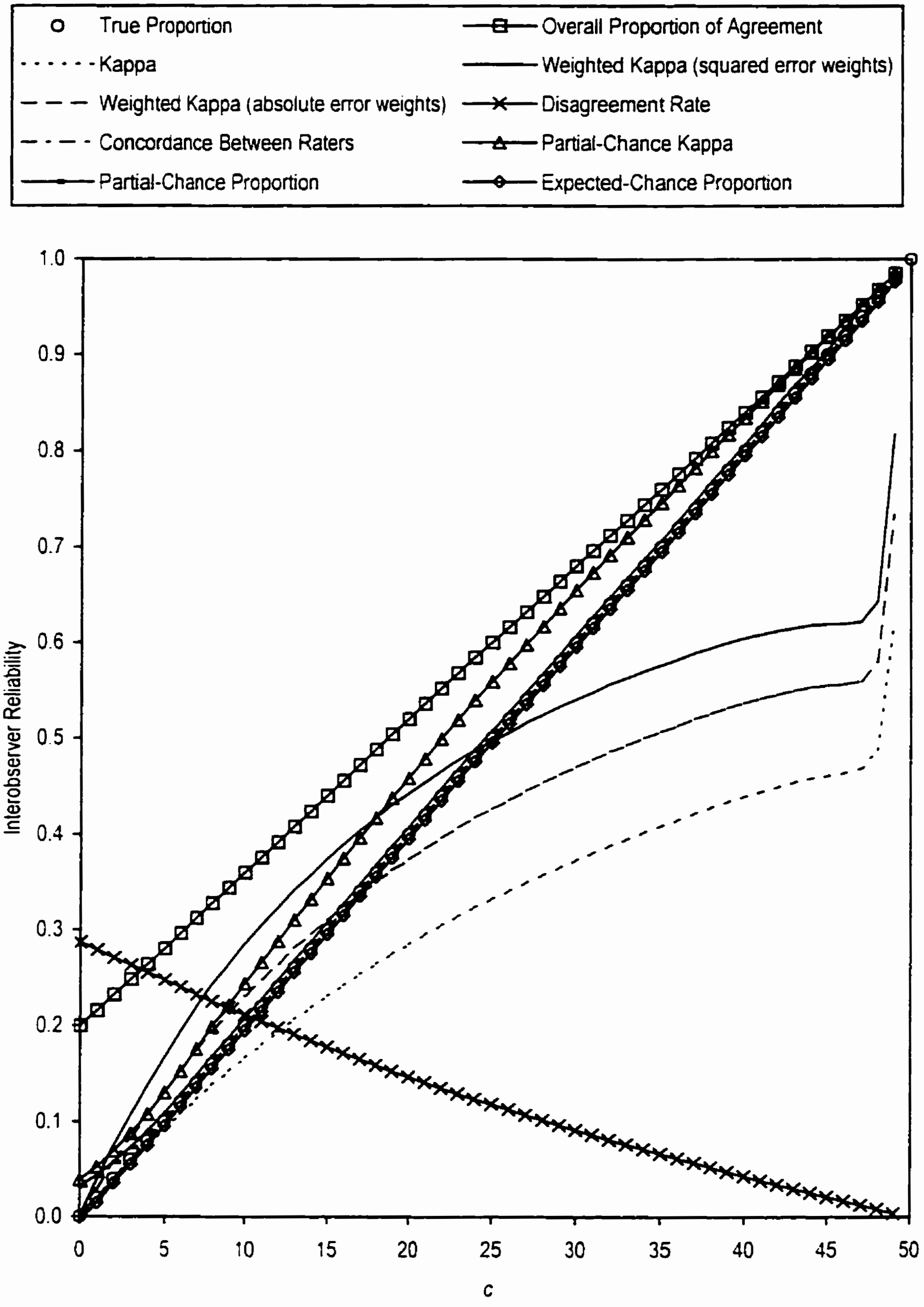


Figure 8: Mean values of the measures of interobserver reliability versus c when n =50, k =5, and d =2. Simulation results are based on the nondegenerate samples.

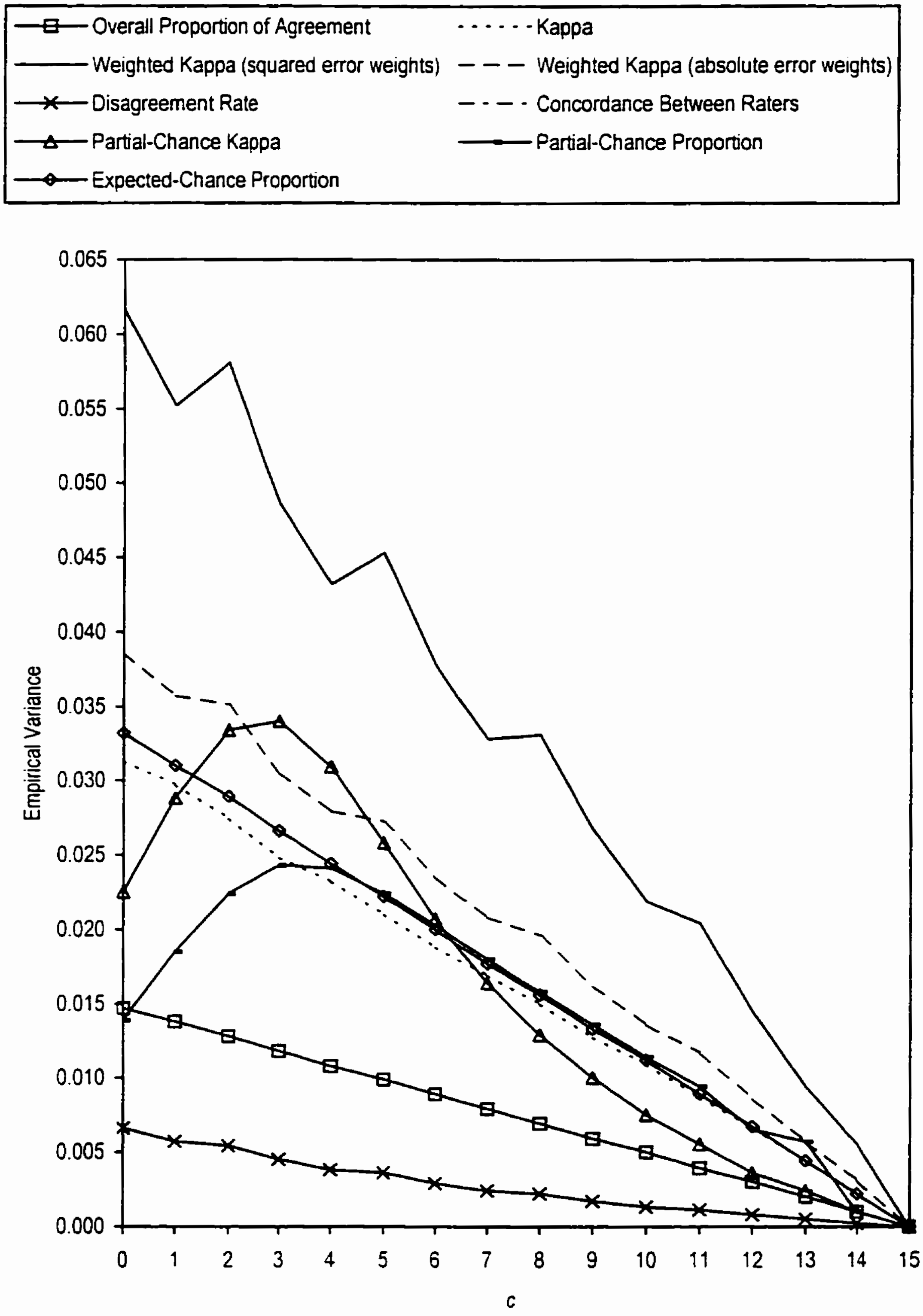


Figure 9: Empirical variances of the measures of interobserver reliability versus c when $n=15$, $k=3$, and $d=1$. Simulation results are based on the nondegenerate samples.

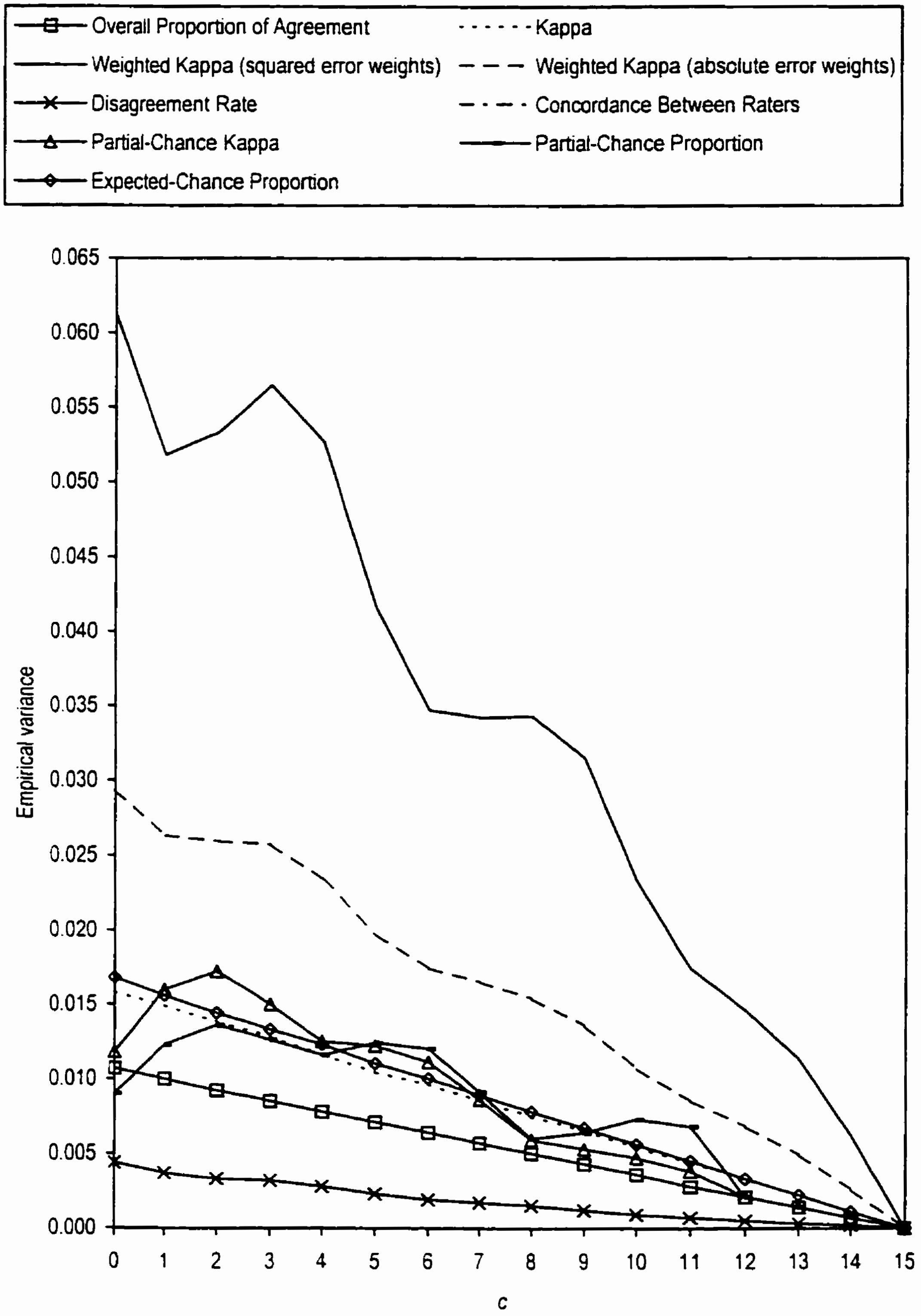


Figure 10: Empirical variances of the measures of interobserver reliability versus c when n=15, k=5, and d=1. Simulation results are based on the nondegenerate samples.

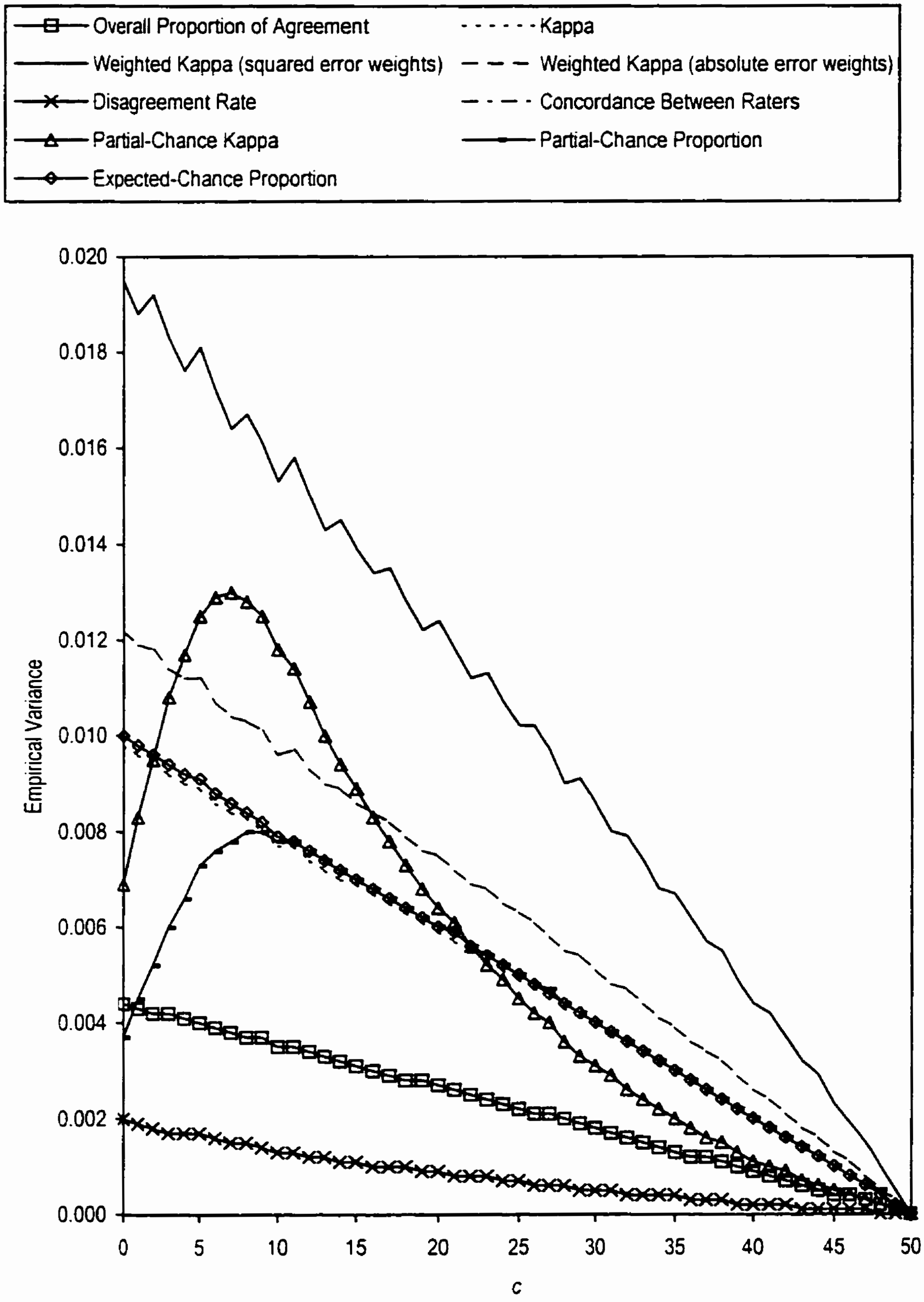


Figure 11: Empirical variances of the measures of interobserver reliability versus c when n =50, k =3, and d =1. Simulation results are based on the nondegenerate samples.

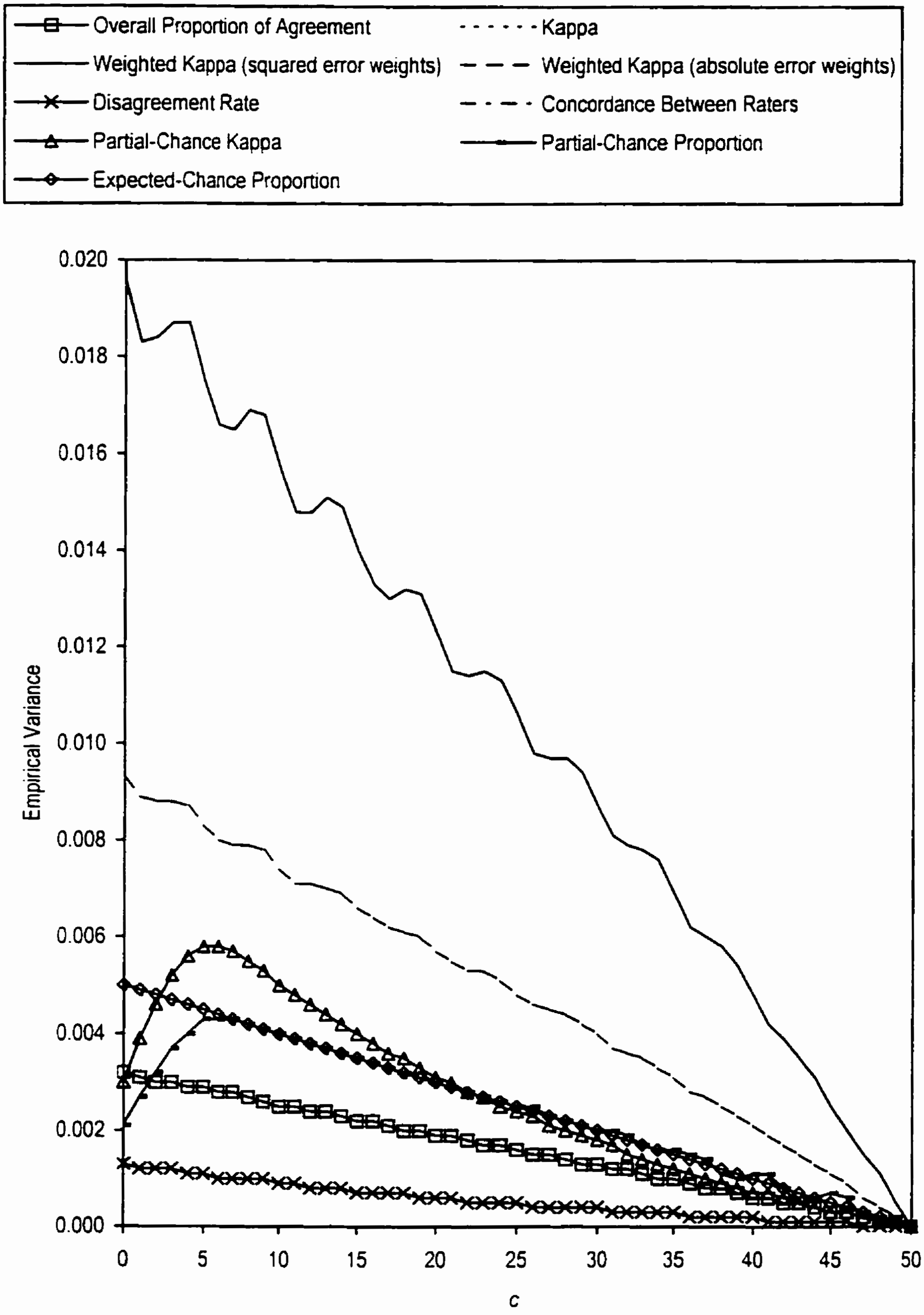


Figure 12: Empirical variances of the measures of interobserver reliability versus c when n =50, k =5, and d =1. Simulation results are based on the nondegenerate samples.

