TOWSTOPIAT, OLGA MICHAEL

MULTIVARIATE MEASURE OF AGREEMENT

The University of Arizona PH.D. 1981

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106

MULTIVARIATE MEASURE OF AGREEMENT

by

Olga Michael Towstopiat

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

1981

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read

the dissertation prepared by Olga Towstopiat

entitled Multivariate Measure of Agreement

and recommend that it be accepted as fulfilling the dissertation requirement

for the Degree of Doctor of Philosophy.

John R. Bergan	3/30/81 Date
Anthony A. Cancell	3/30/81 Date
Shitala P. Mishra	3/30/81 Date
	3/30/81 Date
	3/30/81 Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

John R. Bergan Dissertation Director	3/30/81 Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Olga Towstopiat

ACKNOWLEDGMENTS

The author wishes to express her gratitude to Dr. John Bergan, the dissertation advisor, to Dr. Anthony Cancelli, Dr. Lawrence Aleamoni, Dr. Keith Meredith, and Dr. Shitala Mishra for their support and guidance.

Finally, and most importantly to my parents, whose love and assistance made this work possible.

TABLE OF CONTENTS

TABLE OF CONTENTS--Continued

LIST OF TABLES

ABSTRACT

Reliability issues are always salient as behavioral researchers observe human behavior and classify individuals from criterion-referenced test scores. This has created a need for studies to assess agreement between observers, recording the occurrance of various behaviors, to establish the reliability of their classifications. In addition, there is a need for measuring the consistency of dichotomous and polytomous classifications established from criterion-referenced test scores. The development of several log linear univariate models for measuring agreement has partially met the demand for a probability-based measure of agreement with a directly interpretable meaning. However, multivariate repeated measures agreement procedures are necessary because of the development of complex intrasubject and intersubject research designs.

The present investigation developed applications of the log linear, latent class, and weighted least squares procedures for the analysis of multivariate repeated measures designs. These computations tested the model-data fit and calculated the multivariate measure of the magnitude of agreement under the quasi-equiprobability and quasi-independence models. Applications of these computations were illustrated with real and hypothetical observational data.

It was demonstrated that employing log linear, latent class, and weighted least squares computations resulted in identical multivariate model-data fits with equivalent chi-square values. Moreover, the application of these three methodologies also produced identical measures of the degree of agreement at each point in time and for the multivariate average.

The multivariate methods that were developed also included procedures for measuring the probability of agreement for a single response classification or subset of classifications from a larger set. In addition, procedures were developed to analyze occurrences of systematic observer disagreement within the multivariate tables.

The consistency of dichotomous and polytomous classifications over repeated assessments of the identical examinees was also suggested as a means of conceptualizing criterion-referenced reliability. By applying the univariate and multivariate models described, the reliability of these classifications across repeated testings could be calculated.

The procedures utilizing the log linear, latent structure, and weighted least squares concepts for the purpose of measuring agreement have the advantages of (1) yielding a coefficient of agreement that varies between zero and one and measures agreement in terms of the probability that the observers' judgments will agree, as estimated under a quasi-equiprobability or quasi-independence model, (2) correcting for the proportion of "chance" agreement, and (3) providing a directly interpretable coefficient of "no agreement." Thus, these multivariate procedures may be regarded as a more refined psychometric technology for measuring inter-observer agreement and criterion-referenced test reliability.

CHAPTER 1

INTRODUCTION

Science requires that data be reliable. In the behavioral sciences, quantitative estimates of the consistency, dependability, and predictability of behavioral data are preferred. Such an assessment is necessary prior to the determination of the accuracy or validity of the data. The probability of making valid empirical conclusions based on scientific behavioral research is considered to be a direct function of the reliability of measurement. Reliability issues are always salient as social scientists observe human behavior and as they try to classify individuals. Both direct observation of human behavior and classification of individuals by means of criterion-referenced tests are methods that social scientists use to minimize inferential assumptions and to increase the validity of their assertions about human functioning. The present study will focus on the reliability of human observers' judgments when making direct observations of behavior. Moreover, it will be explained how the procedures used for determining such reliability may be applied to the problems of classifying individuals from criterion-referenced test scores.

Measuring Observer Agreement

Observational studies of human behavior often require the recording of a number of behavioral categories. In addition, most observational studies require the assessment of agreement between observers to establish the reliability of the observations. These reliability measures attempt to document that the behavioral data compiled by the observer are similar to those that would be collected by other trained observers. The measurement procedures involve members of the observational team simultaneously and independently coding identical behaviors emitted by the same experimental subjects throughout the identical experimental sessions. The observers' behavioral protocols are then compared with respect to consistency among observers' judgments. It is imperative that interobserver reliability estimates be assessed since the probability of measuring a change in subject behavior between treatment conditions, is a direct function of the reliability of the observers' judgments.

Observers of human behavior may be predisposed to bias, inattention, inadequate training, subjective judgment, observer feedback, observer drift, and other sources of random and systematic error (Hersen and Barlow, 1976; Kratochwill, 1978). Other sources of error affecting observational measurement may be influenced by the number of behavioral observations and the length of observational periods (Shavelson, 1980). Such multifacted errors become incorporated into the behavioral measurements and affect the reliability of the observers' responses. The extent to which these observers' responses are reproducible or reliable determines the degree and significance of their response agreement. This situation is analogous to the problems that gave rise to classical psychometrics. Human observers correspond to alternate forms of a test, and nominal response data correspond to test scores.

The observers may be school psychologists, the behavioral categories on-task, physical agression, and speaking out, and the units different academic classroom periods; or the observers may be teachers, the categories different types of social skills, and the units small groups of children of various ages, etc. Using these examples it becomes evident that three assumptions must be met for an accurate measure of observer agreement. First, the units (i.e., time or experimental conditions) must be mutually exclusive. Second, the nominal scale behavioral categories must be mutually exclusive and exhaustive. Finally, the observers must respond independently (Cohen, 1960).

In most cases, a criterion for the "correctness" of observer responses is not established, and the observers are equally skilled in coding the behaviors. In addition, there are no assumptions or restrictions involving the distribution of observer judgments across behavioral categories (Cohen, 1960).

Measures such as the percentage of agreement, Cohen's Kappa (Cohen, 1960), and phi have been used to measure observer agreeement, but these coefficients have limitations. For example, the percentage of agreement index compares number of observer agreements and disagreements with a standard. Although this elementary procedure is computationally and interpretatively simple, it is unacceptable because it neglects to consider chance agreement, does not have a meaningful lower bound of acceptability, and does not have a value representing no agreement (Hartmann, 1977). The value for the percentage agreement index also relies excessively on the frequency of the behavior, thereby resulting in evaluations of high frequency behaviors to have inflated high percentage agreement values. Investigators may also notice that acceptable levels of percentage agreement occur when in reality the judges are responding at chance levels. Finally, inclusion of observer agreement frequencies on nonoccurrences of behavior may substantially alter computed percentages of agreement (Yelton, Wildman, and Erickson, 1977).

Categorical or occurrence-nonoccurrence observer agreement data have also been analyzed with correlational statistics such as kappa (k) and phi (Hartmann,1977). Kappa reflects the proportion of agreement after chance agreement is eliminated from consideration. Phi is the product-moment correlation between two observers' dichotomous categorical data. The phi value will approximately equal the kappa value if the marginal frequencies or the sums of the table columns and rows are relatively equal. Percentage agreement, kappa, and phi are unfortunately fraught with several weaknesses. All three statistics allow confounding of random and systematic error (Hartmann, 1977). Disagreements between observers caused by random factors, such as occasional inattention or coding errors, and systematic factors, such as differential operational definitions of target behaviors, cannot be discriminated when the previously mentioned statistics are applied. In addition, these measures of agreement lack a directly interpretable means of assessing the degree of agreement.

Log-Linear Agreement Models

As an alternative, the application of log linear models and the quasi-independence concept for the purpose of measuring observer agreement has the advantages of yielding a probability-based measure of agreement with a directly interpretable meaning, correcting for the proportion of "chance" agreement, and providing a directly meaningful coefficient of "no agreement".

Bergan (1980a, 1980b) has developed procedures for measuring observer agreement using the quasi-independence and quasi-equiprobability concepts. The application of these procedures for measuring observer agreement has several advantages. Use of the quasi-independence or quasi-equiprobability concept yields a coefficient of observer agreement that varies between zero and one and measures agreement in terms of the probability that the observers' judgments will agree, as estimated under a quasi-independence or quasi-equiprobability model. These procedures may also be used to investigate whether or not a single observational category or specific group of categories is a major contributor to the coefficient of agreement. Finally, systematic occurrences of disagreement between observers may be located and measured.

To assess agreement the judgments of the observers are organized into a contingency table. Quasi-equiprobability or quasi-independence among the variables comprising this contingency table is measured by testing the hypothesis that a subset of the contingency table cells are equiprobable or independent. The quasi-equiprobability model is recommended for measuring observer agreement when two or three observers are recording the presence or absence of one specific behavior. The observer agreement measure using the quasi-equiprobability model allows the flexibility of calculating the statistic completely with hand calculations or with the use of various computer programs (Clogg, 1977; Fay and Goodman, 1973). The quasi-independence model is recommended for measuring observer agreement when two or more observers are recording three or more response categories. This analysis includes the calculation of maximum likelihood estimates of expected cell frequencies under the model of quasi-independence which requires an iterative procedure and preferably the use of a high speed digital computer.

A third model that may be considered for the assessment of agreement is the model of symmetry. Given a two-way square contingency table, the model of symmetry disregards the diagonal cells in the table and examines the relationship of the pairs of cells around the diagonal. Under this model, the hypothesis of symmetry tests the equality of the joint probabilities, within each disagreement pair of cells around the diagonal.

Statement of the Problem

The scientific community and research consumers are in need of a statistical "judgmental aid" that condenses a series of human behavioral data from many observers (Kratochwill and Wetzel, 1977). Furthermore, measurements of observer reliability that communicate the accuracy and quality of behavioral data recorded throughout a specific project or experiment are necessary. The increasing development of complex intrasubject and intersubject research designs calls for the construction of multivariate measurement procedures that employ various design facets. Present applied behavioral research designs incorporate design facet combinations of subjects, observers, conditions, behaviors, sessions, and trials. The functional association between reliability and validity requires that data reliability be demonstrated if behavioral differences between treatment conditions are to be shown.

Investigators, such as Bennett (1967, 1968 and 1972), have proposed a multivariate formula for measuring observer agreement across multiple observers and behaviors. Their procedures focused on a chi-square statistic that tested the statistical significance of a multivariate measure of agreement. But, the procedures, such as Bennett's, lacked a means of measuring the degree of agreement among the observers. Although Bennett did provide a measure for the average proportion of "positive findings" (i.e., agreement among observers), his index neglects to consider the contribution of chance agreement.

To meet the need for a multivariate measure of the degree of agreement among observers, the present investigation was designed to develop a multivariate extension of the Bergan (1980a, 1980b) procedure for measuring observer agreement. The study encompassed the construction of a multivariate repeated measures statistical procedure for assessing the magnitude of observer agreement across settings, subjects, behaviors, conditions, observers, and trials.

By focusing on the collection of observer agreement data across different points in time, a repeated measures design was created. Specifically, applications of the log-linear, latent class, and weighted least squares procedures for the analysis of multivariate repeated measures designs were constructed. These methodologies assessed the fit between data and models of quasi-independence and quasi-equiprobability. A multivariate measure of agreement was also calculated. In addition, methods for measuring the probability of observer agreement for a single response classification or subset of classifications from a larger set was constructed. Finally, procedures for analyzing occurrences of systematic observer disagreement were developed. A pretest-posttest-followup intrasubject research design with accompanying data from two observers was provided by the University of Arizona Psychology Department. Observer judgments of the presence of one behavior across two points in time was compiled and a two-way contingency table constructed. In addition to this data, hypothetical data illustrating the judgments of two observers coding three behavioral categories across two points in time was provided. These data were used to test and illustrate the application of the multivariate repeated measures statistical procedures.

Since the observers' judgments are represented as codings of the occurrence-nonoccurrence of target behaviors, it is postulated that such categorical data is similar to the dichotomous mastery and nonmastery classification decisions that occur within the realm of criterion-referenced testing. The consistency of mastery and nonmastery classifications over repeated assessments of the identical subjects has been suggested as an approach to conceptualize criterion-referenced reliability (Huynh, 1976; Subkoviak, 1976; Swaminathan, Hambleton and Algina, 1973; Subkoviak, 1980). The present investigation also demonstrated how the multivariate agreement measures developed can be applied to assess the reliability of dichotomous and polytomous classifications established from criterion-referenced test scores.

The implications drawn from the proposed research suggest techniques for the psychometric validation of procedures for determining inter-observer agreement. The growing sophistication of behavioral assessment procedures mandates that these psychometrically sound measurement principles be established. By increasing the reliability of psychometric procedures, more accurate selection of appropriate treatment strategies by psychologists and educators may occur.

CHAPTER 2

REVIEW OF THE LITERATURE

The following sections will present the various procedures for estimating interobserver reliability along with evidence suggesting why a lack of consensus regarding the "preferred" method exists among applied behavioral researchers. The literature review will specifically discuss the percentage agreement measure, Cohen's kappa, phi, generalization of the kappa coefficient, and multivariate observational measures. An extension of the section on multivariate observational measures will include a discussion on assessing agreement with graphical judgmental aids and with probability based formulas. Following these sections will be a presentation of measures of observer agreement under the models of quasi-independence and quasi-equiprobability, and measures of observer agreement from repeated measurement experiments. Finally, evidence suggesting how multivariate agreement measures may be applied to the computation of reliability indexes for criterion-referenced test scores will be presented.

Percentage Agreement Measure

Although the past two decades have produced a vast array of sophisticated observational technologies, the development of psychometrically refined measurement principles has been languorous. Hartmann (1977) noted that researchers who use behavioral observers have tended not to emphasize that the critical association between validity and reliability is affected by the reliability of the human observer. This problem is masked by researchers' excessive employment of the percentage agreement formula (Kelly, 1977).

All procedures for measuring observer agreement require that the categorical or occurrence-nonoccurrence data be summarized into a multiway contingency table. A two-by-two table, for example, would represent agreement between two observers that the behavior occurred (cell A) and agreement the behavior did not occur (cell D) in the two diagonal cells, respectively. The off-diagonal cells would represent the two types of disagreement; observer A judging the behavior occurred while observer B judging it did not occur (cell B) and vice-versa (cell C).

The most commonly used agreement statistic is percentage agreement (Kelly, 1977). This agreement estimate is defined by dividing the frequency of observer agreements by the total number of observations and multiplying by 100, i.e., (A + D/A + B + C + D) x 100 (Kratochwill and Wetzel, 1977). Hawkins and Dotson (1975) argue that this statistic is inadequate because agreement estimates are inflated when behavioral occurrences are low, thereby incorporating the agreement frequencies that the target behavior was not emitted.

Two additional versions of the percentage agreement procedure have also been reported by Hartmann (1977), Kratochwill and Wetzel (1977), and Hawkins and Dotson (1975). Hartmann (1977) presented a version called the "effective percentage agreement", in which behavioral occurrences may be calculated if the investigation focuses on observer agreement regarding low rates of target behavior occurrences. This procedure reduces the inflationary effects of observers agreeing that the target behavior did not occur. Effective percentage agreement for occurrences is defined by dividing the frequency in agreement cell A by the sum of cells A, B, and C and multiplying that proportion by 100.

