THE INTERACTIVE EFFECTS OF DATA CATEGORIZATION AND NONCIRCULARITY ON THE SAMPLING DISTRIBUTION OF GENERALIZABILITY COEFFICIENTS IN ANALYSIS OF VARIANCE MODELS: AN EMPIRICAL INVESTIGATION

By

HAN JOO EOM

B.P.E and M.P.E., Sung Kyun Kwan University in Korea, 1985
M.P.E., The University of British Columbia in Canada, 1989

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENT FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Interdisciplinary Studies:
Human Kinetics / Educational Psychology / Statistics)

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1993

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature) _____

Department of HUMAN KINETICS

The University of British Columbia
Vancouver, Canada

Date OCTOBER, 15, 1993

DE-6 (2/88)

ABSTRACT

The present study employed Monte Carlo procedures to investigate the effects of data categorization and noncircularity on generalizability (G) coefficients for the one-facet and two-facet fully-crossed balanced designs as well as on the Type I error rates for F tests in repeated measures ANOVA designs. Computer programs were developed to conduct a series of simulations under various sampling conditions. Five independent parameters were considered in the simulations: (a) three levels of repeated measures (3, 5, 7); (b) three G coefficients (.60, .75, .90); (c) three epsilon values (.50, .70, 1.0); (d) three sample sizes (15, 30, 45); and (e) six measurement scales (Continuous, 5-point and 3-point scales with either normal or uniform distribution, and dichotomous).

For the one-facet design, the results of the simulations indicated that categorical data resulted in a considerably smaller G coefficient than for the parent continuous data, especially for a 3-point or less scale. Noncircularity did not introduce any bias to the estimate, but yielded more variable estimates of the G coefficient. The sampling theory of G coefficients with continuous data was fairly robust to a moderate departure from circularity, but somewhat sensitive to severe noncircularity (about 6% for ε = .7 and about 7.2% for ε = .5 of the sample estimates lay in the 5% region of the upper tail). However, it was not adequate for categorical data, especially for a 3-point or less scale. The results of the two-facet design closely paralleled those of the one-facet design in

terms of the effects of categorization, sample size, and population G values. The primary difference in the findings between the two designs was that the sampling theory of G coefficients for the two-facet design, which was developed using Satterthwaite's procedure, was very satisfactory and quite robust to violations of the circularity assumption.

Type I error rates of the F test for continuous data were inflated when the circularity assumption failed, with categorization causing a slight reduction in this inflation. Relationships among the population epsilon, the sample estimate, and the Type I error rates across the 81 simulated conditions revealed the presence of a strong negative relationship between the epsilon estimates and the associated Type I error rates, thus supporting current theory. However, for the $\varepsilon = 1.0$ condition the associated Type I error rates were all close to the nominal level, and the correlation with the estimated epsilon was near zero. Further investigation of the correlations among the sample estimates ($\hat{\varepsilon}$, MS_e, and MS_r) within each population epsilon condition suggested that the inflation in Type I error rates is not, as is commonly assumed, merely a function of the population epsilon value. This led us to question the current practice of utilizing an epsilon-adjusted F test in repeated measures ANOVA designs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

This dissertation is the culmination of many years of study, along with the combined of efforts of many individuals to whom I am indebted. Therefore, I would like to take this opportunity to express my sincere gratitude to all those who have given me continued encouragement and support throughout my graduate education.

My most sincere appreciation is offered to Dr. Robert Schutz, my supervisor. Over the six years that I have studied under his guidance, Dr. Schutz has invested enormous time and effort towards my academic development. He has always provided numerous valuable suggestions and stimulating discussions during all phases of my graduate study, and given me the confidence to pursue this research. Without his guidance and support, this research would not have been completed.

I am grateful to Dr. Robert Conry, who introduced me to Measurement Theory, for his support and helpful advice.

I am truly indebted to Dr. Ian Franks, who served as a committee member for both my Master and Ph.D programs, for sharing his expertise in Sport Analysis and for his continued support throughout my graduate study.

I am deeply grateful to Dr. Michael Schulzer for his teaching in Statistical Consulting and Biostatistics, and his many valuable suggestions during this dissertation.

I would further like to thank Dr. Robert Hindmarch, Dr. Robert Morford, and Mr. Dale Ohman. They made it possible for me to come to Canada and start my graduate education at U.B.C., and have continued to provide support and encouragement to the present time.

I am grateful to my family who have supported me both morally and financially throughout my graduate education, most notably, my mother, brothers, sisters Kwi Ok Eom, Chong Ae Suh, and their families. I am also grateful to my friends who have provided their moral and intellectual support since the beginning of my graduate education. In particular, I thank my faithful friends Sung Won Yoon, Yoon Hak Park, Chi Yong Shin, and Drs, Sun Dong Yoo and In Gyu Kim.

Finally, I gratefully acknowledge and thank Mr. David Chiu for sharing his expertise in computer programming, and for freely devoting so much of his time to develop the simulation programs which made this study possible.

CHAPTER ONE: INTRODUCTION

Anyone who regularly plays a game with objective scoring, whether it be one of physical activities or of mental tasks, is acutely aware of the variability in human performance. This inconsistency in human performance stems from a variety of factors, depending on the nature of the measurement. Among the important factors are subtle variations in physical and mental efficiency of the test taker, uncontrollable fluctuations in external conditions, variations in the specific tasks required of individual examinees, and inconsistencies of those who evaluate examinee performance. Quantification of these sources of error that affect the measurement process constitutes the essence of reliability analysis within the context of classical test theory or generalizability theory.

Generalizability theory (G theory) was proposed by Cronbach and his colleagues (1963, 1972) as an alternative to classical test theory. G theory can be viewed as an extension and liberalization of classical test theory that is achieved primarily through the application of analysis of variance (ANOVA) procedure to measurement data. The use of an appropriate factorial ANOVA model permits one to identify and independently estimate several sources of measurement error, which is regarded as an amorphous quantity in classical test theory. The application of G theory to measurement problems has greatly expanded over the past several years in a wide range of behavioral research. In the educational and psychological literature a number of authors demonstrated the use of G theory

to deal with multiple sources of measurement errors and stressed the advantages of G theory approach to reliability estimation over the classical test theory approach (e.g., Brennan, 1983; Shavelson & Webb, 1991).

The use of G theory is evident in observational studies (e.g., Godbout & Schutz, 1983; Huysamen, 1990; Lomax, 1982; Morgan, 1988; Ulrich, Ulrich, & Branta, 1988). Although these observational studies vary widely in content and method, they all use human observers as a primary source of the necessary measure by classifying and recording certain behaviors of subjects. Particular examples of observational research that are of interest in the present study include the assessment of consistency of judges' ratings on motor behavior as well as individual and team performances in sport competitions, the evaluation of dependability of ratings on social and classroom behaviors in educational settings, and the assessment of diagnostic accuracy of subjects with physical or mental problems in clinical settings. Data collection in any of these examples can be done through the use of a simple checklist or a sophisticated instrument such as a computerized recording system. However, the quantification from "excellent" to "bad" of the qualitative performances, the determination of occurrence or nonoccurrence of a particular behavior, or the classification from "not at all" to "severe" of symptom is solely based on human judgement before any data recording process takes place. Therefore, reliable and consistent judgement is a primary objective, and unreliable observation, particularly in clinical

settings, may have substantial effects on subsequent treatments. In fact, Fleiss (1981, p.192) noted that the single most important source of the discrepancies in the results of similar studies he examined in clinical settings was the unreliability of psychiatric diagnosis.

Researchers employing G theory often tend to place great emphasis on the G coefficient in their interpretation, and report it as a means of summarizing the adequacy of the measurement process. However, the estimation of G coefficients involves some combination of variance component estimates (or mean squares), each of which is somewhat subject to sampling error as well as to the violation of assumptions underlying ANOVA procedures. Consequently, these variance components could collectively produce a considerable amount of error that may affect the estimation of the G coefficients, and this would be especially true for small samples. Therefore, in the absence of information such as a likely range of the estimate which might occur under certain experimental conditions, one would not be sure whether a large value of a G coefficient indicates that the measurement process is reliable, or is merely an overestimate due to sampling bias. The main focus of the present study, therefore, is to investigate and compare the sampling distribution of G coefficients obtained under various simulated sampling conditions.

Overview of test theory

Within the context of classical test theory, a variety of procedures have been developed for estimating different aspects of reliability, for example: calculation of test-retest correlations to estimate the stability of measurements over different testing occasions; correlation of scores obtained from parallel or alternative forms of a test to estimate the equivalence of measurements based on different sets of items; and applications of various formulas to estimate the internal consistency, or homogeneity, of a pool of test items (e.g., see Feldt & Brennan, 1989). Each of these approaches defines the concept of measurement error in somewhat different ways, and thus identifies a different source of error depending on the purpose of the study under consideration. Classical test theory, which is based on the concept of parallel measures (see p.21 for definition), postulates that an observed score on a test can be decomposed into a true score and a single source of random error. As such, any single application of the classical test theory model cannot clearly differentiate among multiple sources of potential error that are inherent in most behavioral measurements. A technical description of reliability estimations in classical test theory is given in chapter II.

To overcome some of the measurement problems underlying classical test theory, Cronbach and his colleagues (1963, 1972) proposed generalizability theory (G theory) as an alternative to classical test theory. Following the pioneering work of the analysis of variance (ANOVA) approach to measurement issues by

Burt (1955), Ebel (1951), Horst (1949), and Hoyt (1941) among others, Cronbach, Rajaratnam, and Gleser (1963) formulated a theoretical model that does not rely upon the restrictive assumptions of classical test theory. By applying the mathematical rationale of Cornfield and Tukey (1956) they reformulated reliability estimation procedures with no assumptions of the equivalence among the conditions. Based on the variance component estimates obtained from the Cornfield and Tukey method, Cronbach et al. (1963) derived a formula for calculating an intraclass correlation coefficient for a composite score in a one-facet design (i.e., a two-way, n_p persons by n_r raters, random effects ANOVA model), which is identical to the reliability coefficient from the Hoyt ANOVA procedure. (The term "facet" is analogous to the ANOVA term "factor", but the subject factor is not considered as a facet in G theory.) Cronbach, Gleser, Nanda, and Rajaratnam (1972) later denoted this coefficient as Ep^2 and named it the coefficient of generalizability. The use of the symbol Ep^2 for the G coefficient was "... intended to imply that a generalizability coefficient is approximately equal to the expected value ... of the squared correlation between observed and universe scores." (Brennan, 1983, p.17). Following the foundation papers by Cronbach et al. (1963) and Gleser, Cronbach, and Rajaratnam (1965), extensive treatments of G theory were documented in a book by Cronbach et al. (1972), and more recently by Brennan (1983), and Shavelson and Webb (1991). In addition, many measurement specialists provided extensive reviews and

pedagogical aspects of G theory (e.g., Brennan & Kane, 1977; Cardinet, Tourneur, & Allal, 1976, 1981; Godbout & Schutz, 1983; Hopkins, 1984; Morrow, 1989; Shavelson & Webb, 1981; Webb, Rowley, & Shavelson, 1988). These authors presented fundamental concepts in G theory with a conceptual framework for estimating the reliability of behavioral measurements in a wide range of educational and psychological research. By demonstrating the applications of various factorial ANOVA procedures, they advocated the use of G theory and stressed the advantages of a multifaceted approach to reliability estimation over the traditional classical test theory approach. The basic concepts and terminology in G theory, along with underlying statistical models, are presented in chapter II.

The application of G theory to measurement problems has greatly expanded over the past several years in a wide range of behavioral research. In the educational and psychological literature a number of authors have used a generalizability approach to deal with measurement issues. Bert (1979) and Mitchell (1979), for example, applied G theory to the estimation of inter- and intra-rater reliability; Gillmore (1983) to problems of program evaluation; Johnson and Bell (1985) to the assessment of survey efficiency; Kane and Brennan (1977) to the assessment of class means; Lane and Sabers (1989) to the evaluation of scoring systems for sample essays; Macready (1983) to diagnostic testing problems; Morrow (1986) to reliability of anthropometric measures; Staybrook and Corno (1979) to the disattenuation of measurement error in path-analytic approaches;

and Violato and Travis (1988) to the assessment of behavior and personality. In addition, the use of G theory has also increased rapidly in observational research. Examples include the evaluation of social and classroom behavior (Huysamen, 1990; Lomax, 1982) in educational settings, the assessment of diagnostic accuracy in clinical settings (Morgan, 1988), and the evaluation of motor behavior and sport performances (Godbout & Schutz, 1983; Looney & Heimerdinger, 1991; Ulrich, Ulrich, & Branta, 1988).

While it is evident in the literature that the number of research papers employing G theory has greatly increased in recent years, most appear to be limited to the applications and flexibilities of the theory in dealing with measurement problems, and very little attention has been given to the presence of sampling errors in the estimated G coefficient. The possible reason for the lack of attention in this area may be that the common sources of G theory (e.g., Brennan, 1983; Crocker & Algina, 1986; Cronbach, et al.,, 1972; Shavelson & Webb, 1991) as well as many review and pedagogical papers on G theory (e.g., Cardinet, Tourneur, & Allal, 1981; Shavelson, Webb, & Rowley, 1989) have paid relatively little attention to the fact that the G coefficient is subject to sampling error as well as the violation of ANOVA assumptions. The major emphasis has been on the concepts of randomness and on the use of ANOVA techniques as a means of obtaining variance component estimates (or observed mean squares).

Sampling error of variance components

G theory makes extensive use of ANOVA procedures to estimate variance components. The estimated variance components serve the basis in G theory for describing and indexing the relative contribution of each source of error and the dependability of a measurement. However, the problems associated with estimating variance components (e.g., negative estimates) have been frequently found in practice, and subsequently several alternative approaches to variance components estimation, such as maximum likelihood estimators, restricted maximum likelihood estimators, nonnegative estimators and Bayesian estimators, have been proposed to deal with such problems for the cases of both balanced and unbalanced experimental designs (e.g., see the reviews by Khuri & Sahai, 1985; Shavelson & Webb, 1981).

Cronbach et al. (1972) earlier suggested that large scale G studies should be conducted to provide accurate and consistent variance component estimates. They further warned that "... the behavioral scientist is on dangerous ground when he employs estimates of components and coefficients from a G study with the usual modest value of n_i and n_j, unless he can confidently make assumptions of equivalence, homoscedasticity, and normality." (p.49). It may be intuitively obvious that in such large scale G studies, the variance component estimates are likely stable. However, there are many situations in which sufficient resources are not available to conduct a large scale preliminary G study. This may be particularly true in ratings, clinical and

observational studies where human judgement provides the necessary measure. Although G theory was not devised particularly with ratings in mind, the use of this method in observational research has rapidly increased, and most published research in these fields involves relatively small samples with a few conditions in each facet (e.g., Booth, Mitchell, & Solin, 1979; Huysamen, 1990; Lane & Sabers, 1989; Looney & Heimerdinger, 1991; Violato & Travis, 1988).

In response to the increased use of G theory with small samples in the literature, Smith (1978, 1982) conducted empirical simulation studies in order to examine the sampling properties of the variance component estimates with 'small' samples (i.e., n_p = 25, 50, or 100 and n_i, n_j = 2, 4, or 8; where n_p = number of subjects, and n_i, n_j = number of conditions in each facet) under several two-facet designs (i.e., crossed and nested designs). His results indicated that variance component estimates based on small samples are very unstable, resulting in discouragingly wide confidence intervals. Bell (1986) and Smith (1982) further showed that the degree of instability in the variance component estimates depends on a combined relationship between the sample sizes, magnitude of variance components, and design configurations. More recently, Marcoulides (1990) also empirically demonstrated that the variance component estimates in the one-facet and two-facet designs (n_p = 25) are sensitive to nonnormal distributional forms. Given these empirical results, it is apparent that the estimation of G coefficients would also be unstable since it is

computed by some combination of estimated variance components. However, no further investigations were attempted in these studies to examine the sampling characteristics of G coefficients. Estimation procedures for variance components are presented in chapter II.

Sampling distribution of G coefficients

Although very little research has focused on inferential properties of G coefficients, a considerable amount of work has been directed toward examining sampling properties and inferential procedures for reliability coefficients (e.g., Feldt, 1965, 1969, 1980; Hakstian & Whalen, 1976; Kraemer, 1981; Kristof, 1963, 1970; Sedere & Feldt, 1976; Woodruff & Feldt, 1986). The investigations in this area have not dealt with G theory per se, but instead have mostly addressed the properties of a form of intraclass correlation coefficients -- Cronbach's coefficient alpha, Kuder-Richardson 20 (KR-20) and the generalized Spearman-Brown formula. These indices are algebraically equivalent to the G coefficient in a one-facet crossed design in generalizability terminology.

In what follows is a brief summary of the inferential procedures for reliability coefficients (e.g., Feldt, 1965; Fleiss & Shrout; 1978; Kristof, 1963) as well as that for the G coefficients for various two-facet designs developed by Schroeder and Hakstian (1990).

Inferential procedures for coefficient alpha. Kristof (1963) derived a sampling theory for reliability estimates and demonstrated a method to apply it to a hypothesis testing. Feldt (1965) also derived similar results based on a two-way random effects ANOVA model and presented a method to construct $(1 - \alpha)$ probability tolerance limits for the sample estimate in terms of an F-distributed quantity. The derived $100(1-\alpha)\%$ tolerance interval for the sample estimate provides the basis for describing the distributional properties of the estimate and can be used to compute any percentile point of the estimate in the distribution for a known population parameter. Feldt (1969, 1980) extended it further to develop inferential techniques for making two independent as well as two dependent sample comparisons for coefficient alpha. Woodruff and Feldt (1986) took an extra step to consider the general case involving K dependent coefficients. Using Paulson's (1942) normalizing transformation for an F-variable, they developed a test statistic that is distributed approximately as a chi-square. This test is essentially an extension of that by Hakstian and Whalen (1976) who developed inferential procedures for testing the equality of k independent alpha coefficients.