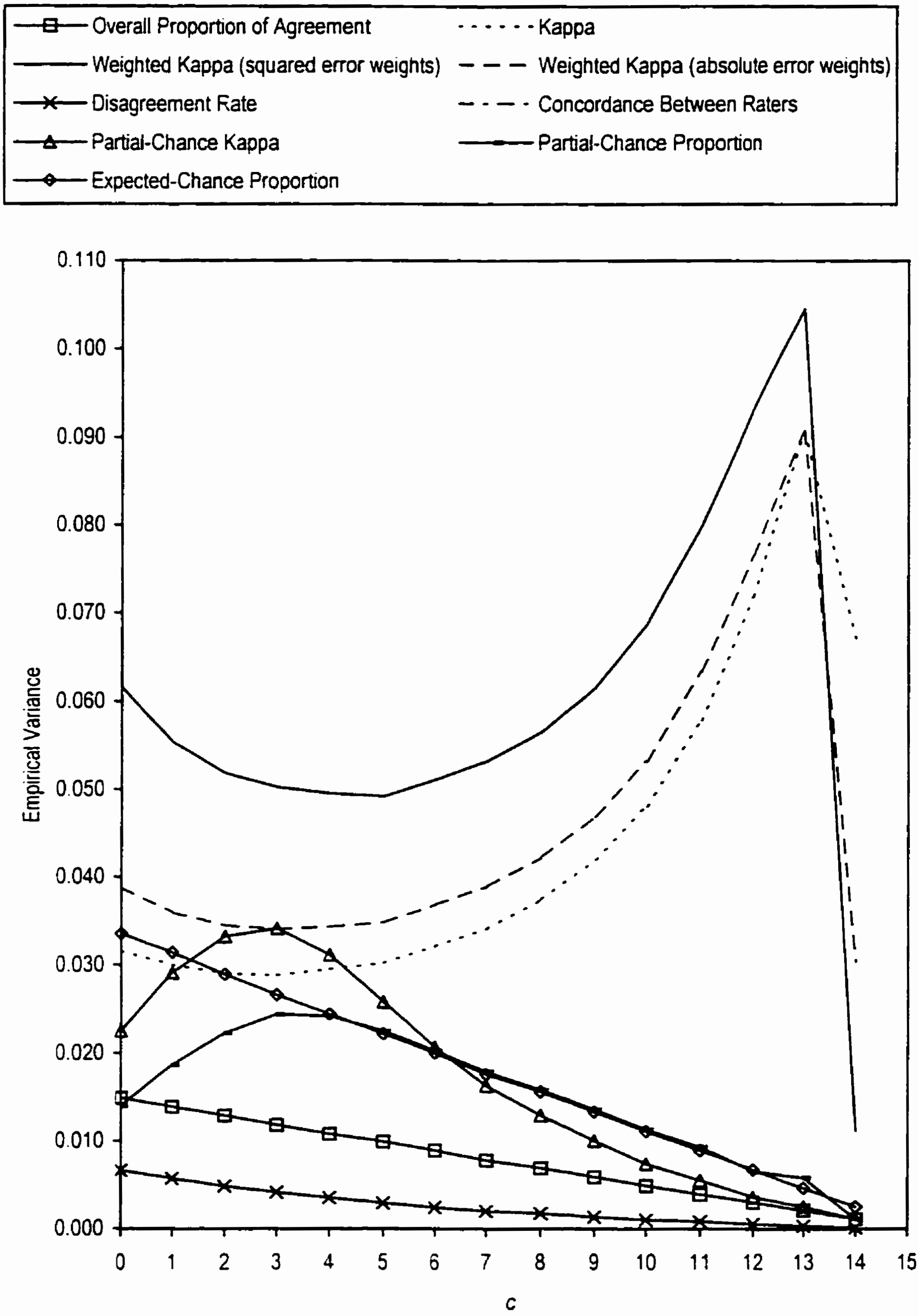


Figure 13: Empirical variances of the measures of interobserver reliability versus c when n=15, k=3, and d=2. Simulation results are based on the nondegenerate samples.

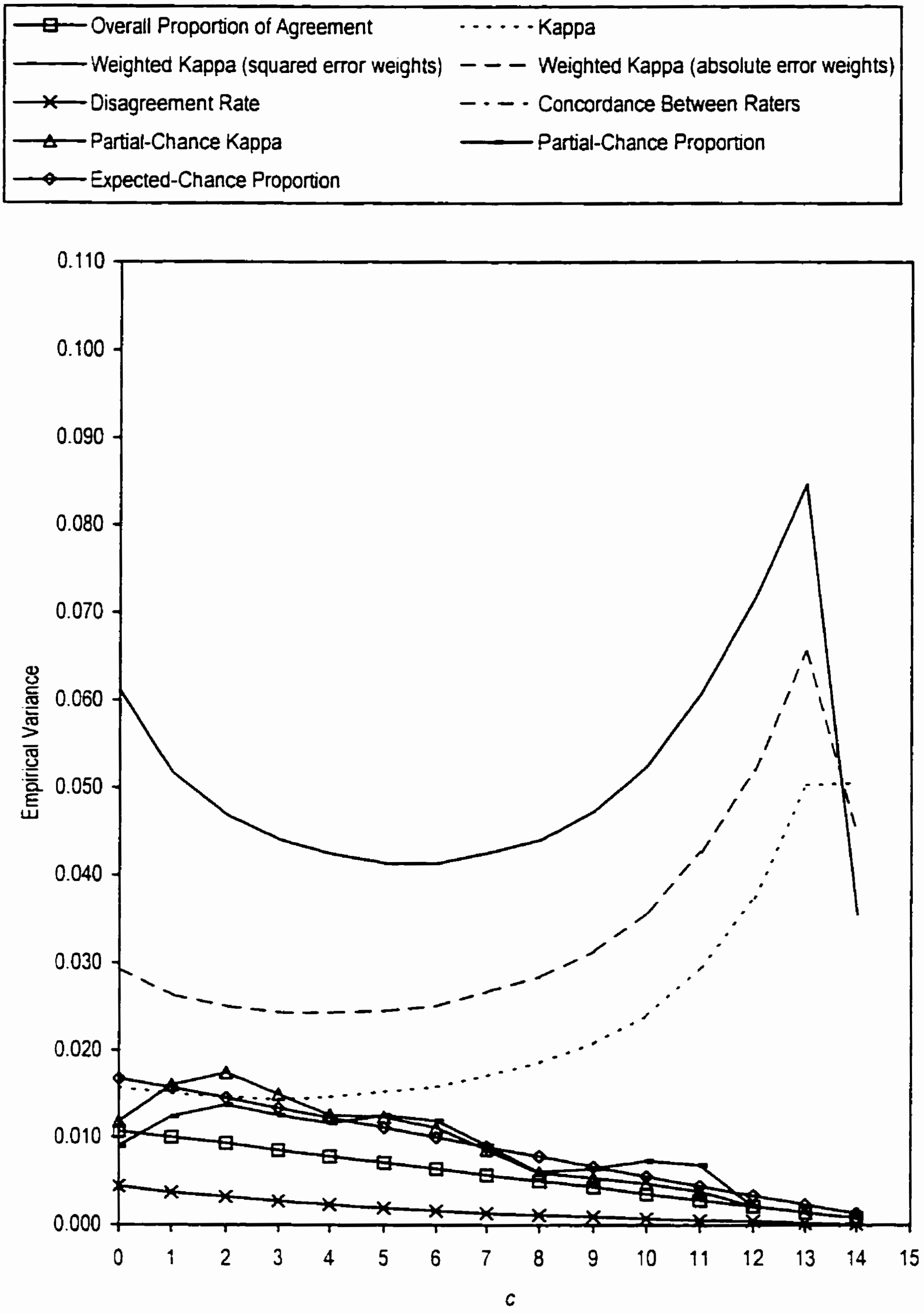


Figure 14: Empirical variances of the measures of interobserver reliability versus c when n=15, k=5, and d=2. Simulation results are based on the nondegenerate samples.

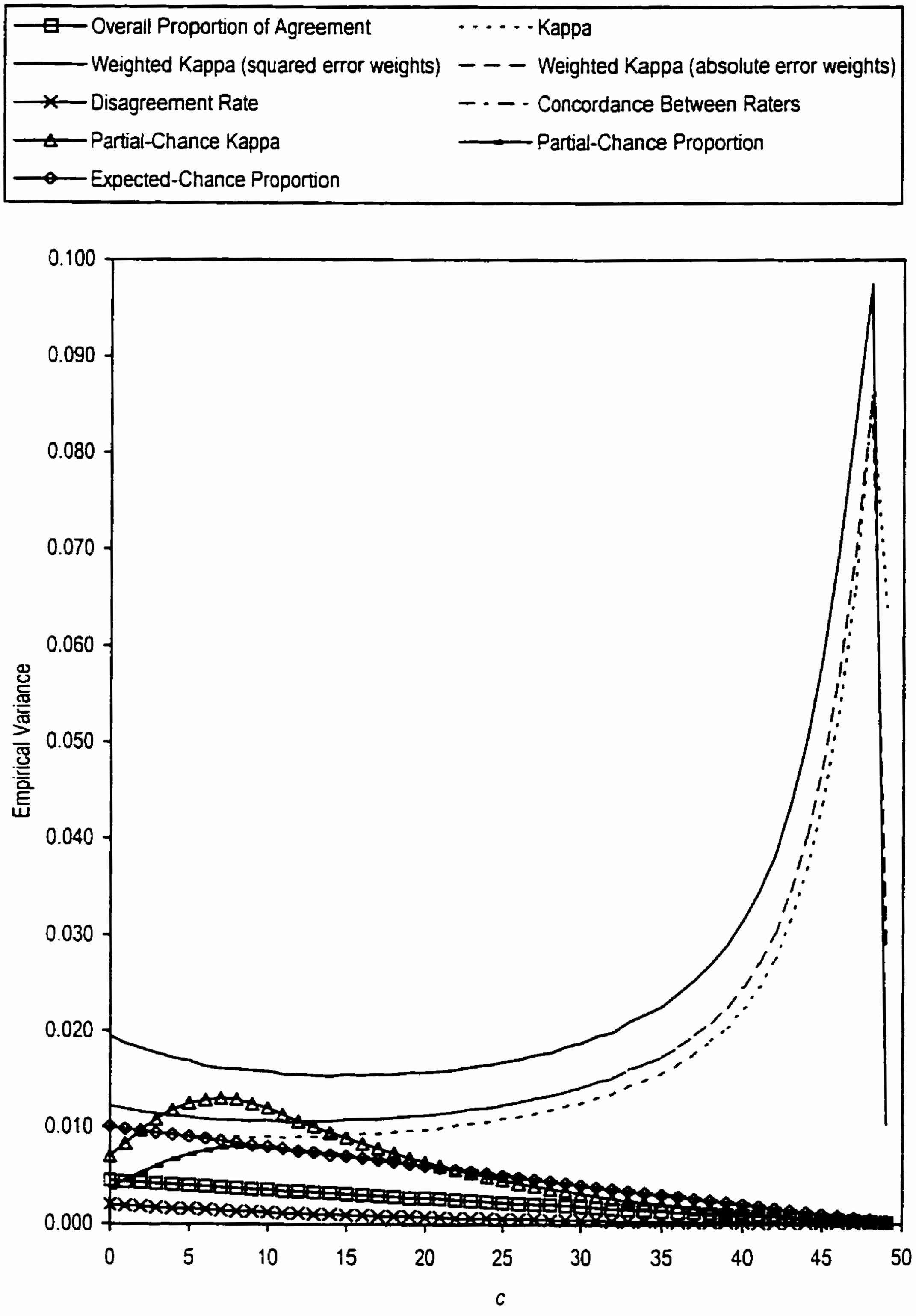


Figure 15: Empirical variances of the measures of interobserver reliability versus c when $n=50$, $k=3$, and $d=2$. Simulation results are based on the nondegenerate samples.

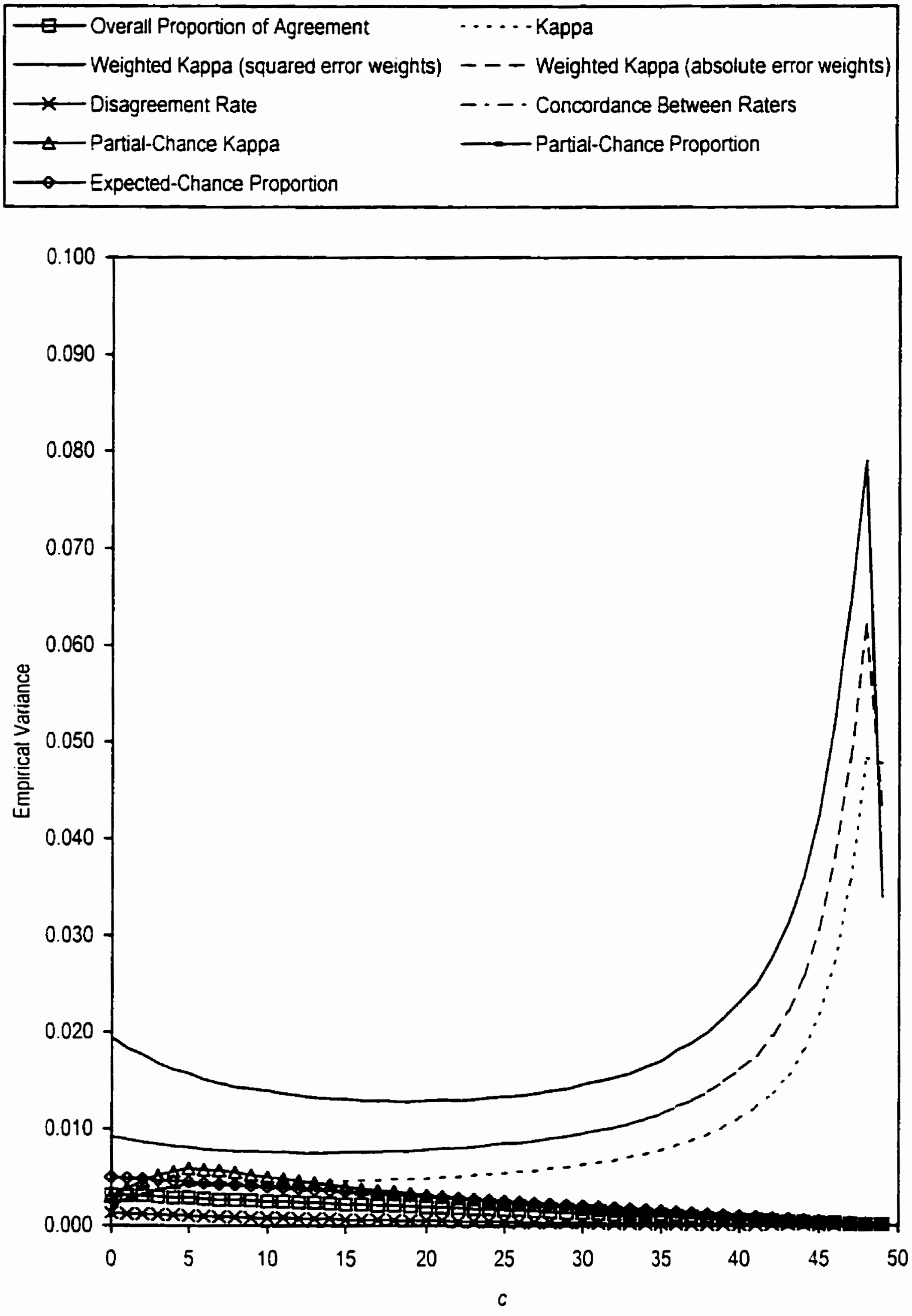


Figure 16: Empirical variances of the measures of interobserver reliability versus c when n=50, k=5, and d=2. Simulation results are based on the nondegenerate samples.