An investigation that focuses on the low nonoccurrence rate of a target behavior may apply the effective percentage agreement formula for nonoccurrence of target behaviors (Hartmann, 1977). This agreement measure is given by dividing the frequency of agreement cell D by the sum of cells B, C, and D and multiplying that proportion by 100. These effective percentage agreement formulas purport to reduce the chance agreements that may arise when different rates of occurrence of target behaviors are presented. "Chance" agreement refers to the expected proportion of agreement that may be attributed to independent-observer judgments. Due to such manifestations, Hartmann recommends under conditions of low rates of target behavior occurrence that nonoccurrence agreement frequencies be discarded to reduce the contribution to agreement by "chance" agreement. Analogously, if high rates of target behaviors are emitted, omission of occurrence agreement frequencies is warranted to prevent excessive inflation of the agreement index. Hartmann (1977) cautions that these percent agreement estimates have additional deficiencies regarding the specific conditions (i.e., specific rate of target behavior occurrence) under which these procedures should be used, the minimum acceptable level of agreement, and the value designating no agreement.

Since an objective method for determining the frequency at which one effective percentage agreement measure should be used over the other does not exist, researchers such as Birkimer and Brown (1979a) have suggested reporting agreement as the average of the occurrence and nonoccurrence agreement scores. Harris and Lahey (1978) countered that suggestion by developing a percent agreement formula that incorporated occurrence agreement weighted by the average rate of nonoccurrence, and nonoccurrence agreement weighted by the average rate of occurrence. Although this procedure attempts to minimize the contribution of chance agreement, it does not remedy the previously mentioned limitations inherent in all the versions of the percent agreement formulas.

Cohen's Kappa and Phi

During the previous attempts to develop a more precise measure of observer agreement, there merged a consensus among most researchers that there must be a demonstration that "obtained" agreement significantly differed from "chance" agreement. The investigators also added that summarizing categorical agreement data within a contingency table allows researchers to use several commonly known postulates to develop reliability measures (Goodman and Kruskal, 1954).

Light (1971) asserted that agreement is a special case of association. Only under the conditions where two observers are judging the occurrence-nonoccurrence of a single behavior is this statement untrue. Given a two-by-two contingency table, agreement measures will equal association measures. Thus, measures based on the chi-square statistic, such as the likelihood ratio chi-square statistic, Pearson chi-square, or phi, may be applied (Light, 1971). However, if more than two behavioral categories are applied, then standard association measures should not be used to measure agreement.

Light argues that the difference between agreement and association is that for two or more observer judgments to agree, they must be categorized in the identical contingency table cells, while for two observer judgments to be associated mandates that one observer's judgment be predictable from information regarding the other's judgment. Therefore, agreement is a special case of association, and the responses in a contingency table may indicate high agreement and high association or even low agreement and high association (Light, 1971).

Historically, investigators have emphasized reliability measures for continuous and ranked data (Ebel, 1951). Recent advances with categorical data (Bergan, 1980a) have allowed more investigators to incorporate categorical variables into their research designs. A breakthrough first came in 1955 when Scott generated an observer agreement measure that corrected for chance agreement and assumed the distribution of response proportions across behavioral categories for the population was known and equal for all observers. Scott's (1955) "coefficient of intercoder agreement" was estimated as:

$$\pi = \frac{t_o - t_e}{1 - t_e}$$

where t_o is the sum of the observed proportion of agreement response pairs, and t_e is the sum of the p_{ii} diagonal cell probabilities. However, p_{ii} was calculated with the assumption of known population marginal probabilities, and on the basis that both judges had marginal distributions of proportions that were equal and identical to the population proportions.

Cohen (1960) commented it would be impossible for investigators to meet Scott's assumptions of symmetric marginals. Therefore, Cohen developed an index, kappa, that made no assumptions regarding equality of the marginal distributions. He defined kappa (k) as:

$$k = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the sum of the observed proportion of cases in the main diagonal of the contingency table and P_e is the sum of the expected proportion of cases in the main diagonal. The expected proportion of cases was calculated by finding the joint probabilities of the observed marginals. Thus, P_o denoted the proportion of cases in which the observers agreed and P_e denoted the proportion of cases for which observer agreement was expected by chance.

The kappa statistic presented in the previous paragraph is analogous to the agreement expression $1 - \frac{d_o}{d_e}$, where d_o is the observed disagreement proportion, and d_e is the expected disagreement proportion, under the hypothesis of random agreement. Based on this expression, kappa can be given by:

$$k = 1 - (1 - \frac{\sum_{i=1}^{c} n_{ii}}{n} / 1 - \frac{\sum_{\substack{i=1 \\ j=1}}^{c} n_{i+} \quad n_{+j}}{n^2},$$

where n_{ii} is the frequency for cell $_{ii}$, n is the total number of responses in the table, and n_{i+} and n_{+j} are the observed marginals in the i^{th} row and j^{th} column, respectively. Thus, kappa may be conceptualized as a ratio of disagreements or measures of distance between two judges. These measures of distance are calculated by counting a series of ones and zeros or the frequency of agreement versus disagreement pairs in the complete set of observer respones (i.e., 2n).

Coefficient k represents the proportion of agreement after chance agreement is removed from consideration. If the observed magnitude of agreement equals the magnitude expected by chance alone, then k will equal zero. Complete agreement between the observers will result in a k value of 1.0. A negative k value will occur if the observed agreement is less than the agreement expected by chance.

Further investigations by Cohen (1968), Everitt (1968), Fleiss, Cohen, and Everitt (1969), and Hubert (1977) produced a weighted kappa coefficient, which allowed the various disagreement categories to be differentially weighted. For the purpose of assessing the magnitude of observer agreement, all disagreements should be regarded as equally serious. Thus, weighted kappa, which places varying emphasis on the different degrees of disagreement, need not be computed.

Procedures have also been developed by Cohen (1960, 1968) and Everitt (1968) for measuring the standard error of the kappa statistic, but their formulas overestimate the variance. Application of their procedure would produce conservative significance tests and confidence intervals. However, Fleiss, Cohen and Everitt (1969) developed a more accurate formula for the estimated large-sample variances. By applying the Fleiss et al. (1969) procedure, accurate tests of the significance of the kappa statistic can be conducted and kappa confidence intervals generated.

Phi, the product-moment correlation between two observers responses on the occurrence-nonoccurrence of a single behavior, is denoted by:

$$\text{Phi} = (AD - BC)/(A+B)(C + D)(A + C)(B + D)^{1/2},$$

where A is the frequency for behavioral occurrence agreement, D is the frequency for behavioral nonoccurrence agreement, and C and B are the disagreement frequencies.

The values for phi range from -1.0 to +1.0, with 0.0 representing no agreement between the two judges responses and +1.0 representing complete agreement between the two judges responses. A comparison of kappa, applied to dichotomous responses from two judges, and phi will show the two indexes to be equal if the marginal proportions for the two judges are the same. When the marginal proportions for the two judges ($P_{1A} \cong P_{1B}$) are within .10 - .20, kappa will have a value within a few hundreths of the phi value (Cohen, 1960). If the marginal proportions are not equal, then the kappa value will be less than the phi value, thereby indicating an inflation of the phi value. Thus, if the marginal proportions are equal, phi may portray the chance corrected proportion of agreement.

Various disadvantages exist with the use of the phi coefficient. If variability between the observers' responses does not exist (variability = 0), then the correlation is undefined. Other problems may be exhibited if the observers' errors are correlated or if the range of scores is not considered (Hartmann, 1977).

Generalizations of the Kappa Coefficient

The application of the kappa statistic is limited to the situation where only two observers are considered and the same two observers code each subjects' behavior (Cohen, 1960; Hubert, 1977). Generalizations have been essential for circumstances where more than two observers are coding behaviors and where the observers are not all judging the same subjects.

Fleiss (1971) generalized the kappa formula for assessing observer agreement among a specific number of judges who have been randomly assigned to observe various subjects. Let N refer to the total number of subjects, n refer to the number of raters observing each subject, and B refer to the number of behavioral categories the observers were coding. Allow i = 1, ... N to denote the subjects and j = 1, ... B to denote the behavioral categories. Let $_{ij}$ refer to the number of observers who judged the i^{th} subject to the j^{th} behavioral category and P_j to be given by:

$$P_j = \frac{1}{Nn} \sum_{t=1}^{N} n_{ij}$$

Define P_j as the proportion of all codings to the j^{th} behavioral category. Summing across all the behavioral categories will demonstrate:

$$\sum_j n_{ij} = n \quad \text{and} \quad \sum_j P_j = 1,$$

The following algorithm calculates the proportion of observer agreement pairs out of all the n(n-1) possible pairs of observer judgments:

$$\bar{P} = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{B} n_{ij}^{2} - Nn\right)$$

If a subject is randomly selected and observed by two randomly chosen raters, the second rater's judgments would agree with the first rater's judgments $\bar{P}$ proportion of the time. Since random chance agreement is to be expected, the mean proportion of agreement must be calculated:

$$\bar{P}_e = \sum_{i=1}^{B} P_j^{2}$$

The value of $1 - \bar{P}_e$ assesses the degree of observer agreement possible following the correction for chance. The degree of agreement observed following chance correction is $\bar{P} - \bar{P}_e$. Therefore, the normalized measure of overall agreement, corrected for chance agreement, equals:

$$K = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}.$$

An equivalent and condensed kappa statistic may be given as:

$$K = \frac{\sum_{i=1}^{N} \sum_{j=1}^{B} n_{ij}^{2} - Nn \quad 1 + (n - 1) \sum_{j=1}^{B} P_j^{2}}{Nn(n - 1)(1 - \sum_{j=1}^{B} P_j^{2})}$$

Fleiss (1971) also added, if the total number of subjects (N) is large, a kappa variance under the hypothesis of no agreement beyond chance may be calculated. Furthermore, under the hypothesis of no agreement following chance correction, K/standard error (k) will, by the central limit theorem, be distributed approximately as a standard normal variate (Fleiss, 1971). The resulting index from K/standard error (k) will allow investigators to infer if the overall observer agreement from ratings for B behavioral categories is significantly greater than chance or not.

Finally, Fleiss (1971) also developed procedures for measuring agreement on a specific behavioral category j. Under the assumption that each subject is judged to belong to one of the B behavioral categories by randomly chosen observers, the agreement measure is defined as the conditional probability that the second observer's assignment to behavioral category j is identical to the first observer's assignment. Variance measures and statistical significance tests of these conditional probabilities may also be calculated if the investigator wishes to generate confidence intervals and test hypotheses that the kappa values are significantly different from zero.

Light (1971) also described a conditional measure of the agreement level, for the purpose of partitioning an overall kappa value into various partial kappas (k_j, j=1, ..., B). This measure, which is based on Coleman's unpublished work, allows one to infer the level of agreement between two raters' judgments for only those subjects which one rater placed into a predetermined behavioral category. Coleman's algorithm does not assume the observers are randomly chosen, as Fleiss (1971) did. Coleman also believed a conditional association measure of k should be based on the sum of agreement and disagreement pairs. The following index k_{pi} measures the level of agreement between two raters for those subjects which the first rater assigned to the j^{th} behavioral category:

$$K_{pj} = 1 - (1 - \frac{n_{+j}}{n_{i+}})/(1 - \frac{n_{+j}}{N}) \; ,$$

where $i = j$. Coleman also generated variance measures and procedures for assessing the statistical significance of K_{pi}, as it differs from zero.

Further generalizations of the kappa statistic by Light (1971) included a measure of the level of agreement among more than two observers. The multiple observer agreement formula is based on the following kappa expression:

$$1 - \frac{d_o}{d_e} \quad ,$$

where d_o is the observed proportion of observer disagreements and d_e the expected proportion of observer disagreements given the n (number of observers) observed marginals. A multiple observer agreement statistic k_m was developed by Light, based on the above expression, along with a measure of the estimated standard error, a procedure for testing the hypothesis that k_m is not significantly different from zero, and a measure of the conditional k_{pj} (i.e., partial kappa) for cases with more than two observers.

The kappa measures described in the previous paragraphs compared the observed overall level of agreement with the expected overall level. Thus, these procedures did not assess the differences between the observed and expected patterns of observer agreement. Light's (1971) alteration of the chi-square test allows an assessment of the pattern of observer agreement between two observers. By collapsing all of the off-diagonal disagreement cells, differences between observed and expected cell frequencies can be eliminated. Light expanded upon this concept and the chi-square statistic and generated a pattern agreement measure A_p that stresses each agreement cell and collapses all the disagreement into one category. In conjunction with the measure of the level of observer agreement, kappa, Light asserts more precise inferences regarding observer judgments may be made. The index A_p is distributed asymptotically, under the hypothesis of random agreement, as chi-square with I degrees of freedom (I X I table). If this "chi-square" A_p index is not statistically significant, then one cannot conclude that the observed agreement pattern is statistically significantly different from the expected pattern. By using the A_p and k indexes, the investigator may infer if the pattern and level of observer agreement are significantly different from the expected pattern and level of agreement.

The previous paragraphs described generalizations of the kappa coefficient for cases where more than two observers were employed and where the observers did not all judge the same subjects. In addition, the kappa variance, specific behavioral agreement measures, kappa partitioning, and pattern agreement measures were discussed. The following section will expand upon these issues and present a measure among multiple observers judging multiple behaviors.

Measures of Multivariate Observations

Bennett's (1967, 1968) initial interest with chi-square tests of observer agreement centered around the development of a measure to test the statistical significance of agreement among multiple observers, coding the presence/absence of a single behavior. The limited application of this statistic encouraged him to develop a multivariate formula to test the statistical significance of agreement among multiple observers coding the presence/absence of multiple behaviors (Bennett, 1972). For example, let there be _n_ observers coding the presence or absence of _B_ correlated behaviors on a sample of _N_ subjects. The binary response patterns of the n observers form a B-dimensional contingency table with observed frequencies.

The null hypothesis of no difference among observers, for all behavioral categories, may be tested by calculating probabilities for each cell and applying a chi-square statistic (degrees of freedom = (n - 1)B). This multiple observer and multiple behavior test of agreement assesses if observer agreement across the behavioral categories is statistically significant or not. Furthermore, by summing the binary responses across behaviors, one can calculate the average proportion of "positive responses" among the observers. Although Bennett (1972) tried to convey the average proportion of "positive responses" as a measure of the degree of agreement, he did not succeed. Chance agreement is not considered at all by Bennett, thereby creating a measure that contains all of the faults that are inherent in a percentage agreement measure. Bennett's chi-square statistic is also not recommended since it only assesses the statistical significance of the agreement among observers and neglects to measure the degree of observer agreement.

Fleiss (1965) commented that if many dichotomous responses regarding a subject are combined to produce an integer, the parametric means of assessing reliability may be applied. However, if a single dichotomous evaluation describes a subject, the model with a normal distribution of errors is not recommended. Given the case where many observers are judging the presence or absence of one behavior across many subjects, Fleiss recommends a nonparametric model to represent the errors underlying the observers' responses. He recommends that Cochran's Q statistic be used for assessing the hypothesis of no systematic differences among the observers' responses. Fleiss stated that if the Q test assessment of the presence of bias among the observers produces evidence that biases are minimal, then it is recommended that the intraclass correlation between all pairs of observer responses on the same subject be computed.

In 1966 Fleiss published an extension of the procedures described in the previous paragraph for cases where multiple observers are judging several behaviors across all subjects. Likelihood ratio tests were designed for testing the hypothesis of no differences among the observer means. If there are a large number of subjects or observers, the resulting statistic has an approximate chi-square distribution. Similar to the single behavior coefficient mentioned in the previous paragraph, Fleiss (1966) also developed a means for measuring the magnitude of errors of measurement (i.e., reliability) by applying the intraclass correlation statistic. The measures developed by Fleiss (1965, 1966) primarily test if the mean response vectors for the observers are significantly different. The issue of the magnitude of agreement among observers is not adequately addressed by Fleiss, because chance agreement is not considered.

Finally, another method for measuring agreement, given the case when multiple observers dichotomously (i.e., 1 or 0 scores) judge a single behavior across many subjects, is to use the Kuder-Richardson formula KR20 (Maxwell and Pilliner, 1968). The KR20 index purports to estimate the correlation between the given totals of subject scores, and the corresponding totals for a parallel group of observers. This index has severe limitations as a measure of agreement, such as underestimating the reliability coefficient when there is heterogeneity among subjects on the behavior of interest.