Inferential procedures for intraclass correlation coefficients. Many of the reliability indices can be viewed as versions of the intraclass correlation, typically a ratio of the variance of interest over the sum of the variance of interest plus error variance. These intraclass correlation coefficients

can give, however, quite different results when applied to the same data, depending on the definition of error variance under a particular experimental design (Bartko, 1976; Shrout & Fleiss, 1979). Fleiss and Shrout (1978), and Shrout and Fleiss (1979) formulated six forms of intraclass correlation coefficients under the one- and two-way ANOVA models and presented guidelines for choosing the appropriate form depending on the intent of the study. They also derived the approximate $100(1-\alpha)\%$ confidence intervals for these intraclass correlation coefficients using Satterthwaite's (1941, 1946) approximation to the F distribution for a composite of mean squares. Kraemer (1981) also demonstrated the procedures for testing the homogeneity of the intraclass correlation coefficients based on the sampling theory by Kristof and Feldt. Recently, Alsawalmeh and Feldt (1992) derived, using Satterthwaite's procedure, an approximate statistical test for the hypothesis that the two independent intraclass coefficients are equal within the context of a two-way random effects ANOVA model.

Inferential procedures for G coefficients. Schroeder and Hakstian (1990) extended the work of Feldt (1965, 1969), and of Hakstian and Whalen (1976) to develop inferential procedures for G coefficients for various two-facet designs. In doing that, they first applied the Satterthwaite's approximation procedure for a composite of independent mean squares, which is involved in the calculation of G coefficients in a two-way design. This allowed them to treat the quantity $(1-\hat{E}\rho^2)/(1-E\rho^2)$, which is

the ratio of two chi-squared variates, as an approximate F-variate. From this, they took a further step to derive the normalized expression of the quantity $(1-\hat{}\rho^2)^{1/3}$ using Paulson's (1942) normalizing transformation, and developed an asymptotic variance expression for the estimate $(1-\hat{}\rho^2)^{1/3}$ by employing the delta method (Rao, 1973, p.387). The resulting variance expressions permit the construction of confidence intervals for a single sample G coefficient and can be applied to develop an inferential procedure for testing the equality of K independent G coefficients under normal theory.

The sampling theory of the G coefficient (including coefficient alpha) and the inferential procedures described above have been developed under conditions in which the underlying ANOVA assumptions are fully met. It is, however, conceivable that "real-world" data will not always fulfill the rigorous underlying assumptions of the models. In addition, unlike test development studies, most observational studies employing G theory rarely involve more than a few conditions (e.g., usually 3 to 5) of each facet, along with rather small or moderate sample sizes. Data collection in these studies are often done in such a way that a rater or a group of raters successively observes the behavior or performance of subjects and numerically codes them using an instrument which yields only a limited number of score values. Therefore, due to the nature of the data, it is quite likely that the violation of ANOVA assumptions will occur in conjunction with limitations of the

score scale when such data are subjected to ANOVA procedures. Considerable research has focused on examining the effects of the number of scale points on the estimated reliability coefficient (e.g., Cicchetti, Showalter, & Tyrer, 1985; Jenkins & Taber, 1977; Lissitz & Green, 1975). In general, studies have indicated that the reliability of a test increases with an increasing number of scale points, but in most cases this increase quickly levels off for anything beyond a 5-point scale. However, these studies were mainly concerned with the magnitude of the estimates under various measurement scales, and no attention was given to the sampling variability of the estimates under violation of ANOVA assumptions.

In principle, G theory is based on random effects repeated measures ANOVA models. The assumption of randomness itself does not carry with it the assumption of normality. Most estimation procedures for variance components and thus mean squares do not require normality. However, when distributional properties of the resulting estimators are of interest, normality is assumed in the distribution of random effects (Searle, Casella, & McCulloch, 1991). In fact, Scheffe' (1959, p.345) earlier demonstrated that non-zero kurtosis seriously affects inferences about variances of random effects, although it has little effect on inferences about means.

In the statistical literature, a number of empirical studies on the effect of violating ANOVA assumptions of normality and homogeneity of variance on Type I error rates have shown that the ANOVA F statistic is generally robust with

respect to moderate departures from these assumptions, especially if sample sizes are equal (e.g., Glass, Peckham, & Sanders, 1972, but see Bradley, 1978). However, ANOVA loses its robustness, especially when the covariance matrix underlying the repeated measures deviates from a certain pattern, referred to as compound symmetry or circularity (see Appendix A for details). Box (1954), for example, has shown that the violation of this assumption can result in more unstable estimates of the mean squares than would be the case if all observations were independent. Subsequent investigations have also suggested that the inflation in the Type I error rates of the F test introduced by violating the circularity assumption was quite substantial in a variety of specific cases (e.g., Collier, Baker, Mandeville, & Hayes, 1967; Gessaroli & Schutz, 1983; Greenhouse & Geisser, 1959; Huynh, 1978; Huynh & Feldt, 1976; Maxwell & Bray, 1986; Stoloff, 1970; Wilson, 1975).

Knowing that violating ANOVA assumptions will result in unstable estimates of observed mean squares, it is expected that this will add additional variability to the estimates of the G coefficient. This would be especially true with a small scale measurement design involving a limited number of score values, and in such cases the usual interpretations for the estimated G coefficient and subsequent generalization over the conditions of the universe could be misleading. In fact, Schroeder and Hakstian (1990) found that G coefficient estimates are highly variable with small samples, especially if the error variances are relatively large. They suggested that researchers should

interpret with caution G coefficients calculated from designs involving small numbers of objects of measurement (i.e., n_p=25) since the population value could vary markedly from the observed estimate, ranging from trivially low to impressively high. However, there is very little published research regarding the extent to which the sampling variability of the estimates and/or the magnitude of the sample estimates will be affected by the violation of ANOVA assumptions, especially the circularity assumption.

Feldt (1965, 1969) provided some empirical evidence to support the proposed sampling theory. He concluded that despite certain violations of ANOVA assumptions inherent in binary data, the empirical distribution corresponds, in general, quite closely to the theoretical one, at least when the number of items equals 80. However, Bay (1973, p.56) found in employing a one-facet (n=30 by k=8) design that the non-zero kurtosis of the true score distribution has substantial influence on the sampling distribution and standard error of reliability estimates, although the influence on the error score is negligible for fairly large k.

By applying Box's work to reliability estimation, Maxwell (1968, p.810) showed analytically that the correlated errors in a (n by k) repeated measures ANOVA model will positively bias the estimate of the reliability coefficient. Recently, Smith and Luecht (1992) empirically investigated the effect of correlated errors on the variance component estimates in a one-facet G study design. Their results showed that serially

correlated errors underestimated the residual variance component and overestimated by a similar amount the person variance component. Although they did not examine explicitly the G coefficient, they noted from these results, that "Together or separately, these biases will result in an overestimation of the computed generalizability coefficient..." (p.232). Smith and Luecht further noted that the effect of serially correlated errors are equally likely to be present in designs employing more than a single facet (p.234). However, there appears to be very little, if any, published research regarding the effects of violating compound symmetry or circularity on an inferential procedure for the G coefficients in either one- or two-facet designs. In fact, Schroeder and Hakstian (1990) are the only ones who stated the assumption of circularity explicitly in their study. In developing inferential procedures for G coefficients in the two-facet designs, they assumed, besides normality, that the covariance matrices for all subsets of first and second facet conditions and their interaction have (local) circularity or sphericity properties, which are necessary to treat the sums of squares in the model as central chi-squared variates. Although they provided some evidence of the insensitivity of the proposed procedures to nonnormality, the effect of noncircularity was not part of their study, but they suggested that " ... the procedures may also be robust with respect to violation of the local circularity assumption, although at this point we have no proof of this." (p. 443).

Purposes of the present study

There is insufficient information in the related literature substantiating the conditions under which the estimated G coefficient would be most variable and the relative extent to which the estimated G coefficient for categorical data will be affected by a particular sampling condition or a combination of simulated conditions, in comparison to their parent continuous data. Empirical work is certainly in need to provide information about the sampling characteristics of G coefficients under various sampling conditions of G study designs. Thus, the major focus of the present study is an empirical investigation of the sampling variability of the estimated G coefficients for both categorical and continuous data under violation of the circularity assumption. In doing so, the empirical distributions of the G coefficient estimates under various simulated sampling conditions are obtained for both one- and two-facet designs, and the variabilities of the estimates are compared across simulated conditions within each design to investigate the precision and accuracy of the sample estimates. Then, the empirical percentages of the sample estimates falling beyond the theoretical limits are compared to the corresponding theoretical values to assess the robustness of the proposed sampling theory. For all simulated conditions, empirical results for both categorical and the parent continuous data are obtained and compared in order to assess the degree of relative bias in sample estimates for categorical data, in comparison to the parent continuous data.

CHAPTER TWO: THEORIES AND MATHEMATICAL DEVELOPMENT

This chapter consists of four sections: classical test theory, generalizability theory, variance component estimation, and sampling theory of G coefficients. First, definitions and estimation procedures for various reliability coefficients are presented. Second, the basic concepts and terminology in G theory are briefly reviewed, along with the statistical models for one- and two-facet fully-crossed designs and the formulations of G coefficients. Third, a brief overview of estimation procedures and variance expression for variance components is given. Lastly, the sampling theory of coefficient alpha (equivalently, a G coefficient for relative decisions in the one-facet crossed design) developed by Feldt (1965) and Kristof (1963) is presented in detail and extended to develop an approximate sampling distribution of the G coefficient for relative decisions in two-facet fully-crossed design.

A. Classical test theory

According to the true-score model in classical test theory, an observed score is viewed as a composite of two components -- a theoretical score (true score) and an error score. In symbols:

[2-1]

$$x = t + e$$

where, x is the observed score; t is the true score; and e is random error.

Fundamental assumptions imposed by classical test theory are:

(a) the true scores are stable over time;

(b) the expected error score is zero: E(e)=0;

(c) the correlation between error score and true score is zero: $r_{te} = 0$ (Ghiselli, Campbell, & Zedeck, 1981).

Consequently, the variance of observed scores is simply the sum of the true and error score variances:

[2-2]
$$\sigma^2_x = \sigma^2_t + \sigma^2_e.$$

Given this, the ratio of the true score variance to observed variance is called the reliability of measure x and can be expressed as:

[2-3]
$$r_x = \sigma^2_t \; / \; \sigma^2_x$$

This ratio can be shown to be equal to the squared correlation between observed and true scores:

[2-4]
$$\begin{aligned} r_{xt} &= \sigma_{xt} \; / \; \sigma_t \; \sigma_x \\ &= \sigma_{(t+e)\,(t)} \; / \; \sigma_t \; \sigma_x \\ &= \sigma_{tt} + \sigma_{(te)} \; / \; \sigma_t \; \sigma_x \qquad (\text{since } \sigma_{te} = 0) \\ &= \sigma^2_t \; / \; \sigma_t \; \sigma_x \\ &= \sigma_t \; / \; \sigma_x. \end{aligned}$$

Thus, squaring the last expression of [2-4] gives the same result as in Equation [2-3], which indicates the degree to which test scores are free from errors of measurement.

Parallel Measures

Since the true-score model includes an unobservable element (true score), in practice, reliability of a measure is often assessed by correlating parallel measurements. By definition (Ghiselli et al., 1981), two measurements are said to be parallel if they have identical true scores and equal variances. As a result, the means and variances of both measures are also equal. In addition, according to the assumption that errors are independent, it follows that errors associated with parallel measures are not correlated among themselves, nor are they correlated with true scores. That is, $r(e_1e_2) = r(e_1t) = r(e_2t) = 0$, where the subscripts 1 and 2 denote parallel measures. Using these, it can be shown that the correlation between parallel measures is an estimate of the reliability of either one of them. Expressing observed scores on each measure as composites of true and error scores, the correlation between two parallel measures is:

[2-5]

$$\begin{aligned} r_{x1x2} &= \sigma(x_1x_2) \;/\; \sigma_{x1}\,\sigma_{x2} \\ &= \sigma[(t+e_1)\;(t+e_2)] \;/\; \sigma_{x1}\,\sigma_{x2} \\ &= \sigma[tt + te_2 + te_1 + e_1e_2] \;/\; \sigma_{x1}\,\sigma_{x2} \end{aligned}$$

Since the last three terms of the numerator equal zero and because $\sigma_{x1} = \sigma_{x2}$, $r_{x1x2} = \sigma^2_t \;/\; \sigma^2_x$ which is consistent with the definition of reliability given earlier. Thus, one can estimate reliability of a measure by administering two parallel measures. However, because of the very restrictive assumptions underlying parallel measures which are rarely met in practice,

the usefulness of this approach is limited. Some researchers have proposed different formulations that may be viewed as variations on the true-score model. Such alternatives are tau-equivalent measures (identical true scores), essentially-tau-equivalent (true scores differ by an constant), and congeneric measures (error variances, true-score variances, and true-score means need not be equal as long as both measure the same phenomenon). These measures are in the order of less restricted assumptions on the parallel measures (Novick & Lewis, 1967; Joreskog, 1971). The comparison among these models in terms of determining which model is better or preferred in estimating reliability can be done by using a computer program, such as LISREL (Joreskog & Sorbom, 1989).

Alternative Form and Test-Retest

Commonly used procedures that require two test administrations to estimate reliability of a measure are the alternative-form method and the test-retest method. The former is used to assess the degree of interchangeability between two alternative forms of a test. It requires a test user to construct two similar forms of a test and administer both forms to the same group of examinees within a very short time period. The correlation coefficient between the two sets of scores is then computed, called the coefficient of equivalence, and taken as an estimate of the test reliability. It is generally suggested that the mean and variance for each form should be quite similar, but ideally the two forms should meet the

condition of parallelism as defined above, if the coefficient of equivalence is to be interpreted as a reliability estimate (Crocker & Algina, 1986).

The test-retest method is used when a test user is interested in how consistently examinees respond to the same measure at different times. Thus, the same measure is readministered to the same group within a certain time period. The correlation coefficient between the two sets of scores, called the coefficient of stability, is taken as an estimate of the test reliability as it indicates the degree of consistency. However, the use of the correlation coefficient in the test-retest method as a measure of reliability has been criticized by several researchers (Carmines & Zeller, 1979; Erikson, 1978; Heise, 1969). The correlation between the test and retest scores of the same measure will inevitably be less than perfect because of the temporal instability of measures taken at multiple points in time and the measurement error. As a result, a simple test-restest correlation is inappropriate to estimate a variable's true reliability as well as the variable's temporal stability unless one can assume that either the underlying variable remains perfectly stable, or the variable is measured with perfect reliability (Erikson, 1978). This point is illustrated analytically in the following. Let x_1 and x_2 be two test scores, then, the correlation between the two scores can be expressed as follows:

[2-6]

$$r_{x1x2} = \sigma_{(x1x2)} \; / \; (\sigma_{x1} \; \sigma_{x2})$$

$$= \sigma_{(t1t2)} \; / \; [(\sigma^2_{t1} + \sigma^2_{e1}) \; (\sigma^2_{t2} + \sigma^2_{e2})]^{1/2}.$$

Substituting for the covariance term in the numerator, $\sigma_{(t1t2)}$, by $(r_{t1t2})(\sigma_{t1} \; \sigma_{t2})$ since $r_{t1t2} = \sigma_{(t1t2)} \; / \; (\sigma_{t1} \; \sigma_{t2})$, yields

[2-7]

$$r_{x1x2} = (r_{t1t2})(\sigma_{t1} \; \sigma_{t2}) \; / \; [(\sigma^2_{t1} + \sigma^2_{e1}) \; (\sigma^2_{t2} + \sigma^2_{e2})]^{1/2}.$$

Furthermore, since the reliability of a variable is the true score variance divided by the true score variance plus the error variance, the equation [2-7] can be rewritten as:

[2-8]

$$r_{x1x2} = (r_{t1t2})(r_1 \; r_2)^{1/2}$$

where, r_j = reliability of x_j, $j = 1,2$.

From the above equation, we find that a simple test-retest correlation is inappropriate as a measure of reliability unless one can assume that the underlying variable remains perfectly stable (i.e., $r_{t1t2} = 1.0$). In addition, it is possible that an obtained low correlation in the test-retest case may not indicate that the reliability of the test is low but may, instead, signify that the underlying theoretical concept itself (true score) has changed (Carmines et al., 1979).

Internal Consistency

Split-Half. The earliest form of the internal consistency approach to the estimation of reliability may be a split-half reliability estimate. The split-half approach may be viewed as a variation of the alternative-form estimate of reliability. The items that comprise a given measure are split in half, and each half is treated as if it were an alternative form for the other, thereby obviating the need to construct two forms of the same measure. A reliability estimate is obtained by correlating scores on the two halves of the measure. In order to estimate the reliability of an original measure that is twice as long as each half, split-half correlations are stepped up by the Spearman-Brown formula. The general Spearman-Brown formula, of which the split-half method is a special case, for the estimation of reliability of a measure is:

[2-9]

$$r_x = \frac{k\ r_{12}}{1 + (k-1)\ r_{12}}$$

where k is the factor by which the instrument is increased or decreased (i.e., k=2 in case of the split-half method); r_x is the estimated reliability of a measure k times longer than the existing one, r_{12}.

Rulon (1939) proposed a simplified procedure for estimating the reliability coefficient by means of split-halves. This method involves the use of difference scores between the half-tests (i.e., d = a - b, a and b being the examinee's score on the first half and the second half of the original test,

respectively). The variance of the difference scores, σ^2_d, is then used as an estimate of the error variance, σ^2_e, in the definitional formula of the reliability coefficient so that: [2-10]

$$r_x = 1 - \frac{\sigma^2_d}{\sigma^2_x}$$

where σ^2_x is the variance of the scores on the total test. The two methods above yield identical results when the variances of the two half-tests are equal. Otherwise, the Spearman-Brown formula yields systematically larger coefficients than Rulon's method. In general, the split-halves method does not yield a unique estimate since there are many possible ways of dividing a test into halves. If a particular way of splitting a measure into halves happens to be an unlucky one, not parallel, it may result in an underestimate or an overestimate of reliability.

Coefficient Alpha. The logical extension of the split-half approach to estimate the reliability of a measure is to split a measure into as many parts as it has items, and thus the arbitrariness of splitting a measure in halves can be avoided. Several approaches to the estimation of the internal-consistency have been formulated based on the assumption that all items are measures of the same underlying attribute. That is, the test is homogeneous in content. A most general form for this approach is known as Cronbach's alpha (1951) which can be computed by the formula:

[2-11]

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\Sigma\sigma^2_i}{\Sigma\sigma^2_i + 2(\Sigma\sigma_{ij})}\right]$$

where; k is the number of items in the measure, and σ^2_i and σ_{ij} are the variance of item i and the covariance of any pair of items i and j (where, i > j), respectively. Cronbach's alpha is a general form of the Kuder-Richardson 20 or KR-20 (1937). Thus, it yields identical results with the KR-20 when items are scored dichotomously. In addition, when all items are standardized, having a mean of zero and a variance of one, it is reduced to the Spearman-Brown formula, replacing a mean of all pairwise inter-item correlations for the r_{12} in Equation [2-9]. Thus, Cronbach's alpha is known as the mean of all possible split-half coefficients of a given measure.