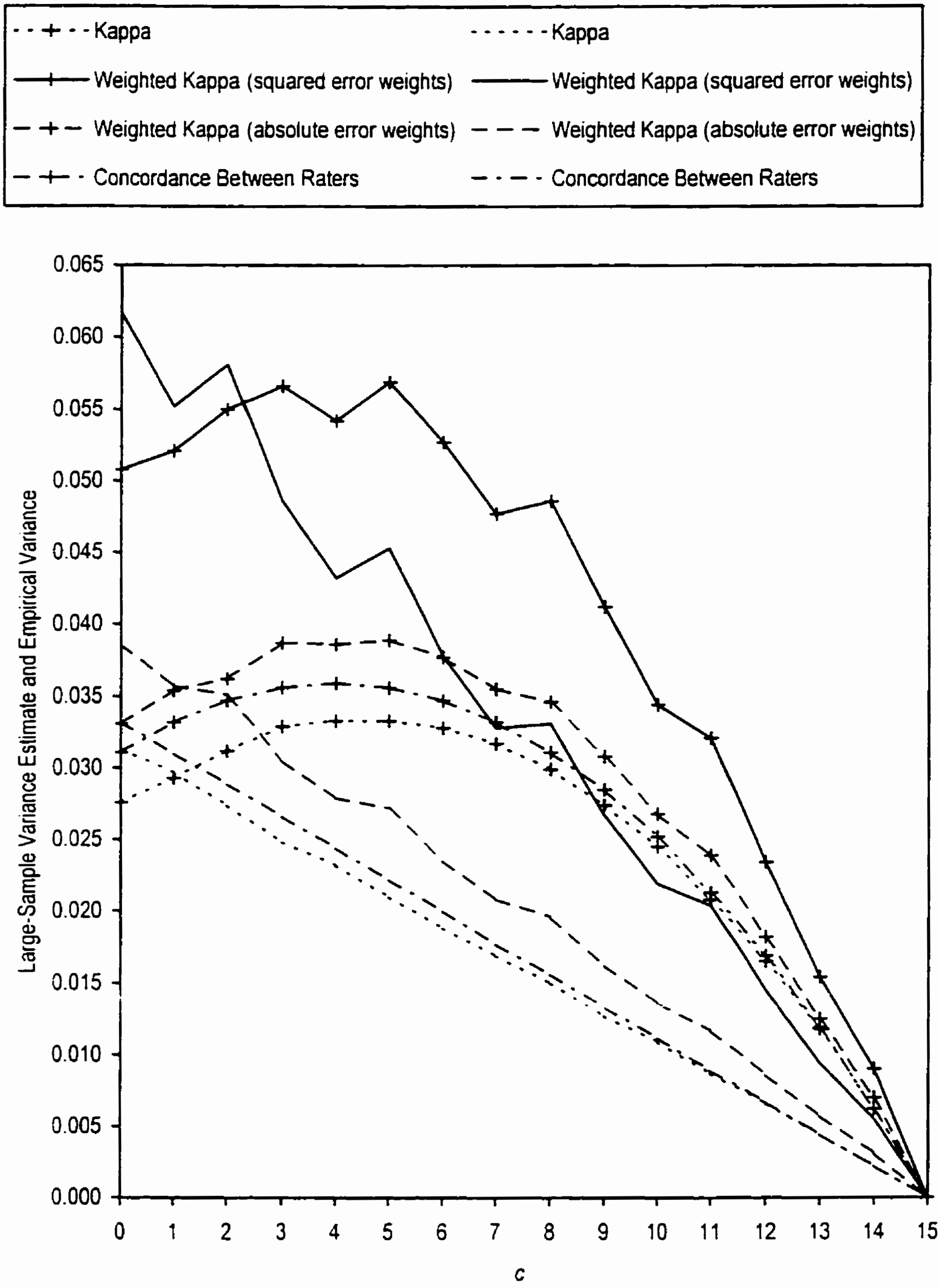


Figure 17: Mean values of the large-sample variance estimates and the empirical variances of provided measures of interobserver reliability versus c when n =15, k=3, and d=1. Simulation results are based on the nondegenerate samples. (unmarked line, empirical variance; marked line [+], mean of the large-sample variance estimates)

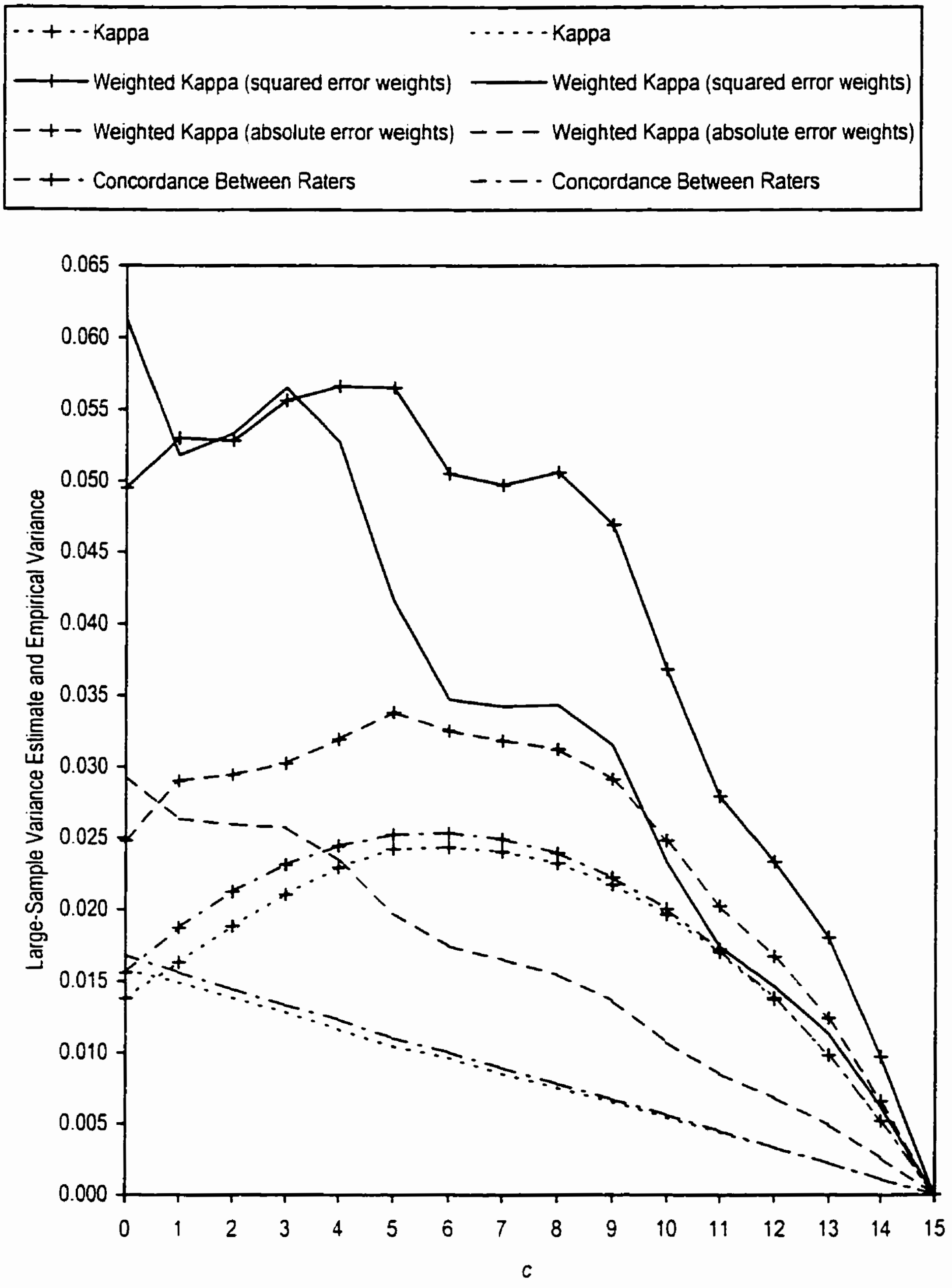


Figure 18: Mean values of the large-sample variance estimates and the empirical variances of provided measures of interobserver reliability versus c when n =15, k=5, and d=1. Simulation results are based on the nondegenerate samples. (unmarked line, empirical variance; marked line [+], mean of the large-sample variance estimates)

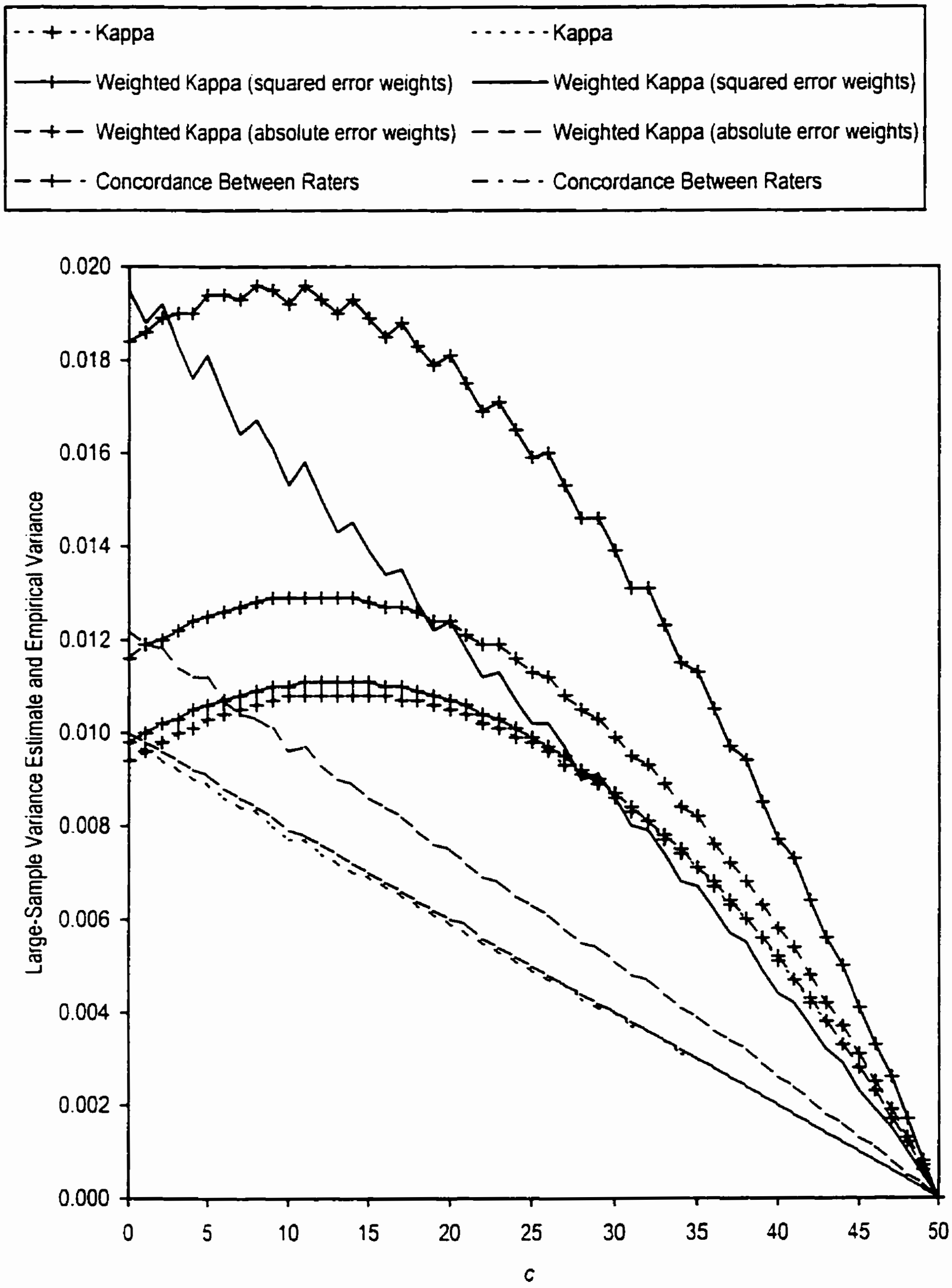


Figure 19: Mean values of the large-sample variance estimates and the empirical variances of provided measures of interobserver reliability versus c when n =50, k=3, and d=1. Simulation results are based on the nondegenerate samples. (unmarked line, empirical variance; marked line [+], mean of the large-sample variance estimates)

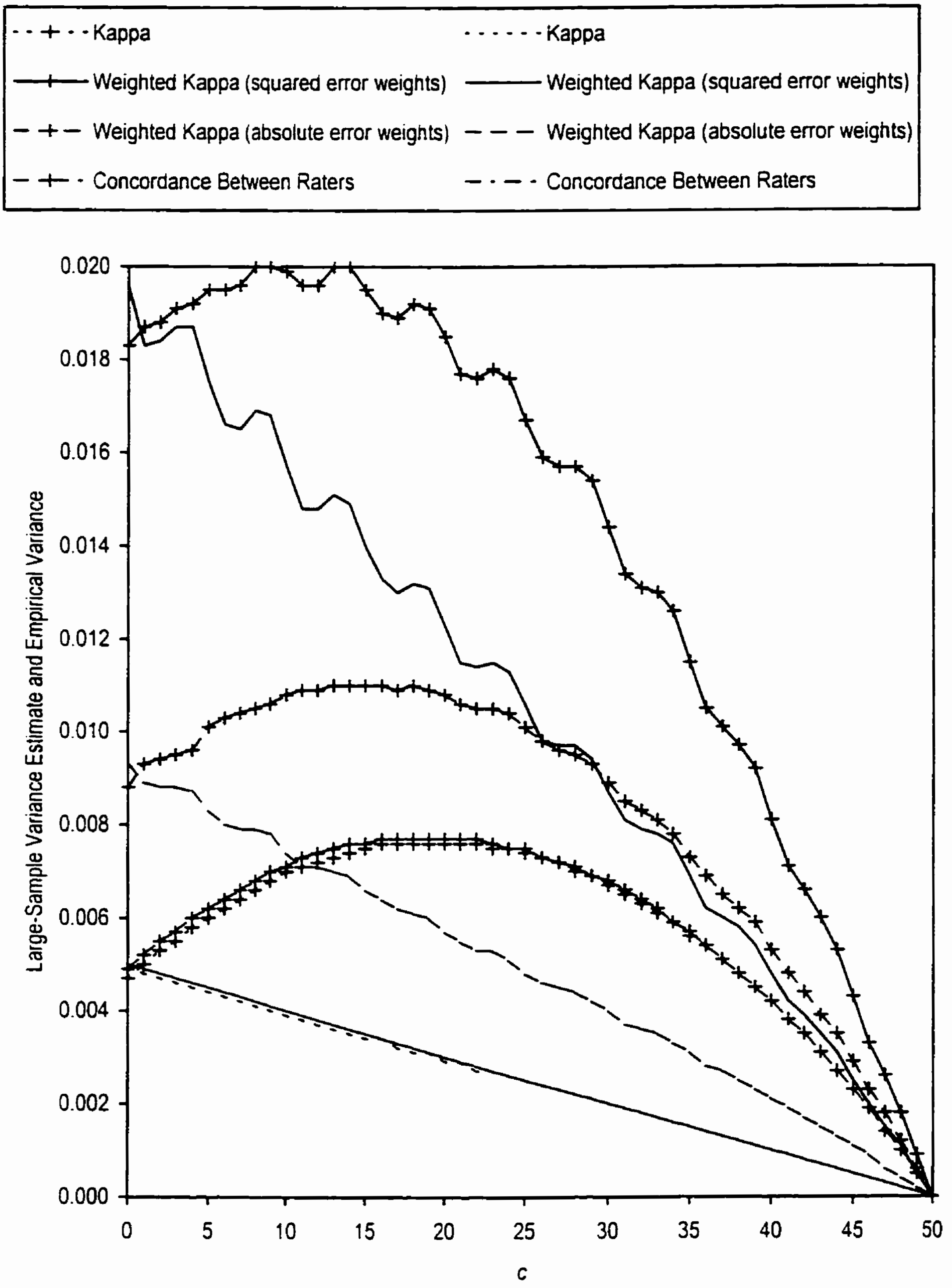


Figure 20: Mean values of the large-sample variance estimates and the empirical variances of provided measures of interobserver reliability versus c when n =50, k=5, and d=1. Simulation results are based on the nondegenerate samples. (unmarked line, empirical variance; marked line [+], mean of the large-sample variance estimates)

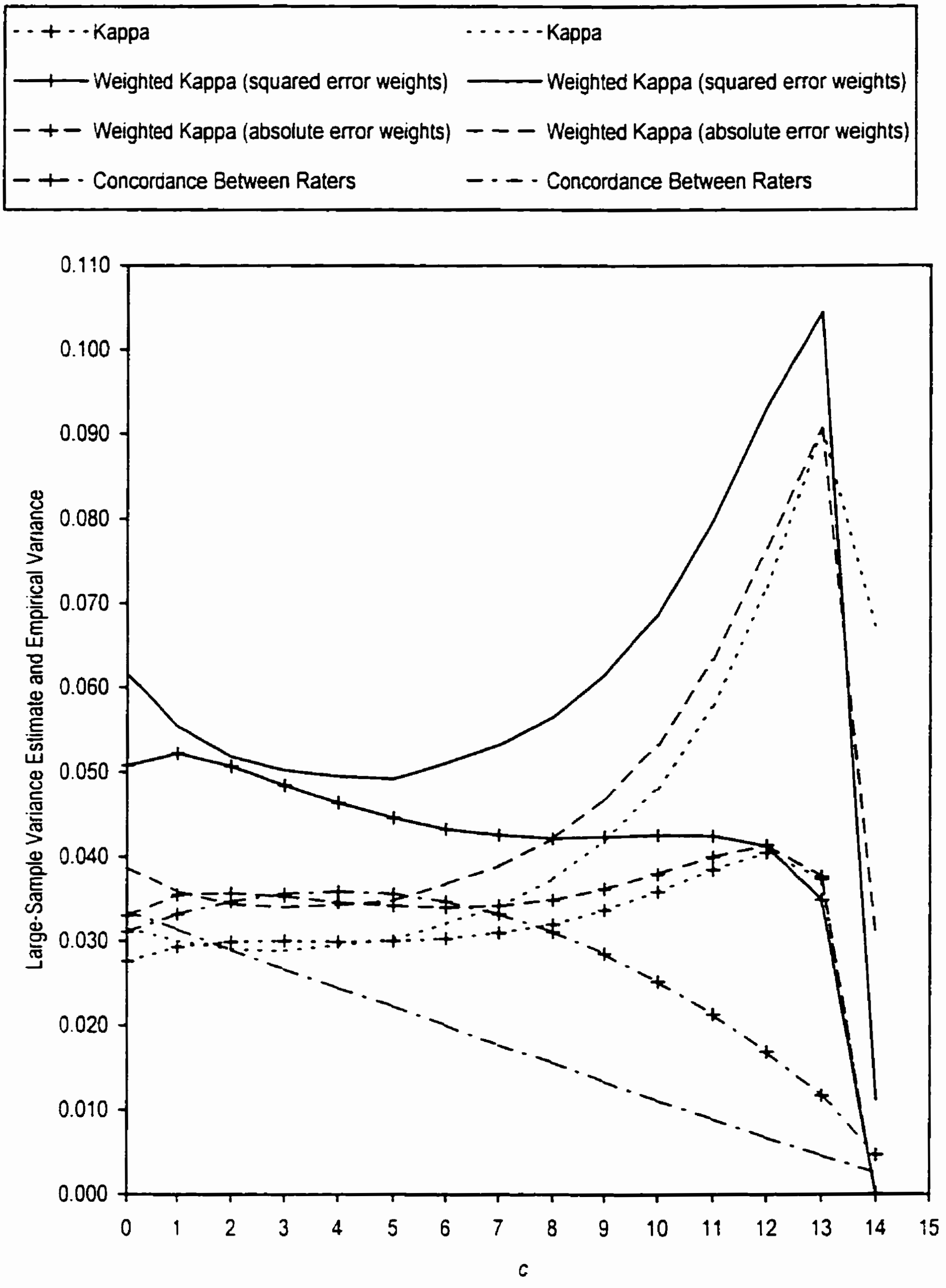


Figure 21: Mean values of the large-sample variance estimates and the empirical variances of provided measures of interobserver reliability versus c when n =15, k=3, and d=2. Simulation results are based on the nondegenerate samples. (unmarked line, empirical variance; marked line [+], mean of the large-sample variance estimates)