Assessing Agreement with Graphical Judgmental Aids

Birkimer and Brown (1979a) proposed that interobserver reliability assessments be presented graphically when data are presented in the form of percentage of trials in which a target behavior is judged to occur by an experimental rater. They asserted that by graphing obtained and chance reliability data the interpretation of the presence of experimental effects would be facilitated. The evaluation of observer agreement and disagreement, according to Birkimer and Brown, requires six steps. First, both observers' data for each reliability check must be graphically illustrated. Second, a percentage of disagreement reliability index must be calculated. Birkimer and Brown's definition of percentage disagreement is as follows:

$$\frac{\text{Number of disagreements}}{\text{Number of intervals or trials}} \times 100.$$

Third, disagreement percentages should be illustrated on the graph as a vertical band, centered halfway between the two observers' target behavior rates. Fourth, Birkimer and Brown recommend calculating the following chance disagreement percentage:

$$\text{Chance percentage} = \frac{(0_1 \text{ X } N_2) + (N_1 \text{ X } 0_2)}{T} \text{ X } 100,$$

where observer 1 judges the occurrence of the target behavior in 0_1 of the intervals and nonoccurrence in N_1 of the intervals, and observer 2 judges the occurrence of the target behavior in 0_2 of the intervals and nonoccurrence in N_2 of the intervals. The total number of intervals is designated as T. Fifth, the chance disagreement percentages should also be illustrated on the graph as a vertical band, centered halfway between the two observers' target behavior rates. This step was recommended so that investigators could compare the obtained percentages with chance disagreement and agreement ranges. Finally, Birkimer and Brown state the conclusion of the presence of an experimental effect requires that an overlap not exist between the disagreement range(s) calculated under different or adjoining experimental phases.

Birkimer and Brown's graphic procedures may seem elegant, but they have potential for encouraging the misinterpretation of bbhavioral data. The application of the disagreement range to express the "believability" of experimental effects may encourage excessive levels of Type II error (Hartmann and Gardner, 1979). Kratochwill

(1979) adds that investigators may regard these disagreement ranges as confidence intervals and conclude that the absence of overlap, between disagreement ranges in adjoining before and after experimental phases, indicates "believable" experimental effects. He adds that even under conditions when experimental effects did not occur the data patterns may be above the disagreement range between the two raters and not be due to observational chance variations. Although demonstrated observer reliability is crucial, evaluation of experimental effects should be by criteria independent of the judges' responses (Kratochwill, 1978; Kratochwill, 1979). Hartmann and Gardner (1979) add that to evaluate experimental effects not only must reliability be considered, but also the data level, slope, and variability.

Ktatochwill (1979) commented that several other problems exist with the Birkimer and Brown method. First, the investigator must question how much "overlap" between the disagreement ranges will be allowed. Second, the investigator must question how many observer reliability checks are necessary to convince the investigator the raters' judgments are consistent. Finally, these procedures add unnecessary complexity to the evaluation of the presence or absence of an experimental effect.

In a subsequent article, Birkimer and Brown (1979b) suggested the application of significance tests to facilitate the analysis of a two-by-two table, representing the judgments of two rates coding the presence or absence of a single behavior. In addition, they generated three simple procedures for measuring the statistical significance of the agreement data. These procedures are strongly criticized by Hartmann and Gardner (1979), who recommend that the reader ignore Birkimer and Brown's (1979b) informal procedures since more accurage agreement measures may be calculated with simple chi-square tests. In addition, by increasing the number of observational intervals, the investigator may greatly increase the probability of high agreement values, using the Birkimer and Brown (1979b) procedures.

Probability-Based Agreement Formula

Another agreement formula reported in the literature is Yelton, Wildman and Erickson's (1977) probability-based agreement formula. That, too, is greatly flawed. Although the Yelton et al. measure purports to take chance agreement into consideration and to measure the probability of two observers agreeing upon the occurrence of a target behavior, their procedures ignore all instances where one observer recorded a behavioral occurrence and the other observer did not record anything. For example, suppose observer A recorded 100 behavioral occurrences and observer B recorded 50 behavioral occurrences. If all of observer B's recordings matched observer A's, then the application of the Yelton et al. probability formula would result in an observer agreement probability index of 1.0. This index would indicate perfect agreement, yet observer B neglected to record 50 behavioral occurrences. Thus, by not considering nonoccurrence judgments, the Yelton et al. formula is seriously flawed.

Observer Agreement under the Models of Quasi-Independence and Quasi-Equiprobability

As mentioned in the previous sections, measures such as the percentage of agreement, Cohen's kappa, and phi have been used to measure observer agreement, but these coefficients have limitations. Fortunately an alternative to these procedures exist. The application of log linear models and the quasi-independence and quasi-equiprobability concepts for the purpose of measuring observer agreement have the advantages of yielding a measure of agreement with a directly interpretable meaning, correcting for the proportion of "chance" agreement, and providing a directly interpretable coefficient of "no agreement".

Bergen (1980a, 1980b) has developed procedures for measuring observer agreement using the quasi-independence and quasi-equiprobability concepts. The application of these procedures for measuring observer agreement has several advantages. Use of the quasi-independence or quasi-equiprobability concept yields a coefficient of observer agreement that varies between zero and one and measures agreement in terms of the probability that the observers' judgments will agree, as estimated under a quasi-independence model. These procedures may also be used to investigate whether or not a single observational category or specific group of categories is a major contributor to the coefficient of agreement. Finally, systematic occurrences of disagreement between observers may be located and measured.

To assess agreement, the judgments of the observers are organized into a contingency table. Quasi-equiprobability or quasi-independence among the variables comprising this contingency table is measured by testing the hypothesis that a subset of the contingency table cells are equiprobable or independent. The quasi-equiprobability model is recommended for measuring observer agreement when two or three observers are recording the presence or absence of one specific behavior. Observer agreement using the equiprobability model may be calculated completely with hand calculations or with the use of various computer programs (Clogg, 1977; Fay and Goodman, 1973). The quasi-independence model is recommended for measuring observer agreement when two or more observers are recording three or more response categories. These measurements of observer agreement require the calculation of maximum likelood estimates of expected cell frequencies under the models of quasi-independence and quasi-equiprobability. To obtain the estimates for each cell in the contingency table, the probability of occurrence of each cell must be calculated. The maximum likelihood estimates of expected cell frequencies are subsequently calculated by multiplying the cell probabilities by N, the number of observations being analyzed. Furthermore, the calculation of maximum likelihood estimates of expected cell frequencies under the model of quasi-independence requires an iterative procedure, thus the use of a high speed digital computer is recommended.

Quasi-equiprobability and quasi-independence models were first developed by Leo Goodman (1975) and applied to the assessment of observer agreement by John Bergan (1980a, 1980b) and Clifford Clogg (1979). Goodman first employed the quasi-independence concept in his response scaling techniques. Bergan (1980a) and Clogg (1979) subsequently introduced the observer agreement measure by adopting a version of Goodman's quasi-independence methods. The following chapter on univariate models will present in detail the quasi-equiprobability and quasi-independence models, along with Bergan's (1980a, 1980b) procedures for measuring observer agreement under the models of quasi-equiprobability and quasi-independence.

Another advantage to employing log-linear models and the independence and equiprobability concepts is the option for computing the variance ($S_{\bar{B}}^{2}$) for the estimated Beta parameters within the log-linear model (Goodman, 1978). The parameters within the log-linear model correspond to the main and interaction effects that describe the categorical data. For each of these estimated Beta parameters, a "standardized value" of the estimate may be computed by dividing each estimated Beta parameter by its estimated standard deviation ($S_{\bar{B}}$). Additionally, each of these standardized Beta values can be tested to see whether the original Beta parameter is nil. Goodman (1978) postulates, given a large sample size, if a specific Beta parameter is nil, then the standardized value of the estimated Beta parameter will be approximately normally distributed with a mean of zero and a variance of one. Finally, variance estimates also allow for the computation of parameter confidence intervals.

Repeated Measurement of Categorical Data

The development of a multivariate extension of the Bergan (1980a, 1980b) procedure for measuring agreement required consideration of the procedures, developed by Grizzle, Starmer, and Koch (1969) and Koch, Landis, Freeman, Freeman, and Lehnen (1977), for analyzing multivariate categorical data obtained from repeated measurement experiments. Following the formulation of specific hypotheses, these researchers developed corresponding test statistics based on the application of the weighted least squares methods. Koch and his associates emphasize generating and testing models and hypotheses based on the weighted regression method because computational procedures are greatly simplified and linear models are more easily conceptualized.

Koch et al. (1977) recommended applying a noniterative procedure for fitting the experimental functions to a linear model, for testing the model-data fit, and for testing hypotheses regarding the linear model parameters. Given a repeated measurement experiment, conceptualize every subject being observed for _D_ points in time or under _D_ different conditions, and rated across _L_ behavioral categories. This design enables $r = L^D$ multivariate response profiles to be generated. If no differences exist among the _D_ points in time or D experimental conditions, then the hypothesis of first order marginal symmetry (homogeneity) will be retained. Specifically, this hypothesis tests the equality of the sums of the row and column pairs (e.g., row 1 ≅ column 1; row 2 ≅ column 2, etc...).

Koch and his colleagues cautioned investigators to keep several issues in mind when working with repeated measurement designs. First, they noted that with repeated measurement designs, there is a lack of independence among responses to the different measurement conditions.

This is caused by "subject effects" that influence measurements for the same subject under two different conditions to have more resemblance than measurements on different subjects under two different conditions. Second, the association of subject response measures across time is affected by the proximity of the measures; measurements closer together in time have greater associations with one another than measurements taken further apart.

Given repeated measurement designs, multiway contingency tables must be constructed. The third issue, according to Koch and his associates, is that with large tables moderate sample sizes may be more easily accommodated with the weighted least squares approach rather than with the maximum likelihood procedures. Weighted least squares estimates are based on first order marginal probabilities or mean scores, whereas maximum likelihood estimators require joint probabilities and observed frequencies. Thus, larger sample sizes are required by the maximum likelihood procedure, in comparison to the least squares method. This need for a larger sample size is further evidence by the maximum likelihood procedures greater sensitivity to zero cell frequencies. The fourth issue, according to Koch and his associates, is that the maximum likelihood procedure requires the manipulation of the entire multiway contingency table. Conversely, the least squares approach only requires the calculation of the first order marginal probabilities and mean scores from the observed data.

Based on the procedures developed by Grizzle et al. (1969) and Koch et al. (1977), Landis and Koch (1977) generated observer measures that test inter-observer bias with the hypothesis of first-order marginal symmetry (homogeneity) and measure inter-observer agreement with a generalized weighted kappa statistic. The advantages and disadvantages of applying the kappa statistic apply with their procedures also. Although the kappa statistic allows the researcher to not be concerned about zero cell frequencies when measuring agreement, a meaningful interpretation of the k statistic is still lacking.

Reliability Measures for Criterion Referenced Tests

Swaminathan, Hambleton, and Algina (1973) Huynh (1976) and Subkoviak (1976) stated the intention for criterion-referenced testing in objective-based instructional sequences is to categorize students into mastery or nonmastery classifications for each objective incorporated into a test. Given a dichotomous classification system of mastery and nonmastery categories, it is justifiable to regard the consistency of mastery-nonmastery decisions for repeated test administrations as a reliability measure. In fact, criterion-referenced test reliability may be regarded as a measure of agreement between mastery-nonmastery classifications assigned in repeated test administrations (Swaminathan et al., 1973). Furthermore, if a test purports to measure several objectives, it is critical to assess the reliability for each subtest assessing a specific objective.

Swaminathan et al. (1973) proposed applying the kappa statistic for measuring the reliability of criterion-referenced tests. By categorizing the students into one of B mutually exclusive mastery classifications during each test administration, they stated it was permissible to apply the kappa statistic to the mastery state data and produce a kappa measure of the proportion of consistent classifications beyond that attributable to chance.

The Swaminathan et al. reliability procedure required two test administrations, whereas the reliability procedures developed by Hunyh (1976), Marshall and Haertel (1976), and Subkoviak (1976) called for only one test administration. Huynh's reliability procedure included the calculation of the mean, standard deviation, Kuder-Richardson 21 coefficient, and alpha parameter, and the determination of a cut-off score. The alpha parameter, along with information regarding the Beta parameter and number of test items, were necessary for the specification of the shape of the joint determination of scores on the first and second test forms. These values were then incorporated into a generalized form of the kappa statistic to produce a value indicative of the proportion of consistent classifications.

Subkoviak (1976) developed an analogous reliability measure, based on one test administration. He substituted assumptions for the second test administration that included (1) the parallel test scores were independently distributed and (2) the observed scores for each person formed a binomial distribution. Given these assumptions, Subkoviak also generated a generalized kappa measure of consistent classifications.

The Marshall-Haertal method (1976) is similarly based on a single administration of a criterion-referenced test and on the assumption that if examinees were retested, the observed score distribution for each examinee would form a binomial distribution. The procedure estimates examinees' scores for a test twice the length of the original test and calculates the proportion of consistent classifications on two tests, from the average of the various split-half estimates.

Although the Swaminathan procedure is simple to compute, it requires two test administrations and inflates the errors of estimation for moderate sample sizes. Conversely, the Huynh, Subkoviak, and Marshall procedures require one administration of the criterion-referenced test and have smaller standard errors for moderate sample sizes. However, these latter three procedures also have biased agreement estimates for short tests, are computationally more complex, and rely on the kappa statistic (i.e., Huynh and Subkoviak).

Decisions regarding mastery classifications and hence the reliability of criterion-referenced tests are influenced by factors such as the way mastery classifications are made, cutting or criterion scores, length of the test, and subject heterogeneity and test score variability. Thus, reliability measures of mastery consistency are situation specific and require accompanying specifications of cutting scores, student skills, and test length (Swaminathan et al., 1973).

Measuring the consistency of dichotomous mastery classifications across repeated testings is analogous to assessing agreement of behavioral assignments across several observers. Therefore, one could suggest that the multivariate measures that were developed for the present study may be applied to assess the reliability of criterion-referenced tests.

CHAPTER 3

UNIVARIATE MODELS OF AGREEMENT

The following sections will discuss the models of independence, quasi-independence, equiprobability, quasi-equiprobability, and symmetry. In addition, model comparison strategies will be discussed as a means of assessing agreement. Following these comparison strategies will be a presentation of procedures for estimating the magnitude of agreement under the models of quasi-independence, quasi-equiprobability, and symmetry. To facilitate the agreement computations under the model of quasi-independence, there will be a discussion of how to apply the Deming-Stephens iterative procedure. Finally, the Kappa statistic is described and recommended to the researcher if certain statistical assumptions cannot be met.

Independence and Quasi-Independence Models

The models of quasi-independence are best illustrated by associating them with the model of independence. If the judgments of two observers, A and B, are organized into a two-dimensional I x I contingency table, the rows in the table will represent the first observer's responses, 1 to I, and the columns will represent the second observer's responses, 1 to I. Cell frequencies within the contingency table are labeled with f's. For instance, f_{22} represents the frequency with which both observers coded the second response category. An inspection of the table will reveal that observer agreement frequencies are represented in the diagonal cells.

A test of independence of responses by two observers, depicted in a two-dimensional table, may be portrayed with the model $\pi_{ii} = \pi^{A}_{i} \times \pi^{B}_{i}$. The symbol π_{ii} represents the probability of occurrence of cell ii, π^{A}_{i} represents the probability of occurrence of variable A at level i and π^{B}_{i} represents the probability of occurrence of variable B at level i. The calculation of maximum likelihood estimates of expected cell frequencies for the test of independence are based on the aforementioned mathematical model. These estimates are computed by multiplying the cell probabilities by the total frequency of observations represented in the table (N). The model under investigation "fits" the data if the maximum likelihood estimates of expected cell frequencies conform closely to the observed cell frequencies. The likelihood-ratio statistic tests the fit of the data and model hypothesizing independence between observer responses.