Intraclass Correlations

Within the context of the variance component model of analysis of variance (ANOVA), the intraclass correlation coefficient is derived from the concept of the statistical dependence between any two observations x_{pi} and $x_{pi'}$, $(i =|= i')$ in the same class (i.e., with the same p) (Scheffe', 1959, p.223). In a two-way array with scores on a test (having i items) for n_p persons, an observed score of person p on item i, x_{pi}, is viewed as the sum of four independent components:

[2-12]

$$x_{pi} = u + (u_p - u) + (u_i - u) + (x_{pi} - u_p - u_i + u)$$
$$= u + p + i + e \ .$$

where: u_p is the universe (true) score of person p (the mean of x_{pi} over all i in the universe); u_i is the mean score on i over all persons in the population; and u is the mean over both the u_p and the u_i (the grand mean). The p, i, and e are considered to be independently distributed random variables with zero means and their own variances. Thus, the model renders itself as an additive model in ANOVA, which forces the variance of the items, σ^2_i, to be equal and covariances or correlations between items to be equal, which leads the covariance matrix to have compound symmetry (Winer, 1971).

Most treatments of reliability based on [2-12] define reliability as the ratio of true score variance (σ^2_p) to observed score variance ($\sigma^2_p + \sigma^2_e$), which is an intraclass correlation coefficient by definition in classical test theory. Using the mean squares and variance components obtained by the ANOVA procedure, the intraclass correlation coefficient for a single item score is:

[2-13]

$$r_1 = \frac{MS_p - MS_e}{MS_p + (n_i - 1)\ MS_e},$$

and that for the mean test score over all items is:

[2-14]

$$r = \frac{MS_p - MS_e}{MS_p}.$$

One of the first attempts (among others being Burt, 1955; Ebel, 1951; Horst, 1949) to use the intraclass correlation as an estimate of reliability appears in Hoyt (1941). Hoyt derived

Equation [2-14] by means of estimating reliability in a Persons by Items design. Hoyt related this formula to the theoretical definition of the reliability by noting that the MS_p represents the observed score variance, and the MS_e represents the error variance in the theoretical reliability expression [i.e., $(\sigma^2_x - \sigma^2_e) / \sigma^2_x$].

Hoyt (1941, p.155) noted that ANOVA procedures do not depend on any particular choice in subdividing the items, and they approximate an average of all the possible correlations that might have been obtained by different ways of assigning items to alternative forms. Therefore, this method of estimating the reliability of a test gives a better estimate than any method based on an arbitrary division of the test into halves or into any other fractional parts. Although Hoyt drew attention to its application only to the case where items are scored dichotomously, Equation [2-14] yields identical results with Equation [2-11], Cronbach's alpha, as well as with KR-20.

Numerous papers on reliability subsequently made use of the ANOVA procedure or the closely related intraclass correlation. Various definitions and procedures were formulated, each defining the measurement error in somewhat different ways, and thus estimating different aspects of reliability depending on the purpose of the study under consideration (e.g., Algina, 1978; Bartko, 1976; Bert, 1979; Fleiss, 1975; Lahey, Downey, & Saal, 1983; Mitchell, 1979; Shrout & Fleiss, 1979).

B. Generalizability (G) theory

It has been shown that reliability can be defined as either the correlation between parallel measures, or the squared correlation between the true score and the observed score, or the ratio of the true score variance to the observed score variance (i.e., $r_{x1x2} = r^2_{tx} = \sigma^2_t/\sigma^2_x$). It also has been shown that different formulations for the estimation of reliability lead essentially to the same results. Although they appear different in form due to different theoretical orientations, they all share the same underlying concepts in classical test theory.

The concept of the parallelism or equivalence of measures has been criticized as a major limitation in classical test theory as it is very difficult to construct, and is often unattainable in practice. As a result, Cronbach, Rajaratnam, and Gleser (1963) proposed generalizability theory as an alternative to classical theory which does not rely upon the restricted assumptions of classical theory. Generalizability theory (hereafter G theory) extends in some aspects the concept of the intraclass correlation to the estimation of reliability. G theory offers a comprehensive set of concepts and estimation procedures for various measurement designs by making use of various factorial designs in ANOVA in which the conditions of observation are classified in several respects.

In the following, concepts and terminology in G theory are presented briefly, followed by statistical models underlying G theory and formulation of G coefficients for the one-facet and

two-facet fully-crossed measurement designs. Throughout the manuscript, the symbols G_1 and G_2, instead of Ep^2, are used to denote the population G coefficient for notational convenience, and $\hat{}G_1$ and $\hat{}G_2$ for the sample estimates. The subscript indicates the number of facets involved in the measurement design.

Basic Concepts and Terminology in G theory

Object of measurement. The object of measurement in G theory is the element of the study about which one wishes to make judgments. In most applications of G theory, persons are the object of measurement, but it can be any population of objects other than persons.

Universes of admissible observation. The universe of admissible observations is defined by all possible combinations of the conditions that theoretically could be included in a study in a G study. The variation of these conditions is central to the study. Thus, G theory requires one to specify a universe of conditions of observation over which s/he wishes to generalize. Related to the concept of universe is the concept of universe score. The universe score is viewed as a mean score for an object of measurement over all conditions in the universe of generalization, like the notion of true score in classical test theory.

G and D studies. G theory draws the distinction between a G study and a decision study (D study). The purpose of the G study is to obtain as much information as possible about the sources of variation in the measurement. Therefore, the G study should ideally define the universe of admissible observations as broadly as possible. The purpose of the D study is to make decisions about the object of measurement. The D study makes use of the information provided by the G study to design the best possible application of the measurement for a particular purpose. In planning a D study, the decision maker defines a universe of generalization over which the scores are to be generalized. As well, using information from the G study about the magnitude of the various sources of measurement error, the decision maker evaluates the effectiveness of alternative designs to optimize reliability (i.e., nested or fixed). In practice, however, the same data are usually used for both G and D studies; in this case the G and D studies are the same.

Facet and Condition. The design of a measurement procedure implies the specification of the sources of error affecting the measurement (e.g., judges, occasions), which are called facets. The term facet is analogous to the factor in ANOVA terminology, but the subject factor is not considered as a facet in G theory. The conditions (cf. levels of a factor in ANOVA) representing these facets usually constitute random sampling from the predefined universe of conditions. Variability in the measurement due to a facet or an interaction among facets is

defined as error variance, whereas the variability among individuals over all objects of measurement is defined as universe score variance (Cronbach et al., 1972, p.15). Brennan (1983, p.16,18) further clarifies that a set of randomly sampled conditions is just one of an infinite number of sets from the universe. Thus in G theory randomly parallel tests, for example, can have different means, and the between-test variance is therefore generally not zero since any test may consist of an especially easy or difficult set of items relative to the entire universe of items.

Relative and absolute decisions. In G theory, how generalizable a measure is depends on how the data will be used in the D study. In the D study one of two kinds of decisions is made. A relative decision is made when the interest attaches to the standing of individuals relative to one another. In contrast, an absolute decision is made when the concern is with how well a person's universe score estimates the universe score for that person, without regard to the performance of others. The variance components contributing to measurement error are somewhat different for the relative and absolute decisions. For the relative decisions those variance components that interact with the object of measurement, and thus influence the relative standing of individuals, contribute to error. For example, in a one-facet Persons-by-Raters design, the systematic disagreement between the raters would not introduce error into the estimation of the person's universe score relative to the average universe

score for all persons. Therefore, the variance component due to the Rater facet does not contribute to error in the relative decision. For the absolute decisions all variance components except the object of measurement contribute to measurement error. These variance components include all interactions and the facet main effects.

Statistical models underlying G theory

One-facet crossed design. Consider a measurement design where, for example, *n* persons are observed by *k* raters in an observational setting, assuming the *n* and *k* are a random sample from a respective population. The resulting scores from such a design can be arranged in a two-way (n x k) array. A two-way (Persons by Raters) random effects ANOVA interaction model can be used to partition observed scores into their effects, which can be written as:

[2-15]
$$x_{ij} = u + p_i + r_j + pr_{ij} + e_{ij}, \quad (i=1,..,n;\ j=1,..,k).$$

Note here that because of a single observation per cell, an extra subscript for the number of entries in each cell is not used. In this model, the p_i is the effect due to Person *i*, r_j is the effect due to Rater *j*, pr_{ij} is the interaction effect between the two, and e_{jk} is the residual error. All the effects in [2-15], except for a grand mean u, are assumed to be random variables with zero means and their own variances, and all pairwise covariances are zero (Searle, Casella, & McCulloch, 1992):

$$E(p_i) = E(r_j) = E(pr_{ij}) = E(e_{ij}) = 0;$$
$$\text{var}(p_i) = E(p^2_i) = \sigma^2_p;$$
$$\text{cov}(p_i, p_i') = E(p_i, p_i') = 0 \quad \text{for all } i =|= i';$$
$$\text{cov}(p_i, r_j) = \text{cov}(p_i, pr_{ij}) = \text{cov}(p_i, e_{ij}) = 0.$$

The symbol E denotes expectation. Similar statements can be made for the remaining terms. Thus, the variance of an observed score can be expressed as:

[2-16]
$$\sigma^2_x = \sigma^2_p + \sigma^2_r + \sigma^2_{pr} + \sigma^2_e.$$

The variance component for persons (σ^2_p), or universe-score variance, for example, can be obtained by taking $E(u_p - u)^2$, which indicates the average (over the population of persons) of the squared deviations of the persons' universe scores from the grand mean (Shavelson & Webb, 1991). Other terms can be defined similarly. The estimates of the variance components are usually obtained in practice by solving the expected mean square expressions from ANOVA procedures. Searle et al. (1992) showed that the ANOVA estimators of variance components with balanced data are unbiased and have the smallest variance of all estimators that are both quadratic functions of the observations and unbiased (p.129). They also noted that the ANOVA estimation of variance components does not demand any normality assumptions for the error term or of the random effects unless the distributional properties are of interest.

The statistical model in [2-15] is what Cronbach et al. (1963) have used to formulate generalizability theory in an

attempt to break away from restrictive assumptions in classical test theory. Following the work of Cornfield and Tukey (1956), Cronbach et al. (1963) derived the expected mean square expressions for each effect in the model. Furthermore, they put the variance components σ^2_{pr} and σ^2_{e} together by stating that "... with one observation per cell, it is impossible to separate the interaction component of variance from the within-cell residual." (p.150). This allowed them to derive a G coefficient that is comparable with a reliability coefficient and coefficient alpha.

Huck (1978) demonstrated the application of Tukey's (1949) 'one degree of freedom for non-additivity' in estimating reliability by decomposing the interaction effect from the error term. However, this method may not be relevant in an observational study. For example, when a group of judges independently rates the behaviors of *n* persons (e.g., class or social activities, live or taped sport performances), each person has a true score which must remain constant across judges. Therefore, any interaction between the judges and subjects in this case should be considered as a consequence of inconsistency among the raters themselves and thus be part of the measurement error. Under this presumption, the *p* x *r* interaction term drops out from the model [2-15]. Thus, the model without the pr_{ij} term renders itself as the two-way additive, rather than nonadditive, ANOVA model with a single observation per cell (Winer, 1971, p.394). Nonetheless, with a single observation per cell the expected mean square expressions

for either the nonadditive model with Cronbach's modification or the additive model are structurally identical, and they are presented in Table 2-1.

Table 2-1

The two-way (Persons by Raters) random effects ANOVA model with a single observation per cell

Source of variability	df	Mean Square	Expected mean squares
Person(p)	n-1	MS_p	$\sigma^2_e + n_r\sigma^2_p$
Rater (r)	k-1	MS_r	$\sigma^2_e + n_p\sigma^2_r$
pr.e (e)	(n-1)(k-1)	MS_e	σ^2_e

Note: the term pr.e reflects a combined effect of pr_{ij} and e_{ij} by following Cronbach's modification in the model [2-15].

A generalizability coefficient for relative decisions is defined as the ratio of the universe score variance to the expected observed score variance, which is a form of intraclass correlation coefficient. The G coefficient, like the reliability coefficient, reflects the proportion of variability in individuals' scores (i.e., the object of measurement) that is systematic. The population G coefficient for the one-facet design, expressed in terms of variance components, is:

[2-17]

$$G_1 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_e/n_r},$$

and expressing the estimate of G_1, $\hat{G}_1$, in terms of the estimated mean squares yields:

[2-18]

$$\hat{G}_1 = \frac{MS_p - MS_e}{MS_p}.$$

This formula is exactly identical to Hoyt's formula for a reliability coefficient presented in the previous section, which in turn gives the same results as the Cronbach's alpha for any metric, and as the KR-20 when items are scored dichotomously. In addition, it can be shown that the formula [2-17] yields the same results as the generalized Spearman-Brown formula. For example, as noted previously, the intraclass correlation is defined as $\sigma_{ii'} / \sigma^2_x = \sigma^2_p / (\sigma^2_e + \sigma^2_p)$, which indicates the degree of statistical dependence between two conditions within the same subject. This follows from the fact that the expected variance of any condition i is defined as: $\sigma^2_i = \sigma^2_e + \sigma^2_p$, and the expected covariance over pairs of conditions within the same subject as: $\sigma(pi, pi') = \sigma_{ii'} = \sigma^2_p$. From this, it may be shown that the expectation of all of the observed variances (mean squares) in the analysis of variance can be expressed in terms of parameters of the variance-covariance matrix. Winer (1971) showed, for example, that

$$MS_{person} = \underline{var} + (n_r - 1)\ \underline{cov}, \text{ and}$$
$$MS_{error} = \underline{var} - \underline{cov}.$$

where, $\underline{var}$ and $\underline{cov}$ = mean of the variances and covariances, respectively, in the n_r by n_r variance-covariance matrix.

Therefore,

$$E(MS_p) = E[\underline{var} + (n_r-1)\ \underline{cov}]$$
$$= \sigma^2_x + (n_r-1)\sigma^2_p$$
$$= \sigma^2_e + \sigma^2_p + n_r\sigma^2_p - \sigma^2_p$$
$$= \sigma^2_e + n_r\sigma^2_p \text{ , and}$$
$$E(MS_e) = E[\underline{var} - \underline{cov}]$$
$$= \sigma^2_x - \sigma^2_p$$
$$= [\sigma^2_e + \sigma^2_p] - \sigma^2_p$$
$$= \sigma^2_e.$$

Applying these terms in Equation [2-17] yields:

[2-19]

$$G_1 = \frac{n_r\ \underline{cov}}{\underline{var} + (n_r-1)\underline{cov}},$$

which is in form identical with the Spearman-Brown formula.

A G coefficient for absolute decisions for the one-facet design is defined as the ratio of the universe score variance to the total observed variance. Brennan and Kane (1977) called this an index of dependability, which can be defined as

[2-20]

$$G_{abs} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_r/n_r + \sigma^2_e/n_r},$$

and expressing the estimate of G_{abs} in terms of obtained mean squares, with some algebra, as:

[2-21]

$$G_{abs} = \frac{MS_p - MS_e}{MS_p + [MS_r - MS_e]/n_p}$$

This coefficient, which takes the variability of the raters into account as error, has been given special attention in observational and clinical studies where an absolute decision is often required (Berk, 1979; Booth et al, 1979; 1983; Brennan & Kane, 1977; Huysamen, 1990; Mitchell, 1979; Lomax, 1982). Since the classical test theory of parallelism indicates that the means across conditions (i.e., raters) are assumed to be equal, the rater effect is assumed to be zero; if not, classical theory cannot formally distinguish between the error variance and the rater variance. However, in G theory, randomly parallel conditions (Brennan, 1983) can have different means, and thus the rater variance is generally not zero. Consequently, there is no equivalent formula in classical reliability theory to the G coefficient for absolute decisions.

Two-facet fully-crossed design. The partitioning of observed scores into their effects and the decomposition of the variance of observed scores into variance components for the separate effects can be easily extended to measurement design with additional facets. Consider, for example, the two-facet, fully-crossed design where *n* persons are observed by *r* raters on *o* different occasions. Persons here are the object of measurement; raters and occasions constitute sources of unwanted variation in the measurement. Persons, raters, and occasions are considered to be randomly sampled from a respective population or universe. The ANOVA model for this design is a three-way (Persons by Occasion by Raters) random effects ANOVA

model with a single observation per cell:

[2-22]

$$x_{ijk} = u + p_i + o_j + r_k + po_{ij} + pr_{ik} + or_{jk} + por_{ijk.e}.$$

In this model, the p_i is the effect due to Person *i* (*i*=1,2,..,n_p), o_j is the effect due to Occasion *j* of the first facet (*j*=1,2,..,n_o), and r_k is the effect due to Rater *k* of the second facet (*k*=1,2,..,n_r). The $por_{ijk.e}$ term reflects a combined effect of the three-way interaction and the residual. The remaining terms in [2-22] represent two-way interaction effects. As in the one-facet design, all the effects in [2-22], except for a grand mean u, are assumed to be random variables with zero means and their own variances, and all pairwise covariances are zero. Normality assumptions of the effects are added when the distributional properties are of interest.

Given the independence of the components in [2-22], the variance of observed scores can be decomposed into variance components for each effect as:

[2-23]

$$\sigma^2_x = \sigma^2_p + \sigma^2_r + \sigma^2_o + \sigma^2_{pr} + \sigma^2_{po} + \sigma^2_{ro} + \sigma^2_{por.e}.$$

For the G coefficient for relative decisions, all variance components representing interaction with the object of measurement (i.e., persons) contribute to the unwanted variations in the measurement. Thus, in the two-facet design, the population G coefficient is defined as:

[2-24]

$$G_2 = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_e}{n_o n_r}},$$

Note that the symbol σ^2_e is used for the $\sigma^2_{por.e}$ term.