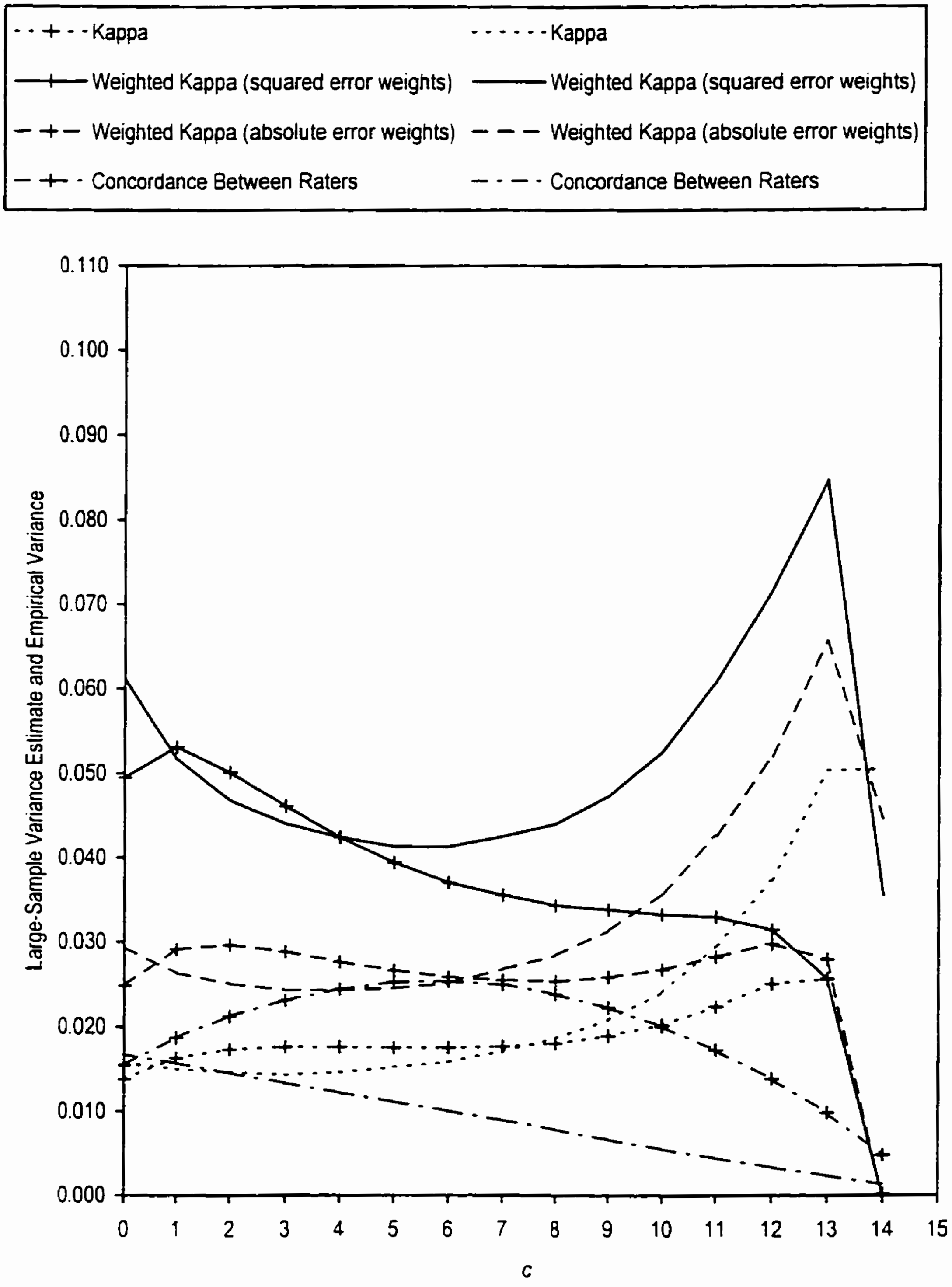


Figure 22: Mean values of the large-sample variance estimates and the empirical variances of provided measures of interobserver reliability versus c when n =15, k=5, and d=2. Simulation results are based on the nondegenerate samples. (unmarked line, empirical variance; marked line [+], mean of the large-sample variance estimates)

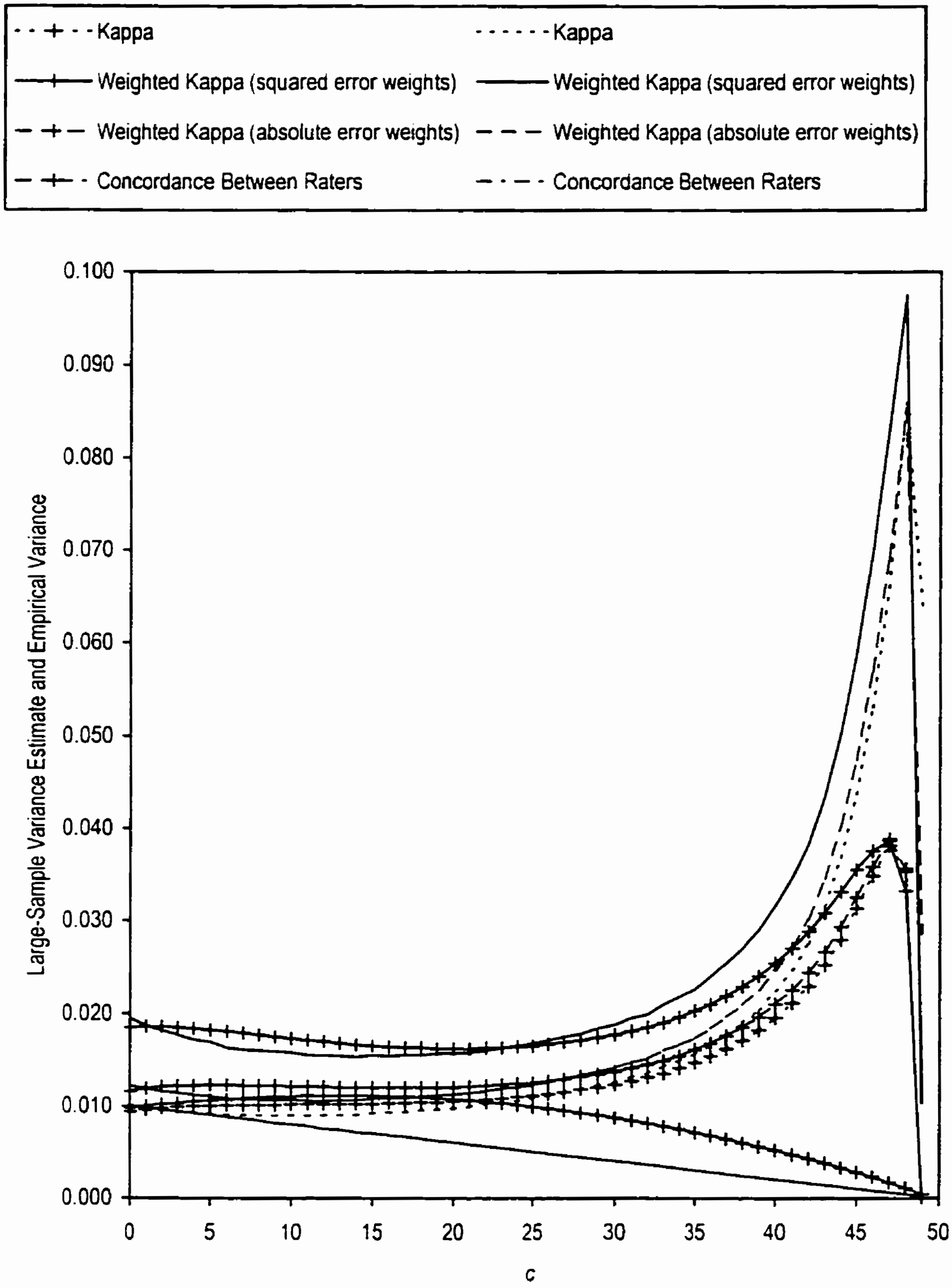


Figure 23: Mean values of the large-sample variance estimates and the empirical variances of provided measures of interobserver reliability versus c when n =50, k=3, and d=2. Simulation results are based on the nondegenerate samples. (unmarked line, empirical variance; marked line [+], mean of the large-sample variance estimates)

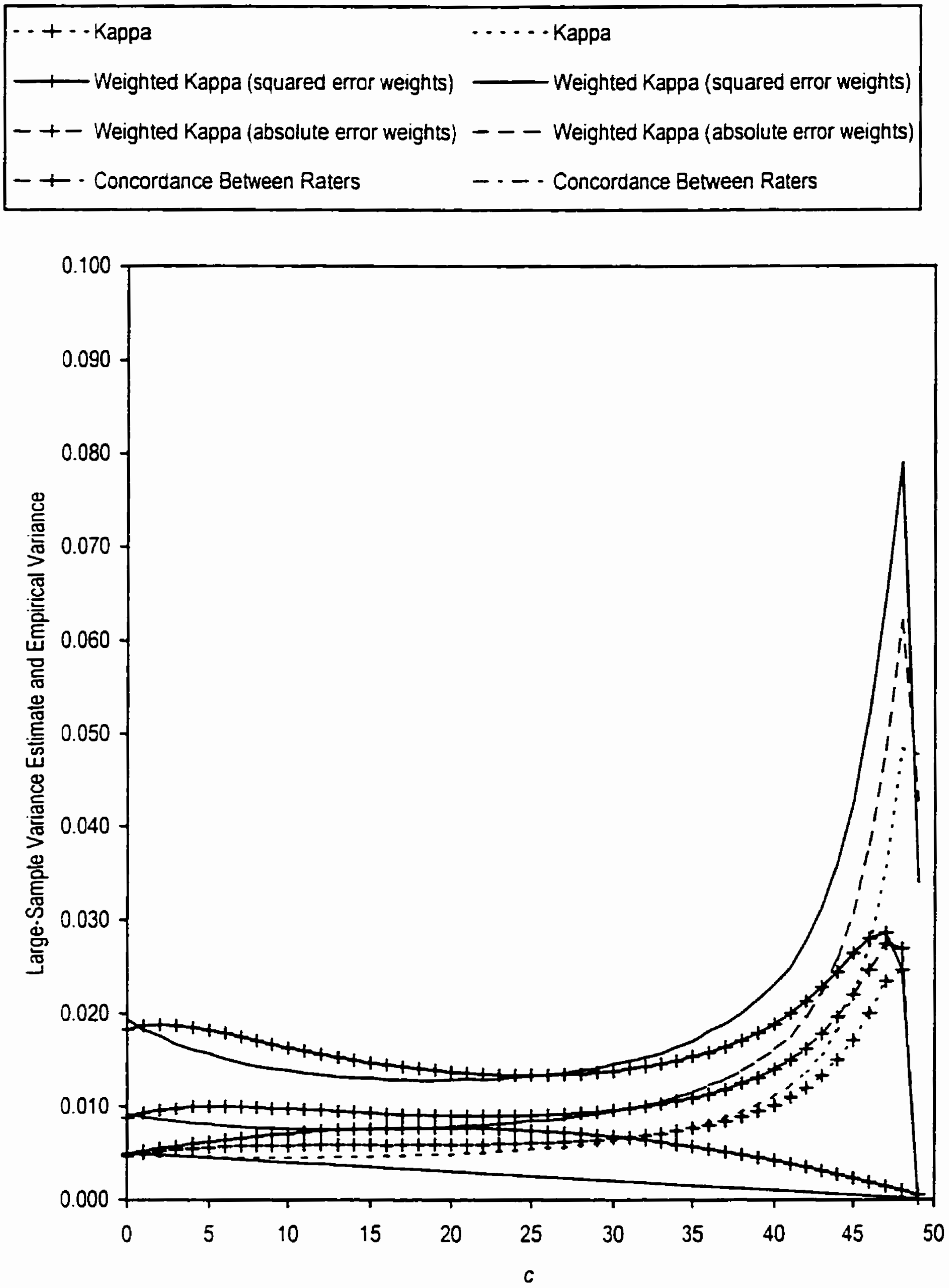


Figure 24: Mean values of the large-sample variance estimates and the empirical variances of provided measures of interobserver reliability versus c when n =50, k=5, and d=2. Simulation results are based on the nondegenerate samples. (unmarked line, empirical variance; marked line [+], mean of the large-sample variance estimates)

Table B-2: Central moments of the null distributions of provided critical ratios when $c = 0$. Simulation results are based on the nondegenerate samples.

Central Moment	Expected Value	$\hat{\kappa}^{a}$	$\hat{\kappa}^{b}$	$\hat{\kappa}_{w}^{ac}$	$\hat{\kappa}_{w}^{bc}$	$\hat{\kappa}_{w}^{ad}$	$\hat{\kappa}_{w}^{bd}$	$\hat{C}_{AB}^{e}$	$\hat{P}_{ec}^{a}$
(i) $n = 15$; $k = 3$; $d = 1$									
Mean	0	-0.0026	-0.0027	-0.0023	-0.0024	-0.0027	-0.0028	-0.0017	-0.1993
Variance	1	0.9954	1.0665	1.0000	1.0714	0.9978	1.0691	0.9955	1.0848
Skewness	0	0.1563	0.1563	0.0028	0.0028	0.1162	0.1162	0.1801	0.1249
Kurtosis	0	-0.1284	-0.1284	-0.2960	-0.2960	-0.2370	-0.2370	-0.0880	-0.0384
(ii) $n = 15$; $k = 5$; $d = 1$									
Mean	0	0.0031	0.0032	0.0036	0.0037	0.0042	0.0044	0.0020	-0.1476
Variance	1	1.0046	1.0763	0.9995	1.0709	1.0033	1.0750	1.0050	1.0902
Skewness	0	0.3383	0.3383	0.0048	0.0048	0.1301	0.1301	0.3829	0.2889
Kurtosis	0	-0.0500	-0.0500	-0.3154	-0.3154	-0.2108	-0.2108	0.0211	-0.0431
(iii) $n = 50$; $k = 3$; $d = 1$									
Mean	0	-0.0054	-0.0054	-0.0024	-0.0025	-0.0039	-0.0040	-0.0060	-0.1081
Variance	1	0.9986	1.0190	1.0006	1.0210	1.0001	1.0205	0.9976	1.0193
Skewness	0	0.0964	0.0964	-0.0124	-0.0124	0.0561	0.0561	0.1011	0.0956
Kurtosis	0	-0.0203	-0.0203	-0.0606	-0.0606	-0.0439	-0.0439	-0.0222	-0.0214
(iv) $n = 50$; $k = 5$; $d = 1$									
Mean	0	-0.0080	-0.0081	-0.0036	-0.0036	-0.0057	-0.0058	-0.0078	-0.0816
Variance	1	0.9955	1.0159	1.0067	1.0273	1.0042	1.0247	0.9955	1.0166
Skewness	0	0.1939	0.1939	-0.0009	-0.0009	0.0702	0.0702	0.2002	0.1872
Kurtosis	0	-0.0204	-0.0204	-0.0821	-0.0821	-0.0531	-0.0531	-0.0219	-0.0265

[a] Exact null variance using the approach of Everitt (1968).

[b] Large-sample null variance using the approach of Fleiss, Cohen and Everitt (1969).

[c] Squared error weights of Fleiss and Cohen (1973).

[d] Absolute error weights of Cicchetti and Allison (1971).

[e] Null variance using the approach of Kupper and Hafner (1989).

Table B-2: *(concluded).*

Central Moment	Expected Value	$\hat{\kappa}^{a}$	$\hat{\kappa}^{b}$	$\hat{\kappa}_{w}^{ac}$	$\hat{\kappa}_{w}^{bc}$	$\hat{\kappa}_{w}^{ad}$	$\hat{\kappa}_{w}^{bd}$	$\hat{C}_{AB}^{e}$	$\hat{P}_{ec}^{a}$
(v) $n = 15$; $k = 3$; $d = 2$									
Mean	0	-0.0018	-0.0019	-0.0037	-0.0039	-0.0030	-0.0031	-0.0027	-0.2007
Variance	1	1.0023	1.0739	0.9993	1.0707	1.0004	1.0719	1.0039	1.0937
Skewness	0	0.1503	0.1503	-0.0083	-0.0083	0.1022	0.1022	0.1716	0.1122
Kurtosis	0	-0.1600	-0.1600	-0.3143	-0.3143	-0.2584	-0.2584	-0.1172	-0.0835
(vi) $n = 15$; $k = 5$; $d = 2$									
Mean	0	-0.0001	-0.0001	0.0053	0.0055	0.0031	0.0033	-0.0010	-0.1507
Variance	1	1.0030	1.0746	0.9989	1.0702	1.0022	1.0738	1.0020	1.0887
Skewness	0	0.3394	0.3394	0.0073	0.0073	0.1305	0.1305	0.3761	0.2824
Kurtosis	0	-0.0716	-0.0716	-0.3281	-0.3281	-0.2440	-0.2440	-0.0271	-0.0793
(vii) $n = 50$; $k = 3$; $d = 2$									
Mean	0	-0.0017	-0.0018	0.0005	0.0005	-0.0005	-0.0005	-0.0019	-0.1041
Variance	1	1.0050	1.0255	0.9998	1.0202	1.0022	1.0227	1.0051	1.0270
Skewness	0	0.1054	0.1054	0.0086	0.0086	0.0753	0.0753	0.1094	0.1018
Kurtosis	0	-0.0109	-0.0109	-0.0860	-0.0860	-0.0597	-0.0597	-0.0085	-0.0054
(viii) $n = 50$; $k = 5$; $d = 2$									
Mean	0	-0.0038	-0.0038	-0.0046	-0.0046	-0.0039	-0.0039	-0.0041	-0.0777
Variance	1	1.0029	1.0233	0.9966	1.0169	0.9966	1.0169	1.0012	1.0224
Skewness	0	0.2038	0.2038	-0.0025	-0.0025	0.0685	0.0685	0.2115	0.1985
Kurtosis	0	-0.0029	-0.0029	-0.0800	-0.0800	-0.0522	-0.0522	0.0093	0.0023

[a] Exact null variance using the approach of Everitt (1968).

[b] Large-sample null variance using the approach of Fleiss, Cohen and Everitt (1969).

[c] Squared error weights of Fleiss and Cohen (1973).

[d] Absolute error weights of Cicchetti and Allison (1971).

[e] Null variance using the approach of Kupper and Hafner (1989).

Table B-3: Empirical tail areas of the null distributions of provided critical ratios when $c = 0$. Simulation results are based on the nondegenerate samples.