Quasi-independence among variables comprising a contingency table is measured by testing the hypothesis that a subset of the contingency table cells are independent (Bishop, Fienberg, and Holland, 1975). By eliminating specific cells from the initial contingency table it is possible to segregate critical cells that account for association between the variables. The actual process of eliminating cells from the contingency table refers to placing structural zeros within the critical cells. Structural zeros are created by constraining expected cell frequencies to be equal to observed values. Setting estimates of expected frequencies equal to observed frequencies achieves this constraint and does not contribute to the value of the likelihood-ratio chi-square statistic. A demonstration of how structural zeros do not contribute to the chi-square value can be shown by applying the following likelihood-radio statistic:

$$X^2_L = 2 \sum (\text{observed}) \log \frac{\text{Observed}}{\text{Expected}}$$

For example, if the diagonal cell's observed and expected values were set equal, the quantity for the portion of the formula log (observed/ expected) would be zero for all of the diagonal cells. Therefore, placing structural zeros in the diagonal cells would eliminate any contribution to the chi square value by the diagonal cells.

To test the hypothesis of quasi-independence it is mandatory that an algorithm called iterative proportional fitting be used to estimate the maximum likelihood expected cell frequencies. This procedure establishes preliminary estimates of the expected values and successively adjusts them until they meet the criterion that the marginal totals for the estimated frequencies is equal to the marginal totals for the observed values. The expected and observed marginal totals, in an incomplete table with structural zeros in the diagonal, will converge only if the following assumption is met:

$$X_{i+} + X_{+i} < N$$

where X_{i+} is the sum of the frequencies in non-structural-zero cells in row i, X_{+i} is the sum of the frequencies in non-structural-zero cells in column i, and N is the sum of the frequencies in all of the non-structural zero cells (Bishop, Feinberg, and Holland, 1975).

Once the maximum likelihood expected frequencies are calculated, the likelihood-ratio chi-square statistic may be used to assess independence among the non-structural zero cells. Degrees of freedom for the model of quasi-independence are determined by subtracting from the total number of contingency table cells the number of cells with structural zeros, one for the sample size constraint, and the number of independent parameters.

Equiprobability and Quasi-Equiprobability Models

The models of quasi-equiprobability are best illustrated by associating them with the model of equiprobability. Equiprobability among the variables comprising a contingency table is measured by testing the hypothesis that the probability of occurrence is identical for all cells in the contingency table. To calculate the estimates of expected cell frequencies ($\hat{F}_{ij}$) under the model of equiprobability, for a 2 x 2 table representing the dichotomous judgments of two observers, the following formula would be applied:

$$\hat{F}_{ij} = \frac{N}{C}$$

The estimated frequency of the cell ($\hat{F}_{ij}$) is denoted with observer A responses which occur at level i (i = 1,2), and observer B responses which occur at level j (j = 1, 2). As previously indicated, N refers to the sample size or total number of responses and C refers to the number of cells being considered. With the equiprobability model, the sample size imposes a single constraint on the estimates of expected cell frequencies. Therefore, a model of equiprobability for a 2 x 2 contingency table would have 3 degrees of freedom.

As with models of quasi-independence, imposing structural zeros within critical cells creates models of quasi-equiprobability which test the hypothesis of equiprobability among remaining response patterns. Estimates of cell frequencies, under models of quasi-equiprobability, are calculated with the follwing formula:

$$\hat{F}_{ij} = \delta_{ij} \, N_s/C_s$$

where $\delta_{ij} = \begin{matrix} 1 \text{ if } (i,j) \, \varepsilon S \\ 0 \text{ otherwise} \end{matrix}$, where $\underline{S}$ is the subset of cells remaining following the insertion of structural zeros and $\underline{C}_s$ equals the number of cells considered within the subset $\underline{S}$.

Symmetry Model

Given a two-way square (I x J) contingency table, the model of symmetry disregards the diagonal cells in the table and examines the relationship of the pairs of cells around the diagonal. The model focuses on the joint probabilities, P_{ij}, where $i \neq j$. Under this model, the hypothesis of symmetry can be tested, stated in terms of the joint probabilities as H_0: $P_{ij} = P_{ji}$, for all $i \neq j$.

With the model of symmetry, structural zeros are imposed on the diagonal cells and the estimated expected cell values are calculated with the following formula:

$$\hat{F}_{ij} = \frac{X_{ij} + X_{ji}}{2} \qquad \text{where } i \neq j$$

$$= X_{ij} \qquad \text{where } i = j.$$

The model of symmetry, as defined by the expression $\hat{F}_{ij} = \hat{F}_{ji}$ for all $i = j$, implies marginal homogeneity. The model of marginal homogeneity is denoted by the expression $F_{i+} = F_{+j}$, where $i = 1$ to I and $j = 1$ to J. The plus (+) sign signifies that the frequencies are summed across the columns and rows, respectively, following the elimination of the diagonal frequencies. Thus, the model of marginal homogeneity asserts the sum of the first row equals the sum of the first column and each subsequent row and column sum equal one another.

The degrees of freedom under the model of symmetry are calculated by subtracting the number of diagonal cells and number of pairs of off-diagonal cells from the total number of cells in the table. Therefore, a model of symmetry for a 3 x 3 table would have 3 degrees of freedom. By applying the likilihood-ratio chi-square statistic the hypothesis of symmetry may be tested.

Assessing Agreement by Comparing Models of Independence and Quasi-Independence

Since information regarding agreement by two observers is located within the diagonal cells in the contingency table, disagreement in the table may be assessed with a chi-square test of quasi-independence with diagonal cells deleted (Bishop, Feinberg, and Holland, 1975). By applying the chi-square test of independence, which measures agreement and disagreement, a baseline model can be formed. Statistical tests measuring the significance of the diagonal cells contribution to agreement may be conducted by subtracting the chi-square values for assorted hierarchical tests of quasi-independence from the chi-square value for the test of independence. Goodman (1975) defined two models as hierarchically related if the subordinate model possessed all of the constraints of the superordinate model in addition to one or more further constraints. For instance, the model of independence is hierarchically related to a model of quasi-independence with the diagonal cells deleted. With hierarchical models the superordinate model implies the subordinate model. Therefore, the model of independence implies the model of quasi-independence. If the model of independence fits the data (i.e., has a statistically nonsignificant chi-square value), then the model of quasi-independence would also fit the data.

The advantage of the likelihood-ratio chi-square statistic lies in its ability to be partitioned exactly into independent component chi-squares and summed to achieve the overall contingency table chi-square and degrees of freedom (Cochran, 1954). This property allows the independence model and chi-square to be partioned into component chi-squares such as the chi-square for the test of quasi-independence and the chi-square indicating the difference between the independence and quasi-independence values. Subtracting the chi-square value and related degrees of freedom for a test of quasi-independence with all the diagonal cells eliminated from the chi-square value and related degrees of freedom for the test of independence would provide a chi-square value that would measure if the diagonal cells provide a significant contribution to the association in the contingency table.

By applying the aforementioned model comparison procedures, the specific chi-square contribution of a single agreement cell or subset of agreement cells can be assessed. An experimeter wishing to investigate the contribution of each agreement category represented in the diagonals of a 3 x 3 table may accomplish this by using several different quasi-independence models and compare them with the independence model. For example, the investigator could set up three quasi-independence models each ruling out one of the diagonal cells f_{11}, f_{22}, and f_{33}, respectively. Each of the chi-square values for these independence models could be subtracted from the chi-square value for the independence model to test if the specific cell provided a significant contribution to model-data fit. If a model of quasi-independence ruled out a single cell such as f_{11}, and the difference between the chi-square values for the quasi-independence and independence models had a value of 3.84 (critical value for 1 degree of freedom) or larger than the contribution of that cell to agreement would be statistically significant. If the difference chi-square value was less than 3.84 then the investigator could not conclude that the observers' judgments agreed for the first behavioral category, regardless of the number of agreement frequencies in the f_{11} cell.

Investigators may also find that an off-diagonal disagreement cell provides a significant contribution to the overall chi-square value. A case of systematic disagreement between observers may occur if observer A codes a specific behavior in the first category and observer B codes that behavior in the second category. To test if a significant association between the observers' responses exists, a quasi-independence model may be developed which places structural zeros in the hypothesized cell or cells denoting systematic disagreement. The chi-square value for the quasi-independence model is subtracted from the chi-square value for the independence model to test the statistical significance of the association.

Assessing Agreement by Comparing Models of Equiprobability and Quasi-Equiprobability

The model comparison procedures described in the previous paragraphs may also be used with equiprobability and quasi-equiprobability models. By using the model of equiprobability as a baseline and constructing a model of quasi-equiprobability with the agreement diagonal cells deleted, an investigator can test the improvement in model-data fit provided by the agreement cells and find the difference between the models' chi-square values.

Two other hierarchical quasi-equiprobability models may also be constructed by ruling out the f_{11} or f_{22} cells, respectively. A hierarchical sequence of models could then be established. A model of equiprobability (three degrees of freedom) would be hierarchical to the two models of quasi-equiprobability with a structural zero in the f_{11} or f_{22} cell (two degrees of freedom each), respectively. The two previously mentioned models would in turn be hierarchically related to the agreement quasi-equiprobability model with structural zeros in the f_{11} and f_{22} cells (one degree of freedom). Differences may be computed between the f_{11} or f_{22} models and agreement model (f_{11} and f_{22} cells deleted) to determine the improvement in model-data fit afforded by the respective models.

Estimating the Magnitude of Agreement under the Model of Quasi-Independence

A quasi-independence model and the maximum likelihood estimates of probabilities for agreement and disagreement may be used for the computation of the degree of agreement between observers. Goodman's (1975) work with response scaling and Bergan (1980a) have demonstrated that from models of quasi-independence, with structural zeros in the cells representing agreement, maximum likelihood probability estimates may be calculated for the agreement cells. In addition, the off-diagonal or disagreement cells may also have a probability estimate computed. The precision of the probability estimates is based on the model fitting the data. Therefore, a chi-square value for a quasi-independence model must be statistically non-significant to indicate an appropriate model-data fit. Given a 3 x 3 table with observations for three behavioral categories, the maximum likelihood estimates representing agreement and disagreement would be expressed in four classifications. The first three classifications would represent each of the diagonal cells, respectively. The fourth classification would represent the six cumulative off-diagonal disagreement cells.

Goodman (1975) assumed that the off-diagonal observer disagreement responses (i.e., cells without structural zeros) were independent. He also assumed the expected response pattern for each diagonal cell with a structural zero had a probability of 1. Given these assumptions, the following formula computes the maximum likelihood estimate for the probability that observers' A and B responses would be represented within the disagreement category:

$$\hat{\pi}_o = \hat{\pi}^{AB}_{ij} / \hat{\pi}^{\bar{A}}_{io} \hat{\pi}^{\bar{B}}_{jo} \qquad (3.1)$$

where $\hat{\pi}_o$ is the estimated probability of disagreement between observers, $\hat{\pi}^{AB}_{ij}$ is the estimated probability of a disagreement response F_{ij} $(i \neq j)$ for both observers; $\hat{\pi}^{A}_{io}$ is the conditional probability of observer A emitting response i, assuming the observers' ij response pattern denotes a disagreement between the observers; and $\hat{\pi}^{B}_{jo}$ is the conditional probability of observer B emitting response j. The probability of a specific agreement category t is expressed with the following maximum likelihood estimate:

$$\hat{\pi}_t = p_{ij} - \hat{\pi}_0 \hat{\pi}^{\bar{A}}_{io} \hat{\pi}^{\bar{B}}_{jo}$$

where p_{ij} is the observed proportion of a specific observer agreement category t as designated in the ij cell, $\hat{\pi}^{A}_{io}$ is the maximum likelihood estimate of observer A's response i given the disagreement category 0 and $\hat{\pi}^{B}_{jo}$ is the maximum likelihood estimate of observer B's response j given the disagreement category.

In formula (3.1) the $\bar{\pi}^{AB}_{ij}$ value is calculated by dividing the response pattern ij expected cell frequency ($\bar{F}_{ij}$) by the total number of observer responses (N). The computation of the $\bar{\pi}^{A}_{io}$ and $\bar{\pi}^{B}_{jo}$ values require the expected cell frequencies and use of the following formula for polytomous variables:

$$\bar{\pi}^{A}_{io} = \Omega^{A}_{i}/i'o\left(\sum_{i=1}^{I} \Omega^{A}_{i}/i'0\right)$$

where $\Omega^{A}_{i}/i'o$ represents the odds of an i disagreement response to an i' disagreement response, by observer A. These odds are obtained from the following estimated expected cell frequencies:

$$\Omega^{A}_{i}/i'o = \hat{F}_{ij}/\hat{F}_{ij'}$$

where ij and ij' are disagreement response patterns.

Goodman's work with response scaling and models of quasi-idenpendence demonstrated that the probability for the agreement (i.e., structural zero cells) and disagreement (i.e., non-structural zero cells) categories add to one. Therefore, the estimated proportion of the sum of the agreement cells equals one minus the probability for the disagreement cells. By placing structural zeros in the agreement/diagonal cells within a contingency table signifying the response distribution of two observers, an index of the magnitude of observer agreement can be developed. The following formula connotes the magnitude of observer agreement as the estimated probability that judgments from two observers will occur in one of the agreement categories ($\bar{\pi}_{A}$):

$$\hat{\pi}_A = 1 - \hat{\pi}_0$$

where $\hat{\pi}_o$ is the estimated probability that a pair of judgments from the observers will occur in the disagreement category.

Estimating the Magnitude of Agreement under the Model of Quasi-Equiprobability

Bergan (1980b) has also developed procedures to measure the magnitude of observer agreement under the model of quasi-equiprobability. As with the measure of observer agreement under the model of quasi-independence, the calculation of the probability of observer agreement under the model of quasi-equiprobability requires that one first compute the estimated probability of observer disagreement. Observer disagreement is evidenced if the pair of judgments occur in the disagreement category designated by the off-diagonal cells. As defined under the model of quasi-independence, the estimated probability of disagreement (π_o), under the model of quasi-equiprobability is expressed by:

$$\hat{\pi}_o = \hat{\pi}^{AB}_{ij} / \hat{\pi}^{\bar{A}}_{io}\ \hat{\pi}^{\bar{B}}_{jo}$$

where $\hat{\pi}^{AB}_{ij}$ is the estimated probability of a pair of observer judgments in the f_{12} or f_{21} off-diagonal cell, $\hat{\pi}^{\bar{A}}_{io}$ is the estimated conditional probability of observer A emitting response i given the disagreement category, where $\hat{\pi}^{\bar{B}}_{jo}$ is defined in a parallel fashion.

Given the model of quasi-equiprobability, the obervers' responses are dichotomously categorized. Such categorization results in the assignment of the value of .5 to $\hat{\pi}^{A}_{io}$ and $\hat{\pi}^{B}_{jo}$. The formula for the estimated probability of disagreement reduces to the folloiwng expression:

$$\hat{\pi}_o = \hat{\pi}^{AB}_{ij}/.25$$

where $\hat{\pi}^{AB}_{ij}$ is estimated by $\hat{F}_{12}/N$ or $\hat{F}_{21}/N$.

Under the model of quasi-equiprobability the probability of observer agreement $\hat{\pi}_A$ is similarly defined as:

$$\hat{\pi}_A = 1 - \hat{\pi}_o$$

As under the model of quasi-independence, it is possible to measure the probability of a specific agreement category t under the model of quasi-equiprobability. Since there are only two diagonal cells (f_{11} and f_{22}) and the disagreement category (expresed as 0), t will range from 0 to 2. The estimated probability for a specific agreement category f_{ij} (f_{11} or f_{22}) is expressed by:

$$\hat{\pi}_t = P_{ij} - \hat{\pi}_o\ (.25)$$

where $\hat{\pi}_t$ is the estimated probability of agreement category t (1 or 2) and P_{ij} is the observed proportion of a specific observer category t. By subtracting the expression $\hat{\pi}_0$ (.25) from P_{ij} the value $\hat{\pi}_t$ is corrected for chance agreement.