The estimation of variance components in [2-24] can lead to an estimate of G_2, and it is usually done by ANOVA procedures. Table 2-2 presents the expected mean square expressions for this design from the ANOVA procedure. These expressions can be solved to obtain estimates of each variance component. It is, however, more convenient to express the estimate of G_2, $\hat{G}_2$, in terms of the observed mean squares as:

[2-25]

$$\hat{G}_2 = \frac{MS_p - MS_{po} - MS_{pr} + MS_e}{MS_p}$$

A G coefficient for absolute decisions for the two-facet design can be defined as:

[2-26]

$$G_{abs} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_r}{n_r} + \frac{\sigma^2_o}{n_o} + \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{ro}}{n_r n_o} + \frac{\sigma^2_e}{n_r n_o}}$$

The estimate of G_{abs} could be expressed in terms of observed mean squares presented in Table 2-2, but its complexity in form still remains the same.

Table 2-2

The three-way (Persons by Occasions by Raters) random effects ANOVA model with a single observation per cell

Source	VC	MS	*E*MS
Person(p)	σ^2_p	MS_p	$\sigma^2_e + n_r\sigma^2_{po} + n_o\sigma^2_{pr} + n_r n_o\sigma^2_p$
Rater(r)	σ^2_r	MS_r	$\sigma^2_e + n_p\sigma^2_{ro} + n_o\sigma^2_{pr} + n_p n_o\sigma^2_r$
Occasion(o)	σ^2_o	MS_o	$\sigma^2_e + n_p\sigma^2_{ro} + n_r\sigma^2_{po} + n_p n_r\sigma^2_o$
pr	σ^2_{pr}	MS_{pr}	$\sigma^2_e + n_o\sigma^2_{pr}$
po	σ^2_{po}	MS_{po}	$\sigma^2_e + n_r\sigma^2_{po}$
ro	σ^2_{ro}	MS_{ro}	$\sigma^2_e + n_p\sigma^2_{ro}$
pro.e (e)	σ^2_e	MS_e	σ^2_e

Note: VC = variance components; MS = mean squares; and *E*MS = expected mean squares.

C. Variance component estimation

In most cases a variance component estimate is computed using some linear combination of available mean squares (MS) divided by a constant:

[2-27]
$$\hat{\sigma}^2_i = [\Sigma\ a_i MS_i] \ / \ c_i$$

where; $a_i = \pm 1$, $MS_i = 1,2,...m^{th}$ mean square, and c_i is the constant associated with the variance component σ^2_i.

Under normality and independence assumptions of the random effects model with balanced design, it is known that $f_i(MS_i)/ EMS_i$ is distributed as a chi-square (χ^2) with f_i degrees of

freedom, and MS's are independent of one another (Searle, 1971, p.409). Therefore, the sampling variability of variance component estimates can be considered as a linear combination of χ^2-variables. For any mean square in the model,

[2-28]

$$MS_i = \frac{EMS_i \ \chi^2}{f_i}$$

Thus,

[2-29]

$$var(MS_i) = \frac{(EMS_i)^2 \ var(\chi^2_f)}{f_i}$$

$$= \frac{2(EMS_i)^2}{f_i} \quad \text{since } var(\chi^2_f) = 2f$$

where, the symbol E denotes expectation, and 'var' denotes variance. Therefore,

[2-30]

$$var(\hat{\sigma}^2_i) = \frac{1}{c^2_i} \ \Sigma \ var \ (MS_i)$$

$$= \frac{2}{c^2_i} \ \Sigma \ \frac{(EMS_i)^2}{f_i}$$

It is more convenient to express above derivations in a general form using the matrix notations (e.g., Brennan, 1983, Searle, 1971; Winer, 1971). Let $\mathbf{m}$ = a k-by-1 column vector of mean squares in the design, having the same order as $\boldsymbol{\sigma}^2$, the vector of variance components in the model. Suppose $\mathbf{P}$ is such that

$$E(\mathbf{m}) = \mathbf{P}\boldsymbol{\sigma}^2.$$

$\mathbf{P}$ is a k by k (nonsignular) matrix of coefficients of the variance components in the expected mean square expressions for the model. Then, the ANOVA estimator of σ^2 is $\hat{\sigma}^2$, obtained from $\mathbf{m} = \mathbf{P}\hat{\sigma}^2$ as

[2-31]
$$\hat{\sigma}^2 = \mathbf{P}^{-1}\,\mathbf{m}$$

which is a k by 1 column vector whose elements are unbiased, because (Searle et al., 1992)

[2-32]
$$E(\hat{\sigma}^2) = \mathbf{P}^{-1}\,E(\mathbf{m}) = \mathbf{P}^{-1}\mathbf{P}\sigma^2 = \sigma^2.$$

From [2-31], the variance of $\hat{\sigma}^2$ can be expressed as:

[2-33]
$$\mathrm{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\,\mathrm{var}(\mathbf{m})\,\mathbf{P}^{-1\prime}.$$

And with the last expression of [2-29], it can be rewritten as:

[2-34]
$$\mathrm{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\,[2(E\mathrm{MS}_i)^2/f_i]\,\mathbf{P}^{-1\prime}$$

Although mean squares in the balanced design are distributed independently of one another, the estimated variance components are themselves subject to sampling variability. Furthermore, two estimated variance components are generally not uncorrelated, unless there are no common mean squares used in estimating the two variance components. Assuming a multivariate normal distribution for the score effects, the variance-covariance matrix associated with the estimated variance components in $\hat{\sigma}^2$ is

[2-35]
$$\mathbf{V} = \mathbf{P}^{-1}\,\mathbf{D}\,\mathbf{P}^{-1\prime}$$

where **D** is a k by k diagonal matrix containing the variance of MS_i expressed in Equation [2-29]. Equation [2-35] is a theoretical expression for the population values, and thus, in practice, the estimate (**^V**) of **V** can be obtained by replacing observed mean squares [i.e., $2(MS_i)^2/f_i$] in the diagonal of **D** for the corresponding expected values. However, Searle (1971, p.417) showed that the elements in **^V** are biased. He showed that the unbiased estimate of **V** can be obtained by replacing $(f_i + 2)$ for f_i in the diagonal of **D**. For example, by definition, $var(MS) = E(MS^2) - (EMS)^2$, from this it follows that

[2-36]
$$E(MS^2) = (EMS)^2 + var(MS)$$
$$= (EMS)^2 + [2(EMS)^2 / f]$$
$$= [(f + 2)(EMS)^2] / f.$$

Therefore, the unbiased estimator of $2(EMS)^2/f$ is $2(MS^2)/(f+2)$.

The square root of the diagonal elements in the **^V** is the estimated standard error of the variance components, which may be used to construct a confidence interval of interest. The variance components are essentially linear functions of mean squares, and the exact distributional properties of a composite of mean squares are too complicated to be of practical utility (e.g., Burdick & Graybill, 1988; Fleiss, 1971). However, Satterthwaite (1941, 1946) suggested that the sampling distribution of a linear combination of mean squares can be approximated by the F distribution and recommended the use of chi-square distribution in which the number of degrees of freedom is chosen so as to provide good agreement between the two.

Several methods of approximation procedures for conducting confidence intervals for variance components have been proposed, and these (about ten approximation procedures including Satterthwaite's) are empirically compared by Boardman (1974). More recent work on variance component analysis and on the confidence intervals for a linear combination or a ratio of variance components in both balanced and unbalanced random models is thoroughly reviewed and presented in Burdick and Graybill (1988), and Khuri and Sahai (1985). In addition, two bibliographies on this subject, Sahai (1979) and Sahai, Khuri, and Kapadia (1985), provide a comprehensive coverage of variance components and other related topics.

In practice, the variance components are usually estimated through an ANOVA procedure. As is almost always the case with real-world data, the estimation for mean squares, and thus for variance components, from the sample data are always subject to sampling error. This is particularly true with small samples. Smith (1978, 1982), for example, conducted empirical studies to investigate the sampling error of variance component estimates based on small samples with a few conditions for each facet. His results revealed that the variance component estimates are unstable and sometimes negative. He further noted that the confidence intervals for variance component estimates with a small number of facet conditions are discouragingly wide.

Negative variance estimates are not uncommon in practice (e.g., Khuri & Sahai, 1985; Shavelson & Webb, 1981; Verdooren, 1982). Several methods have been proposed to treat such a

negative estimate (e.g., Brennan, 1983; Cronbach et al., 1972; Searle, 1971). The most common method, among others, is setting the negative estimates to zero and carrying the zero through wherever that variance components enters the expected mean square of another variance component. Another option is setting negative estimates to zero, but using the negative estimate wherever that variance components enters the expected mean squares of another variance component.

As an alternative approach to variance component estimation, Shavelson and Webb (1981) reviewed a Bayesian approach, but they concluded that this approach is not well enough developed to have widespread applicability. More recently, Marcoulides (1990) empirically examined the performances of restricted maximum Likelihood estimation (RMLE) and compared it with ANOVA estimates. In most cases of his simulations, RMLE provided estimates for the variance components that are more stable and closer to the true parameter than those from the least square estimation of ANOVA. He also found that ANOVA estimates were more sensitive to nonnormal distributional form and produced consistently incorrect estimates to a greater degree than RMLE. Only in balanced data sets from normal distributions did the two methods perform similarly. However, he concluded that although the sampling variability of RMLE estimates is smaller than that for ANOVA procedure, it is still quite sizeable, and unfortunately RMLE does not completely solve the problem of large sampling variability.

D. Sampling theory of G coefficients

As shown in the previous sections, the formulation of the G coefficient for relative decisions in the one-facet design is exactly identical to Hoyt's formula for a reliability coefficient, which in turn gives the same result as Cronbach's alpha. Therefore, although the sampling theory derived by Feldt and Kristof was for coefficient alpha, we use it here in the context of a G coefficient for the one-facet design. We then extend Feldt's approach to develop an approximate sampling distribution of the G coefficient for the two-facet fully-crossed design using Satterthwaite's (1946) approximation procedure. The reason for using Satterthwaite's procedure in the present study is that it is the most commonly used method and generally works well in many applications for constructing confidence intervals on the sum or ratio of variance components. Some researchers reported that it provides somewhat liberal intervals under certain conditions and suggested modified or new procedures, but the complexities of the proposed methods are overwhelming (e.g., Birch, Burdick, & Ting, 1990; Burdick & Graybill, 1988). An additional reason for using Satterthwaite's procedure lies in its simplicity. In addition, several simulation studies have found that the quasi F is an acceptable approximation to the conventional F as long as the total degrees of freedom are relatively large (e.g., Davenport & Webster, 1973; Gaylor & Hopper, 1969). Hudson and Krutchkoff (1968) also found that when the total number of observations was 64 or greater no negative values of the quasi F were observed out of

2000 simulations.

Following the presentation of the sampling theory of G coefficients, the amount of bias in the sample estimate for the one-facet design derived by Kristof (1963) is presented. Additionally the derivation of an unbiased estimator for the population G coefficient is presented and extended further to the two-facet design. A subsequent modification is made in the sampling distribution expressions for the unbiased estimator. Then, variance expressions for the estimated G coefficient for the one- and two-facet designs are presented. Finally, the application of this theory is illustrated to construct a 100(1-α)% confidence interval for the population parameter as well as a (1-α) probability tolerance interval for the sample estimate.

One-facet Design. In deriving the sampling distribution of the estimated G coefficient for relative decisions, we start with rearranging Equation [2-18] as:

[2-37]

$$\hat{G}_1 = \frac{MS_p - MS_e}{MS_p} = 1 - \frac{1}{MS_p / MS_e}.$$

From the above expressions, it is evident that the sampling distribution of $\hat{G}_1$ can be defined by derivation of the sampling distribution of MS_p/MS_e. Suppose SS is a sum of squares on f degrees of freedom, and MS is the corresponding mean squares. Under the normality assumptions the quantity $SS/E(MS) = fMS/E(MS)$ is distributed as a chi-squared variable with f degrees of freedom (Searle, Casella, & McCulloch, 1992, p.131).

From this relationship, it can be shown that

[2-38]

$$\frac{f_p MS_p}{E(MS_p)} = \chi^2 \quad \text{with } df = f_p.$$

Thus,

$$\frac{MS_p}{n_r\sigma^2_p + \sigma^2_e} = \chi^2/f_p \text{ with } df = f_p = (n_p-1).$$

Similarly,

[2-39]

$$\frac{MS_e}{\sigma^2_e} = \chi^2/f_e \quad \text{with } df = f_e = (n_p-1)(n_r-1).$$

According to Craig's theorem (1938), these chi-squares are independent of one another. In addition, ratios of two independent χ^2-variables, each divided by its degrees of freedom, have F-distributions (Searle, et al., 1992, p.465). Therefore, the ratio of the two quantities in [2-38] and [2-39]; namely,

[2-40]

$$\frac{MS_p \;/\; (n_r\sigma^2_p + \sigma^2_e)}{MS_e \;/\; \sigma^2_e}$$

is distributed as a central F with $f_p=(n_p-1)$ and $f_e=(n_p-1)(n_r-1)$ degrees of freedom. Rearrangement of this ratio yields:

[2-41]

$$\frac{MS_p}{MS_e} \cdot \frac{\sigma^2_e}{n_r\sigma^2_p + \sigma^2_e} = F(n_p-1),(n_p-1)(n_r-1).$$

By denoting the ratio MS_p/MS_e as F_{obs} meaning the "observed

F", and the $(n_r\sigma^2_p + \sigma^2_e)/\sigma^2_e$, which equals EMS_p/EMS_e, as F_{pop} meaning the "population F", Equation [2-41] may be rewritten as:

[2-42]

$$F_{obs}\ (1/F_{pop}) = F(n_p-1),(n_p-1)(n_r-1).$$

Since $F_{obs} = 1/(1-\hat{}G_1)$ and $F_{pop} = 1/(1-G_1)$, Equation [2-42] can be rewritten as:

[2-43]

$$(1 - G_1) \ / \ (1 - \hat{}G_1) = F(n_p-1),(n_p-1)(n_r-1).$$

Feldt (1965, p.362) noted that regardless of whether the variance component σ^2_p be zero or be greater than zero, the ratio in [2-43] is distributed as a central F with (n_p-1) and $(n_p-1)(n_r-1)$ degrees of freedom. This sampling property of $\hat{}G_1$ in [2-43] can be used to derive a variance expression for $\hat{}G_1$ as well as to define a critical region in inferential applications for an unknown population parameter.

Two-facet Design. Following procedures similar to those demonstrated above, we have extended Feldt's approach to develop an approximate sampling distribution of G coefficients for the two-facet design. First, the formula for the population G coefficient in Equation [2-24] can be rearranged as:

[2-44]

$$G_2 = 1 - \frac{1}{\dfrac{\sigma^2_e + n_r\sigma^2_{po} + n_o\sigma^2_{pr} + n_on_r\sigma^2_p}{\sigma^2_e + n_r\sigma^2_{po} + n_o\sigma^2_{pr}}}.$$

Similarly, Equation [2-25] can be rewritten as:

[2-45]

$$\hat{G}_2 = 1 - \frac{1}{MS_p \,/\, (MS_{po} + MS_{pr} - MS_e)}.$$

It can be noticed that the ratio of the expected mean square expressions in [2-44] has the proper structural requirements for the F statistic for the test of $\sigma^2_p = 0$. Therefore, as in the one-facet design, the sampling distribution of $\hat{G}_2$ can be defined by derivation of the sampling distribution of $[MS_p/(MS_{po} + MS_{pr} - MS_e)]$ in [2-45]. However, because the estimated mean squares in the denominator of [2-45] involves a composite of different sources of variation, the sampling distribution of this ratio (i.e., quasi F ratio) is not the usual F distribution, and the exact distributional properties of this ratio are too complicated to be of practical utility (e.g., see Burdick & Graybill, 1988). Consequently, Satterthwaite (1941, 1946) suggested that the sampling distribution of the quasi F ratio can be approximated by the usual F distribution and recommended to use a chi-square distribution in which the number of degrees of freedom is chosen so as to provide good agreement between the two. For example, if W is a linear combination of independent mean squares with v_1, v_2, ...v_k degrees of freedom,

$$W = a_1MS_1 + a_2MS_2 + \ldots + a_kMS_k,$$

where a_i is ± 1), then the quantity $fW/E(W)$ is approximately distributed as a chi-squared variable with f degrees of freedom, where f is given by

$$f = \frac{(\Sigma\ a_i\ EMS_i)^2}{\Sigma\ [(a_i\ EMS_i)^2\ /\ v_i]} \qquad (i = 1,2,\ldots k).$$

Applying Satterthwaite's approximation to Equation [2-45], we see that the sampling distribution of $\hat{G}_2$ can be defined by considering the quantity $[MS_p\ /\ (MS_{po} + MS_{pr} - MS_e)]$. From the one-facet case, we know that the quantity $(n_p-1)MS_p/EMS_p$ is distributed as a chi-squared variate with (n_p-1) degrees of freedom. Furthermore, the composite of mean squares (i.e., $MS_{po} + MS_{pr} - MS_e$) is approximately distributed as a chi-squared variate with f_a degrees of freedom (the subscript $_a$ denotes 'adjusted') given by:

[2-46]

$$f_a = \frac{(EMS_{po} + EMS_{pr} - EMS_e)^2}{\frac{EMS^2_{po}}{(n_p-1)(n_o-1)} + \frac{EMS^2_{pr}}{(n_p-1)(n_r-1)} + \frac{EMS^2_e}{(n_p-1)(n_o-1)(n_r-1)}}$$

Therefore, the ratio of these two chi-squared variates in [2-47] below is approximately distributed as an F-variate with degrees of freedom equal to $f_1 = (n_p-1)$ and f_a:

[2-47]

$$\frac{MS_p\ /\ (\sigma^2_e + n_r\sigma^2_{po} + n_o\sigma^2_{pr} + n_on_r\sigma^2_p)}{(MS_{po} + MS_{pr} - MS_e)\ /\ (\sigma^2_e + n_o\sigma^2_{pr} + n_r\sigma^2_{po})}.$$

Rearranging the terms in [2-47] yields the following:

[2-48]

$$\frac{MS_p}{MS_{pr} + MS_{po} - MS_e} \cdot \frac{EMS_{po} + EMS_{pr} - EMS_e}{EMS_p}.$$

To be consistent with the terminology used in the one-facet design, we again denote the first term and the reciprocal of the second term of [2-48] as F_{obs} and F_{pop}, respectively, and rewrite the expression as:

[2-49]
$$F_{obs}\ (1/F_{pop}) \ \tilde{=}\ F\ (f_1,\ f_a).$$

Since $F_{obs} = 1 / (1 - \hat{}G_2)$ and $F_{pop} = 1 / (1 - G_2)$, the equation [2-49] can be rewritten as:

[2-50]
$$(1 - G_2) / (1 - \hat{}G_2) \ \tilde{=}\ F\ (f_1,\ f_a).$$

This expression describes the distributional property of $\hat{}G_2$, which is precisely the same in form as in [2-43], except for the degrees of freedom of the denominator, which involves the Satterthwaite's procedure.