Interval	Expected Prop.	$\hat{\kappa}^{a}$	$\hat{\kappa}^{b}$	$\hat{\kappa}_{w}^{ac}$	$\hat{\kappa}_{w}^{bc}$	$\hat{\kappa}_{w}^{ad}$	$\hat{\kappa}_{w}^{bd}$	$\hat{C}_{AB}^{e}$	$\hat{P}_{ec}^{a}$
(i) $n = 15$; $k = 3$; $d = 1$									
$Z \leq -2.576$	0.005	0.0023	0.0029	0.0028	0.0042	0.0015	0.0026	0.0025	0.0074
$Z \leq -1.96$	0.025	0.0210	0.0228	0.0236	0.0279	0.0198	0.0246	0.0197	0.0362
$Z \geq 1.96$	0.025	0.0298	0.0330	0.0238	0.0281	0.0274	0.0327	0.0309	0.0298
$Z \geq 2.576$	0.005	0.0067	0.0084	0.0027	0.0041	0.0052	0.0070	0.0086	0.0051
(ii) $n = 15$; $k = 5$; $d = 1$									
$Z \leq -2.576$	0.005	0.0000	0.0001	0.0025	0.0038	0.0016	0.0024	0.0000	0.0019
$Z \leq -1.96$	0.025	0.0127	0.0198	0.0226	0.0271	0.0195	0.0242	0.0000	0.0368
$Z \geq 1.96$	0.025	0.0333	0.0387	0.0240	0.0286	0.0280	0.0324	0.0180	0.0220
$Z \geq 2.576$	0.005	0.0087	0.0111	0.0028	0.0040	0.0058	0.0074	0.0180	0.0061
(iii) $n = 50$; $k = 3$; $d = 1$									
$Z \leq -2.576$	0.005	0.0037	0.0042	0.0048	0.0052	0.0038	0.0043	0.0053	0.0054
$Z \leq -1.96$	0.025	0.0232	0.0246	0.0258	0.0271	0.0238	0.0250	0.0285	0.0288
$Z \geq 1.96$	0.025	0.0263	0.0275	0.0240	0.0254	0.0260	0.0271	0.0218	0.0218
$Z \geq 2.576$	0.005	0.0057	0.0062	0.0043	0.0047	0.0054	0.0058	0.0048	0.0049
(iv) $n = 50$; $k = 5$; $d = 1$									
$Z \leq -2.576$	0.005	0.0017	0.0020	0.0047	0.0050	0.0038	0.0042	0.0012	0.0033
$Z \leq -1.96$	0.025	0.0193	0.0200	0.0257	0.0267	0.0236	0.0247	0.0190	0.0198
$Z \geq 1.96$	0.025	0.0297	0.0307	0.0250	0.0263	0.0262	0.0275	0.0299	0.0299
$Z \geq 2.576$	0.005	0.0072	0.0076	0.0045	0.0048	0.0058	0.0063	0.0058	0.0058

[a] Exact null variance using the approach of Everitt (1968).

[b] Large-sample null variance using the approach of Fleiss, Cohen and Everitt (1969).

[c] Squared error weights of Fleiss and Cohen (1973).

[d] Absolute error weights of Cicchetti and Allison (1971).

[e] Null variance using the approach of Kupper and Hafner (1989).

Table B-3: *(concluded).*

Interval	Expected Prop.	$\hat{\kappa}^{a}$	$\hat{\kappa}^{b}$	$\hat{\kappa}_w^{ac}$	$\hat{\kappa}_w^{bc}$	$\hat{\kappa}_w^{ad}$	$\hat{\kappa}_w^{bd}$	$\hat{C}_{AB}^{e}$	$\hat{P}_{ec}^{a}$
(v) $n = 15$; $k = 3$; $d = 2$									
$Z \leq -2.576$	0.005	0.0022	0.0027	0.0030	0.0044	0.0014	0.0026	0.0024	0.0076
$Z \leq -1.96$	0.025	0.0212	0.0233	0.0240	0.0280	0.0201	0.0248	0.0203	0.0376
$Z \geq 1.96$	0.025	0.0298	0.0330	0.0225	0.0264	0.0265	0.0317	0.0304	0.0291
$Z \geq 2.576$	0.005	0.0066	0.0081	0.0028	0.0039	0.0049	0.0065	0.0083	0.0051
(vi) $n = 15$; $k = 5$; $d = 2$									
$Z \leq -2.576$	0.005	0.0001	0.0001	0.0027	0.0039	0.0016	0.0024	0.0000	0.0020
$Z \leq -1.96$	0.025	0.0121	0.0195	0.0223	0.0267	0.0187	0.0230	0.0000	0.0363
$Z \geq 1.96$	0.025	0.0336	0.0385	0.0232	0.0280	0.0283	0.0331	0.0177	0.0219
$Z \geq 2.576$	0.005	0.0089	0.0111	0.0028	0.0038	0.0050	0.0065	0.0177	0.0063
(vii) $n = 50$; $k = 3$; $d = 2$									
$Z \leq -2.576$	0.005	0.0038	0.0042	0.0043	0.0046	0.0035	0.0038	0.0052	0.0054
$Z \leq -1.96$	0.025	0.0234	0.0247	0.0245	0.0259	0.0230	0.0240	0.0286	0.0288
$Z \geq 1.96$	0.025	0.0273	0.0285	0.0254	0.0266	0.0275	0.0287	0.0228	0.0229
$Z \geq 2.576$	0.005	0.0061	0.0064	0.0041	0.0046	0.0054	0.0059	0.0052	0.0052
(viii) $n = 50$; $k = 5$; $d = 2$									
$Z \leq -2.576$	0.005	0.0020	0.0022	0.0045	0.0050	0.0037	0.0041	0.0016	0.0036
$Z \leq -1.96$	0.025	0.0191	0.0196	0.0243	0.0257	0.0227	0.0238	0.0186	0.0193
$Z \geq 1.96$	0.025	0.0305	0.0314	0.0240	0.0252	0.0260	0.0271	0.0306	0.0306
$Z \geq 2.576$	0.005	0.0078	0.0083	0.0044	0.0048	0.0055	0.0060	0.0064	0.0064

[a] Exact null variance using the approach of Everitt (1968).

[b] Large-sample null variance using the approach of Fleiss, Cohen and Everitt (1969).

[c] Squared error weights of Fleiss and Cohen (1973).

[d] Absolute error weights of Cicchetti and Allison (1971).

[e] Null variance using the approach of Kupper and Hafner (1989).