Latent Agreement

The agreement models of quasi-independence and quasi-equiprobability may also be conceptualized as latent class models of agreement (Bergan, 1980a; Clogg, 1979). Latent class models assert observable variables measure latent or unobservable characteristics, and are taken as indicators of that characteristic's value. Each latent variable is comprised of latent classes, representing each agreement category within the diagonal of the contingency table and a latent class representing disagreement within the off-diagonal cells. For example, a latent class model of quasi-equiprobability would contain three latent classes, one for each of the two agreement categories within the diagonal and one for the disagreement category.

The first latent class assumes that:

$$\pi^{\bar{A}X}_{11} = \pi^{\bar{B}X}_{11} = 1,$$

where $\pi^{\bar{A}X}_{11}$ is the probability of observer A responding in category 1 given the observation is contained in the first latent class of latent variable X and $\pi^{\bar{B}X}_{11}$ is the probability of observer B responding in category 1 given the observation is contained in the first latent class of latent variable X. The first latent class portrays agreement in the f_{11} cell. The second latent class declares that:

$$\pi^{\bar{A}X}_{22} = \pi^{\bar{B}X}_{22} = 1,$$

and portrays agreement in the f_{22} cell. The third latent class represents disagreement within the two off-diagonal cells. Given the model of quasi-equiprobability, the disagreement latent class asserts that:

$$\pi^{\bar{A}X}_{10} = \pi^{\bar{B}X}_{10} = .5 \quad .$$

For both the models of quasi-equiprobability and quasi-independence, the estimated probability $\hat{\pi}^{X}_{t}$ of a specific latent class t in the latent variable X is estimated with the following equation:

$$\hat{\pi}^{X}_{t} = \sum_{ij} \hat{\pi}^{ABX}_{ijt}$$

where $\hat{\pi}^{ABX}_{ijt}$ is the estimated joint probability of observer A's response i, observer B's response j, and latent class t. By applying Goodman's (1974) iterative procedure the estimated joint probability $\hat{\pi}^{ABX}_{ijt}$ may be computed. These estimates may also be generated by using Clogg's (1977) computer program which computes estimates of expected cell frequencies and $\hat{\pi}^{X}_{t}$ for various latent class models using the iterative procedure.

Applying the Model of Quasi-Equiprobability with Large Contingency Tables

The quasi-equiprobability and quasi-independence procedures described in the previous sections were primarily recommended for conditions where two observers' responses were categorized across two (quasi-equiprobability) or three or more (quasi-independence) behavioral categories. Bergan's (1980b) algorithms for measuring observer agreement using the quasi-equiprobability model, representing two behavioral categories, may be calculated completely with hand calculations or with the use of various computer programs (Clogg, 1977; Fay and Goodman, 1973). The calculation of maximum likelihood estimates of expected cell frequencies under the model of quasi-independence requires an iterative procedure and preferably the use of a high-speed digital computer. If access to a computer is not available, if the off-diagonal contingency table cells (i.e., for 3 x 3 or larger tables) have many zero frequencies and/or the model does not fit the data the models and procedures developed by the author and described in the following paragraphs are recommended.

The model of quasi-equiprobability may be conceptualized with the following expression:

$$P_{ij} = \begin{matrix} \pi_k + \pi_o \pi^A_i \pi^B_j & k \rightarrow (i,j) \text{ diagonal cell} \\ \pi_o \pi^A_i \pi^B_j & \text{otherwise} \end{matrix}$$

where P_{ij} is the observed probability for cell ij, π_k is the parameter for the respective diagonal cells, π_o represents measurement error within each cell, π^A_i represents the parameter for rows 1 to i, and π^B_j represents the parameter for columns 1 to j. Under the model of quasi-equiprobability, the parameter π^A_i equals 1/I since all the alpha (α) values, 1 to I, are equivalent. In addition, the parameter π^{B}_{j} equals 1/J since all the Beta (β) values, 1 to J, are equivalent. A 2 x 2 quasi-equiprobability model may then be considered with the parameters given by:

$$\pi_1 + \pi_o\tfrac{1}{4};\ \pi_2 + \pi_o\tfrac{1}{4};\ \pi_o\tfrac{1}{4};\ \pi_o\tfrac{1}{4}$$

where the first two expressions symbolize the two diagonal cells and the second two expressions reflect the two off-diagonal cells.

The expressions in the previous sentence may be rewritten and incorporated into the following condensed table:

Diagonal Cell 1	Diagonal Cell 2	Off-Diagonal Cells
$\pi_1 + \frac{1}{4}\pi_o$	$\pi_2 + \frac{1}{4}\pi_o$	$\frac{1}{2}\pi_o$

with the three cell probabilities summing to equal one. In this model all off-diagonal cells are combined into one cumulative cells. Thus, larger contingency tables may also be analyzed under the model of quasi-equiprobability. A 3 x 3 table, with two observers judging three behavioral categories, could be condensed into the table given by:

Cell f_{11}	Cell f_{22}	Cell f_{33}	Cumulative Off-Diagonal Cells
$\pi_1 + \frac{1}{9}\pi_o$	$\pi_2 + \frac{1}{9}\pi_o$	$\pi_3 + \frac{1}{9}\pi_o$	$\frac{6}{9}\pi_o$

where each of the six off-diagonal cells are hypothesized to have the same probability of occurrence.

A general table may be established by:

R

$\pi_1 + \frac{1}{K}\pi_o$	$\pi_2 + \frac{1}{K}\pi_o$	$\pi_3 + \frac{1}{K}\pi_o$	......	$\frac{K - R}{K}\pi_o$

where R is equal to the number of cells in the diagonal and K is equal to the total number of cells in the two-way table $({}_R{}^2)$. Furthermore, all of the probabilities in the above table sum to one. Maximum likelihood estimates under the model of quasi-equiprobability may be generated and incorporated into the algorithms for the estimated probability of agreement and disagreement. Maximum likelihood estimates of probabilities of agreement within each of the agreement categories may be generated by:

$$\underset{(i=j)}{\pi_{ij}} = \frac{(K - R)\ X_{ij} - X_o}{(K - R)\ N}$$

where K is the total number of cells in the two-way table, R is the number of cells in the diagonal, X_{ij} is the observed frequency within a single diagonal cell $_{ij}$, X_o is the total number of off-diagonal frequencies, and N is the total number of judgments in the complete two-way table (agreements and disagreements). Thus a maximum likelihood probability estimate may be computed for every agreement category within the table. Furthermore, a probability estimate of disagreement may be calculated by:

$$\pi_o = \frac{KX_o}{(K - R)N} = \frac{R}{R - 1} \left(\frac{X_o}{N}\right) .$$

The probabilities for each of the agreement categories (π_{ij}) and the disagreement category (π_o) sum to one. Therefore, the probabilities for the agreement cells may be summed and indicate the magnitude of agreement in a directly interpretable fashion.

Simple Observer Agreement Calculations

When an investigator does not have access to a computer and wishes to hand calculate the magnitude of observer agreement and/or has a set of data with many zero frequencies in the off-diagonal cells, the following procedures will provide the most accurate measures. First, set up a two-way contingency table and test the fit of the quasi-equiprobability model described in the previous section. If the model fits the data, also apply the procedures for measuring the magnitude of observer agreement recommended in the previous section. The quasi-equiprobability model is recommended as the first model to be tested because it has the most degrees of freedom and least amount of structure imposed by the parametric model. If the quasi-equiprobability model does not fit the data the model of symmetry should be tested next. If this model fits, then Bergan's (1980a) procedures for computing the magnitude of observer agreement may be used. An unacceptable model-data fit would next warrant testing the model of quasi-independence. To derive the expected cell frequencies under the model of quasi-independence, an iterative process must be applied. The following section will describe the Demming-Stephens iterative fitting program (Feinberg, 1978) for obtaining expected cell frequencies with hand calculations.

The three models mentioned in the previous paragraph impose varying constraints and estimate different off-diagonal parameters. In the case of a 3 x 3 table with three behavioral categories, the following expressions illustrate the various ways the models estimate the off-diagonal cell probabilities;

<u>Quasi-equiprobability</u>

5 degrees of freedom

$$P_{ij}\begin{cases}\frac{1}{6} & (i,j)\epsilon S\\ 0 & (i,j)\not\epsilon S\end{cases}$$

<u>Symmetry</u>

3 degrees of freedom

$$P_{ij}\begin{cases}\alpha_i\alpha_j & (i,j)\epsilon S\\ 0 & (i,j)\not\epsilon S\end{cases}$$

<u>Quasi-independence</u>

1 degree of freedom

$$P_{ij}\begin{cases}\alpha_i\beta_j & (i,j)\epsilon S\\ 0 & (i,j)\not\epsilon S\end{cases}$$

where S signifies the subset of off-diagonal cells. The expression for the model of quasi-equiprobability asserts the off-diagonal cells have the same probability of occurrence. The probability is determined by only the number of off-diagonal cells. Symmetry implies the joint probabilities for the pairs of cells around the diagonal are equiprobabile. For instance, cell f_{12} is equiprobable with cell f_{21} and cell f_{13} is equiprobabile with cell f_{31}. Each pair of off-diagonal cells imposes an additional constraint upon the data. Cell probabilities under the model of quasi-independence are expressed as the product of factors that are functions of two marginals.

Iterative Computations of Expected Frequencies

The estimated expected frequencies under the model of quasi-independence must be computed by an algorithm called iterative proportional fitting. In the Deming-Stephens (Feinberg, 1978) algorithm, preliminary estimates of the expected values are made, then successively adjusted until they meet the criterion that the marginal totals for the estimated frequencies equal the marginal totals for the observed values. Therefore:

$$\hat{F}_{i+} = f_{i+} \text{ and } \hat{F}_{+j} = f_{+j}$$

for all i and j in the subset of off-diagonal cells. Since the diagonal cells contain structural zeros under the agreement model of quasi-independence, the frequency summations are only across non-structural-zero cells.

Let $\hat{F}_{ij}$ equal the expected frequency of the (i,j)th cell, with X_{ij} equal to the observed frequency. Let us also assume that X_{i+} or $\hat{F}_{i+}$, etc. ..., refer to the summation across only non-structural-zero cells. The initial start values within the table are denoted as $\hat{F}^{(o)}{}_{ij}$ and the subsequent Kth iteration as $\hat{F}^{(K)}{}_{ij}$.

To hasten the iterative process, rather than insert values of one within the table for start values it is recommended that a proportion of the cell frequencies be estimated from the marginal values and used as the start values $(\hat{F}^{(o)}{}_{ij})$. The procedure sequentially fixes one set of marginal values and allows the other set of marginals to vary. To calculate the proportional start values, consider a 3 x 3 table with the diagonal cells deleted. Begin by fixing the column marginals and allowing the row marginals to vary. Cell $\hat{F}^{(o)}{}_{21}$ is estimated with specific marginal frequencies from the original table:

$$\hat{F}^{(o)}{}_{21} = X_{+1}\left(\frac{X_{2+}}{X_{2} + X_{3+}}\right)$$

Similarly cells $\hat{F}^{(o)}{}_{31}$ and $\hat{F}^{(0)}{}_{13}$ are estimated by:

$$\hat{F}^{(o)}{}_{31} = X_{+1}\left(\frac{X_{3+}}{X_{2+} + X_{3+}}\right)$$

$$\hat{F}^{(0)}{}_{13} = X_{+2}\left(\frac{X_{1+}}{X_{1+} + X_{3+}}\right)$$

Once the start values are calculated, the algorithm proceeds in a two-step manner, with:

$$\hat{F}_{ij}{}^{(K + 1)} = \frac{\hat{F}_{ij}{}^{(K)}\ X_{i+}}{\hat{F}_{i+}{}^{(K)}} .$$

and

$$\hat{F}_{ij}{}^{(K + 2)} = \frac{\hat{F}_{ij}{}^{(K + 1)}\ X_{+j}}{\hat{F}_{+j}{}^{(K + 1)}}$$

The procedure alternates between fixing the row and column marginals. During the iterative process the improvement in the subsequent fits come from the estimated marginals getting nearer to the observed marginals. Following several large estimation leaps initially, the convergence process slows down with smaller estimation changes. Thus, once the investigator gets past the initial estimation leaps, a "reasonable" approximation may be expected without considerable iteration.

Summary and Conditions for the Use of Kappa

The previous sections described a variety of methods for measuring observer agreement. In summary, the quasi-equiprobability model is recommended for measuring observer agreement when two or three observers are recording the presence or absence of one specific behavior. Observer agreement using the equiprobability model may be calculated completely with hand calculations or with the use of various computer programs (Clogg, 1977; Fay and Goodman, 1973). The quasi-independence model is recommended for measuring observer agreement when two or more observers are recording three or more response categories. The analysis includes the calculation of maximum likelihood estimates of expected cell frequencies under the model of quasi-independence which requires an iterative procedure and the use of a high speed digital computer.

If the investigator does not record the presence or absence of a single behavior, have access to a computer to test quasi-independence, or meet the assumptions for quasi-independence, then it is recommended that the fit of the large quasi-equiprobability model, symmetry model, and quasi-independence model may be tested successively. The procedures for hand calculating the measure of observer agreement for the quasi-equiprobability, symmetry, and quasi-independence models were described in the previous sections.

An investigator may encounter situations where all the model-data fits are unacceptable or where there is such an excessive number of zero frequencies in the off-diagonal cells that maximum likelihood estimates of expected cell frequencies cannot be calculated. If such instances occur with the investigator's two-way tables, it is recommended that Cohen's (1960) Kappa (K) coefficient be calculated to measure observer agreement.

Coefficient K represents the proportion of agreement after chance agreement is removed from consideration:

$$K = \frac{P_o - P_c}{1 - P_c}$$

with $P_o = \sum_i P_{ij}$ ($i = j$, only)

and $P_c = \sum_i P_{i+} P_{+j}$ (for $i = j$ cells, only).

Let P_o be the sum of the observed proportion of cases in the main diagonal, P_c be the sum of the expected proportion of cases in the main diagonal, and P_{i+} and P_{+j} be the marginal proportions in the ith row and jth column, respectively.

P_o is not an acceptable measure of agreement, because some agreement cases could be expected to occur in the main diagonal by chance.

Therefore, to correct for chance occurrences P_c was subtracted from P_o in the formula. The value for P_c may be regarded as the sum of the proportions expected under the model of independence. Since P_c depends on marginal totals, it is normalized by dividing by $1 - P_c$, the maximum value possible for $P_o - P_c$ given marginal totals P_{i+} and P_{+j}.

If the observed magnitude of agreement equals the magnitude expected by chance alone then K will equal zero. Complēte agreement between the observers will result in a Kappa of 1.0. A negative Kappa value will occur if the observed agreement is less than the agreement expected by chance.

CHAPTER 4

MULTIVARIATE MODELS OF AGREEMENT

Repeated measurement experiments incorporate data from groups of experimental units each of which is assessed under two or more distinct conditions. Thus, studies which classify each individual with respect to each categorical variable at several successive points in time may be characterized as repeated measurement designs. The data obtained from such designes may be conceptually arranged in multidimensional contingency tables.

The following sections will present a variety of procedures for analyzing categorical data obtained from repeated measurement designs. This is the first such work of this kind. The first three procedures use the quasi-equiprobability concept to measure model-data fit and observer agreement under conditions when two observers are recording the presence or absence of one specific behavior at two points in time. The first two procedures use log linear and latent class models with maximum likelihood statistical methods to estimate and test the model parameters. The third procedure applies the weighted least squares statistical methods to test the hypothesis of equi-probability between specific contingency table marginal values. The fourth and fifth procedures employ the quasi-independence concept to measure model-data fit and degree of observer agreement under conditions when two observers are recording three response categories at two points in time. The fourth procedure uses the maximum likelihood methods to estimate and test the parameters within this 3 x 3 x 2 design. The fifth procedure uses the weighted least squares method to test the hypothesis of independence among the marginal values. Finally, in a concluding section, a discussion is presented about how these procedures may be employed to assess reliability of classifications established from criterion-referenced test scores.