Bias of the sample estimates

In this section the accuracy of the estimator for the one-facet design is examined, and the desired unbiased estimator is subsequently presented based on the work of Kristof (1963). The results are directly extended to the two-facet design.

One-facet design. To show that the estimator $\hat{}G_1$ is biased, we begin with Equation [2-43]. The reciprocal of the ratio in [2-43] is also distributed as F with degrees of freedom reversed, namely:

[2-51]
$$(1 - \hat{}G_1) / (1 - G_1) = F(n_p\text{-}1)(n_r\text{-}1),(n_p\text{-}1).$$

Let $f_1 = (n_p-1)(n_r-1)$ and $f_2 = (n_p-1)$. Denoting the expected value of $\hat{G}_1$ by $E(\hat{G}_1)$, it follows that

[2-52]

$$E(\hat{G}_1) = 1 - (1-G_1)\ E[F(f_1,f_2)].$$

Since the expected value of the F distribution is $[f_2 / (f_2 - 2)]$ (Winer, 1971, p.832), substituting this for $E[F(f_1,f_2)]$ in [2-52] yields:

[2-53]

$$E(\hat{G}_1) = 1 - (1-G_1)\ \left[\frac{f_2}{f_2 - 2}\right].$$

Replacing (n_p-1) for f_2, with some simplication, yields

[2-54]

$$E(\hat{G}_1) = \frac{G_1(n_p-1) - 2}{n_p - 3}.$$

From [2-54], it is apparent that $E(\hat{G}_1)$ is not equal to the population parameter G_1, except for the unrealistic case $G_1 = 1$. Thus, $\hat{G}_1$ is biased and tends to underestimate the population parameter G_1. In addition, this bias becomes larger for a smaller population parameter and is independent of n_r, the number of levels of the facet. Kristof (1963, p.232) presented the desired unbiased estimator of G_1, $\hat{G}u_1$, in relation to $\hat{G}_1$ as follows:

[2-55]

$$\hat{G}u_1 = \frac{\hat{G}_1(n_p-3) + 2}{n_p - 1}.$$

From this the sampling distribution of unbiased estimator, $\hat{G}u_1$, can be easily derived by replacing $[\hat{G}u_1(n_p-1) - 2]/(n_p-3)$ for $\hat{G}_1$ in the denominator of the equation [2-51], with some simplication, as:

[2-56]

$$\frac{n_p - 3}{n_p - 1} \cdot \frac{1 - G_1}{1 - \hat{G}u_1} = F(n_p-1),(n_p-1)(n_r-1).$$

Two-facet design. As can be expected from the structural similarity in the sampling distribution between $\hat{G}_1$ and $\hat{G}_2$ in the previous section, the unbiased estimator of G_2, $\hat{G}u_2$, is precisely the same as that in the one-facet design. That is, the amount of bias is the same for both $\hat{G}_1$ and $\hat{G}_2$. Furthermore, it is also independent of the Satterthwaite's adjusted degrees of freedom, namely:

[2-57]

$$\hat{G}u_2 = \frac{\hat{G}_2(n_p-3) + 2}{n_p - 1}.$$

Replacing $[\hat{G}u_2(n_p-1) - 2]/(n_p-3)$ for $\hat{G}_2$ in the denominator of Equation [2-50], the sampling distribution of the unbiased estimator $\hat{G}u_2$ can be derived as:

[2-58]

$$\frac{n_p - 3}{n_p - 1} \cdot \frac{1 - G_2}{1 - \hat{G}u_2} \simeq F\ (f_1,\ f_a).$$

Variance expression for the sample estimates

The variability of the distribution of the estimates is an important aspect for assessing the performance of the estimator and also can be used as a means for comparing the variabilities of the theoretical and empirical distributions. Therefore, we now derive the variance expression of $\hat{G}_1$ and $\hat{G}_2$, using the properties of the F distribution.

One-facet design. In deriving the variance expression for $\hat{G}_1$, we start with Equation [2-51], namely:

$$(1 - \hat{G}_1) \; / \; (1 - G_1) \; = F(f_1, f_2)$$

where, $f_1 = (n_p-1)(n_r-1)$ and $f2 = (n_p-1)$. From this it follows that

[2-59]
$$\hat{G}_1 = 1 - (1-G_1) \; F(f_1, f_2).$$

Thus, the variance expression for $\hat{G}_1$, $var(\hat{G}_1)$, can be written as:

[2-60]
$$var(\hat{G}_1) = var[1 - (1-G_1) \; F(f_1, f_2)]$$

$$= (1-G_1)^2 \; var[F(f_1, f_2)].$$

The variance of the F distribution is given in Winer (1971, p.832) as:

[2-61]
$$var[F(f_1, f_2)] = \frac{2 \; f_2^2 \; (f_1 + f_2 - 2)}{f_1 (f_2 - 2)^2 (f_2 - 4)}.$$

Replacing $f_1 = (n_p-1)(n_r-1)$ and $f_2 = (n_p-1)$ in [2-61] yields:

[2-62]

$$\text{var}[F(f_1,f_2)] = \frac{2(n_p-1)^2[(n_p-1)(n_r-1) + (n_p-1) - 2]}{(n_p-1)(n_r-1)[(n_p-1) - 2]^2[(n_p-1) - 4]}.$$

Therefore, the theoretical variance expression of $\hat{G}_1$, with some simplication, can be written as:

[2-63]

$$\text{var}(\hat{G}_1) = (1-G_1)^2 \left\{\frac{2(n_p-1)[n_r(n_p-1) - 2]}{(n_r-1)(n_p-3)^2(n_p-5)}\right\}.$$

And, from the relationship between $\hat{G}_1$ and $\hat{G}u_1$ shown in [2-55], we find that the variance expression of $\hat{G}u_1$, in relation to $\hat{G}_1$, is:

$$\text{var}(\hat{G}u_1) = [(n_p-3)/(n_p-1)]^2 \text{ var}(\hat{G}_1).$$

It is apparent from Equation [2-63] that the variability will increase as G_1 becomes smaller for a fixed n_p and n_r. For example, the theoretical standard deviation of $\hat{G}_1$ in a simulated condition with $G_1 = .60$, $n_p = 30$, and $n_r = 5$ would be $.1349 = [(.0625)(.1138)]^{1/2}$, which is considerably larger than $.0337 = [(.01)(.1138)]^{1/2}$ for $G_1 = .90$ with the same n_p and n_r. The estimate of the $\text{var}(\hat{G}_1)$, in practice, can be obtained by substituting an $\hat{G}_1$ for G_1 in the calculation.

Two-facet design. The variability of the distribution of $\hat{G}_2$ also follows precisely the same structure as that in the one-facet design, except for the associated degrees of freedom. Following the same steps, starting from the reciprocal of Equation [2-50], i.e., $(1-\hat{G}_2) / (1-G_2) \simeq F(f_a, f_1)$, we derive

the variance expression of $\hat{G}_2$ as:

[2-64]

$$\text{var}(\hat{G}_2) = \text{var}[1 - (1 - G_2)\, F(f_a, f_1)]$$
$$= (1 - G_2)^2\, \text{var}[F(f_a, f_1)].$$

where, f_a is defined in [2-46], and $f_1 = (n_p - 1)$.

The variance of the F distribution with f_a and $f_1 = (n_p - 1)$ degrees of freedom is:

[2-65]

$$\text{var}[F(f_a, f_1)] = \frac{2(n_p-1)^2[f_a + (n_p-1) - 2]}{f_a[(n_p-1) - 2]^2[(n_p-1) - 4]}.$$

Thus, theoretical variance expression of $\hat{G}_2$, with some simplication, can be written as:

[2-66]

$$\text{var}(\hat{G}_2) = (1-G_2)^2 \left\{\frac{2(n_p-1)^2\,[f_a + n_p - 3]}{f_a(n_p-3)^2\,(n_p-5)}\right\}.$$

and that of $\hat{G}u_2$ as:

$$\text{var}(\hat{G}u_2) = [(n_p-3)/(n_p-1)]^2\, \text{var}(\hat{G}_2).$$

For example, the theoretical standard deviation of $\hat{G}_2$ in a simulated two-facet condition with $G_2 = .90$, $n_p = 30$, $n_o = 3$, $n_r = 5$, and $f_a = 76.91$ (from Equation [2-46] with EMS_{po}=66, EMS_{pr}=55, EMS_e=31) would be $.0353 = [(.01)(.1246)]^{1/2}$. The estimate of the $\text{var}(\hat{G}_2)$, in practice, can be obtained by substituting the sample estimates, $\hat{G}_2$ and $\hat{f}_a$, for G_2 and f_a in the calculation.

Inferential application of the sampling theory

The sampling distributions presented previously are now applied to construct a $100(1-\alpha)\%$ tolerance interval for the sample estimate for a given population G value as well as to establish a $100(1-\alpha)\%$ confidence interval for an unknown population G value.

One-facet design. From Equation [2-43] we know that the ratio $(1-G_1)/(1-\hat{G}_1)$ is distributed as an F with (n_p-1) and $(n_p-1)(n_r-1)$ degrees of freedom. From this, we can construct a $100(1-\alpha)\%$ tolerance interval for the sample estimate as follows. If the probability (P) of $[F_L < F < F_U] = 1 - \alpha$, then we may rewrite this as:

[2-67]

$$1 - \alpha = P \left[F_L < \frac{1-G_1}{1-\hat{G}_1} < F_U \right]$$

where, F_L is the lower $\alpha/2$ percentage point and F_U is the upper $(1-\alpha/2)$ percentage point of the F distribution with degrees of freedom (n_p-1) for the numerator and $(n_p-1)(n_r-1)$ for the denominator. Further manipulations of the inequality relationship above yields:

[2-68]

$$1 - \alpha = P \left[1/F_L > \frac{1-\hat{G}_1}{1-G_1} > 1/F_U \right]$$

$$= P \left[\frac{1-G_1}{F_L} > 1 - \hat{G}_1 > \frac{1-G_1}{F_U} \right]$$

$$= P \left[1 - \frac{1-G_1}{F_L} < \hat{G}_1 < 1 - \frac{1-G_1}{F_U} \right].$$

Following the same steps and denoting the quantity $(n_p-3)/(n_p-1)$ by M in Equation [2-56], we can also derive a $100(1-\alpha)\%$ tolerance interval for the unbiased sample estimate as:

[2-69]

$$1 - \alpha = P\left[F_L < \frac{M(1-G_1)}{1-\hat{}Gu_1} < F_U \right]$$

$$= P\left[1 - \frac{M(1-G_1)}{F_L} < \hat{}Gu_1 < 1 - \frac{M(1-G_1)}{F_U} \right].$$

The last expression of [2-68], which describes a $(1-\alpha)$ probability statement for the sample estimates, provides the basis for describing the sampling distribution of $\hat{}G_1$. Consider, for example, a simulated measurement condition with $G_1 = .90$, $n_p = 30$, and $n_r = 5$. The lower and upper limits for a 90% tolerance interval for the $\hat{}G_1$, since $F_{L(29,116)} = .5882$ and $F_{U(29,116)} = 1.5653$ from F distribution, would be:

$$1 - (.10/.5882) < \hat{}G_1 < 1 - (.10/1.5653)$$

$$= .8300 < \hat{}G_1 < .9361.$$

If we obtain 2000 sample estimates from the above condition in the simulation, we expect 5% of them to fall beyond either the lower or upper limit and the remaining 90% within the two limits. Moreover, the empirical percentage of the unbiased sample estimates falling beyond either limit in the last expression of [2-69] would be identical since both limits of the tolerance interval for the unbiased sample estimates will be

shifted upwards correspondingly.

A 100(1-α)% confidence interval for an unknown population G coefficient also can be constructed similarly by manipulating the inequality relationship, starting from Equation [2-67]. The resulting formula for a 100(1-α)% confidence interval for G_1 is:

[2-70]

$$1 - \alpha = P[\ 1 - (1-\hat{}G_1)F_U < G_1 < 1 - (1-\hat{}G_1)F_L\ ],$$

and a 100(1-α)% confidence interval using an unbiased estimator is:

[2-71]

$$1 - \alpha = P\left[\ 1 - \frac{(1-\hat{}Gu_1)\ F_U}{M} < G_1 < 1 - \frac{(1-\hat{}Gu_1)\ F_U}{M}\ \right].$$

The empirical percentage of 2000 confidence intervals that fail to include the population G value in either lower or upper direction would be essentially the same as the empirical percentage of 2000 sample estimates that fall beyond either limit of the tolerance interval. Therefore, using either tolerance interval or confidence interval approach, it is possible to assess and compare the sampling behavior of the estimated G coefficients under various simulated sampling conditions.

Two-facet design. The procedures presented for the one-facet design are directly extended to the two-facet design. Since the resulting equations for the two-facet design are precisely the same in form as those in the one-facet design, only the final equations for the tolerance interval and

confidence interval are presented below. A 100(1-α)% tolerance interval for the sample estimate in the two-facet design is:

[2-72]

$$1 - \alpha = P\left[\, 1 - \frac{1-G_2}{F_L} < \hat{G}_2 < 1 - \frac{1-G_2}{F_U} \,\right],$$

and that for unbiased estimator is:

[2-73]

$$1 - \alpha = P\left[\, 1 - \frac{M(1-G_2)}{F_L} < \hat{G}u_2 < 1 - \frac{M(1-G_2)}{F_U} \,\right].$$

A 100(1-α)% confidence interval for the population G_2 is:

[2-74]

$$1 - \alpha = P[\, 1 - (1-\hat{G}_2)F_U < G_2 < 1 - (1-\hat{G}_2)F_L \,],$$

and that for an unbiased estimator is:

[2-75]

$$1 - \alpha = P\left[\, 1 - \frac{(1-\hat{G}u_2)\ F_U}{M} < G_2 < 1 - \frac{(1-\hat{G}u_2)\ F_U}{M} \,\right].$$

It should be noted that to construct the lower and upper limits for a 100(1-α)% tolerance interval for the sample estimate, the denominator degrees of freedom, f_a, for determining the critical F value should be calculated based on expected mean squares. However, to establish a 100(1-α)% confidence interval for an unknown population G_2, the f_a is estimated using the observed mean squares, instead of expected mean squares. In both cases, the degrees of freedom f_a is in general fractional as it involves the Satterthwaite's approximation. Thus, it can be rounded off, in practice, to the nearest integer. In the present study, we obtained an exact critical F value using a

fractional f_a by referring to F-inverse function in the International Mathematical and Statistical Library (IMSL, 1991). Nevertheless, as in the one-facet design, either the tolerance or confidence interval approach can be used to assess and compare the sampling behavior of the estimated G coefficients under various simulated sampling conditions in the two-facet.

CHAPTER THREE: METHODS AND PROCEDURES

Overview of problems and simulation conditions

As shown in the previous chapters, G theory is formulated based on a form of repeated measures ANOVA models, and a sample estimate of G coefficient is calculated based on mean squares obtained from the ANOVA procedure. Moreover, the sampling theory of G coefficients was developed under the conditions in which the underlying ANOVA assumptions are fully met. However, knowing that violating compound symmetry or circularity assumption in repeated measures ANOVA resulted in more variable mean squares, (which in turn inflated Type I error rates in the F test) it is somewhat conceivable that more-variable mean squares would produce particularly a small or a large estimated G coefficient when circularity assumption failed.

There is, however, virtually no research that has investigated the effect of noncircularity on the magnitude of the estimated G coefficient, nor on the robustness of the sampling theory of G coefficient under the violation of compound symmetry or circularity assumption. To systematically examine the effect of noncircularity on the estimated G coefficient, the present study employed Monte Carlo procedures.

In defining the simulation conditions of interest, we first considered the characteristics of population epsilon (and of its estimate). As shown in Appendix B, Box's (1954) epsilon (ε), which is a measure of the degree of departure from compound symmetry or circularity, is a function of the variances and covariances in the population matrix. When the matrix fulfills

compound symmetry or circularity condition, $\varepsilon = 1.0$; otherwise, $\varepsilon < 1.0$, with a minimum of $1/(k-1)$, where k = the number of repeated measures. In other words, a matrix whose covariance elements are all equal, yields epsilon equal to unity; whereas a matrix whose covariance elements are further away from each other, produces a lower epsilon (given that the variances are constant).

Since each covariance element influences on the calculation of epsilon, there are numerous different patterns of the covariance matrix that can yield the same epsilon value. Therefore, by systematically varying the average of the covariance elements (i.e., different population G values) in the matrix for a given epsilon value we will be able to investigate whether the effect of noncircularity is the same, regardless of the covariance structure as long as the covariance matrices produce an equal epsilon value. Furthermore, by constructing covariance matrices with a varying degree of noncircularity for a given population G value, we can compare the effect of noncircularity on the sample estimate of G coefficient as well as a possible interaction between the levels of the two population parameters (i.e., ε and G).

Since Box's work, epsilon estimate (i.e., $\hat{\varepsilon}$ due to Greenhouse and Geisser, 1959; or $\tilde{\varepsilon}$ due to Huyhn and Feldt, 1976) has been widely used to correct a possible positive bias in the F test. However, despite its wide use, the sampling properties of epsilon are unknown and rarely investigated. It can be speculated though that a sample covariance matrix is

always expected to exhibit some degrees of violation of the condition, even though the population matrix does not. In addition, because of the theoretical lower limits on ε, it is more likely that an epsilon estimate ($\hat{\varepsilon}$) for a population $\varepsilon < 1.0$ (say, $\varepsilon = .50$) would be smaller in a measurement design with a large number of k than that with a small number of k. Therefore, by varying the number of repeated measures (k), we will be able to examine the sampling behavior of the estimated G coefficient in relation to the magnitude of epsilon estimate for a given population G and ε, as well as across different G and ε values.