APPENDIX C

SIMULATION PROGRAM

```
C  MIRS.FOR - Measures of Interobserver Reliability Simulation
C
C
C  Variable dictionary
C
C  IDUM     :Random number generator seed (must be a negative integer)
C  RAN2     :Function (see for explanation)
C  N        :Sample size of subjects
C  LC       :Number of levels of C
C  CA(_)    :Levels of C
C  C        :Perfect knowledge based (nonguessed) agreements
C  K        :Categories of classification
C  D        :Distribution of C
C             1-C is spaced evenly along the main diagonal
C             2-C observes 1st category of classification (prevalence)
C  SIMS     :Number of simulations
C  I        :Counter
C  J        :Counter
C  OUTMAT   :Indicator for output simulated matrix
C  CC       :Parameter combination loop counter
C  TP       :True proportion
C  A        :Simulation loop counter
C  DEG      :Number of degenerate samples per parameter combination
C  PA       :Overall proportion of agreement
C  SPA      : sum
C  SPA2     : sum of squares
C  PASV     : empirical variance
C  KC       :Cohen's kappa
C  SKC      : sum
C  SKC2     : sum of squares
C  KCSV     : empirical variance
C  VKC      : large-sample estimate of nonnull variance
C  SVKC     : sum of VKC
C  V0KC     : large-sample estimate of null variance
C  SUM1     : sum of terms in calculation of V0KC
C  EVKC     : exact null variance
C  ZEKC     : Z ratio using exact null variance
C  SZEKC(_) : sums of powers of ZEKC
C  BZEKC(_) : central moments of ZEKC
C  ZKC      : Z ratio using large-sample null variance estimate
C  SZKC(_)  : sums of powers of ZKC
C  BZKC(_)  : central moments of ZKC
C  PZEKC(_) : 1-tailed areas using exact null variance
             :
             .
```

```
C     PZKC(_)  : 1-tailed areas using large-sample null variance estimate
C     KW1      :Weighted kappa (squared error weights)
C     SKW1     : sum
C     SKW12    : sum of squares
C     KW1SV    : empirical variance
C     VKW1     : large-sample estimate of nonnull variance
C     SVKW1    : sum of VKW1
C     V0KW1    : large-sample estimate of null variance
C     EVKW1    : exact null variance
C     ZEKW1    : Z ratio using exact null variance
C     SZEKW1(_): sums of powers of ZEKW1
C     BZEKW1(_): central moments of ZEKW1
C     ZKW1     : Z ratio using large-sample null variance estimate
C     SZKW1(_) : sums of powers of ZKW1
C     BZKW1(_) : central moments of ZKW1
C     PZEKW1(_): 1-tailed areas using exact null variance
C     PZKW1(_) : 1-tailed areas using large-sample null variance estimate
C     KW2      :Weighted kappa (absolute error weights)
C     SKW2     : sum
C     SKW22    : sum of squares
C     KW2SV    : empirical variance
C     VKW2     : large-sample estimate of nonnull variance
C     SVKW2    : sum of VKW2
C     V0KW2    : large-sample estimate of null variance
C     EVKW2    : exact null variance
C     ZEKW2    : Z ratio using exact null variance
C     SZEKW2(_): sums of powers of ZEKW2
C     BZEKW2(_): central moments of ZEKW2
C     ZKW2     : Z ratio using large-sample null variance estimate
C     SZKW2(_) : sums of powers of ZKW2
C     BZKW2(_) : central moments of ZKW2
C     PZEKW2(_): 1-tailed areas using exact null variance
C     PZKW2(_) : 1-tailed areas using large-sample null variance estimate
C     DR       :The disagreement rate
C     NUM      : numerator in calculation of DR
C     DEN      : denominator in calculation of DR
C     SDR      : sum
C     SDR2     : sum of squares
C     DRSV     : empirical variance
C     CAB      :The concordance between raters
C     SCAB     : sum
C     SCAB2    : sum of squares
C     CABSV    : empirical variance
C     V0CAB    : large-sample estimate of null variance
C     VCAB     : large-sample estimate of nonnull variance
C     SVCAB    : sum of VCAB
C     ZCAB     : Z ratio using large-sample null variance estimate
C     SZCAB(_) : sums of powers of ZCAB
C     BZCAB(_) : central moments of ZCAB
C     PZCAB(_) : 1-tailed areas using large-sample null variance estimate
C     GSTAR    :Estimated number of guessed observations in N subjects
C     KPC      :Partial-Chance Kappa (equal weights)
C     SKPC     : sum
C     SKPC2    : sum of squares
C     KPCSV    : empirical variance
C     PPC      :Partial-Chance Proportion
C     SPPC     : sum
```

```
C  SPPC2     : sum of squares
C  PPCSV     : empirical variance
C  PEC       :Expected-Chance Proportion
C  SPEC      : sum
C  SPEC2     : sum of squares
C  PECSV     : empirical variance
C  EVPEC     : exact null variance
C  ZEPEC     : Z ratio using exact null variance
C  SZEPEC(_): sums of powers of ZEPEC
C  BZEPEC(_): central moments of ZEPEC
C  PZEPEC(_): 1-tailed areas using exact null variance
C  W(_,_,2)  :K+1 x K+1 weight matrix, 3rd-dimension is an indicator:
C             1-Squared error weights of Fleiss and Cohen (1973)
C             2-Absolute error weights of Cicchetti and Allison (1971)
C  FLAG      :Flags a degenerate sample
C  NRAN      :Number of guessed observations
C  IX1       :Guessed observation from Rater A
C  IX2       :Guessed observation from Rater B
C  IX(_,2)   :N x 2 table of subjects' classifications, subjects'-matrix
C  JCNT      :Counter
C  IC(_,_)   :K+1 x K+1 table of simulated obs., observations-matrix
C  DEGVAL    :Remote Block (see for explanation)
C  TO        :Total number of observed agreements
C  TC        :Total number of chance-expected agreements
C  V0TO      :Exact null variance of TO
C  SUMV      :Sum of variances used in calculation of V0TO
C  SUMC      :Sum of covariances used in calculation of V0TO
C  TOW1      :Wtd(1) total number of observed agreements
C  TCW1      :Wtd(1) total number of chance-expected agreements
C  V0TOW1    :Exact null variance of TOW1
C  SUMVW1    :Sum of variances used in calculation of V0TOW1
C  SUMCW1    :Sum of covariances used in calculation of V0TOW1
C  TOW2      :Wtd(2) total number of observed agreements
C  TCW2      :Wtd(2) total number of chance-expected agreements
C  V0TOW2    :Exact null variance of TOW2
C  SUMVW2    :Sum of variances used in calculation of V0TOW2
C  SUMCW2    :Sum of covariances used in calculation of V0TOW2
C  S         :Counter
C  T         :Counter
C  ENDCEL    :Indicator for searching for end of cells
C  SIMS2     :Number of non-degenerate samples

C  Declare variables

      INTEGER IDUM,LC,CA(51),SIMS,D,OUTMAT,CC,N,C,K,A,NRAN,IX1,INT,
     *IX2,IX(50,2),JCNT,I,J,IC(6,6),TO,ABS,SUMV,SUMC,S,T,ENDCEL,SUM1,
     *DEG,FLAG,NUM,GSTAR,SIMS2,PZEKC(4),PZKC(4),PZEKW1(4),PZKW1(4),
     *PZEKW2(4),PZKW2(4),PZCAB(4),PZEPEC(4)

      REAL RAN2

      DOUBLE PRECISION W(6,6,2),TC,V0TO,TOW1,TOW2,TCW1,TCW2,SUMVW1,
     *SUMVW2,SUMCW1,SUMCW2,V0TOW1,V0TOW2,DEN,MAX,DSQRT,TP,PA,SPA,SPA2,
     *PASV,KC,EVKC,VKC,V0KC,ZEKC,ZKC,SKC,SKC2,SVKC,SZEKC(4),SZKC(4),
     *BZEKC(4),BZKC(4),KCSV,KW1,EVKW1,VKW1,V0KW1,ZEKW1,ZKW1,SKW1,SKW12,
     *SVKW1,SZEKW1(4),SZKW1(4),BZEKW1(4),BZKW1(4),KW1SV,KW2,EVKW2,VKW2,
     *V0KW2,ZEKW2,ZKW2,SKW2,SKW22,SVKW2,SZEKW2(4),SZKW2(4),BZEKW2(4),
     *BZKW2(4),KW2SV,DR,SDR,SDR2,DRSV,CAB,VCAB,V0CAB,ZCAB,SCAB,SCAB2,
     *SVCAB,SZCAB(4),BZCAB(4),CABSV,KPC,SKPC,SKPC2,KPCSV,PPC,SPPC,
     *SPPC2,PPCSV,PEC,EVPEC,ZEPEC,SPEC,SPEC2,SZEPEC(4),BZEPEC(4),PECSV

C  Open existing output files

      OPEN(1,FILE='C:\FORTRAN\MIRS\MIRS_MAT.TXT',STATUS='OLD')
      OPEN(2,FILE='C:\FORTRAN\MIRS\MIRS_ME1.TXT',STATUS='OLD')
      OPEN(3,FILE='C:\FORTRAN\MIRS\MIRS_ME2.TXT',STATUS='OLD')
      OPEN(4,FILE='C:\FORTRAN\MIRS\MIRS_VA1.TXT',STATUS='OLD')
      OPEN(7,FILE='C:\FORTRAN\MIRS\MIRS_VA2.TXT',STATUS='OLD')
      OPEN(8,FILE='C:\FORTRAN\MIRS\MIRS_BZ1.TXT',STATUS='OLD')
      OPEN(9,FILE='C:\FORTRAN\MIRS\MIRS_BZ2.TXT',STATUS='OLD')
      OPEN(10,FILE='C:\FORTRAN\MIRS\MIRS_BZ3.TXT',STATUS='OLD')
      OPEN(11,FILE='C:\FORTRAN\MIRS\MIRS_BZ4.TXT',STATUS='OLD')
      OPEN(12,FILE='C:\FORTRAN\MIRS\MIRS_DEG.TXT',STATUS='OLD')
      OPEN(13,FILE='C:\FORTRAN\MIRS\MIRS_PRO.TXT',STATUS='OLD')

C  Output simulation program introduction

      WRITE(*,10005)
      WRITE(1,10005)
10005 FORMAT(/,'********************************************************
     ****',/,'mirs.for - Measures of Interobserver Reliability Simulatio
     *n',/,'***********************************************************'
     *)

C  Read in parameters for simulation

      IDUM=0
      WHILE (IDUM.GE.0) DO
        WRITE(*,10010)
10010   FORMAT(/,'RANDOM NUMBER GENERATOR SEED (INTEGER<0):')
        READ *,IDUM
      END WHILE
      WRITE(*,10020)
10020 FORMAT(/,'N [SAMPLE SIZE OF SUBJECTS]:')
      READ *,N
      WRITE(*,10023)
10023 FORMAT(/,'LEVELS OF C [PERFECT KNOWLEDGE BASED (NONGUESSED) AGREE
     *MENTS]:')
      READ *,LC
      DO 10024 I=1,LC
        WRITE(*,10025) I
10025   FORMAT(' ','  LEVEL ',I2,' OF C:')
        READ *,CA(I)
10024 CONTINUE
      WRITE(*,10026)
10026 FORMAT(/,'K [CATEGORIES OF CLASSIFICATION ON A GIVEN SCALE]:')
      READ *,K
      WRITE(*,10036)
10036 FORMAT(/,'SIMULATIONS:')
      READ *,SIMS
      D=0
      WHILE (D.LT.1.OR.D.GT.2) DO
        WRITE(*,10042)
10042   FORMAT(/,'DISTRIBUTION OF C? [1:EVENLY SPACED, 2:PREVALENT]')
```

```
        READ *,D
      END WHILE
      OUTMAT=0
      WHILE (OUTMAT.LT.1.OR.OUTMAT.GT.2) DO
        WRITE(*,10043)
10043   FORMAT(/,'OUTPUT SIMULATED MATRIX? [1:YES, 2:NO]')
        READ *,OUTMAT
      END WHILE


C  Output simulation parameters to output file MIRS_MAT.TXT

      WRITE(1,10041) IDUM,SIMS,D,K,N,(CA(I),I=1,LC)
10041 FORMAT(/,'IDUM:  ',I11,/,'SIMS:',6X,I6,/,'    D:',11X,I1,/,'    K:',
     *11X,I1,/,'    N:',10X,I2,/,'    C:',10X,10(I2,2X))

C  Output column headings to various output files

      WRITE(2,20053)
20053 FORMAT(/,'C_TP_PA_KC_KW1_KW2_DR')
      WRITE(3,20054)
20054 FORMAT(/,'C_TP_CAB_KPC_PPC_PEC')
      WRITE(4,20055)
20055 FORMAT(/,'C_PASV_KCSV_VKC_KW1SV_VKW1_KW2SV_VKW2')
      WRITE(7,20056)
20056 FORMAT(/,'C_DRSV_CABSV_VCAB_KPCSV_PPCSV_PECSV')
      DO 20057 I=1,4
        WRITE(I+7,20058) I
20058   FORMAT(/,I1,'_C_ZEKC_ZKC_ZEKW1_ZKW1_ZEKW2_ZKW2_ZCAB_ZEPEC')
20057 CONTINUE
      WRITE(12,20059)
20059 FORMAT(/,'C_DEG')
      WRITE(13,20060)
20060 FORMAT(/,'C_Tail_ZEKC_ZKC_ZEKW1_ZKW1_ZEKW2_ZKW2_ZCAB_ZEPEC')

C  Parameter combination loop commences

      DO 4000 CC=1,LC
      C=CA(CC)
      IF (C.GT.N) GO TO 4000

C  Initialize variables per parameter combination

      DEG=0
      SPA=0.0D0
      SPA2=0.0D0
      SKC=0.0D0
      SKC2=0.0D0
      SVKC=0.0D0
      SKW1=0.0D0
      SKW12=0.0D0
      SVKW1=0.0D0
      SKW2=0.0D0
      SKW22=0.0D0
      SVKW2=0.0D0
      SDR=0.0D0
      SDR2=0.0D0
      SCAB=0.0D0
      SCAB2=0.0D0
      SVCAB=0.0D0
      SKPC=0.0D0
      SKPC2=0.0D0
      SPPC=0.0D0
      SPPC2=0.0D0
      SPEC=0.0D0
      SPEC2=0.0D0
      DO 20062 I=1,4
        SZEKC(I)=0.0D0
        SZKC(I)=0.0D0
        SZEKW1(I)=0.0D0
        SZKW1(I)=0.0D0
        SZEKW2(I)=0.0D0
        SZKW2(I)=0.0D0
        SZCAB(I)=0.0D0
        SZEPEC(I)=0.0D0
20062 CONTINUE
      DO 20063 I=1,4
        PZEKC(I)=0
        PZKC(I)=0
        PZEKW1(I)=0
        PZKW1(I)=0
        PZEKW2(I)=0
        PZKW2(I)=0
        PZCAB(I)=0
        PZEPEC(I)=0
20063 CONTINUE

C  Construct weight matrix

      DO 10110 I=1,K
        DO 10110 J=1,K
          W(I,J,1)=1.0D0-(((I-J)**2)/(((K-1)**2)*1.0D0))
          W(I,J,2)=1.0D0-((ABS(I-J))/(K-1.0D0))
10110 CONTINUE

C  True Proportion

      TP=C/(N*1.0D0)

C  Simulation loop commences

      DO 10050 A=1,SIMS
        WRITE(*,10051) A,SIMS,D,K,N,C
10051   FORMAT(' ',I6,'  OF  ',I6,'      D=',I1,'  K=',I1,'  N=',I2,'  C
     *=',I2)
        FLAG=0

C  Construct subjects'-matrix with guessed observations

        NRAN=N-C
        WHILE (NRAN.GT.0) DO
          IX1=1+INT(K*RAN2(IDUM))
          IX2=1+INT(K*RAN2(IDUM))
          IX(NRAN,1)=IX1
          IX(NRAN,2)=IX2
          NRAN=NRAN-1
        END WHILE

C  Fill in balance of subjects'-matrix with knowledge based agreements

        IF (D.EQ.1) THEN
          JCNT=1
          DO 10070 I=N-C+1,N
            IX(I,1)=JCNT
            IX(I,2)=JCNT
            IF (JCNT.GE.K) JCNT=0
            JCNT=JCNT+1
10070     CONTINUE
        ELSE
          DO 10071 I=N-C+1,N
            IX(I,1)=1
            IX(I,2)=1
10071     CONTINUE
        END IF

C  Construct observations-matrix and marginal totals

        DO 10095 I=1,K+1
          DO 10095 J=1,K+1
            IC(I,J)=0
10095   CONTINUE
        DO 10100 I=1,N
          IC(IX(I,1),IX(I,2))=IC(IX(I,1),IX(I,2))+1
10100   CONTINUE
        DO 10105 I=1,K
          DO 10105 J=1,K
            IC(I,K+1)=IC(I,K+1)+IC(I,J)
            IC(K+1,I)=IC(K+1,I)+IC(J,I)
10105   CONTINUE

C  Total number of agreements observed and chance-expected

        TO=0
        TC=0.0D0
        DO 410 I=1,K
          TO=TO+IC(I,I)
          TC=TC+(IC(I,K+1)*IC(K+1,I))/(N*1.0D0)
  410   CONTINUE

C  Check for degenerate sample

        IF (TC.EQ.N*1.0D0) THEN
          WRITE (1,55001)
55001     FORMAT (/, 'DEGENERATE SAMPLE SINCE TC.EQ.N*1.0D0')
          EXECUTE DEGVAL
          GO TO 10054
        END IF

C  Exact null variance of TO and TOW

        SUMV=0
        SUMC=0
        DO 411 I=1,K
          SUMV=SUMV+IC(I,K+1)*(N-IC(I,K+1))*IC(K+1,I)*(N-IC(K+1,I))
          DO 411 J=I+1,K
            SUMC=SUMC+IC(I,K+1)*IC(J,K+1)*IC(K+1,I)*IC(K+1,J)
  411   CONTINUE
        V0TO=(SUMV+2*SUMC)/((N**2)*(N-1.0D0))

C  Check for degenerate sample

        IF (V0TO.EQ.0.0D0) THEN
          WRITE (1,55002)
55002     FORMAT (/, 'DEGENERATE SAMPLE SINCE V0TO.EQ.0.0D0')
          EXECUTE DEGVAL
          GO TO 10054
        END IF

        SUMVW1=0.0D0
        SUMVW2=0.0D0
        SUMCW1=0.0D0
        SUMCW2=0.0D0
        DO 412 I=1,K
          DO 412 J=1,K
            SUMVW1=SUMVW1+(W(I,J,1)**2)*IC(I,K+1)*(N-IC(I,K+1))*IC(K+1,J
     *)*(N-IC(K+1,J))
            SUMVW2=SUMVW2+(W(I,J,2)**2)*IC(I,K+1)*(N-IC(I,K+1))*IC(K+1,J
     *)*(N-IC(K+1,J))
            S=I
            T=J
            ENDCEL=0
            WHILE (ENDCEL.EQ.0) DO
              IF (T+1.LE.K) THEN
                T=T+1
              ELSE IF (S+1.LE.K) THEN
                S=S+1
                T=1
              ELSE
                ENDCEL=1
              END IF
              IF (ENDCEL.EQ.0) THEN
                IF (I.EQ.S.AND.J.NE.T) THEN
                  SUMCW1=SUMCW1-W(I,J,1)*W(S,T,1)*IC(I,K+1)*IC(K+1,J)*IC
     *(K+1,T)*(N-IC(I,K+1))
                  SUMCW2=SUMCW2-W(I,J,2)*W(S,T,2)*IC(I,K+1)*IC(K+1,J)*IC
     *(K+1,T)*(N-IC(I,K+1))
                ELSE IF (J.EQ.T.AND.I.NE.S) THEN
                  SUMCW1=SUMCW1-W(I,J,1)*W(S,T,1)*IC(I,K+1)*IC(S,K+1)*IC
     *(K+1,J)*(N-IC(K+1,J))
                  SUMCW2=SUMCW2-W(I,J,2)*W(S,T,2)*IC(I,K+1)*IC(S,K+1)*IC
     *(K+1,J)*(N-IC(K+1,J))
                ELSE
                  SUMCW1=SUMCW1+W(I,J,1)*W(S,T,1)*IC(I,K+1)*IC(S,K+1)*IC
     *(K+1,J)*IC(K+1,T)
                  SUMCW2=SUMCW2+W(I,J,2)*W(S,T,2)*IC(I,K+1)*IC(S,K+1)*IC
     *(K+1,J)*IC(K+1,T)
                END IF
              END IF
            END WHILE
412     CONTINUE
        V0TOW1=(SUMVW1+2*SUMCW1)/((N**2)*(N-1))

C  Check for degenerate sample

        IF (V0TOW1.EQ.0.0D0) THEN
          WRITE (1,55003)
55003     FORMAT (/, 'DEGENERATE SAMPLE SINCE V0TOW1.EQ.0.0D0')
          EXECUTE DEGVAL
          GO TO 10054
        END IF

        V0TOW2=(SUMVW2+2*SUMCW2)/((N**2)*(N-1))

C  Check for degenerate sample

        IF (V0TOW2.EQ.0.0D0) THEN
          WRITE (1,55004)
55004     FORMAT (/, 'DEGENERATE SAMPLE SINCE V0TOW2.EQ.0.0D0')
          EXECUTE DEGVAL
          GO TO 10054
        END IF

C  Large-sample est. of the null and nonnull variance of Kappa

        VKC=0.0D0
        SUM1=0
        DO 431 I=1,K
          VKC=VKC+IC(I,I)*((N*(N-TC)-(IC(K+1,I)+IC(I,K+1))*(N-TO))**2)
          SUM1=SUM1+IC(I,K+1)*IC(K+1,I)*((N-IC(K+1,I)-IC(I,K+1))**2)
          DO 431 J=1,K
            IF (I.NE.J) THEN
              VKC=VKC+((N-TO)**2)*IC(I,J)*((IC(K+1,I)+IC(J,K+1))**2)
              SUM1=SUM1+IC(I,K+1)*IC(K+1,J)*((IC(K+1,I)+IC(J,K+1))**2)
            END IF
431     CONTINUE
        V0KC=(SUM1-(N*TC)**2)/((N**3)*((N-TC)**2))
        VKC=(VKC-N*((TO*TC+N*(TO-2*TC))**2))/((N**2)*((N-TC)**4))

C  Check for degenerate sample

        IF (V0KC.EQ.0.0D0) THEN
          WRITE (1,55005)
55005     FORMAT (/, 'DEGENERATE SAMPLE SINCE V0KC.EQ.0.0D0')
          EXECUTE DEGVAL
          GO TO 10054
        END IF

C  Construct weight matrix marginals (weighted averages of the weights)

        DO 10115 I=1,K
          W(I,K+1,1)=0.0D0
          W(I,K+1,2)=0.0D0
          W(K+1,I,1)=0.0D0
          W(K+1,I,2)=0.0D0
10115   CONTINUE
```

```
        DO 10120 I=1,K
          DO 10120 J=1,K
            W(I,K+1,1)=W(I,K+1,1)+(W(I,J,1)*IC(K+1,J))/(N*1.0D0)
            W(I,K+1,2)=W(I,K+1,2)+(W(I,J,2)*IC(K+1,J))/(N*1.0D0)
            W(K+1,I,1)=W(K+1,I,1)+(W(J,I,1)*IC(J,K+1))/(N*1.0D0)
            W(K+1,I,2)=W(K+1,I,2)+(W(J,I,2)*IC(J,K+1))/(N*1.0D0)
10120   CONTINUE

C  Weighted total number of agreements observed and chance-expected

        TOW1=0.0D0
        TOW2=0.0D0
        TCW1=0.0D0
        TCW2=0.0D0
        DO 600 I=1,K
          DO 600 J=1,K
            TOW1=TOW1+W(I,J,1)*IC(I,J)
            TOW2=TOW2+W(I,J,2)*IC(I,J)
            TCW1=TCW1+(W(I,J,1)*IC(I,K+1)*IC(K+1,J))/N
            TCW2=TCW2+(W(I,J,2)*IC(I,K+1)*IC(K+1,J))/N
  600   CONTINUE

C  Large-sample est. of the null and nonnull variance of Weighted Kappa

        VKW1=0.0D0
        VKW2=0.0D0
        V0KW1=0.0D0
        V0KW2=0.0D0
        DO 704 I=1,K
          DO 704 J=1,K
            VKW1=VKW1+IC(I,J)*((W(I,J,1)*(N-TCW1)-(W(I,K+1,1)+W(K+1,J,1)
     *)*(N-TOW1))**2)
            VKW2=VKW2+IC(I,J)*((W(I,J,2)*(N-TCW2)-(W(I,K+1,2)+W(K+1,J,2)
     *)*(N-TOW2))**2)
            V0KW1=V0KW1+IC(I,K+1)*IC(K+1,J)*((W(I,J,1)-W(I,K+1,1)-W(K+1,
     *J,1))**2)
            V0KW2=V0KW2+IC(I,K+1)*IC(K+1,J)*((W(I,J,2)-W(I,K+1,2)-W(K+1,
     *J,2))**2)
704     CONTINUE
        VKW1=(VKW1-((TOW1*TCW1+N*(TOW1-2*TCW1))**2)/N)/((N-TCW1)**4)
        VKW2=(VKW2-((TOW2*TCW2+N*(TOW2-2*TCW2))**2)/N)/((N-TCW2)**4)
        V0KW1=(V0KW1-TCW1**2)/(N*((N-TCW1)**2))

C  Check for degenerate sample

        IF (V0KW1.EQ.0.0D0) THEN
          WRITE (1,55006)
55006     FORMAT (/, 'DEGENERATE SAMPLE SINCE V0KW1.EQ.0.0D0')
          EXECUTE DEGVAL
          GO TO 10054
        END IF

        V0KW2=(V0KW2-TCW2**2)/(N*((N-TCW2)**2))

C  Check for degenerate sample

        IF (V0KW2.EQ.0.0D0) THEN
```

```
          WRITE (1,55007)
55007     FORMAT (/, 'DEGENERATE SAMPLE SINCE V0KW2.EQ.0.0D0')
          EXECUTE DEGVAL
          GO TO 10054
        END IF

C  Exact null variance of Kappa and Weighted Kappa

        EVKC=V0TO/((N-TC)**2)
        EVKW1=V0TOW1/((N-TCW1)**2)
        EVKW2=V0TOW2/((N-TCW2)**2)

C  Overall Proportion of Agreement

        PA=TO/(N*1.0D0)

C  Kappa and Weighted Kappa

        KC=(TO-TC)/(N-TC)
        ZEKC=KC/DSQRT(EVKC)
        ZKC=KC/DSQRT(V0KC)
        KW1=(TOW1-TCW1)/(N-TCW1)
        ZEKW1=KW1/DSQRT(EVKW1)
        ZKW1=KW1/DSQRT(V0KW1)
        KW2=(TOW2-TCW2)/(N-TCW2)
        ZEKW2=KW2/DSQRT(EVKW2)
        ZKW2=KW2/DSQRT(V0KW2)

C  The Disagreement Rate

        NUM=0
        DEN=0.0D0
        DO 500 I=1,K
          DO 500 J=1,K
            NUM=NUM+IC(I,J)*(ABS(I-J))
            DEN=DEN+IC(I,J)*MAX(((I+J)/2.0D0)-1,K-((I+J)/2.0D0))
  500   CONTINUE
        DR=NUM/(2*DEN)

C  The Concordance Between Raters

        CAB=(K*PA-1)/(K-1)
        VCAB=((K**2)*PA*(1-PA))/(((K-1)**2)*N)
        V0CAB=1.0D0/(N*(K-1))