Quasi-Equiprobability and Maximum Likelihood Methods

A repeated measurement design that encompasses two observers categorizing the presence or absence of a single behavior for each subject at two points in time creates a 4 x 4 contingency table. Table 1 illustrates the format for constructing a contingency table from such a repeated measures design. The table incorporates the four response variations of the two observers at time I and time II. The rows in the table represent the time I judgments of the two observers, A and B. For example, category 11 indicates the two observers judged that the behavior occurred at time I and category 12 denotes that observer A judged that the behavior occurred but observer B judged that the behavior did not occur at time I. The columns in the table depict the response variations of the same two observers, referred to as C and D, at time II. Table 1 also contains two cells with .5 frequencies (cells 1221 and 2112). Originally these cells contained a frequency of zero, but .5 was inserted in order to allow subsequent log transformations of the cell values.

Table 1. Observed cell frequencies, expected cell frequencies, and expected frequency formulas under the model that $\overline{AB}$ is quasi-equiprobable, $\overline{CD}$ is quasi-equiprobable, and the saturated model is true.

		Time II				
Observer	C	1	1	2	2	Observed Marginal Values
	D	1	2	1	2	
Observer A	B					
**** 1	1	77** $f_{ijk\ell}$	3 $\frac{f_{ijk\ell}}{f_{..k\ell}}\frac{s^*}{2}$ (2.8421)***	4 $\frac{f_{ijk\ell}}{f_{..k\ell}}\frac{s}{2}$ (4.2533)	22 $f_{ijk\ell}$	106
1	2	2 $\frac{f_{ijk\ell}}{f_{ij..}}\frac{r}{2}$ (2.1818)	1 $\frac{f_{ijk\ell}}{f_{ij..}f_{..k\ell}}\frac{r}{2}\frac{s}{2}$ (1.0335)	.5 $\frac{f_{ijk\ell}}{f_{ij..}f_{..k\ell}}\frac{r}{2}\frac{s}{2}$ (.5775)	2 $\frac{f_{ijk\ell}}{f_{ij..}}\frac{r}{2}$ (2.1818)	5.5
Time I 2	1	2 $\frac{f_{ijk\ell}}{f_{ij..}}\frac{r}{2}$ (1.8462	.5 $\frac{f_{ijk\ell}}{f_{ij..}f_{..k\ell}}\frac{r}{2}\frac{s}{2}$ (.4372)	1 $\frac{f_{ijk\ell}}{f_{ij..}f_{..k\ell}}\frac{r}{2}\frac{s}{2}$ (.9774)	3 $\frac{f_{ijk\ell}}{f_{ij..}}\frac{r}{2}$ (2.7694)	6.5
2	2	39 $f_{ijk\ell}$	5 $\frac{f_{ijk\ell}}{f_{..k\ell}}\frac{s}{2}$ (4.7368)	3 $\frac{f_{ijk\ell}}{f_{..k\ell}}\frac{s}{2}$ (3.1765)	61 $f_{ijk\ell}$	108
Observed Marginal Values		120	9.5	8.5	88	226 Total

* $r = f_{ij..} + f_{i'j'..}$

$s = f_{..k\ell} + f_{..k'\ell'}$

** Observed frequencies

*** Expected frequencies within parentheses

**** An observer's response of 1 signifies the behavior occurred. A 2 response indicates the behavior did not occur.

Since the four corner cells, 1111, 1122, 2211, and 2222 portray observer agreement at each of the two points in time, it may be asserted that a model assessing equiprobability between specific marginals may test observer agreement at the two time periods. Specifically, the agreement model constrains the observed and expected row 1 and row 4 marginal values to be equal, constrains the observed and expected column 1 and column 4 marginal values to be equal, assesses equiprobability between the row 2 and row 3 marginal values, and equiprobability between the column 2 and column 3 marginal values. By constraining the previously mentioned marginal values, the rows and columns (i.e., 12 and 21) representing disagreement between the observers may be investigated. Furthermore, the process of constraining the four observed and expected marginal values to be equal also constrained the observed and expected cell values in the four agreement corner cells to be equal. Thus, a quasi-equiprobability model that measures the remaining observer disagreement cells was created.

This log linear agreement model asserts that the joint variable $\bar{A}\bar{B}$ is quasi-equiprobabile, the joint variable $\bar{C}\bar{D}$ is quasi-equiprobable, and that the saturated model is true. In essence this quasi-equiprobability model states that an agreement model may be primarily formulated from the marginal values. Furthermore, in the following paragraphs it is demonstrated that the X^2 value for the 4 x 4 agreement table will equal the sum of the X^2 values for the two 2 x 2 tables measuring quasi-equiprobability among the row marginals and column marginals, respectively.

The analysis of repeated measures designs focuses on the marginal values. This emphasis on the marginal values transpired because model testing procedures based on maximum likelihood statistical procedures may become quite complex in some cases. Although this issue on how to employ maximum likelihood procedures with repeated measures designs of agreement seemed difficult initially, it should be recognized that hierarchical sets of models may be developed with models that not only incorporate marginals but also models that rely on marginals exclusively. Thus, maximum likelihood procedures utilizing marginals have been developed.

The model asserting the joint variable $\bar{A}\bar{B}$ is equiprobable, the joint variable $\bar{C}\bar{D}$ is equiprobable, and that the saturated model is true is a model that subsumes two hierarchical models that involve marginals. The test that the joint variable $\bar{A}\bar{B}$ is equiprobable sums across the C and D cells and examines the joint $\bar{A}\bar{B}$ marginals. Conversely, the test that the joint variable $\bar{C}\bar{D}$ is equiprobable sums across the A and B cells and examines the joint $\bar{C}\bar{D}$ marginals. The saturated model component allows the reproduction of the large initial table without adding to the large table's chi-square value. This occurs because the chi-square value for the saturated model equals zero. Thus, by adding the two chi-square values for the joint variables ($\bar{A}\bar{B}$ and $\bar{C}\bar{D}$) equiprobability test and the chi-square for the saturated model, the chi-square for the full table may be generated.

The maximum likelihood expected cell frequency formulas within Table 1 are for the model that holds $\overline{AB}$ is the quasi-equiprobable, $\overline{CD}$ is quasi-equiprobable, and that the saturated model is true. These expected cell frequencies and their accompanying X^2 formulas were obtained by applying the following algorithmic procedures.

<u>Saturated model</u>:

$$\hat{F}_{ijk\ell}{}^{(0)} = f_{ijk\ell}$$

<u>$\overline{AB}$ is equiprobable</u>:

$$\text{expected frequency } \hat{F}_{ij..}{}^{(1)} = \begin{cases} \frac{(f_{ij..} + f_{i'j'..})}{2} & \text{for } (i,j)(i',j') \\ & \varepsilon S \\ f_{ij..} \quad \text{otherwise} & \end{cases}$$

The joint variable $\overline{AB}$ is generated by summing across the C and D cells. Similarly, the observed frequency $f_{ij..}$ is produced by summing over the k and ℓ cells. The prime values refer to opposite cell values. For example, if $f_{ij..}$ refers to cell 12, then $f_{i'j'.}$ refers to cell 21. Finally, the notation εS represents a member of set S. An equiprobability test between the 12 and 21 marginals would include only the 12 and 21 marginals within set S. Thus, the marginals 11 and 22 would not be included in set S. The notation below for the joint variable $\overline{CD}$ is similarly defined.

<u>$\overline{CD}$ is equiprobable</u>:

$$\text{expected frequency } \hat{F}_{..k\ell}{}^{(2)} = \begin{cases} \frac{(f_{..k\ell} + f_{..k'\ell'})}{f_{..k\ell} \quad \text{otherwise}} \text{for } i,j)(i',j')\varepsilon R \end{cases}$$

The three chi-square formulas below test the three separate components of the model; the saturated model is true, the joint variable $\overline{AB}$ is equiprobable, and the joint variable $\overline{CD}$ is equiprobable. The notation ℓn refers to the natural log.

$$X^2 H_0 = 2 \sum f_{ijk\ell} \ \ell n \left[f_{ijk\ell}/f_{ijk\ell}\right]$$

$$X^2 H_1 = 2 \sum (f_{ij11} + \ldots + f_{ij22}) \ \ln \left[f_{ij..}/\hat{F}_{ij..}\right]$$

$$X^2 H_2 = 2 \sum (f_{11k\ell} + \ldots + f_{22k\ell}) \ \ln \left[f_{..k\ell}/\hat{F}_{..k\ell}\right]$$

The following chi-square formula sums the above three chi-squares and obtains the chi-square formula for testing the values in the large table.

$$\begin{aligned} X^2 H_3 = {} & 2 \sum f_{ijk\ell} \ell n \left[f_{ijk\ell}/f_{ijk\ell}\right] + \\ & 2 \sum f_{ijk\ell} \ell n \left[f_{ij..}/\hat{F}_{ij..}\right] + \\ & 2 \sum f_{ijk\ell} \ell n \left[f_{..k\ell}/\hat{F}_{..k\ell}\right] = \\ & 2 \sum f_{ijk\ell} \ell n \ f_{ijk\ell} - \ell n \ f_{ijk\ell} + \ell n \ f_{ij..} - \\ & \ell n \ \hat{F}_{ij..} + \ell n \ f_{..k\ell} - \ell n \ \hat{F}_{..k\ell} = \\ & \ell n \left[f_{ijk\ell} \ f_{ij..} \ f_{..k\ell}/f_{ijk\ell} \ \hat{F}_{ij..} \ \hat{F}_{..k\ell}\right] \end{aligned}$$

The following computations for the members of the S and R sets incorporate the equiprobability restrictions between the 12 and 21 marginals.

$$\text{for S and R} = \ell n \left[f_{ijk\ell} \ f_{ij..} \ f_{..k\ell}/f_{ijk\ell}(\hat{F}_{ij..} + f_{i'j'..})/2 \quad (f_{..k\ell} + f_{..k'\ell'})/2\right]$$

for neither S or R = $\ell n \left[f_{ijk\ell} / f_{ijk\ell}\right]$

$$\hat{F}_{ijk\ell}^{(3)} \text{ for S and R} = \frac{(f_{ijk}\ (f_{ij..} + f_{i'j'..})/\ 2)\ (f_{..k\ell} + f_{..k'\ell'})/2}{f_{ij..}\ f_{..k\ell}}$$

$$\hat{F}_{ijk\ell} \text{ for neither S or R} = f_{ijk\ell}$$

$$\hat{F}_{ijk\ell} \quad \text{for S but not R} = \frac{(f_{ijk\ell}\ (f_{ij..} + f_{i'j'..})/2}{f_{ij..}}$$

The formulas for the maximum likelihood expected cell frequencies under the 4 x 4 agreement model were applied to data obtained from a geriatric study conducted by the University of Arizona Psychology Department. The observed frequencies within Table 1 were tabulated by two observers recording geriatric patients' dressing skills during pretest and posttest experimental sessions. Table 1 shows the observed and expected cell frequencies obtained from these data.

The X^2 value for the model-data fit is equal to .1488 (degrees of freedom = 2, p>.05). Thus, the joint $\overline{AB}$ and $\overline{CD}$ quasi-equiprobability and saturated model fits the data well. However, it may be demonstrated that this chi-square value may be calculated by just utilizing the marginal values, rather than the entire 4 x 4 table. For instance, Table 2 shows that the four time I marginals may be subsumed into a 2 x 2 table and the four time II marginals may be incorporated into a second 2 x 2 table. A test of quasi-equiprobability within each of the

2 x 2 tables, with the 11 and 22 cells deleted, resulted in chi-square values of .08 and .06, respectively; thereby, summing to a chi-square value of .14.

Table 2. Observed cell frequencies for the time I and time II marginals.

		Time I Observer B				Time II Observer D	
	Response	1	2		Response	1	2
Observer	1	106	5.5	Observer	1	120	9.5
A	2	6.5	108	C	2	8.5	88

In summary, it has been demonstrated that the chi-square value from the 4 x 4 table for the model that $\overline{AB}$ is quasi-equiprobable, $\overline{CD}$ is quasi-equiprobable, and that the saturated model is true equals the sum of the two chi-square values from the 2 x 2 tables for the model of quasi-equiprobability among the 12 and 21 marginal cells at time I and II. Thus, to measure the model-data fit and the degree of observer agreement, the researcher may choose to consider only the eight marginal values obtained from the large 4 x 4 table for his/her investigation and obtain identical results.

Latent Class Approach to Multivariate Agreement

Latent class models may also be used to conceptualize multivariate agreement. Latent class models assert that observable variables measure latent or unobservable characteristics and are taken as indicators of that characteristic's value. The latent class method that is recommended assesses the model-data fit and degree of observer agreement by only considering the eight marginal values obtained from the previously described 4 x 4 table. Following the construction of the two 2 x 2 tables from the time I and time II marginals, latent classes may be generated.

Latent Class Model of Agreement

The latent class model of agreement is represented by the following equation:

$$\hat{\pi}^{X}_{t} = \sum_{t=1}^{T} \hat{\pi}^{ABOX}_{ijkt} \qquad (4.1)$$

where $\hat{\pi}^{X}_{t}$ is the estimated probability of a specific latent class t in the latent variable X and $\hat{\pi}^{ABOX}_{ijkt}$ is the estimated joint probability of observer A's response i, observer B's response j, time k, and latent class t.

In addition, the following formula connotes the magnitude of observer agreement as the estimated probability that judgments from two observers will occur in one of the agreement categories ($\hat{\pi}_{A}$):

$$\hat{\pi}_A = 1 - (\hat{\pi}_3^{\,X} + \hat{\pi}_6^{\,X}) \tag{4.2}$$

where $\hat{\pi}_3^{\,X}$ and $\hat{\pi}_6^{\,X}$ are the estimated probability that a pair of judgments from the observers will occur in the disagreement category (previously defined as $\hat{\pi}_o$).

The latent class model of agreement contains three latent classes for each of the two time periods, one for each of the two agreement categories within the diagonal and one for the disagreement category. This results in a total of six latent classes. The following restrictions are imposed in order to produce the six latent classes.

The first latent class assumes that:

$$\pi^{\bar{A}X}_{11} = \pi^{\bar{B}X}_{11} = \pi^{\bar{O}X}_{11} = 1$$

where $\pi^{\bar{A}X}_{11}$ is the probability of observer A responding in category 1 given the observation is contained in the first latent class of latent variable X, $\pi^{\bar{B}X}_{11}$ is the probability of observer B responding in category 1 given the observation is contained in the first latent class of latent variable X, and $\pi^{\bar{O}X}_{11}$ is the probability of the category 1 observation occurring at time I given that the observation is contained in the first latent class of latent variable X. The first latent class represents agreement in the f_{11} cell within the time I table. The second latent class states that:

$$\pi^{\bar{A}X}_{22} = \pi^{\bar{B}X}_{22} = \pi^{\bar{O}X}_{12} = 1$$

and depicts agreement in the f_{22} cell within the time I table. The third latent class portrays disagreement within the two off-diagonal cells at time I. Under the model of quasi-equiprobability, the disagreement latent class assumes that:

$$\pi^{\bar{A}X}_{i3} = \pi^{\bar{B}X}_{j3} = .5, \ \pi^{\bar{O}X}_{13} = 1$$

The remaining three latent classes represent identical cells but within the time II table. They are described as follows:

$$\pi^{\bar{A}X}_{14} = \pi^{\bar{B}X}_{14} = \pi^{\bar{O}X}_{24} = 1$$

$$\pi^{\bar{A}X}_{25} = \pi^{\bar{B}X}_{25} = \pi^{\bar{O}X}_{25} = 1$$

$$\pi^{\bar{A}X}_{i6} = \pi^{\bar{B}X}_{j6} = .5, \ \pi^{\bar{O}X}_{26} = 1$$

Under the model of quasi-equiprobability, the estimated probability $\hat{\pi}^{X}_{t}$ of a specific latent class t in the latent variable X is deduced from the equation:

$$\hat{\pi}^{X}_{t} = \sum_{t=i}^{T} \hat{\pi}^{ABOX}_{ijkt}$$

These estimated joint probabilities and latent class probabilities may be computed by applying Clifford Clogg's (1977) computer program which uses the iterative process.