Another consideration in defining the simulation conditions of the present study stemmed from Cohen's (1983) work on the effect of dichotomization. Cohen showed that dichotomization of two continuous variables at their respective means resulted in considerable amount of loss of measurement information, which reduced the original correlation by a factor of .637. The implication of Cohen's work suggests that when categorization is performed on continuous data generated from a population matrix with heterogeneous covariances, we would expect that it alters the structure of the covariance matrix since it would probably reduce the magnitude of the covariances disproportionately, depending on the size of each covariance element. As a result, the categorization not only reduces the magnitude of covariances (or correlations), and thus G coefficient as well, but also affects the degree of noncircularity in the resulting covariance matrix for categorical data. Therefore, by incorporating data

categorization into the simulations, we can examine the relative bias introduced into the estimated G due to categorization, in comparison to the parent continuous data. Moreover, this will allow us to investigate the extent to which the effect of data categorization interacts with the other simulation conditions mentioned above.

As in any simulation study which deals only with a restricted range of the populations of interest, the present investigation was also limited to the G coefficients for relative decisions in the single-facet and two-facet fully-crossed balanced designs. To investigate the magnitude of the estimated G coefficient and its empirical sampling distribution, under various sampling conditions, we incorporated five population parameters into simulations for the one-facet design, and some of the conditions were further incorporated into the two-facet design in simulation II. The following presents the sampling conditions simulated in both designs, each followed by detailed explanations for specific procedures implemented in the simulations.

Simulation I: One-facet design

Sampling Populations

Five independent variables were incorporated into the one-facet design. There were: (1) three different levels of the facet, k = 3, 5, and 7; (2) three generalizability coefficients, G_1 = .60, .75, and .90; (3) three epsilon values, ε = .5, .7,

and 1.0 (the first two were approximated values); (4) three sample sizes, n = 15, 30, and 45; and (5) six measurement scales, continuous (C), normal 5-point (N5), uniform 5-point (U5), normal 3-point (N3), uniform 3-point (U3), and uniform 2-point (U2). The preceding combination of design conditions resulted in 486 data sets simulated (3 x 3 x 3 x 3 x 6).

Although the range of the parameter values considered above would not be sufficient to represent all that could exist in practice, these values were the most commonly observed conditions in published studies. For example, Smith (1978) noted that observational studies employing G theory usually involve about 3-5 levels (or conditions) for each facet. Thus, the range of 3-7 levels of each facet will reflect most situations. The values of the G coefficients were chosen to represent the range from a poor (.60) to a good (.90) indication of a measurement process. With respect to the covariance structure, epsilon (Box, 1954) was used as a measure of departure from the compound symmetry or circularity conditions. The values of epsilon were set to represent severe violation (.50), moderate (.70), and no violation (1.0) (e.g., Huynh, 1978). The range of response categories (i.e., a 2- to 5-point scales) was chosen by considering the results of empirical studies on the effect of the number of scale points on the reliability estimation (e.g., Cicchetti, Showalter, & Tyrer, 1985; Jenkins & Taber, 1977; Lissitz & Green, 1975). These studies demonstrated in general that the reliability of a test quickly levels off for anything beyond a 5-point scale.

Finally, the range of sample sizes was chosen to represent most practical situations in observational research with a rather small or moderate sample size. For example, Smith (1978) considered sample sizes of 25, 50, and 100 as small, moderate, and moderate-to-large samples in his simulation study.

We would like to point out here that the condition of heterogeneity of variance across the levels of k, (i.e., a ratio of .6 to 1.4 among the variances), was originally incorporated into the simulation in order to investigate the effect of unequal variances on the sample estimates of interest. However, the results from preliminary simulations with k = 5 showed that there were no noticeable differences in the estimates between the equal and unequal variances conditions. Thus, these conditions were eliminated from the present study.

Parameter Specification

The model used for data generation in the single-facet design was the two-way (Persons by Raters) random effects ANOVA model. Under the usual assumptions, the total variance in the design is expressed as:

[3-1]
$$\sigma^2_x = \sigma^2_p + \sigma^2_r + \sigma^2_e.$$

From this, a population G coefficient for relative decisions can be defined as:

[3-2]
$$G_1 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_e/n_r}.$$

Alternatively, it also can be expressed in terms of parameters of the population covariance matrix constructed over persons as: [3-3]

$$G_1 = \frac{k\ \underline{cov}}{\underline{var} + (k-1)\underline{cov}}, \qquad \text{(where, } k = n_r\text{)}.$$

Since the term σ^2_r is not included in the calculation of G coefficient, it was not included (i.e., set to zero) in the data generation. In addition, the expected variance σ^2_x, the mean of the diagonal elements of the population covariance matrix, was set to 100. It follows from the formula [3-3] that the average of the covariances, and thus the error variance (i.e., $\underline{var}$-$\underline{cov}$), can be defined for a fixed value of the population G coefficient and a fixed level of the repeated measurements, k. Therefore, these specifications allowed for the generation of a data set from a k-variate normal distribution with a particular covariance matrix as input in the simulations.

A computer program written on Turbo PASCAL was developed to automate computational procedures involved in constructing population matrices of interest. With specified population G and k values, the computer program first constructs a population covariance matrix with an epsilon equal to unity (i.e., all covariances are equal). The procedure was repeated to obtain nine matrices, three matrices under each of k = 3, 5, and 7. Second, each of nine matrices was further manipulated by varying the range of the covariances, while keeping the average constant, in such a way that the resulting population matrix possesses a desired epsilon value, either a moderate (.70) or a

low value (.50) of epsilon. In addition, the pattern of covariances in these matrices was so arranged that it resembled a simplex structure (i.e., the covariance element decreases as one moves away from the main diagonal) or a form of moving average model (i.e., the covariance element away from the main diagonal becomes zero).

Table 3-1.

Characteristics of population parameters in the covariance matrices used for data generation in the one-facet design

k	G	Epsilon			σ^2_p	σ^2_e
		Low	Medium	High		
3	.90 :	.5268	.7051	1.00	75	25
	.75 :	.5395	.7051	1.00	50	50
	.60 :	.5389	.7088	1.00	33.33	66.67
5	.90 :	.5001	.6930	1.00	64.29	35.71
	.75 :	.5179	.7061	1.00	37.5	62.5
	.60 :	.5047	.7066	1.00	23.08	76.92
7	.90 :	.5069	.7024	1.00	56.25	43.75
	.75 :	.5080	.7046	1.00	30	70
	.60 :	.5135	.6993	1.00	17.65	82.35

Unfortunately, there are an infinite number of ways to make an epsilon less than unity for a given G and k because each covariance element influences the epsilon computations. Although the computations of epsilon were fully automated with the computer program, the process to obtain epsilon values close to each other across conditions, while keeping the covariance matrix positive definite, was not an easy task. Because an exact method of doing so was not available, at least to the

researcher at this time, it was based on trial-and-error attempts. Eventually, there were a total of 27 population covariance matrices generated, each defined by a combination of three sampling parameters, G, ε, and k (see Table 3-1), which met the desired conditions. (All 27 population covariance matrices are presented in Appendix B.)

Data Generation

A FORTRAN 77 program under UNIX at the University of British Columbia was developed to conduct the simulations. Simulations were performed in the following order; first, each population covariance matrix was used as input to generate a population (N=90000) of continuous data from a k-variate normal distribution with a mean vector of 50 for each of k levels using a subroutine DRNMVN in the International Mathematical and Statistical Library (IMSL, 1991). Second, the simulated population continuous data were independently transformed into a Likert-scale with respective scale proportions of normal and uniform distributions as shown in Table 3-2. To be more precise, the mean (m) and standard deviation (s) of a simulated population continuous data (N=90000) were first computed. Then, the computed m and s, along with a normal deviate score (z) that corresponds to a specified scale proportion, were used to compute a cut-off point (c) for each measurement scale by the formula: $c = zs + m$. The rationale for doing the data transformation in this manner was based on an assumption that the population of the observed Likert-scale data has an

underlying continuous metric with normal (or uniform) distribution characteristics. Both continuous and dichotomous data served as baselines in comparing the performance of G coefficients obtained from the simulations. Third, within each of the simulated population data sets (N=90000), a sample size of 15, 30, and 45 per analysis was sequentially selected (for n=15 every third line of the data was selected). Finally, analyses were performed in order to obtain the following sample estimates; mean square, G coefficient, epsilon, confidence interval, F ratio, correlation between the estimates, and other seemingly necessary information. All computations were done with a double-precision FORTRAN 77 routines. The IMSL subroutines used in the FORTRAN program for the one-facet were: DRNMVN, RNSET, UMACH, DCHFAC, DCORVC, DFDF. For all data sets simulated, 2000 replications were performed for each sample size.

Table 3-2

Scale proportions of transformed data

Distribution	Scale Point	Scale Proportions (%)[a]				
Normal	C	Continuous				
	3	27	46	27		
	5	11	24	31	24	11
Uniform	2	50	50			
	3	33	33	33		
	5	20	20	20	20	20

[a] The scale proportions for the normal distributions are based on Cox's (1957) 'optimum grouping' scale proportions.

Simulation II: Two-facet design

Design specification

The focus of the investigation in the simulation II was on the two-facet (3 Occasions by 5 Raters) random effects model. The choice of this design and its dimensions was rather arbitrary, but based on practical considerations in observational studies. The two-facet, fully-crossed design has received extensive treatments in most published texts and tutorial papers (e.g., Brennan, 1983; Cardinet et al., 1976; Shavelson & Webb, 1991) because of its broad range of application in practice. Particularly, the two-facet design has been applied to a wide range of behavioral research to deal with measurement problems since modifications of the design (e.g., treating a facet as a fixed or nested facet) can easily lead to the formulation of G coefficients for various decision studies (e.g., Brennan, 1983; Cronbach et al., 1972). In addition, unlike educational or psychological test development studies, most observational studies employing G theory rarely involve a large number of conditions of a facet due to practical constraints in real world. Therefore, we felt that the 3 by 5 two-facet random effects model would adequately reflect a typical observational study in practice.

Sampling Populations

The same four independent variables that were used in the one-facet design, except for the variation of the levels of the facet, were incorporated into the two-facet design. There were:

3 G values, 3 epsilon values, 3 sample sizes, and 6 measurement scales. The combination of these parameter conditions resulted in 162 data sets simulated (3 x 3 x 3 x 6), each having 2000 replications.

Parameter Specification

The model used for data generation in the two-facet design was the three-way (Persons by Occasions by Raters) random effects ANOVA model presented in chapter II. Under the usual assumptions, the total variance in the design is expressed as:

[3-4]
$$\sigma^2_x = \sigma^2_p + \sigma^2_o + \sigma^2_r + \sigma^2_{po} + \sigma^2_{pr} + \sigma^2_{or} + \sigma^2_e.$$

From this, a G coefficient for relative decisions for this design can be defined as:

[3-5]
$$G_2 = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{pr}}{n_r} + \frac{\sigma^2_e}{n_o n_r}}.$$

Since the terms σ^2_r, σ^2_o, and σ^2_{ro} are not included in the calculation of the G coefficient, they were not considered (i.e., set to zero) in the simulations. In addition, the expected variance σ^2_x, the diagonal elements of the population covariance matrix, was set to 100 in all matrices (as in simulation I). As in the one-facet design, these specifications allowed for the generation of a data set from a multivariate normal distribution with a particular covariance matrix as input.

With respect to the construction of the population covariance matrices that would yield the desired population G values in the simulation II, we defined and expressed the magnitude of variance components in terms of the variances and covariances of the population matrix (see Brennan, 1983, p.99). Consider, for example, a two-facet, fully-crossed design with 3 and 5 levels of the Occasion and Rater factors, respectively. Table 3-3 illustrates a covariance matrix of 15 x 15 partitioned into nine 5 x 5 submatrices. The 5 x 5 submatrix Σ_{ij} contains the covariances among the five levels of the R factor under each level of the O factor. Therefore, an overall 15 x 15 matrix in Table 3-3 consists of three types of covariances: (a) one has levels of R in common, but different levels of O, and these are the diagonal elements of submatrices Σ_{12}, Σ_{23}, and Σ_{13} [i.e., $\mathrm{cov}(O_iR_j, O_{i'}R_j]$; (b) the second type has levels of O in common, but different levels of R, and these are the off-diagonal elements of submatrices Σ_{11}, Σ_{22}, and Σ_{33} [i.e., $\mathrm{cov}(O_iR_j, O_iR_{j'}]$; and (c) the third type has neither levels of O nor levels of R in common, and these are the off-diagonal elements of submatrices Σ_{12}, Σ_{23}, and Σ_{13} [i.e., $\mathrm{cov}(O_iR_j, O_{i'}R_{j'}]$. These three types of covariances are directly related to the variance components in the model as described in the second half of Table 3-3. Therefore, for a fixed level of the facets, an appropriate specification of the variance component values will lead to the construction of a population covariance matrix with a desired G coefficient.

For each population G value, the variance components σ^2_p and σ^2_e were first determined by solving the equations [3-4] and [3-5] simultaneously under the conditions $\sigma^2_{po} = \sigma^2_{pr} = 0$ and $\sigma^2_o = \sigma^2_r = \sigma^2_{or} = 0$. Under these constraints, the formulas [3-4] and [3-5] are precisely the same in form as those in the one-facet design: $\sigma^2_x = \sigma^2_p + \sigma^2_e = 100$ and $G_2 = [\sigma^2_p / (\sigma^2_p + \sigma^2_e/n_o n_r)]$, respectively. The obtained value for σ^2_p and σ^2_e from these formulations is a minimum and a maximum value, respectively, for a given G, n_o and n_r in a sense that any value beyond this limit would result in a negative value for σ^2_{po} and/or σ^2_{pr}. Since we wanted a positive value, instead of zero, for the variance components for σ^2_{po} and σ^2_{pr}, otherwise the resulting covariance matrix would exhibit a compound symmetry, further computations were done by increasing the minimum value of σ^2_p by a specified amount until the value of σ^2_e became negative. All computations were fully automated by a computer program, written on Turbo PASCAL for the two-facet design, which calculates and prints out a set of suitable variance component values. When a selection is made, the subsequent routines of the program construct and print a 15 x 15 covariance matrix on the computer screen on which further manipulations can be made. This procedure was repeated to determine a covariance matrix for each of three population G values. The resulting three matrices exhibit the local circularity property for all three terms (Occasion, Rater, and their interaction).

Table 3-3

Partitioned covariance matrix for the two-facet design

Level	O_1	O_2	O_3
O_1	Σ_{11}	Σ_{12}	Σ_{13}
O_2	Σ_{21}	Σ_{22}	Σ_{23}
O_3	Σ_{31}	Σ_{32}	Σ_{33}

Note: Each Σ_{ij} is a 5 x 5 matrix containing the covariances of the five levels of the R factor.

a) σ^2_x equals the average of the diagonal elements of Σ_{11}, Σ_{22}, and Σ_{33}.

b) $(\sigma^2_p + \sigma^2_{po})$ equals the average of the diagonal elements of Σ_{12}, Σ_{23}, and Σ_{13}.

c) $(\sigma^2_p + \sigma^2_{pr})$ equals the average of the off-diagonal elements of Σ_{11}, Σ_{22}, and Σ_{33}.

d) σ^2_p equals the average of the off-diagonal elements of Σ_{12}, Σ_{23}, and Σ_{13}.

e) $\sigma^2_e = a - (b - d) - (c - d) - d$.

The next step involves the construction of a covariance matrix in such a way that one or more factors deviate with some degrees from the local circularity conditions. In doing so, each of the three matrices were further manipulated by varying the pattern and/or range of relevant covariance elements in a submatrix, while holding the mean constant. The matrix was again so arranged that the pattern of the resulting matrix resembles a simplex structure in an appropriate submatrix, while

retaining it as a positive definite matrix. After each modification in the covariance element(s), the computer program computed three epsilons from orthonormally transformed submatrices: (1) a (n_o-1) x (n_o-1) matrix for Occasion; (2) a (n_r-1) x (n_r-1) matrix for Rater; and (3) a $(n_o-1)(n_r-1)$ x $(n_o-1)(n_r-1)$ matrix for the Occasion by Rater interaction term. This process was again based on trial-and-error attempts and repeated until a desired epsilon value for each term was obtained. There were a total of 9 population covariance matrices, each with a desired G value and three epsilon values. Table 3-4 presents the characteristics of population conditions, along with variance components related to each matrix. (All nine population covariance matrices are presented in Appendix B.)

Table 3-4

Population characteristics of the nine covariance matrices

Matrix	ε	G:	Variance Component				Epsilon		
			σ^2_p	σ^2_{po}	σ^2_{pr}	σ^2_e	O	R	OR
1	NONE	.90:	54	7	8	31	1.0	1.0	1.0
2	O/OR	.90:	54	7	8	31	.6729	1.0	.6752
3	R/OR	.90:	54	7	8	31	1.0	.6810	.4495
4	NONE	.75:	34	17	18	31	1.0	1.0	1.0
5	O/OR	.75:	34	17	18	31	.6543	1.0	.6742
6	R/OR	.75:	34	17	18	31	1.0	.6810	.5673
7	NONE	.60:	22	24	23	31	1.0	1.0	1.0
8	O/OR	.60:	22	24	23	31	.6552	1.0	.6542
9	R/OR	.60:	22	24	23	31	1.0	.6566	.5149

Note: O = Occasion, R = Rater, and OR = the interaction

As seen in Table 3-4, within each population G value, there were three covariance matrices, each having a different condition of violation of local circularity: (1) none of the three terms violates the local circularity (NONE); (2) only the Occasion and OR interaction terms violate the local circularity (O/OR); and (3) only the Rater and OR interaction terms violate the local circularity (R/OR).

Note that the variance component for σ^2_e is kept constant across all matrices. This was done on purpose. The reason was that because it is the same measurement design used across conditions, it is unlikely that the magnitude of the unmeasured and random sources of variation confounded with three-way interaction would vary, assuming that a person's trait remains constant over the conditions of the facets. Thus, any reduction in the G value would be more attributable to the inconsistency of ratings among raters themselves within an occasion and/or to the variations in ratings from one occasion to another, rather than due to a certain time interval or situational interferences in a measurement setting. Although the magnitude of variance components were chosen somewhat arbitrarily, we felt that their relative proportion may adequately reflect a common measurement setting in observational studies.