        ZCAB=CAB/DSQRT(V0CAB)

C  Partial-Chance Kappa (Equal Weights; p=1/K), Partial-Chance Proportion

        GSTAR=INT(((N-TO)*K)/(K-1.0D0))
        IF (GSTAR.GT.N) GSTAR=N
        KPC=(N-GSTAR)/(2.0D0*N-GSTAR-TO)
        PPC=(N-GSTAR)/(N*1.0D0)

C  Expected-Chance Proportion

        PEC=CAB-(1.0D0/(N*(K-1)))
        EVPEC=((K**2)*V0TO)/((N*(K-1))**2)
        ZEPEC=PEC/DSQRT(EVPEC)

C  Output simulated observations-matrix and measures per simulation
C
C  Note: This matrix has been formatted for 2<=K<=5, and may not output
C        satisfactorily when K is larger.

10054   CONTINUE
        IF (OUTMAT.EQ.1) THEN
          WRITE(1,10055) A,SIMS,D,K,N,C
10055     FORMAT(/,'SIMULATION    SIMS      D    K    N     C',/,4X,I6,2
     *X,I6,6X,I1,4X,I1,3X,I2,4X,I2)
          WRITE(1,10121)
10121     FORMAT(/,30X,'Rater B',/,'Rater A')
          SELECT (K) FROM
          CASE 2
            WRITE(1,10122) (I,I=1,K),'Marginal'
10122       FORMAT(/,15X,2(6X,I1),3X,A8)
          CASE 3
            WRITE(1,10123) (I,I=1,K),'Marginal'
10123       FORMAT(/,15X,3(6X,I1),3X,A8)
          CASE 4
            WRITE(1,10124) (I,I=1,K),'Marginal'
10124       FORMAT(/,15X,4(6X,I1),3X,A8)
          CASE 5
            WRITE(1,10125) (I,I=1,K),'Marginal'
10125       FORMAT(/,15X,5(6X,I1),3X,A8)
          END SELECT
          DO 10135 I=1,K
            WRITE(1,10140) I,(IC(I,J),J=1,K+1)
            WRITE(1,10141) ((IC(I,J)*1.0D0)/N,J=1,K+1)
            WRITE(1,10142) (W(I,J,1),J=1,K+1)
            WRITE(1,10143) (W(I,J,2),J=1,K+1)
10140       FORMAT(/,'  ',I2,4X,'Cij',10X,7(I3,4X))
10141       FORMAT(' ',7X,'Pij',7X,7(F6.4,1X))
10142       FORMAT(' ',7X,'Wij(1)',4X,7(F6.4,1X))
10143       FORMAT(' ',7X,'Wij(2)',4X,7(F6.4,1X))
10135     CONTINUE
          WRITE(1,10144)
          WRITE(1,10145) (IC(K+1,J),J=1,K),N
          WRITE(1,10146) ((IC(K+1,J)*1.0D0)/N,J=1,K)
          WRITE(1,10150) (W(K+1,J,1),J=1,K)
          WRITE(1,10155) (W(K+1,J,2),J=1,K)
10144     FORMAT(/,3X,'Marginal')
10145     FORMAT(' ',7X,'C.j',10X,7(I3,4X))
10146     FORMAT(' ',7X,'P.j',7X,6(F6.4,1X))
10150     FORMAT(' ',7X,'W.jBAR(1)',1X,6(F6.4,1X))
10155     FORMAT(' ',7X,'W.jBAR(2)',1X,6(F6.4,1X))
          IF (FLAG.EQ.1) THEN
            WRITE(1,10064)
10064       FORMAT(/,1X,'***** DEGENERATE SAMPLE *****')
          ELSE
            WRITE(1,10065) TP,PA
10065       FORMAT(//,1X,'True Proportion        ',7X,F7.4,/,' Overall P
     *rop. of Agreement',4X,F7.4)
            WRITE(1,420) KC,EVKC,V0KC,VKC
420         FORMAT(1X,'Kappa',25X,F7.4,3X,'Exact Null Var[Kc]',11X,F7.4,
     */,' L.S.Est. Null Var[Kc]',9X,F7.4,3X,'L.S.Est. Var[Kc]',13X,F7.4)
            WRITE(1,421) KW1,EVKW1,V0KW1,VKW1
421         FORMAT(1X,'Wtd.(1) Kappa',17X,F7.4,3X,'Exact Null Var[Kw1]',
     *10X,F7.4,/,' L.S.Est. Null Var[Kw1]',8X,F7.4,3X,'L.S.Est. Var[Kw1]
     *',12X,F7.4)
            WRITE(1,422) KW2,EVKW2,V0KW2,VKW2
422         FORMAT(1X,'Wtd.(2) Kappa',17X,F7.4,3X,'Exact Null Var[Kw2]',
     *10X,F7.4,/,' L.S.Est. Null Var[Kw2]',8X,F7.4,3X,'L.S.Est. Var[Kw2]
     *',12X,F7.4)
            WRITE(1,501) DR,CAB,V0CAB,VCAB,KPC,PPC,PEC,EVPEC
501         FORMAT(1X,'Disagreement Rate',13X,F7.4,/,' Concordance Betwe
     *en Raters',4X,F7.4,/,' L.S. Null Var[CAB]',12X,F7.4,3X,'L.S.Est.Va
     *r[CAB]',13X,F7.4,/,' Partial-Chance Kappa (p=1/K)',3X,F7.4,/,' Par
     *tial-Chance Prop',12X,F7.4,/,' Expected-Chance Proportion',1X,
     *F7.4,3X,'Exact Null Var[PEC]',10X,F7.4)
          END IF
        END IF

C  Aggregate measures and associated variables

        SPA=SPA+PA
        SPA2=SPA2+PA**2
        SKC=SKC+KC
        SKC2=SKC2+KC**2
        SVKC=SVKC+VKC
        SKW1=SKW1+KW1
        SKW12=SKW12+KW1**2
        SVKW1=SVKW1+VKW1
        SKW2=SKW2+KW2
        SKW22=SKW22+KW2**2
        SVKW2=SVKW2+VKW2
        SDR=SDR+DR
        SDR2=SDR2+DR**2
        SCAB=SCAB+CAB
        SCAB2=SCAB2+CAB**2
        SVCAB=SVCAB+VCAB
        SKPC=SKPC+KPC
        SKPC2=SKPC2+KPC**2
        SPPC=SPPC+PPC
        SPPC2=SPPC2+PPC**2
        SPEC=SPEC+PEC
        SPEC2=SPEC2+PEC**2
        DO 502 I=1,4
          SZEKC(I)=SZEKC(I)+ZEKC**I
          SZKC(I)=SZKC(I)+ZKC**I
          SZEKW1(I)=SZEKW1(I)+ZEKW1**I
          SZKW1(I)=SZKW1(I)+ZKW1**I
          SZEKW2(I)=SZEKW2(I)+ZEKW2**I
          SZKW2(I)=SZKW2(I)+ZKW2**I
          SZCAB(I)=SZCAB(I)+ZCAB**I
          SZEPEC(I)=SZEPEC(I)+ZEPEC**I
502     CONTINUE
        IF (ZEKC.LE.-1.96) THEN
          PZEKC(2)=PZEKC(2)+1
          IF (ZEKC.LE.-2.576) THEN
            PZEKC(1)=PZEKC(1)+1
          END IF
```

```
ELSE IF (ZEKC.GE.1.96) THEN
  PZEKC(3)=PZEKC(3)+1
  IF (ZEKC.GE.2.576) THEN
    PZEKC(4)=PZEKC(4)+1
  END IF
END IF
IF (ZKC.LE.-1.96) THEN
  PZKC(2)=PZKC(2)+1
  IF (ZKC.LE.-2.576) THEN
    PZKC(1)=PZKC(1)+1
  END IF
ELSE IF (ZKC.GE.1.96) THEN
  PZKC(3)=PZKC(3)+1
  IF (ZKC.GE.2.576) THEN
    PZKC(4)=PZKC(4)+1
  END IF
END IF
IF (ZEKW1.LE.-1.96) THEN
  PZEKW1(2)=PZEKW1(2)+1
  IF (ZEKW1.LE.-2.576) THEN
    PZEKW1(1)=PZEKW1(1)+1
  END IF
ELSE IF (ZEKW1.GE.1.96) THEN
  PZEKW1(3)=PZEKW1(3)+1
  IF (ZEKW1.GE.2.576) THEN
    PZEKW1(4)=PZEKW1(4)+1
  END IF
END IF
IF (ZKW1.LE.-1.96) THEN
  PZKW1(2)=PZKW1(2)+1
  IF (ZKW1.LE.-2.576) THEN
    PZKW1(1)=PZKW1(1)+1
  END IF
ELSE IF (ZKW1.GE.1.96) THEN
  PZKW1(3)=PZKW1(3)+1
  IF (ZKW1.GE.2.576) THEN
    PZKW1(4)=PZKW1(4)+1
  END IF
END IF
IF (ZEKW2.LE.-1.96) THEN
  PZEKW2(2)=PZEKW2(2)+1
  IF (ZEKW2.LE.-2.576) THEN
    PZEKW2(1)=PZEKW2(1)+1
  END IF
ELSE IF (ZEKW2.GE.1.96) THEN
  PZEKW2(3)=PZEKW2(3)+1
  IF (ZEKW2.GE.2.576) THEN
    PZEKW2(4)=PZEKW2(4)+1
  END IF
END IF
IF (ZKW2.LE.-1.96) THEN
  PZKW2(2)=PZKW2(2)+1
  IF (ZKW2.LE.-2.576) THEN
    PZKW2(1)=PZKW2(1)+1
  END IF
ELSE IF (ZKW2.GE.1.96) THEN
  PZKW2(3)=PZKW2(3)+1
```

```
          IF (ZKW2.GE.2.576) THEN
            PZKW2(4)=PZKW2(4)+1
          END IF
        END IF
        IF (ZCAB.LE.-1.96) THEN
          PZCAB(2)=PZCAB(2)+1
          IF (ZCAB.LE.-2.576) THEN
            PZCAB(1)=PZCAB(1)+1
          END IF
        ELSE IF (ZCAB.GE.1.96) THEN
          PZCAB(3)=PZCAB(3)+1
          IF (ZCAB.GE.2.576) THEN
            PZCAB(4)=PZCAB(4)+1
          END IF
        END IF
        IF (ZEPEC.LE.-1.96) THEN
          PZEPEC(2)=PZEPEC(2)+1
          IF (ZEPEC.LE.-2.576) THEN
            PZEPEC(1)=PZEPEC(1)+1
          END IF
        ELSE IF (ZEPEC.GE.1.96) THEN
          PZEPEC(3)=PZEPEC(3)+1
          IF (ZEPEC.GE.2.576) THEN
            PZEPEC(4)=PZEPEC(4)+1
          END IF
        END IF

C  Simulation loop terminates

10050 CONTINUE

C  Results of all simulations per parameter combination
C
C  Note: Results are based on the non-degenerate samples, SIMS2. A value
C        of -9.9999 will result when SIMS2 prevents the calculation from
C        being performed. For the 3rd/4th moment of selected critical
C        ratios, a value of -8.8888 will result in order to prevent the
C        square root of a negative value from being performed, this
C        negative value arising because of rounding error in DOUBLE
C        PRECISION arithmetic.

      SIMS2=SIMS-DEG

C  Mean value of measures
C  Mean value of large-sample nonnull variance estimates
C  First moment of selected critical ratios

      IF (SIMS2.GT.0) THEN
        PA=SPA/SIMS2
        KC=SKC/SIMS2
        VKC=SVKC/SIMS2
        BZEKC(1)=SZEKC(1)/SIMS2
        BZKC(1)=SZKC(1)/SIMS2
        KW1=SKW1/SIMS2
        VKW1=SVKW1/SIMS2
        BZEKW1(1)=SZEKW1(1)/SIMS2
        BZKW1(1)=SZKW1(1)/SIMS2
        KW2=SKW2/SIMS2
        VKW2=SVKW2/SIMS2
        BZEKW2(1)=SZEKW2(1)/SIMS2
        BZKW2(1)=SZKW2(1)/SIMS2
        DR=SDR/SIMS2
        CAB=SCAB/SIMS2
        VCAB=SVCAB/SIMS2
        BZCAB(1)=SZCAB(1)/SIMS2
        KPC=SKPC/SIMS2
        PPC=SPPC/SIMS2
        PEC=SPEC/SIMS2
        BZEPEC(1)=SZEPEC(1)/SIMS2

C  Empirical variance of measures
C  Second moment of selected critical ratios

        IF (SIMS2.GT.1) THEN
          PASV=(SPA2-SIMS2*(PA**2))/(SIMS2-1)
          KCSV=(SKC2-SIMS2*(KC**2))/(SIMS2-1)
          BZEKC(2)=(SZEKC(2)-SIMS2*(BZEKC(1)**2))/(SIMS2-1)
          BZKC(2)=(SZKC(2)-SIMS2*(BZKC(1)**2))/(SIMS2-1)
          KW1SV=(SKW12-SIMS2*(KW1**2))/(SIMS2-1)
          BZEKW1(2)=(SZEKW1(2)-SIMS2*(BZEKW1(1)**2))/(SIMS2-1)
          BZKW1(2)=(SZKW1(2)-SIMS2*(BZKW1(1)**2))/(SIMS2-1)
          KW2SV=(SKW22-SIMS2*(KW2**2))/(SIMS2-1)
          BZEKW2(2)=(SZEKW2(2)-SIMS2*(BZEKW2(1)**2))/(SIMS2-1)
          BZKW2(2)=(SZKW2(2)-SIMS2*(BZKW2(1)**2))/(SIMS2-1)
          DRSV=(SDR2-SIMS2*(DR**2))/(SIMS2-1)
          CABSV=(SCAB2-SIMS2*(CAB**2))/(SIMS2-1)
          BZCAB(2)=(SZCAB(2)-SIMS2*(BZCAB(1)**2))/(SIMS2-1)
          KPCSV=(SKPC2-SIMS2*(KPC**2))/(SIMS2-1)
          PPCSV=(SPPC2-SIMS2*(PPC**2))/(SIMS2-1)
          PECSV=(SPEC2-SIMS2*(PEC**2))/(SIMS2-1)
          BZEPEC(2)=(SZEPEC(2)-SIMS2*(BZEPEC(1)**2))/(SIMS2-1)

C  Third moment of selected critical ratios

          IF (SIMS2.GT.2) THEN

            IF (BZEKC(2).GT.0.0D0) THEN
              BZEKC(3)=((SZEKC(3)-3*BZEKC(1)*SZEKC(2)+3*(BZEKC(1)**2)*
     *SZEKC(1)-SIMS2*(BZEKC(1)**3))*SIMS2)/((SIMS2-1)*(SIMS2-2)*
     *((DSQRT(BZEKC(2)))**3))
            ELSE
              BZEKC(3)=-8.8888
            END IF
            IF (BZKC(2).GT.0.0D0) THEN
              BZKC(3)=((SZKC(3)-3*BZKC(1)*SZKC(2)+3*(BZKC(1)**2)*
     *SZKC(1)-SIMS2*(BZKC(1)**3))*SIMS2)/((SIMS2-1)*(SIMS2-2)*
     *((DSQRT(BZKC(2)))**3))
            ELSE
              BZKC(3)=-8.8888
            END IF
            IF (BZEKW1(2).GT.0.0D0) THEN
              BZEKW1(3)=((SZEKW1(3)-3*BZEKW1(1)*SZEKW1(2)+3*
     *(BZEKW1(1)**2)*SZEKW1(1)-SIMS2*(BZEKW1(1)**3))*SIMS2)/((SIMS2-1)*
     *(SIMS2-2)*((DSQRT(BZEKW1(2)))**3))
```

```
            ELSE
              BZEKW1(3)=-8.8888
            END IF
            IF (BZKW1(2).GT.0.0D0) THEN
              BZKW1(3)=((SZKW1(3)-3*BZKW1(1)*SZKW1(2)+3*(BZKW1(1)**2)*
     *SZKW1(1)-SIMS2*(BZKW1(1)**3))*SIMS2)/((SIMS2-1)*(SIMS2-2)*
     *((DSQRT(BZKW1(2)))**3))
            ELSE
              BZKW1(3)=-8.8888
            END IF
            IF (BZEKW2(2).GT.0.0D0) THEN
              BZEKW2(3)=((SZEKW2(3)-3*BZEKW2(1)*SZEKW2(2)+3*
     *(BZEKW2(1)**2)*SZEKW2(1)-SIMS2*(BZEKW2(1)**3))*SIMS2)/((SIMS2-1)*
     *(SIMS2-2)*((DSQRT(BZEKW2(2)))**3))
            ELSE
              BZEKW2(3)=-8.8888
            END IF
            IF (BZKW2(2).GT.0.0D0) THEN
              BZKW2(3)=((SZKW2(3)-3*BZKW2(1)*SZKW2(2)+3*(BZKW2(1)**2)*
     *SZKW2(1)-SIMS2*(BZKW2(1)**3))*SIMS2)/((SIMS2-1)*(SIMS2-2)*
     *((DSQRT(BZKW2(2)))**3))
            ELSE
              BZKW2(3)=-8.8888
            END IF
            IF (BZCAB(2).GT.0.0D0) THEN
              BZCAB(3)=((SZCAB(3)-3*BZCAB(1)*SZCAB(2)+3*(BZCAB(1)**2)*
     *SZCAB(1)-SIMS2*(BZCAB(1)**3))*SIMS2)/((SIMS2-1)*(SIMS2-2)*
     *((DSQRT(BZCAB(2)))**3))
            ELSE
              BZCAB(3)=-8.8888
            END IF
            IF (BZEPEC(2).GT.0.0D0) THEN
              BZEPEC(3)=((SZEPEC(3)-3*BZEPEC(1)*SZEPEC(2)+3*(BZEPEC(1)**
     *2)*SZEPEC(1)-SIMS2*(BZEPEC(1)**3))*SIMS2)/((SIMS2-1)*(SIMS2-2)*
     *((DSQRT(BZEPEC(2)))**3))
            ELSE
              BZEPEC(3)=-8.8888
            END IF

C  Fourth moment of selected critical ratios

            IF (SIMS2.GT.3) THEN

              IF (BZEKC(2).GT.0.0D0) THEN
                BZEKC(4)=(((SZEKC(4)-4*BZEKC(1)*SZEKC(3)+6*(BZEKC(1)**2)
     **SZEKC(2)-4*(BZEKC(1)**3)*SZEKC(1)+SIMS2*(BZEKC(1)**4))*SIMS2*
     *(SIMS2+1))/((SIMS2-1)*(SIMS2-2)*(SIMS2-3)*((DSQRT(BZEKC(2)))**4)))
     *-((3.0D0*((SIMS2-1)**2))/((SIMS2-2)*(SIMS2-3)))
              ELSE
                BZEKC(4)=-8.8888
              END IF
              IF (BZKC(2).GT.0.0D0) THEN
                BZKC(4)=(((SZKC(4)-4*BZKC(1)*SZKC(3)+6*(BZKC(1)**2)*
     *SZKC(2)-4*(BZKC(1)**3)*SZKC(1)+SIMS2*(BZKC(1)**4))*SIMS2*
     *(SIMS2+1))/((SIMS2-1)*(SIMS2-2)*(SIMS2-3)*((DSQRT(BZKC(2)))**4)))
     *-((3.0D0*((SIMS2-1)**2))/((SIMS2-2)*(SIMS2-3)))
              ELSE
```

```
                BZKC(4)=-8.8888
              END IF
              IF (BZEKW1(2).GT.0.0D0) THEN
                BZEKW1(4)=(((SZEKW1(4)-4*BZEKW1(1)*SZEKW1(3)+6*
     *(BZEKW1(1)**2)*SZEKW1(2)-4*(BZEKW1(1)**3)*SZEKW1(1)+SIMS2*
     *(BZEKW1(1)**4))*SIMS2*(SIMS2+1))/((SIMS2-1)*(SIMS2-2)*(SIMS2-3)*
     *((DSQRT(BZEKW1(2)))**4)))-((3.0D0*((SIMS2-1)**2))/((SIMS2-2)*
     *(SIMS2-3)))
              ELSE
                BZEKW1(4)=-8.8888
              END IF
              IF (BZKW1(2).GT.0.0D0) THEN
                BZKW1(4)=(((SZKW1(4)-4*BZKW1(1)*SZKW1(3)+6*(BZKW1(1)**2)
     **SZKW1(2)-4*(BZKW1(1)**3)*SZKW1(1)+SIMS2*(BZKW1(1)**4))*SIMS2*
     *(SIMS2+1))/((SIMS2-1)*(SIMS2-2)*(SIMS2-3)*((DSQRT(BZKW1(2)))**4)))
     *-((3.0D0*((SIMS2-1)**2))/((SIMS2-2)*(SIMS2-3)))
              ELSE
                BZKW1(4)=-8.8888
              END IF
              IF (BZEKW2(2).GT.0.0D0) THEN
                BZEKW2(4)=(((SZEKW2(4)-4*BZEKW2(1)*SZEKW2(3)+6*
     *(BZEKW2(1)**2)*SZEKW2(2)-4*(BZEKW2(1)**3)*SZEKW2(1)+SIMS2*
     *(BZEKW2(1)**4))*SIMS2*(SIMS2+1))/((SIMS2-1)*(SIMS2-2)*(SIMS2-3)*
     *((DSQRT(BZEKW2(2)))**4)))-((3.0D0*((SIMS2-1)**2))/((SIMS2-2)*
     *(SIMS2-3)))
              ELSE
                BZEKW2(4)=-8.8888
              END IF
              IF (BZKW2(2).GT.0.0D0) THEN
                BZKW2(4)=(((SZKW2(4)-4*BZKW2(1)*SZKW2(3)+6*(BZKW2(1)**2)
     **SZKW2(2)-4*(BZKW2(1)**3)*SZKW2(1)+SIMS2*(BZKW2(1)**4))*SIMS2*
     *(SIMS2+1))/((SIMS2-1)*(SIMS2-2)*(SIMS2-3)*((DSQRT(BZKW2(2)))**4)))
     *-((3.0D0*((SIMS2-1)**2))/((SIMS2-2)*(SIMS2-3)))
              ELSE
                BZKW2(4)=-8.8888
              END IF
              IF (BZCAB(2).GT.0.0D0) THEN
                BZCAB(4)=(((SZCAB(4)-4*BZCAB(1)*SZCAB(3)+6*(BZCAB(1)**2)
     **SZCAB(2)-4*(BZCAB(1)**3)*SZCAB(1)+SIMS2*(BZCAB(1)**4))*SIMS2*
     *(SIMS2+1))/((SIMS2-1)*(SIMS2-2)*(SIMS2-3)*((DSQRT(BZCAB(2)))**4)))
     *-((3.0D0*((SIMS2-1)**2))/((SIMS2-2)*(SIMS2-3)))
              ELSE
                BZCAB(4)=-8.8888
              END IF
              IF (BZEPEC(2).GT.0.0D0) THEN
                BZEPEC(4)=(((SZEPEC(4)-4*BZEPEC(1)*SZEPEC(3)+6*(BZEPEC(1
     *)**2)*SZEPEC(2)-4*(BZEPEC(1)**3)*SZEPEC(1)+SIMS2*(BZEPEC(1)**4))*
     *SIMS2*(SIMS2+1))/((SIMS2-1)*(SIMS2-2)*(SIMS2-3)*((DSQRT(BZEPEC(2))
     *)**4)))-((3.0D0*((SIMS2-1)**2))/((SIMS2-2)*(SIMS2-3)))
              ELSE
                BZEPEC(4)=-8.8888
              END IF

            ELSE
              BZEKC(4)=-9.9999
              BZKC(4)=-9.9999
              BZEKW1(4)=-9.9999
              BZKW1(4)=-9.9999
              BZEKW2(4)=-9.9999
              BZKW2(4)=-9.9999
              BZCAB(4)=-9.9999
              BZEPEC(4)=-9.9999
            END IF
          ELSE
            DO 503 I=3,4
              BZEKC(I)=-9.9999
              BZKC(I)=-9.9999
              BZEKW1(I)=-9.9999
              BZKW1(I)=-9.9999
              BZEKW2(I)=-9.9999
              BZKW2(I)=-9.9999
              BZCAB(I)=-9.9999
              BZEPEC(I)=-9.9999
503         CONTINUE
          END IF
        ELSE
          PASV=-9.9999
          KCSV=-9.9999
          KW1SV=-9.9999
          KW2SV=-9.9999
          DRSV=-9.9999
          CABSV=-9.9999
          KPCSV=-9.9999
          PPCSV=-9.9999
          PECSV=-9.9999
          DO 504 I=2,4
            BZEKC(I)=-9.9999
            BZKC(I)=-9.9999
            BZEKW1(I)=-9.9999
            BZKW1(I)=-9.9999
            BZEKW2(I)=-9.9999
            BZKW2(I)=-9.9999
            BZCAB(I)=-9.9999
            BZEPEC(I)=-9.9999
504       CONTINUE
        END IF
      ELSE
        PA=-9.9999
        KC=-9.9999
        VKC=-9.9999
        KW1=-9.9999
        VKW1=-9.9999
        KW2=-9.9999
        VKW2=-9.9999
        DR=-9.9999
        CAB=-9.9999
        VCAB=-9.9999
        KPC=-9.9999
        PPC=-9.9999
        PEC=-9.9999
        PASV=-9.9999
        KCSV=-9.9999
        KW1SV=-9.9999
        KW2SV=-9.9999
        DRSV=-9.9999
        CABSV=-9.9999
        KPCSV=-9.9999
        PPCSV=-9.9999
        PECSV=-9.9999
        DO 505 I=1,4
          BZEKC(I)=-9.9999
          BZKC(I)=-9.9999
          BZEKW1(I)=-9.9999
          BZKW1(I)=-9.9999
          BZEKW2(I)=-9.9999
          BZKW2(I)=-9.9999
          BZCAB(I)=-9.9999
          BZEPEC(I)=-9.9999
505     CONTINUE
      END IF

C  Output results of all simulations per parameter combination

      WRITE(2,3000) C,TP,PA,KC,KW1,KW2,DR
3000  FORMAT(' ',I2,6('_',F8.4))
      WRITE(3,3002) C,TP,CAB,KPC,PPC,PEC
3002  FORMAT(' ',I2,5('_',F8.4))
      WRITE(4,3004) C,PASV,KCSV,VKC,KW1SV,VKW1,KW2SV,VKW2
3004  FORMAT(' ',I2,7('_',F8.4))
      WRITE(7,3006) C,DRSV,CABSV,VCAB,KPCSV,PPCSV,PECSV
3006  FORMAT(' ',I2,6('_',F8.4))
      DO 3008 I=1,4
        WRITE(I+7,3009) I,C,BZEKC(I),BZKC(I),BZEKW1(I),BZKW1(I),
     *BZEKW2(I),BZKW2(I),BZCAB(I),BZEPEC(I)
3009    FORMAT(' ',I1,'_',I2,8('_',F8.4))
3008  CONTINUE
      WRITE(12,3011) C,DEG
3011  FORMAT(' ',I2,'_',I6)
      DO 3015 I=1,4
        IF (SIMS2.GT.0) THEN
          WRITE(13,3016) C,I,PZEKC(I)*1.0D0/SIMS2,PZKC(I)*1.0D0/SIMS2,
     *PZEKW1(I)*1.0D0/SIMS2,PZKW1(I)*1.0D0/SIMS2,PZEKW2(I)*1.0D0/SIMS2,
     *PZKW2(I)*1.0D0/SIMS2,PZCAB(I)*1.0D0/SIMS2,PZEPEC(I)*1.0D0/SIMS2
3016      FORMAT(' ',I2,'_',I1,8('_',F8.4))
        ELSE
          WRITE(13,3017) C,I
3017      FORMAT(' ',I2,'_',I1,8('_-9.9999'))
        END IF
3015  CONTINUE

C  Parameter combination loop terminates

4000  CONTINUE

      STOP

C  Remote block DEGVAL
C
C  Flags a degenerate sample and counts the number of degenerate samples
C  per parameter combination. Assigns a value of 0 to the measures and
C  associated variables per simulation when executed, as the results are
```

```
C  based on non-degenerate samples.

      REMOTE BLOCK DEGVAL
        FLAG=1
        DEG=DEG+1
        PA=0.0D0
        KC=0.0D0
        VKC=0.0D0
        KW1=0.0D0
        VKW1=0.0D0
        KW2=0.0D0
        VKW2=0.0D0
        DR=0.0D0
        CAB=0.0D0
        VCAB=0.0D0
        KPC=0.0D0
        PPC=0.0D0
        PEC=0.0D0
        ZEKC=0.0D0
        ZKC=0.0D0
        ZEKW1=0.0D0
        ZKW1=0.0D0
        ZEKW2=0.0D0
        ZKW2=0.0D0
        ZCAB=0.0D0
        ZEPEC=0.0D0
      END BLOCK

      END

      REAL FUNCTION RAN2(IDUM)
      INTEGER IDUM,IM1,IM2,IMM1,IA1,IA2,IQ1,IQ2,IR1,IR2,NTAB,NDIV
      REAL AM,EPS,RNMX
      PARAMETER (IM1=2147483563,IM2=2147483399,AM=1./IM1,IMM1=IM1-1,
     *IA1=40014,IA2=40692,IQ1=53668,IQ2=52774,IR1=12211,IR2=3791,
     *NTAB=32,NDIV=1+IMM1/NTAB,EPS=1.2E-7,RNMX=1.-EPS)
C  Long period (>2*10^18) random number generator of L'Ecuyer with Bays-
C  Durham shuffle and added safeguards. Returns a uniform random deviate
C  between 0.0 and 1.0 (exclusive of the endpoint values). Call with
C  IDUM a negative integer to initialize; thereafter, do not alter IDUM
C  between successive deviates in a sequence. RNMX should approximate
C  the largest floating value that is less than 1.
      INTEGER IDUM2,J,K,IV(NTAB),IY
      SAVE IV,IY,IDUM2
      DATA IDUM2/123456789/, IV/NTAB*0/, IY/0/
      IF (IDUM.LE.0) THEN
        IDUM=MAX(-IDUM,1)
        IDUM2=IDUM
        DO 4005 J=NTAB+8,1,-1
          K=IDUM/IQ1
          IDUM=IA1*(IDUM-K*IQ1)-K*IR1
          IF (IDUM.LT.0) IDUM=IDUM+IM1
          IF (J.LE.NTAB) IV(J)=IDUM
4005    CONTINUE
        IY=IV(1)
      END IF
      K=IDUM/IQ1
```

```
IDUM=IA1*(IDUM-K*IQ1)-K*IR1
IF (IDUM.LT.0) IDUM=IDUM+IM1
K=IDUM2/IQ2
IDUM2=IA2*(IDUM2-K*IQ2)-K*IR2
IF (IDUM2.LT.0) IDUM2=IDUM2+IM2
J=1+IY/NDIV
IY=IV(J)-IDUM2
IV(J)=IDUM
IF (IY.LT.1) IY=IY+IMM1
RAN2=MIN(AM*IY,RNMX)
RETURN
END
```