An Example

The geriatric data from the 4 x 4 table was used to test the fit of the latent class model and measure latent agreement. As previously mentioned, this quasi-equiprobability model only incorporated the eight marginal values, four at each of the two points in time. The likelihood ratio chi-square value obtained from assessing the model-data fit was .15 (degrees of freedom = 2, $p > .05$). Except for rounding error, this chi-square value is identical to the chi-square values obtained from the previous quasi-equiprobability tests measuring the 4 x 4 table ($X^2 = .1448$) and the two 2 x 2 marginal tables ($X^2 = .14$).

The following six latent class probabilities were also calculated by Clogg's latent structure computer program:

$$\hat{\pi}^{X}_{1} = .2233, \; \hat{\pi}^{X}_{2} = .2277, \; \hat{\pi}^{X}_{3} = .0489$$

$$\hat{\pi}^{X}_{4} = .2478, \; \hat{\pi}^{X}_{5} = .1766, \; \hat{\pi}^{X}_{6} = .0756.$$

These latent class probabilities show that the three latent class probabilities for time I sum to .5 and the three latent class probabilities for time II sum to .5. Thus, these probabilities demonstrate that the large table may be subdivided into two 2 x 2 tables incorporating the marginals at each point in time. The degree of observer agreement at time I, for the latent class model, was computed by summing π^{X}_{1} and π^{X}_{2} and multiplying by two ((.2233 + .2277) x 2), which resulted in a .90 agreement value. Disagreement at time I was obtained by multiplying π^{X}_{3} by two (.0489 x 2) which yielded a .10 disagreement value.

Similar procedures also produced the time II degree of agreement value of .8488 and disagreement value of .1512. The average degree of agreement for both time periods is .8755, which was derived by adding the first, second, fourth and fifth latent class probabilities of agreement at both periods of time.

Log Linear Model Comparison

Since it was demonstrated above, with the log linear model, that X^2 calculations involving only the eight marginal values will equal computations encompassing the entire 4 x 4 table, one may also postulate that the degree of observer agreeement, under the log linear model, may also be measured by only considering the eight marginal values. The degree of observer disagreement may be assessed by first calculating the expected cell frequencies, under the model of quasi-equiprobability, within each of the two 2 x 2 tables (Bergan, 1980b). At time I, observer disagreement (π_o) equaled .0973 (i.e., 5.5/226/.25) and at time II observer disagreement equaled .1504 (i.e., 8.5/226/.25). An average of the two disagreement values equals .1239. Consequently, observer agreement at time I equaled .9027, at time II equaled .8496 and the average of the two time values equaled .8761. Thus, a comparison of the degree of agreement values calculated under the log linear model of quasi-equiprobability and the latent class agreement model finds them to be identical, except for rounding error.

Quasi-Equiprobability and the Weighted Least Squares Method

The weighted least squares procedure, developed by Grizzle, Koch, and their associates, analyzes multivariate categorical data obtained from repeated measurement experiments. This noniterative procedure fits the experimental functions to a linear model, tests the model-data fit, and estimates the parameters underlying the linear regression model. In essence, this procedure creates a categorical data analogue to interval data methodologies such as linear multiple regression.

For the purpose of assessing the degree of agreement between the two observers, coding the presence/absence of one behavior at two points in time, the 4 x 4 table must be considered again. The hypothesis directed at the degree of observer disagreement investigated the first-order marginal distributions of the response profiles and constrained corresponding probabilities. Agreement between the two observers at the two points in time was investigated by fitting the linear regression model assessing equiprobability between the row 2 and 3 marginals and equiprobability between the column 2 and 3 marginals. As with the maximum likelihood method, the marginals for row 1, row 4, column 1, and column 4 were constrained.

Landis, Stanish, and Koch's (1976) GENCAT II computer program utilizes the following asymptotic regression model:

$$E_A(R) = \mu - X\beta \qquad (4.3)$$

where $E_A(R)$ signifies asymptotic expectation, β is a vector of unknown parameters that will be estimated, and X is a design matrix representing the relationship among the components of μ with regard to the independent variables of interest. The estimated β (Beta) parameters for the marginal equiprobability test correspond to the row and column parameters. This model's goodness of fit and the hypotheses regarding linear combinations of the parameters are assessed with Q statistics. If the sample sizes are large then the Q statistics have approximate chi-square distributions.

The GENCAT II computer program was used to test the linear model-data fit, estimate the underlying model parameters, and compute cell probabilities and expected cell frequencies via weighted least squares methodologies. Table 3 shows the computed expected cell frequencies for the 4 x 4 table, following log transformations of the cell values. A comparison of the Table 3 weighted least squares expected cell frequencies and the Table 1 maximum likelihood expected cell frequencies will show both sets of expected values to be equal.

The hypothesis testing equiprobability between the row 2 and row 2 marginals and equiprobability between the column 2 and 3 marginals resulted in a minimum modified (i.e., Wald Statistic) chi-square value of .1750 (degrees of freedom = 2, $p > .05$). A further test partialed out this chi-square value and obtained a chi-square value of .0909 (degrees of freedom = 1, $p > .05$) for the test of equiprobability between the row 2 and 3 marginals and a chi-square value of .0588 (degrees of freedom = 1, $p > .05$) for the test of equiprobability

Table 3. Observed cell frequencies and expected cell frequencies via weighted least squares computations.

				Time II			
		Observer C	1	1	2	2	Observed Marginal Values
		D	1	2	1	2	
	Observer A	B					
	1	1	77 (77.0434)	3 (2.8024)*	4 (4.2262)	22 (22.0124)	106
Time I	1	2	2 (2.1696)	1 (1.0170)	.5 (.5650)	2 (2.1696)	5.5
	2	1	2 (1.8306)	.5 (.4294)	1 (.9718)	3 (2.7346)	6.5
	2	2	39 (38,9302)	5 (4.7008)	3 (3.1866)	61 (61.0426)	108
Observed Marginal Values			120	9.5	8.5	88	226 Total

*Expected cell frequencies are within parentheses.

between the column 2 and 3 marginals. Given that the column and row equiprobability model fit the data well, it was appropriate to conduct an additional test to investigate if the row and column parameters were equal. An advantage of the weighted least squares procedure is that this test may be conducted. Maximum likelihood procedures are not recommended for this test because complex computations using Lagrange multipliers would be necessary. The weighted least squares test produced a chi-square value of 1.5376 (degrees of freedom = 1, p > .05), thereby indicating the marginal distributions were the same for the two time periods.

Following the confirmation that the linear regression model fit the data well, the degree of observer agreement and latent class probabilities were calculated. However, these calculations require that the marginal values be incorporated within two 2 x 2 tables, as described in the previous section. Table 4 illustrates the observed frequencies and weighted least squares estimated cell probabilities for each of the tables representing a specific time. Latent class probabilities may also be computed using the weighted least squares estimated probabilities and the following computational algorithm for the agreement latent classes ($\overset{W}{\pi}_t$):

$$E\{p\} = \overset{w}{\pi}_t = (P_{ij} - \overset{w}{\pi}_o \overset{w\bar{A}}{\pi}_{io} \overset{w\bar{B}}{\pi}_{jo}) /2$$

where P_{ij} is the observed proportion of a specific observer agreement category t designated in the ij cell, $\overset{w}{\pi}_o$ is the probability of disagreement, $\pi^{w\bar{A}}_{io}$ is the observer A's response i given the disagreement category 0 and $\pi^{w\bar{B}}_{jo}$ is the estimate of observer B's response j given the disagreement category. The latent class probability for the disagreement categories is obtained by dividing π^{w}_{o} by two. Since the six latent classes must sum to one, it was also necessary to divide the above latent class quantities by two. This enables the first three latent classes for time I to sum to .5 and the last three latent classes for time II to sum to .5. Thus, the six latent classes for the full 4 x 4 table can sum to one.

Table 4. Observed cell frequencies and estimated cell probabilities using the weighted least squares computations.

		Time I				Time II	
		Observer B				Observer D	
		1	2			1	2
Observer A	Response 1	106 (.4711)*	5.5 (.0242)	Observer C	Response 1	120 (.5333)	9.5 (.0376)
	2	6.5 (.0242)	108 (.4799)		2	8.5 (.0376)	88 (.3911)

*Estimated cell probabilities are within parentheses.

The degree of observer disagreement calculated for time I equaled .0968 (i.e., $\frac{.0242}{(.5)(.5)}$) and for time II equaled .1504 (i.e., $\frac{.0376}{.25}$). This produced an average degree of disagreement of .1236 for the two time periods. The six latent classes for the two time periods are as follows:

$$\overset{w}{\pi}_1 = .2235,\ \overset{w}{\pi}_2 = .2279,\ \overset{w}{\pi}_3 = .0484,\ \overset{w}{\pi}_4 = .2479,\ \overset{w}{\pi}_5 = .1768,\ \overset{w}{\pi}_6 = .0752.$$

A comparison of the latent class probabilities produced by the weighted least squares procedure and the previously described maximum likelihood procedure will show both sets of latent class probabilities to be identical, except for rounding error.

Finally, a comparison of the degree of observer agreement values produced by the log linear maximum likelihood procedure, latent class maximum likelihood methodology, and weighted least squares computations, shows all three procedures generated identical agreement probabilities at time I (.90), time II (.85), and for the average of the two times (.88).

Quasi-Independence and Maximum Likelihood Method

Given conditions when two observers are observing more than one behavior, it is recommended that the quasi-independence concept be employed to assess observer disagreement. The following paragraphs will extend the previously described maximum likelihood algorithms to accommodate a quasi-independence model. A repeated measurement design that includes two observers categorizing three behavioral categories at two points in time creates a 9 x 9 contingency table. Table 5 illustrates the format for constructing a contingency table from the aforementioned design. At each of the two points in time the two observers may evidence nine response variations. The rows in the table represent the response variations at time I and the columns represent the response variations at time II.

Since the cells 1111, 1122, 1133, 2211, 2222, 2233, 3311, 3322, and 3333 portray observer agreement at each of the two time periods, it may be postualted that a model assessing independence between specific marginals may test observer agreement at the two time periods. More specifically, the agreement model constrains the observed and expected row 1, row 5, and row 9 marginal values to be equal, constrains the observed and expected column 1, column 5, and column 9 marginal values to be equal, assesses independence between the row 2, 3, 4, 6, 7, and 8 marginal values, and assesses independence between the column 2, 3, 4, 6, 7, and 8 marginal values. By constraining the aforementioned marginal values, the rows and columns representing disagreement between the two observers may be investigated. In addition, the process of constraining the six marginals also constrains the observed and expected cell values in the nine (i.e., 1111, 1122, 1133, 2211, 2222, 2233, 3311, 3322, and 3333) agreement cells to be equal. Thus, a quasi-independence model is created that measures the remaining observer disagreement cells. The test of marginal quasi-independence also uses the model of agreement that was presented within formula 4.1 (i.e., $\bar{\pi}^{X}_{t} = \sum_{t=1}^{T} \bar{\pi}^{ABQX}_{ijkt}$). In addition to this model, the magnitude

Table 5. Observed cell frequencies, observed marginal values, and expected marginal values under the model of marginal quasi-independence.

		Time II									
	Observer C	1	1	1	2	2	2	3	3	3	Observed/Expected Marginal Values
	D	1	2	3	1	2	3	1	2	3	
	Observer A B										
	1 1	44	1	3	2	25	2	3	2	32	114 (113.68)
	1 2	2	2	3	1	1	3	2	1	1	16 (17.18)
	1 2	1	3	2	4	1	5	2	3	4	25 (23.88)
Time I	2 1	1	3	1	3	2	2	5	1	1	19 (18.95)
	2 2	31	2	3	1	29	4	1	4	37	112 (111.70)
	2 3	1	4	3	1	2	1	1	1	5	19 (19.58)
	3 1	2	0	3	1	1	2	3	2	1	15 (16.43)
	3 2	5	1	1	2	3	1	1	4	5	23 (22.12)
	3 3	50	1	4	3	19	1	4	2	41	125 (124.65)
Observed Marginal Values		137	17	23	18	83	21	33	20	127	468 Total
Expected Marginal Values		(136.99)	(17.25)	(22.32)	(17.96)	(82.92)	(21.22)	(21.91)	(20.33)	(126.88)	

of agreement is similarly defined with the agreement measure within formula 4.2 (.e., $\bar{\pi}_A = 1 - \pi_o$, where π_o is the sum of the disagreement latent classes).

The analysis of repeated measures designs, employing the concept of quasi-independence, also focused on the use of the marginal values. As with the set of quasi-equiprobability models, hierarchical sets of models may be developed with models that not only incorporate marginals but also models that use marginals exclusively.

The agreement model asserting variables A and B are quasi-independent, variables C and D are quasi-independent, and that the saturated model is true is a model that subsumes two hierarchical models that involve marginals. The test that variables A and B are quasi-independent sums across the C and D cells. Conversely, the test that the variables C and D are quasi-independent sums across the A and B cells. The saturated model component allows the reproduction of the large initial table without adding to the large table's chi-square value. Thus, by adding the three chi-square values for the A and B quasi-independence test, C and D quasi-independence test, and the saturated model test, the chi-square for the full table may be generated. The formulas for generating the maximum likelihood expected cell frequencies and the X^2 values for this model are presented below.

Saturated model

$$\hat{F}_{ijk}{}^{(0)} = f_{ijk\ell}$$

A and B are quasi-independent

$\hat{F}_{ij}^{(1)} = \delta_{ij}\, a_i b_j$ for i = 1, ...I; j = 1, ...J, where a and b are parameters for the test of independence.

$$\delta_{ij} = \begin{cases} 1 \text{ for cells } (i,j) \text{ in } S \\ 0 \text{ otherwise} \end{cases}$$

C and D are quasi-independent

$\hat{F}_{..k\ell}^{(2)} = \delta_{k\ell}\, a_k b_\ell$ for k = 1, ...K; ℓ = 1, ...L, where

$$\delta_{ij} = \begin{cases} 1 \text{ for cells } (k,\ \ell) \text{ in } S \\ 0 \text{ otherwise} \end{cases}$$

Chi-square for the Saturated model

$$X^2 H_0 = 2 \sum f_{ijk\ell} \ell n \left[f_{ijk\ell} / f_{ijk\ell} \right]$$

Chi-squares for the two quasi-independence tests (A with B and C with D):

$$X^2 H_1 = 2 \sum (f_{ij11} + \dots + f_{ij22}) \quad \ell n \left[f_{ij..} \ / \ \hat{F}_{ij..} \right]$$

$$X^2 H_2 = 2 \sum (f_{11k\ell} + \dots + f_{22k\ell}) \quad \ell n \left[f_{..k\ell} / \hat{F}_{..k\ell} \right]$$

Chi-square for the total table:

$$X^2 H_3 = 2 \sum f_{ijk\ell}\ \ell n \left[f_{ijk\ell} / f_{ijk\ell} \right] + 2 \sum f_{ijk\ell}\ \ell n \left[f_{ij..} / \hat{F}_{ij..} \right] +$$

$$2 \sum f_{ijk\ell}\ \ell n \left[f_{..k\ell} / \hat{F}_{..k\ell} \right] - =$$

$$2 \sum f_{ijk\ell}\ \ell n\ f_{ijk\ell} - \ell n\ f_{ijk\ell} + \ell n\ f_{ij...} - \ell n\ \hat{F}_{ij..} +$$

$$\ell n\ f_{..k\ell} - n\ \hat{F}_{..k\ell} = =$$

$$\ell n \left[f_{ijk\ell}\ f_{ij..}\ \ f_{..k\ell} / \ f_{ijk\ell}\ \hat{F}_{ij..}\ \ \hat{F}_{..k\ell} \right]$$

Computation of the {a} and {b} estimates under the model of quasi-independence requires the application of the following iterative process.