Data Generation

The nine population covariance matrices were used as input in the simulation program for the two-facet written on FORTRAN 77 under UNIX in order to generate a population (N=90000) of

continuous data from a multivariate normal distribution using an IMSL subroutine. The same procedure as in the one-facet design was used to transform each of the 9 simulated population data independently into a Likert-scale form. This resulted in 54 data sets, from each of which a sample size of 15, 30, and 45 per analysis was sequentially selected. For all data sets simulated, 2000 replications were performed, and the following sample estimates computed from each analysis; mean square, G coefficient, epsilon, confidence interval, quasi F ratio, Satterthwaite's degrees of freedom, correlation between the estimates, and other seemingly necessary information. All computations were done with a double-precision FORTRAN 77 routines. The IMSL subroutines used in the FORTRAN program for the two-facet were: DRNMVN, RNSET, UMACH, DCHFAC, DCORVC, DFDF, DFIN.

CHAPTER FOUR: RESULTS AND DISCUSSION

In this chapter the results of Monte Carlo simulations for the one-facet and two-facet designs are presented. First, the effect of categorization on the G coefficient based on simulated population data (N=90000) across the simulated conditions is examined. It was presumed that the simulated population data set was large enough to describe the population characteristics of the G coefficient across the simulated conditions, and thus would serve as a baseline for investigating the sampling behavior of estimated G coefficients. Second, sample estimates of the G coefficient, and the empirical variability of these estimates, are compared across simulated sampling conditions to investigate the degree of precision and efficiency of the sample estimate. Third, the empirical percentage of the sampling distribution of the estimates are compared to the corresponding theoretical values to assess the adequacy and robustness of the sampling theory of the G coefficient. Finally, the empirical percentage of Type I error rates of a conventional F test (in the one-facet design) and quasi F ratios (in the two-facet design) for the 'trial effect' are presented and compared to previous findings in related literature. The primary purpose of this component of the study was to provide verification of the simulation procedures used in the present study. In relation to the Type I error rates, the empirical performance of epsilon is examined, and its effect on the Type I error rate across categorical scales is discussed.

Simulation I: One-facet design

A. Calculated population G coefficient (G_{cp})

Table 4-1 summarizes the effects of categorization of continuous data on the G coefficient across the levels of ε, G, and k. The values in each line in Table 4-1 were based on a unique simulated population of N=90000 (i.e., 27 different simulated populations). The G coefficient calculated on the simulated population data was named the calculated population G (G_{cp}), in order to distinguish it from the population G defined in the simulation conditions.

As can be seen in Table 4-1, the G_{cp} under the continuous data (C) was virtually identical with the corresponding population G value, and consistently so across the levels of ε and k. However, under the categorical scales, the G_{cp} decreased considerably as the scale approached a uniform 2-point scale. For example, the G_{cp} was reduced by about .02-.03 from C to a 5-point scale (N5 and U5), .06-.07 from C to a 3-point scale (N3 and U3), and .11-.13 from C to a uniform 2-point scale (U2). This decrease was consistent across the 3 levels of G values in terms of absolute magnitudes. However, relative to the magnitude of G, the reduction in G_{cp} was greater for the two smaller G values. A similar trend, but a slightly less reduction, was shown in G_{cp} across the categorical scales for k = 5 and k = 7.

There appears to be no effect of heterogeneity of covariance (ε) on G_{cp} for continuous data as the G_{cp} was virtually identical across the levels of ε within the same G

Table 4-1

The effect of categorization of continuous data on the G coefficient (G_{cp}) with simulated population data (N=90000)

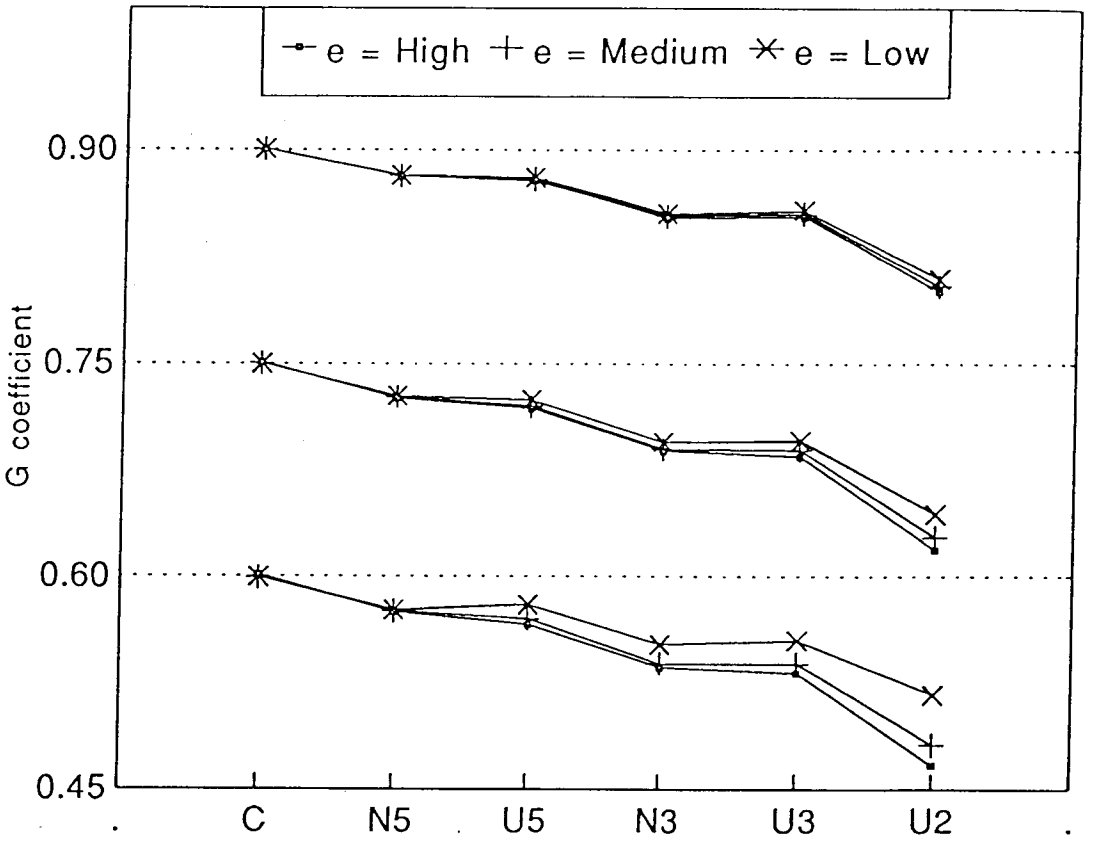
k	G	ε:	C	N5	U5	N3	U3	U2	(C-U2)
3	.90	H:	.9006	.8743	.8737	.8376	.8421	.7802	.1204
		M:	.9006	.8749	.8750	.8395	.8447	.7875	.1131
		L:	.9006	.8750	.8764	.8448	.8479	.8000	.1006
	.75	H:	.7511	.7206	.7139	.6765	.6740	.6020	.1491
		M:	.7509	.7219	.7181	.6807	.6813	.6155	.1354
		L:	.7509	.7234	.7237	.6879	.6915	.6386	.1123
	.60	H:	.6010	.5718	.5612	.5256	.5232	.4558	.1452
		M:	.6008	.5735	.5710	.5345	.5349	.4823	.1185
		L:	.6007	.5788	.5849	.5558	.5603	.5357	.0650
5	.90	H:	.9006	.8813	.8783	.8518	.8527	.8007	.0999
		M:	.9006	.8816	.8795	.8535	.8551	.8044	.0962
		L:	.9005	.8819	.8805	.8552	.8574	.8095	.0910
	.75	H:	.7510	.7256	.7181	.6886	.6847	.6186	.1324
		M:	.7503	.7265	.7200	.6900	.6892	.6274	.1229
		L:	.7500	.7272	.7247	.6948	.6958	.6438	.1062
	.60	H:	.6011	.5749	.5659	.5355	.5317	.4672	.1339
		M:	.5991	.5756	.5700	.5385	.5384	.4816	.1175
		L:	.5993	.5763	.5801	.5520	.5548	.5165	.0828
7	.90	H:	.9004	.8838	.8803	.8589	.8580	.8121	.0883
		M:	.9003	.8844	.8812	.8600	.8598	.8144	.0859
		L:	.9002	.8843	.8822	.8610	.8614	.8175	.0827
	.75	H:	.7502	.7279	.7197	.6931	.6893	.6280	.1222
		M:	.7498	.7276	.7213	.6950	.6922	.6334	.1164
		L:	.7494	.7285	.7257	.6989	.6996	.6537	.0957
	.60	H:	.6000	.5752	.5660	.5387	.5323	.4714	.1286
		M:	.6000	.5765	.5713	.5416	.5400	.4849	.1151
		L:	.5979	.5772	.5781	.5491	.5547	.5158	.0821

Note: The three conditions of ε were denoted by H, M, and L to represent the values 1.0, .70, and .50, respectively.

value, and consistently so across the levels of k. However, it can be noticed by examining the changes in G_{cp} across the levels of both G and ε that the effects of categorization interact with the population G and ε conditions. Consider, for example, the changes in G_{cp} across the categorical scales for k = 5. It is evident that the G_{cp} with a higher epsilon became gradually smaller than that with a lower epsilon as the scale approaches U2. For instance, the difference in G_{cp} between the C and U2 scales was the largest for the highest ε value for all three levels of G, as shown in the last column of Table 4-1. Furthermore, this difference appeared to be more apparent for the two smaller G conditions. The nature of this interaction can be more clearly comprehended from a graph shown in Figure 4-1, which illustrates the pattern of change in G_{cp} from C to U2 scales across the levels of ε for k = 5. One further note from Figure 4-1 is that there was little difference in G_{cp} between the normal and uniform distributions as long as it has the same number of categories (i.e., 3-point and 5-point scale). This trend was consistent across the levels of k as well. Although the results for k = 3 and k = 7 are not explicitly discussed here, they were virtually identical, in terms of trends, to that for k = 5. From these results it is clear that the transformation of continuous variables into categorical scales resulted in a substantial reduction in G_{cp}, especially for a 3-point or less scale.

Figure 4-1

The effect of categorization on the G coefficient (G_{cp}) with simulated population data (k = 5, N = 90000)



As far as the effect of categorization is concerned, these results were not surprising, and in fact showed somewhat expected trends. However, the interactive effects between the categorization, and the population G and ε conditions on the performance of G_{cp} needs further clarification. Consider first the effect of categorization. Cohen (1983), for example, investigated the effect of a 2-point scale on the effect size and power in correlational research. He demonstrated that the

dichotomization of correlated continuous variables results in the systematic loss of a considerable amount of measurement information, which consequently produces a significant reduction in the original bivariate correlation. He further noted, based on Peters and Van Voorhis (cited in Cohen, 1983), that when two variables X and Y with a correlation (r) are both dichotomized at their means, the resulting correlation is reduced to .637(r). Although the work of Cohen was based on the cost of dichotomization in the context of a bivariate correlation, the general concept can be applied to describe the performance of G_{cp} under the categorical scales, especially for the U2 scale. For example, as described in the previous chapter, G_1 can be expressed using an alternative formula (and its standardized version) as:

[4-1]

$$G_1 = \frac{k\,\underline{cov}}{\underline{var} + (k-1)\underline{cov}} = \frac{k\,\underline{r}}{1 + (k-1)\,\underline{r}}.$$

It is evident from this formula that a decrease in the magnitude of $\underline{cov}$ or $\underline{r}$ brought about by categorization produces a drop in the resulting G coefficient (i.e., G_{cp}) for a fixed value of k. Consider, for example, that the G_{cp} = .6280 under the U2 scale with k = 7, G = .75, and ε = H (i.e., 1.0) in Table 4-1. From the above equation it can be easily shown that the $\underline{r}$ must be .1943 in order for the G_{cp} under U2 scale to be equal to .6280. In the corresponding population covariance matrix, the covariances and variances were all set to 30 and 100, respectively (or $\underline{r}$ = .3 for the correlation matrix), which

yielded the G_1 = .75 and ε = 1.0. Thus, the average correlation ($\underline{r}$) was reduced from .30 to .1943 as a result of dichotomization of the k variables, which is very close to Cohen's predicted value of .637(.3) = .1911. Although the aforementioned discussion may not be mathematically explicit in describing the performance of G_{cp} under all categorical scales used in the present study, it does provide a general justification for the effect of categorization on the G coefficient.

With respect to the interactive effects between the categorization, G and ε conditions on the performance of G_{cp}, it appears that the pattern of G_{cp} shown in Table 4-1 across the simulated conditions was a result of the population characteristics defined in the population covariance matrix. As described in chapter III, the construction of the population covariance matrix was done in such a way that it produced a desired G and ε value. Particularly, the covariances in a population matrix with an $\varepsilon < 1.0$ varied widely in size from their mean, and this was even more pronounced for a combination of a large k and a small G value. When the categorization was applied to the data generated from the population covariance matrix with all covariances being equal (i.e., ε = 1.0), it would affect all covariances similarly and thus reduce them about the same amount. On the other hand, it is probable that the same procedure for the population matrix with an $\varepsilon < 1.0$ could have a smaller effect, in terms of the size of reduction, on those elements which were already so low (close to zero) than it would have on larger covariances. Therefore, in general, the

population covariance matrix with a combination of the lowest ε and the smallest G values would be relatively less affected by these transformations, and thus result in a higher G_{cp} (i.e., less reduction). In addition, this phenomenon would be more apparent if a crude grouping interval (i.e., dichotomy) is used in categorization. The last column of Table 4-1, which presents the difference in G_{cp} between C and U2 scales, reflects these aforementioned population characteristics. In all cases, the least reduction in G_{cp} occurs in the G = .60 and ε = L conditions. With these population characteristics of the G coefficient under the simulated conditions in mind, we now examine the sampling characteristics of the estimated G coefficient.

B. Estimated G coefficient ($\hat{G}_1$)

The results in this section are examined for the effects of the five independent parameters; categorization, ε, G, n, and k, on the estimated G coefficient, which was calculated based on the observed mean squares as: $\hat{G}_1 = (MS_p - MS_e) / MS_p$. Additionally, the empirical variabilities of $\hat{G}_1$ are compared across the sampling conditions as well as to the corresponding theoretical value. The results for the 482 conditions, each with 2000 replications, are summarized and are presented in Tables 4-2, 4-3, and Figure 4-2.

Table 4-2

The mean of $\hat{G}_1$, averaged over the levels of n and k, and the calculated population G coefficient (G_{cp}) across the six scales

G	ε:	C	N5	U5	N3	U3	U2	(C-U2)
.90	H:	.8911	.8705	.8693	.8392	.8418	.7856	.1055
	M:	.8912	.8709	.8706	.8407	.8438	.7902	.1010
	L:	.8912	.8710	.8717	.8436	.8464	.7975	.0937
.75	H:	.7268	.7008	.6953	.6615	.6586	.5871	.1397
	M:	.7268	.7022	.6978	.6639	.6638	.5967	.1301
	L:	.7267	.7028	.7026	.6691	.6717	.6185	.1082
.60	H:	.5622	.5349	.5277	.4941	.4892	.4208	.1414
	M:	.5627	.5375	.5351	.4996	.4993	.4406	.1221
	L:	.5625	.5392	.5446	.5139	.5181	.4827	.0798
G	ε:	G_{cp}, averaged over the levels of k						
.90	H:	.9005	.8798	.8774	.8494	.8509	.7977	.1028
	M:	.9005	.8803	.8786	.8510	.8532	.8021	.0984
	L:	.9004	.8804	.8797	.8537	.8556	.8090	.0914
.75	H:	.7508	.7247	.7172	.6861	.6827	.6162	.1346
	M:	.7503	.7253	.7198	.6886	.6876	.6254	.1249
	L:	.7501	.7264	.7247	.6939	.6956	.6454	.1047
.60	H:	.6007	.5740	.5644	.5333	.5291	.4648	.1359
	M:	.6000	.5752	.5708	.5382	.5378	.4829	.1171
	L:	.5993	.5774	.5810	.5523	.5566	.5227	.0766

Table 4-2 presents the mean of $\hat{G}_1$ across the levels of ε and G under the six scales, averaged over the levels of n and k, as well as the corresponding results for G_{cp} for comparative purposes. The results showed, in general, that the pattern of $\hat{G}_1$ across the simulated conditions is very similar to that shown for G_{cp}, except for one aspect, that is, the magnitude of bias in $\hat{G}_1$. The results also indicate that there was no effect

of heterogeneity of covariance on the magnitude of $\hat{G}_1$ -- note that the $\hat{G}_1$ values under the continuous data were virtually identical among the three levels of ε within the same G value. Furthermore, they gradually decreased as the scale approached U2, but with no appreciable difference between normal and uniform distributions under the same number of categories. This trend in $\hat{G}_1$ closely resembled that of G_{cp}. Additionally, as can be seen in the last column of Table 4-2, the decrease in $\hat{G}_1$ from C to U2 scales was larger for $\varepsilon = 1.0$ within the same G value, and it was more apparent for the two smaller G conditions.

With respect to the bias in $\hat{G}_1$, as shown in chapter II, the amount of bias is independent of the number of repeated measures (k), but increases with decreasing G_1 (see Equation 4-2).