BIBLIOGRAPHY

Bell, M. J., Ternberg, J. L., Feigin, R. D., Keating, J. P., Marshall, R., Barton, L., and Brotherton, T. (1978). Neonatal Necrotizing Enterocolitis: Therapeutic Decisions Based Upon Clinical Staging. *Annals of Surgery*, **187**, 1-7.

Born, J. D., Hans, P., Albert, A., and Bonnal, J. (1987). Interobserver Agreement in Assessment of Motor Response and Brain Stem Reflexes. *Neurosurgery*, **20**, 513-517.

Bhattacharyya, G. K. and Johnson, R. A. (1977). *Statistical Concepts and Methods*. Wiley, New York.

Byrt, T., Bishop, J., and Carlin, J. B. (1993). Bias, Prevalence and Kappa. *Journal of Clinical Epidemiology*, **46**, 423-429.

Cicchetti, D. V. and Allison, T. (1971). A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings. *American Journal of EEG Technology*, **11**, 101-109.

Cicchetti, D. V. and Fleiss, J. L. (1977). Comparison of the Null Distributions of Weighted Kappa and the C Ordinal Statistic. *Applied Psychological Measurement*, **1**, 195-201.

Cicchetti, D. V. (1981). Testing the Normal Approximation and Minimal Sample Size Requirements of Weighted Kappa When the Number of Categories is Large. *Applied Psychological Measurement*, **5**, 101-104.

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

Cohen, J. A. (1968). Weighted Kappa: Nominal Scale Agreement With Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, **70**, 213-220.

Everitt, B. S. (1968). Moments of the Statistics Kappa and Weighted Kappa. *The British Journal of Mathematical and Statistical Psychology*, **21**, 97-103.

Feinstein, A. R. and Cicchetti, D. V. (1990). High Agreement But Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*, **43**, 543-549.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large-Sample Standard Errors of Kappa and Weighted Kappa. *Psychological Bulletin*, **72**, 323-327.

Fleiss, J. L. and Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, **33**, 613-619.

Gordon, N. S., Fois, A., Jacobi, G., Minns, R. A., and Seshia, S. S. (1983). The Management of the Comatose Child. *Neuropediatrics*, **14**, 3-5.

Graham, P. and Jackson, R. (1993). The Analysis of Ordinal Agreement Data: Beyond Weighted Kappa. *Journal of Clinical Epidemiology*, **46**, 1055-1062.

Hall, J. N. (1974). Inter-rater Reliability of Ward Rating Scales. *British Journal of Psychiatry*, **125**, 248-255.

Huttenlocher, P. R. (1972). Reye's Syndrome: Relation of Outcome to Therapy. *The Journal of Pediatrics*, **80**, 845-850.

Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Houghton Mifflin, Boston.

Kliegman, R. M. and Fanaroff, A. A. (1984). Necrotizing Enterocolitis. *The New England Journal of Medicine*, **310**, 1093-1103.

Knuth, D. E. (1981). *Seminumerical Algorithms*, 2nd edition, Vol. 2 of The Art of Computer Programming. Addison-Wesley, Reading, MA.

Koran, L. M. (1975a). The Reliability of Clinical Methods, Data and Judgments. *The New England Journal of Medicine*, **293**, 695-701.

Koran, L. M. (1975b). The Reliability of Clinical Methods, Data and Judgments. *The New England Journal of Medicine*, **293**, 642-646.

Kupper, L. L. and Hafner, K. B. (1989). On Assessing Interrater Agreement for Multiple Attribute Responses. *Biometrics*, **45**, 957-967.

L'Ecuyer, P. (1988). Efficient and Portable Combined Random Number Generators. *Communications of the ACM*, **31**, 742-774.

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

Markus J. B., Somers, S., Franic, S. E., Moola, C., and Stevenson, G. W. (1989). Interobserver Variation in the Interpretation of Abdominal Radiographs. *Radiology*, **171**, 69-71.

Mata, A. G. and Rosengart, R. M. (1980). Interobserver Variability in the Radiographic Diagnosis of Necrotizing Enterocolitis. *Pediatrics*, **66**, 68-71.

McNemar, Q. (1962). *Psychological Statistics*, 3rd edition. Wiley, New York.

Park, S. K. and Miller, K. W. (1988). Random Number Generators: Good Ones Are Hard to Find. *Communications of the ACM*, **31**, 1192-1201.

Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J., and Cheney, F. W. (1990). Measuring Interrater Reliability Among Multiple Raters: An Example of Methods for Nominal Data. *Statistics in Medicine*, **9**, 1103-1115.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd edition. Cambridge University Press, New York.

Rogot, E. and Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Diseases*, **19**, 991-1006.

Seshia, S. S., Seshia, M. M. K., and Sachdeva, R. K. (1977). Coma in Childhood. *Developmental Medicine and Child Neurology*, **19**, 614-628.

Shinar, D., Gross, C. R., Hier, D. B., Caplan, L. R., Mohr, J. P., Price, T. R., Wolf, P. A., Kase, C. S., Fishman, I. G., Barwick, J. A., and Kunitz, S. C. (1987). Interobserver Reliability in the Interpretation of Computed Tomographic Scans of Stroke Patients. *Archives of Neurology*, **44**, 149-155.

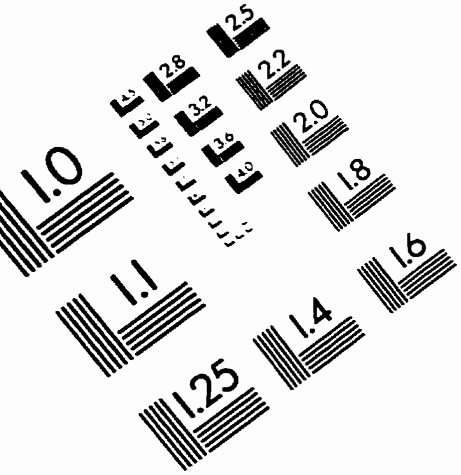
Simpson, D. and Reilley, P. (1982). Pediatric Coma Scale. *Lancet*, **2**, 450.

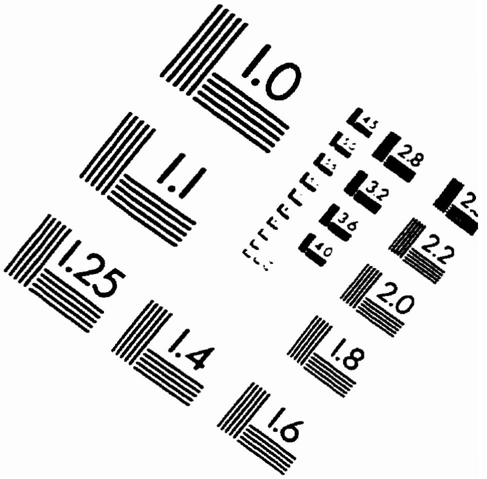
Solari, A., Filippini, G., Gagliardi, L., Bevilacqua, L., Amantini, A., Giuliana, G., Messina, C., Rossi, G., Savettieri, G., Tredici, G., and Duca, P. G. (1989). Interobserver Agreement in the Diagnosis of Multiple Sclerosis. *Archives of Neurology*, **46**, 289-292.

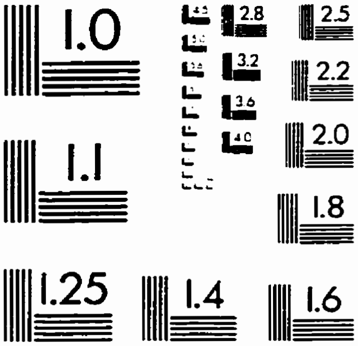
Teasdale, G., Knill-Jones, R., and Van der Sande, J. (1978). Observer Variability in Assessing Impaired Consciousness and Coma. *Journal of Neurology, Neurosurgery, and Psychiatry*, **41**, 603-610.

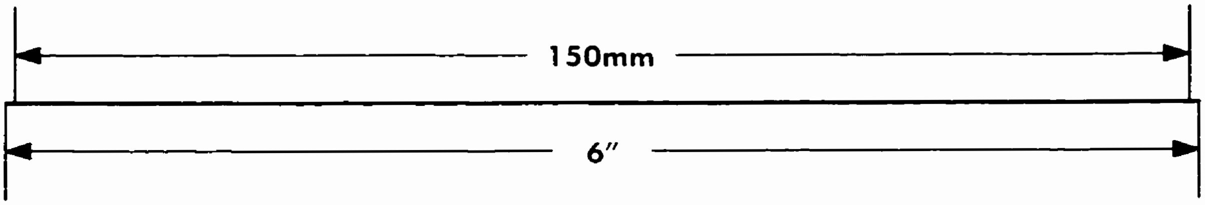
Yager, J. Y., Johnston, B., and Seshia, S. S. (1990). Coma Scales in Pediatric Practice. *American Journal of Diseases of Children*, **144**, 1088-1091.

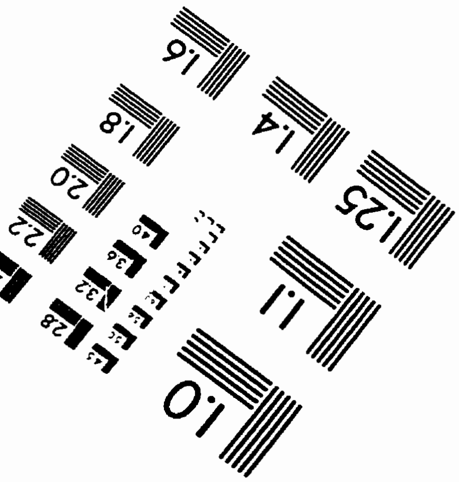
IMAGE EVALUATION TEST TARGET (QA-3)

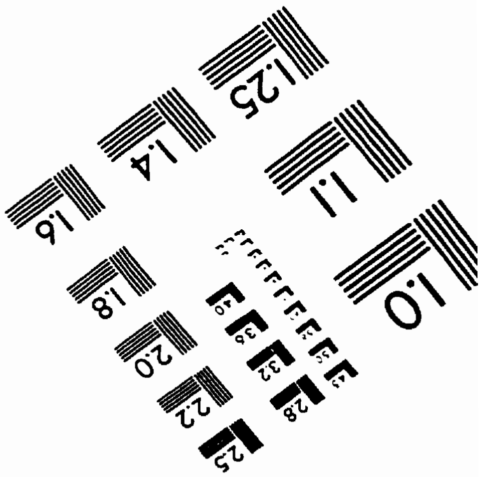












APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989