Set $b_j^{(0)} = 1$ for j = 1, ..., J, and proceed with the iterative process, as previously described, by setting

$$a^{(v)}_{i} = \frac{x_{i\ldots}}{\Sigma_{j}\delta_{ij..}b_{j}^{(v-1)}} \qquad \text{for } i = 1, \ldots, I \text{ and}$$

$$b^{(v)}_{j} = \frac{x_{j\ldots}}{\Sigma\delta_{ij}\, a_{i}^{(v)}} \qquad \text{for } j = 1, \ldots, J$$

Following the v^{th} cycle, compute the $\bar{F}$ vaues with the following formula:

$$F^{(2v)}_{ij} = \delta_{ij}\, a^{(v)}_{i}\, b^{(v)}_{j} \qquad \text{for all } (i,j)\ .$$

Alternatively, if a researcher has a repeated measures contingency table that meets the maximum likelihood assumption of $X_{i+} + X_{+i} < N$, where X_{i+} is the sum of the frequencies in non-structural zero cells in row i, X_{+i} is the sum of the frequencies in non-structural zero cells in column i, then the full table may be considered as a univariate model. In other words, if the lack of excessive zero cell frequencies allows the researcher to meet the above assumption then procedures utilizing the univariate quasi-independence model may be employed.

As described in the previous chapter, the procedures using the univariate quasi-independence model included the insertion of structural zeros in cells representing agreement (e.g., cell 1111) and the application of the iterative proportional fitting algorithm for the computation of expected cell frequencies. Following the calculation of expected cell frequencies, the probabilitiy of a specific agreement category _t_ was generated with the formula:

$$\hat{\pi}_t = P_{ij} - \hat{\pi}_o \hat{\pi}^{A}_{io} \hat{\pi}^{\bar{B}}_{jo}.$$

If the above maximum likelihood assumption is met then the univariate quasi-independence model procedure may be applied to any size table (i.e., 4 x 4 or 9 x 9). Given the 4 x 4 repeated measures table, the cells 1111 and 2222 denote pure agreement at both time periods and the cells 1122 and 2211 represent agreement within time periods but not across time; the remaining cells are considered disagreement cells. The calculation of the probabilities (i.e., $\hat{\pi}_t$) for agreement cells 1111 and 2222 would enable the researcher to determine the degree of pure agreement across time. Furthermore, the computation of the probabilities for agreement cells 1122 and 2211 would allow the researcher to obtain information regarding agreement within time periods but not across time. The sum of these four cell probabilities would yield the overall degree of agreement during the two observational time periods.

Quasi-Independence and the Weighted Least Squares Method

The weighted least squares procedure was also applied to fit the experimental functions within the 9 x 9 table to a linear model, test the fit of the quasi-independence model, estimate the parameters underlying the linear regression model, and compute the expected marginal frequencies. To test the quasi-independence model, the 1, 5, and 9 row marginal values were constrained and the 1, 5, and 9 column marginal values were constrained. Independence was then assessed between the 2, 3, 4, 6, 7, and 8 row marginal values and between the 2, 3, 4, 6, 7, and 8 column marginal values. This process enabled the marginals representing disagreement between the two observers to be investigated.

The asymptotic regression model presented in formula 4.3 (i.e., $E(R) = \mu = X\beta$), for the quasi-equiprobability test, is also used for the quasi-independence test. However, the β parameters for the quasi-independence test would reflect independence among the marginals, rather than equiprobability.

The weighted least squares marginal quasi-independence test measures the observer by time association. This association is examined by testing the log odds ratios within the table. Under the hypothesis of quasi-independence, if the observer by time measure of association reflects independence then the log odds will equal zero (Koch, personal communication).

The hypothetical data in Table 5 was used to test the aforementioned marginal quasi-independence model. The observed frequencies and model based estimates of the marginals, in Table 5, were constructed by the GENCAT II computer program. By using these marginal estimates a goodness of fit test for the marginal quasi-independence model was conducted. The weighted least squares fit for the linear functions produced a chi-square value of 2.4081 (degrees of freedom = 2, $p > .05$); thereby providing a good model-data fit,

To generate expected cell frequencies for the model, it is recommended by Gary Koch (personal communication) that the iterative proportional fitting algorithm be applied to adjust the observed cell frequencies to agree with the fitted margins produced by the GENCAT II analysis. If this procedure is applied the zero cell frequencies in the contingency table should be changed to 0.5. G. Koch also suggested that if a researcher encounters a zero marginal value the investigator should eliminate the row or column which is involved from consideration and analyze the table which remains. Alternatively, the researcher may add 1/2 to every cell and then proceed to multiply every cell by the original sample size divided by the original sample size plus 1/2 the number of cells; however, any of these strategies regarding the ways to handle zero marginal values would be ad hoc.

Reliability of Criterion-Referenced Test Categorizations

Since the observers' behavioral occurrence-nonoccurrence judgments are dichotomous codings, it is postulated that such categorical data is identical to the dichotomous mastery and nonmastery classification decisions that are generated by criterion-referenced testing. Given a dichotomous classification system of mastery and nonmastery categories, it is reasonable to consider the consistency of mastery-nonmastery decisions from repeated test administrations as a reliability measure. Moreover, criterion-referenced test reliability may be defined as a measure of agreement between mastery-nonmastery classifications produced from repeated test administrations.

Various univariate and multivariate models may be employed to assess reliability of mastery classifications. Given the case where students were dichotomously categorized during two test administrations would warrant constructing a 2 x 2 table, measuring the fit of the univariate quasi-equiprobability model, and assessing the degree of agreement. An assessment of student mastery of a behavioral domain utilizing two instruments at two points in time would justify constructing a 4 x 4 multivariate table, testing the fit of the multivariate quasi-equiprobability model, and computing the degree of agreement at both points in time. More specifically, the mastery-nonmastery scores produced by two instruments measuring the same behavioral domain would create four possible response patterns (i.e., 11, 12, 21, 22) during the first administration of the tests. The second administration of the tests would also generate the same four response patterns. Thus, the administration of two tests at two points in time would require the construction of a 4 x 4 multivariate table for the data analysis.

Criterion referenced test scores may also be classified within polytomous categories. For example, the classification of student scores on two similar instruments at two points in time would create a multivariate design. Additionally, if the researcher divides the students' scores from each instrument into three categories a 9 x 9 multivariate table would be created.

In summary, to measure the reliability of criterion-referenced test classifications across repeated testings would warrant the application of the univariate and multivariate models described in the previous sections. By inserting the frequency of test classifications into an appropriate contingency table and following the identical model testing and degree of agreement computations presented in the previous sections, a classification reliability measure may be produced. Additionally, reliability measures may be partialed out to determine the degree of agreement at each point in time, for specific behavioral categories, and for specific assessment instruments.

CHAPTER 5

DISCUSSION

The present investigation developed applications of the log linear, latent class, and weighted least squares procedures for the analysis of multivariate repeated measures designs. These computations tested the model-data fit and calculated the multivariate measure of the magnitude of agreement under the quasi-equiprobability and quasi-independence models.

It was demonstrated that employing log linear, latent class, and weighted least squares computations resulted in identical multivariate model-data fits with equivalent chi-square values. Moreover, the application of these three methodologies also produced identical measures of the degree of agreement at each point in time for the multivariate average.

The present investigation also clarified the conditions for applying the quasi-equiprobability as well as the quasi-independence model. If the investigators examined observer agreement on only one behavioral response at two or more points in time, then the quasi-equiprobability model was recommended. However, if the researcher was investigating observer agreement on more than one behavioral response at two or more points in time, the the quasi-independence model was chosen.

The multivariate methods that were developed also included procedures for measuring the probability of agreement for a single response classification or subset of classifications from a larger set. It may also be added, to analyze occurrences of systematic observer disagreement within the multivariate tables, the investigator need only apply the systematic disagreement procedures described within the chapter on univariate models.

The consistency of criterion referenced test classifications over repeated assessments of the identical examinees was also suggested as a means of conceptualizing criterion-referenced reliability. By applying the univariate and multivariate models described in the previous chapters, the reliability of these dichotomous and polytomous classifications across repeated testings could be calculated. To accomplish this, the frequency of test classifications must be inserted into an appropriate table followed by the identical model testing and degree of agreement methods previously described.

The procedures utilizing the log linear, latent structure, and weighted least squares concepts for the purpose of measuring agreement have the advantages of 1) yielding a coefficient of agreement that varies between zero and one and measures agreement in terms of the probability that the observers' judgments will agree, as estimated under a quasi-equiprobability or quasi-independence model, 2) correcting for the proportion of "chance" agreement, and 3) providing a directly interpretable coefficient of "no agreement". Thus, these multivariate procedures may be regarded as a more refined psychometric technology for measuring inter-observer agreement and criterion-referenced test reliability.

The multivariate agreement methodologies that were developed may be applied to complex intrasubject and intersubject research designs incorporating design facet combinations of observers, conditions, sessions, behaviors, and subjects. In addition, these reliability measurements may be applied to compute the accuracy and consistency of behavioral data recorded throughout a specific project or experiment.

Although the present research developed multivariate agreement procedures utilizing the quasi-equiprobability and quasi-independence models, it may behoove behavioral researchers to investigate the possibility of employing other models, such as the model of symmetry, for measuring reliability. Finally, further research applying these multivariate procedures to various student classifications from criterion-referenced test scores is necessary to validate the multiple purposes of this technology.

REFERENCES

Bennett, B. Tests of hypotheses concerning matched samples. Journal Royal Statistical Society, 1967, 29, 268-274.

Bennett, B. Note on X^2 tests for matched samples. Journal Royal Statistical Society, 1968, 30, 368-370.

Bennett, B. Measures for clinicians' disagreements over signs. Biometrics, 1972, 28, 607 - 612.

Bergan, J. Measuring observer agreement using the quasi-independence concept. Journal of Educational Measurement, 1980a, 17, 59-68.

Bergan, J. A quasi-equiprobability model for measuring observer agreement. Journal of Educational Statistics, 1980b, 4, 366-376.

Birkimer, J., and J. Brown. A graphical judgmental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. Journal of Applied Behavior Analysis, 1979a, 12, 522-533.

Birkimer, J., and J. Brown. Back to basics: Percentage agreement measures are adequate, but there are easier ways. Journal of Applied Behavior Analysis, 1979b, 12, 535-543.

Bishop, Y., S. Fienberg, and P. Holland. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press, 1975.

Clogg, C. Unrestricted and Restricted Maximum Likelihood Latent Structure Analysis: A Manual for Users (Working Paper No. 1977-09). Unpublished manuscript, Pennsylvania State University, 1977.

Clogg, C. Some latent structure models for the analysis of Likert-type data. Social Science Research, 1979, 8, 287-301.

Cochran, W. Some methods for strengthening the common X^2 tests, Biometrics, 1954, 10, 417-451.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. Psychological Bulletin, 1968, 70, 213-230.

Coleman, J. Measure of concordance or consensus between members of social groups. Unpublished manuscript, Johns Hopkins University, 1966.

Ebel, R. Estimation of the reliability of ratings. Psychometrica, 1951, 16, 407-424.

Everitt, B. Moments of the statistics Kappa and weighted Kappa. British Journal of Mathematical and Statistical Psychology, 1968, 21, 97-103.

Fay, R., and L. A. Goodman. ECTA program description. Unpublished manuscript, University of Chicago, Chicago, Illinois, 1973.

Feinberg, S. The Analysis of Cross-Classified Categorical Data. Cambridge, Mass.: MIT Press, 1978.

Fleiss, J. Estimating the accuracy of dichotomous judgments. Psychometrica, 1965, 30, 469-478.

Fleiss, J. Assessing the accuracy of multivariate observations. American Statistical Association Journal, 1966, 61, 403-412.

Fleiss, J. Measuring nominal scale agreement among many raters. Psychological Bulletin, 1971, 76, 378-382.

Fleiss, J., J. Cohen and B. Everitt. Large sample standard errors of Kappa and weighted Kappa. Psychological Bulletin, 1969, 72, 323-327.

Goodman, L. A new model for scaling response patterns: An application of the quasi-independence concept. Journal of the American Statistical Association, 1975, 70, 755-768.

Goodman, L. Analyzing Qualitative/Categorical Data. Cambridge, Massachusetts: ABT Associates, 1978.

Goodman, L., and W. Kruskal. Measures of association for cross-classifications. Journal of American Statistical Association, 1954, 49, 732-764.

Grizzle, J., C. Starmer and G. Koch. Analysis of categorical data by linear models. Biometrics, 1969, 25, 489-504.

Hambleton, R., and M. Novick. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, F., and B. Lahey. A method for combining occurrence and nonoccurrence interobserver agreement scores. Journal of Applied Behavior Analysis, 1978, 11, 523-527.

Hartmann, D. Considerations in the choice of interobserver reliability estimates. Journal of Applied Behavior Analysis, 1977, 10, 103-116.

Hartmann, D., and W. Gardner. On the not so recent invention of interobserver reliability statistics. Journal of Applied Behavior Analysis, 1979, 12, 559-560.

Hawkins, R., and V. Dotson. Reliability scores that delude: An Alice in Wonderland Trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp and G. Semb (Eds.), Behavior analysis: Areas of research and application. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

Hersen, M., and D. Barlow. Single Case Experimental Designs. Strategies for Studying Behavior Change in the Individual. New York: Pergamon Press, 1976.

Hubert, L. Kappa revisited. Psychological Bulletin, 1977, 84, 289-297.

Huynh, M. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Kelly, M. A review of the observational data-collection and reliability procedures reported in the Journal of Applied Behavior Analysis. Journal of Applied Behavior Analysis, 1977, 10, 97-101.

Koch, G., University of North Carolina, Chapel Hill, Dept. of Biostatistics, Personal Communication, 1980.

Koch, G., J. Landis, J. Freeman, D. Freeman and R. Lehnen. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics, 1977, 33, 133-158.

Kratochwill, T. (Ed.), Single Subject Research: Strategies for Evaluating Change. New York: Academic Press, 1978.

Kratochwill, T. Just because it's reliable doesn't mean it's believable. Journal of Applied Behavior Analysis, 1979, 12, 553-557.

Kratochwill, T., and R. Wetzel. Observer agreement, credibility, and judgment: Some considerations in presenting observer agreement data. Journal of Applied Behavior Analysis, 1977, 10, 133-139.

Landis, J., and G. Koch. The measurement of observer agreement for categorical data. Biometrics, 1977, 33, 159-174.

Landis, J., W. Stanish and G. Koch. A computer program for the generalized chi-square analysis of categorical data using weighted least squares to compute Wald statistics (GENCAT). Unpublished manuscript, University of Michigan, 1976.

Light, R. Measures of agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin, 1971, 76, 365-377.

Marshall, J., and E. Haertel. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Unpublished manuscript, University of Wisconsin, 1976.

Maxwell, A., and A. Pilliner. Deriving coefficients of reliability and agreement for ratings. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 105-116.

Scott, W. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-324.

Shavelson, R., and N. Webb. Generalizability theory: 1973-1980. Unpublished manuscript, University of California, Los Angeles, 1980.

Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Subkoviak, M. Decision-consistency approaches. In R. Berk (Ed.), Criterion-referenced Measurement. Baltimore: Johns Hopkins University Press, 1980.

Swaminathan, H., R. Hambleton and J. Algina. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1973, 11, 263-267.

Yelton, A., B. Wildman, and M. Erickson. A probability-based formula for calculating interobserver agreement. Journal of Applied Behavior Analysis, 1977, 10, 127-131.