[4-2]

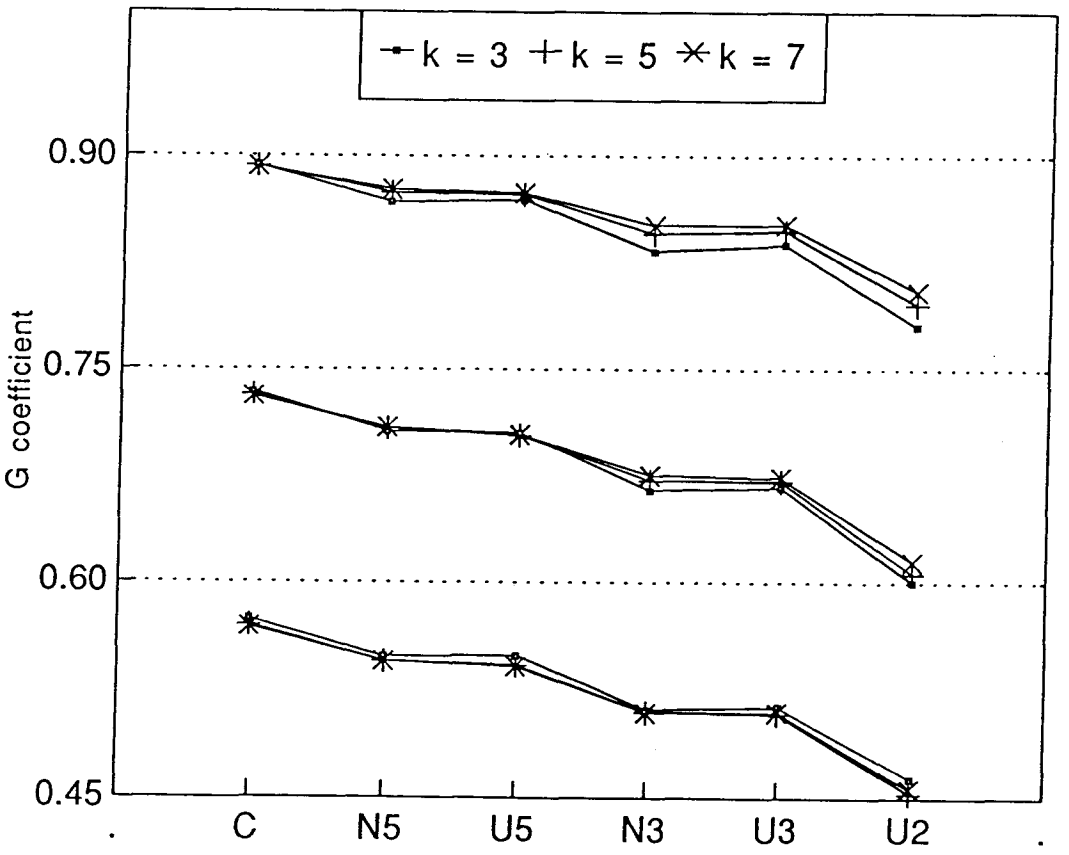
$$E(\hat{G}_1) = \frac{G_1(n_p-1) - 2}{n_p-3}$$

Figure 4-2 illustrates the pattern of $\hat{G}_1$ for the three levels of k across the categorical scales for some selected conditions (i.e., n = 30, averaged over the levels of ε). Note that the sample estimate of the G coefficient was biased, and the amount of bias in $\hat{G}_1$ was larger for the two smaller G values. However, it was virtually identical, within the same G value, for the levels of k under the continuous data. Although the amount of bias in theory is independent of k, it was slightly smaller for the two larger k under the categorical

scales. This trend was due to the population characteristics discussed previously in relation to G_{cp}, rather than sampling ones.

Figure 4-2

The mean of $\hat{G}_1$ (n = 30, averaged over the levels of ε)



Since the dimension of k does not affect the amount of bias in $\hat{G}_1$ (and as is shown later, it also has a relatively small effect on the variance expression for $\hat{G}_1$), we investigate the performance of $\hat{G}_1$ and its empirical variability more thoroughly with k = 5 across simulated conditions. Table 4-3 presents the mean and standard deviation of $\hat{G}_1$ across all simulated conditions for k = 5, each entry being based on 2000 replications. Note first that $\hat{G}_1$ increases with increasing n, reflecting the reduction in the negative bias of $\hat{G}_1$ with larger sample sizes. The results in Table 4-3 support the general statements made in relation to Table 4-2, that is, the pattern of $\hat{G}_1$ across the categorical scales was repetitively demonstrated, regardless of sample sizes. Since the interactive effects between the categorization, and the population G and ε conditions on the pattern of $\hat{G}_1$ across the categorical scales seemed to be consistent, the sampling behavior of the G coefficient across simulated conditions is examined after the effect of categorization has been partialled out. Each entry in Table 4-4 is a difference between the calculated population G (G_{cp}) shown in Table 4-1 and the estimated G_1 ($\hat{G}_1$) values in Table 4-3. Note that the difference was greater for a smaller G value, but consistent across the levels of ε and the categorical scales within the same G value. Therefore, it may be safe to state that the amount of bias is relatively constant across all six scales within the same G value, regardless of the conditions of heterogeneity of covariances, though it seems to be slightly larger for the U2 scale.

Table 4-3

The mean (standard deviation) of $\hat{G}_1$ (k = 5, 2000 replications)

n	G	ε:	C	N5	U5	N3	U3	U2
15	.90	H:	.8848(.0574)	.8657(.0646)	.8643(.0654)	.8340(.0775)	.8374(.0762)	.7809(.1088)
		M:	.8850(.0585)	.8651(.0670)	.8660(.0649)	.8361(.0799)	.8383(.0776)	.7841(.1055)
		L:	.8849(.0601)	.8659(.0670)	.8667(.0662)	.8375(.0808)	.8413(.0773)	.7885(.1034)
	.75	H:	.7110(.1434)	.6868(.1540)	.6810(.1542)	.6489(.1622)	.6463(.1669)	.5689(.2135)
		M:	.7109(.1471)	.6883(.1532)	.6841(.1549)	.6479(.1715)	.6491(.1695)	.5795(.2055)
		L:	.7106(.1528)	.6864(.1639)	.6863(.1634)	.6523(.1766)	.6540(.1744)	.6003(.1989)
	.60	H:	.5370(.2303)	.5087(.2456)	.5054(.2428)	.4733(.2453)	.4645(.2621)	.3946(.3109)
		M:	.5338(.2445)	.5109(.2481)	.5081(.2484)	.4764(.2556)	.4739(.2586)	.4068(.2985)
		L:	.5367(.2512)	.5097(.2665)	.5162(.2581)	.4864(.2641)	.4890(.2620)	.4433(.2873)
30	.90	H:	.8929(.0340)	.8739(.0393)	.8717(.0396)	.8441(.0469)	.8457(.0470)	.7909(.0665)
		M:	.8930(.0351)	.8742(.0397)	.8729(.0396)	.8455(.0476)	.8477(.0478)	.7957(.0642)
		L:	.8931(.0366)	.8747(.0410)	.8741(.0408)	.8471(.0484)	.8501(.0484)	.8005(.0646)
	.75	H:	.7314(.0862)	.7060(.0936)	.7000(.0940)	.6688(.1017)	.6653(.1044)	.5956(.1245)
		M:	.7312(.0877)	.7082(.0931)	.7016(.0945)	.6697(.1039)	.6704(.1046)	.6031(.1277)
		L:	.7315(.0899)	.7086(.0973)	.7057(.0982)	.6745(.1044)	.6758(.1067)	.6197(.1248)
	.60	H:	.5695(.1381)	.5429(.1473)	.5351(.1452)	.5025(.1546)	.4980(.1610)	.4318(.1757)
		M:	.5690(.1385)	.5452(.1453)	.5402(.1470)	.5060(.1545)	.5054(.1581)	.4453(.1759)
		L:	.5701(.1422)	.5455(.1499)	.5500(.1477)	.5199(.1556)	.5218(.1563)	.4821(.1692)
45	.90	H:	.8960(.0251)	.8770(.0291)	.8746(.0296)	.8473(.0352)	.8487(.0354)	.7951(.0512)
		M:	.8960(.0259)	.8773(.0297)	.8759(.0300)	.8489(.0365)	.8510(.0362)	.7992(.0498)
		L:	.8959(.0271)	.8776(.0306)	.8769(.0309)	.8505(.0374)	.8533(.0366)	.8042(.0503)
	.75	H:	.7391(.0638)	.7142(.0698)	.7076(.0711)	.6766(.0784)	.6733(.0799)	.6051(.0962)
		M:	.7385(.0658)	.7153(.0699)	.7096(.0712)	.6776(.0804)	.6778(.0794)	.6140(.0974)
		L:	.7381(.0692)	.7155(.0744)	.7137(.0755)	.6820(.0831)	.6838(.0824)	.6310(.0939)
	.60	H:	.5818(.1031)	.5558(.1106)	.5478(.1114)	.5159(.1190)	.5120(.1225)	.4458(.1343)
		M:	.5800(.1067)	.5569(.1107)	.5521(.1123)	.5182(.1213)	.5189(.1207)	.4607(.1351)
		L:	.5805(.1113)	.5570(.1169)	.5619(.1154)	.5314(.1245)	.5352(.1211)	.4960(.1298)

Table 4-4

The difference between the calculated population G (G_{cp}) and the estimated G ($\hat{G}_1$) values for k = 5 and n = 30

G	ε :	C	N5	U5	N3	U3	U2
.90	H:	.0077	.0074	.0066	.0077	.0070	.0098
	M:	.0076	.0074	.0066	.0080	.0074	.0087
	L:	.0074	.0072	.0064	.0081	.0073	.0090
	mean:	.0076	.0073	.0065	.0079	.0072	.0092
.75	H:	.0196	.0196	.0181	.0198	.0194	.0230
	M:	.0191	.0183	.0184	.0203	.0188	.0243
	L:	.0185	.0186	.0190	.0203	.0200	.0241
	mean:	.0191	.0188	.0185	.0201	.0194	.0238
.60	H:	.0316	.0320	.0308	.0330	.0337	.0354
	M:	.0301	.0304	.0298	.0325	.0330	.0363
	L:	.0292	.0308	.0301	.0321	.0330	.0344
	mean:	.0303	.0311	.0302	.0325	.0332	.0354

With respect to the sampling variability of $\hat{G}_1$, which is one of the main focuses in the present study, it is quite clear from the variance expression of $\hat{G}_1$ in Equation [4-3] that the variability of $\hat{G}_1$ is larger for smaller G values within the same sample size, but decreases with increasing sample size (see also Table 4-3).

[4-3]

$$\text{var}(\hat{G}_1) = (1-G_1)^2 \frac{2(n_p-1)[k(n_p-1) - 2]}{(k-1)(n_p-3)^2(n_p-5)}.$$

The empirical results in Table 4-3 show that with the continuous data the empirical standard deviations for the ε = 1.0 condition was practically identical to its theoretical counterpart (see

Table 4-5), but was larger than the theoretical value for the ε = M and L conditions (e.g., the empirical standard deviation for the condition with n = 15, G = .90, and k = 5 equals .0574, as compared to the corresponding theoretical value of .0575, as shown in Table 4-5). This pattern was consistent across the levels of n and G. This result indicates that heterogeneity of covariance exerted a small but reliable effect on the sampling variability of $\hat{G}_1$. The cause for this increase can be attributed to the effect of ε on the observed mean squares. As can be seen in Table 4-6, the standard deviations of MS_p were virtually identical for the three levels of ε within the same G value, and close to their population counterpart. On the other hand, the variability of MS_e increased for the $\varepsilon < 1.0$ conditions. Therefore, given that the variability of MS_p was stable across the levels of ε, it is clear that the larger variability in $\hat{G}_1$ for the two lower ε conditions was mainly due to the larger sampling variability in MS_e.

Table 4-5

The theoretical standard deviation of $\hat{G}_1$ for k = 5

G :	n = 15	30	45
.90 :	.0575	.0337	.0261
.75 :	.1437	.0843	.0652
.60 :	.2300	.1349	.1021

Note: The theoretical standard deviation was obtained by taking the square root of Equation [4-3]

Table 4-6

The mean (standard deviation) of the observed mean square for persons (MSp) and error (MSe) for some selected conditions (k = 5, n = 30, continuous data only)

G	ε:	MS_{person}	MS_{error}
.90	H:	360.34 (95.37)	35.85 (4.77)
	M:	359.72 (95.38)	35.78 (5.62)
	L:	359.47 (95.40)	35.72 (6.59)
	Pop:	367.16 (93.79)	35.71 (4.69)
.75	H:	251.58 (66.47)	62.79 (8.35)
	M:	250.73 (66.14)	62.74 (9.77)
	L:	250.12 (65.72)	62.59 (11.47)
	Pop:	250.00 (65.65)	62.50 (8.21)
.60	H:	193.21 (50.98)	77.31 (10.27)
	M:	191.91 (50.17)	77.12 (12.02)
	L:	192.22 (50.37)	77.13 (14.36)
	Pop:	192.32 (50.51)	76.92 (10.10)

Note: The 'Pop' values are population mean squares defined in the simulation and their theoretical standard deviation calculated by:

$$[\ 2(EMS_i)^2 / df_i\ ]^{1/2}.$$

With respect to the sampling variability of $\hat{G}_1$ under categorical scales, it is evident from Table 4-3 that the empirical variabilities increased considerably as the scale approached U2. These results were somewhat expected when considering the fact that the magnitude of the population G value (G_{cp}) has the largest relative impact on the variance expression of $\hat{G}_1$. Since the G_{cp} value for the categorical

scales, particularly for a 3-point or less scale, was already substantially lower than its population counterpart, its sample estimate inevitably produced a larger standard deviation. A comparison of the pattern of changes in the variability of $\hat{G}_1$ among the three levels of ε conditions within the same n and G shows some interesting trends. For example, the empirical standard deviation for the ε = 1.0 condition was the largest under the continuous data, but a reverse pattern was shown under the U2 scale. These rather complicated trends in the variability of $\hat{G}_1$ across the categorical scales appeared to be a result of the interactive effects among population conditions discussed earlier in relation to G_{cp}.

Additional insight for the trend in the variability of $\hat{G}_1$ across the categorical scales may be obtained by examining the sampling variability of $\hat{G}_1$ after the effect of categorization has been partialled out. To do so, the theoretical standard deviations were calculated using G_{cp}, instead of using G_1, for some selected conditions (k = 5 and n = 30) and are presented in Table 4-7. As can be seen in Table 4-7, the theoretical standard deviations of $\hat{G}_1$ for continuous data were very similar for the three ε conditions. However, as the scale approached U2, the theoretical standard deviation for the ε = H condition became larger than the other two conditions within the same G, and the difference among the three ε conditions was quite noticeable especially under the U2 scale for G = 60. These results were somewhat expected when considering that the categorization resulted in a smaller G_{cp} for the ε = H

condition, which lead to a larger theoretical variability of $\hat{G}_1$. When the empirical standard deviations for the three ε conditions in Table 4-3 were compared to their corresponding theoretical values in Table 4-7 across the categorical scales, it was apparent that the effect of heterogeneity of covariance on the sampling variability of $\hat{G}_1$ became smaller as the scale approached U2. This appears to be due to the sampling characteristics of epsilon -- the epsilon estimates under categorical scales are considerably larger than those for continuous data, as is shown later in section D. In general, the pattern of changes in empirical standard deviations for the categorical scales in Table 4-3 appears to reflect the trend shown in Table 4-7. Taken together, these results suggest two things. First, a larger standard deviation of $\hat{G}_1$ for the categorical scales and for the ε = H condition was due to the interactive effects among population conditions (i.e., types of scales, G and ε). Second, the sampling variability of $\hat{G}_1$ became less sensitive to the heterogeneity of covariance for the categorical scales, especially for a 3-point or less scale.

In summary, the results reported in this section indicate that the sample estimate of $\hat{G}_1$ was biased, especially for the two smaller G values. However, the amount of bias became trivial for the condition with $G \geq .75$ and $n \geq 30$. The heterogeneity of covariance did not have any effect on the magnitude of $\hat{G}_1$, but they did have some positive effects on the sampling variability of the estimate, especially for continuous data. Although its effect was not large, it would be large

enough to lead to more variable estimates of the G coefficient (rather than to bias the estimate). This would result in too many large estimates of the G coefficient (as well as too many small ones). Lastly, the effect of categorization on the G coefficient in terms of population characteristics was consistent in the sample estimates. Thus, the G coefficient under the categorical scales was seriously underestimated, which resulted in a large sampling variability of $\hat{G}_1$, especially for a 3-point or less scale. These results led us to question the adequacy of the sampling theory of G_1 with the categorical scales.

Table 4-7

The theoretical standard deviations for some selected conditions ($k = 5$ and $n = 30$), calculated by using G_{cp}, instead of G_1

G	ε :	C	N5	U5	N3	U3	U2
.90	H:	.0338	.0403	.0413	.0503	.0500	.0677
	M:	.0338	.0402	.0409	.0498	.0492	.0664
	L:	.0338	.0401	.0406	.0492	.0484	.0647
.75	H:	.0846	.0932	.0957	.1058	.1071	.1295
	M:	.0848	.0929	.0951	.1053	.1056	.1266
	L:	.0849	.0927	.0935	.1037	.1033	.1210
.60	H:	.1355	.1444	.1474	.1578	.1591	.1810
	M:	.1362	.1441	.1461	.1567	.1568	.1761
	L:	.1361	.1439	.1426	.1522	.1512	.1642

C. Empirical sampling distribution of $\hat{G}_1$

The characteristics of the sampling variability of $\hat{G}_1$ across the simulated conditions reported in the previous section are directly related to the properties of the sampling distribution of $\hat{G}_1$ in this section. Therefore, with these characteristics of the variability of $\hat{G}_1$ in mind, we now examine the resultant empirical sampling distributions of $\hat{G}_1$ under the simulated conditions and compare them to the theoretical ones in order to assess the precision of the sample estimate and robustness of the sampling theory of the G coefficient. The main focus in this section is on the effect of heterogeneity of covariance on the performance of $\hat{G}_1$.

As shown in chapter II, to describe and evaluate the empirical sampling distribution of $\hat{G}_1$ we can use either the confidence interval approach by establishing the $100(1-\alpha)\%$ limits for the population G_1 value or the tolerance interval approach by constructing the theoretical $100(1-\alpha)\%$ limits for the estimated G coefficient, namely:

[4-4]

$$[\ \text{Lower limit} < \hat{G}_1 < \text{Upper limit}\ ]$$

$$\left[\ 1 - \frac{1 - G_1}{F_L} < \hat{G}_1 < 1 - \frac{1 - G_1}{F_U}\ \right]$$

where, the F_L and F_U are the critical F values, corresponding to the lower $(\alpha/2)$ and the upper $(1-\alpha/2)$ percentage points from the F distribution with degrees of freedom (n_p-1) for the numerator and $(n_r-1)(n_p-1)$ for the denominator (note that $n_r = k$).

The empirical percentage of the confidence intervals failing to include the population G value is essentially the same as empirical percentage of the estimated G coefficients falling beyond the limits of the tolerance interval. Therefore, we use the latter (using Equation [4-4]) approach to present the results.

We first conducted a chi-square goodness of fit test (Gibbons, 1985) between the empirical sampling distribution and the theoretical F distribution in order to assess and evaluate the adequacy of sampling theory of G coefficients. An empirical sampling distribution of $\hat{G}_1$ with k = 5, n = 15, and G_1 = .75 (6000 replications) was obtained for each of the three population epsilon conditions (ε = 1.0, .70, and .50). Using Equation [4-4] a theoretical tolerance limit of $\hat{G}_1$ for each of the following eleven percentiles of F distribution: 1.0th, 2.5th, 5th, 10th, 25th, 50th, 75th, 90th, 95th, 97.5th, and 99th, was computed. The empirical proportion of estimated G coefficients that fell below each of these theoretical tolerance limits was tabulated for each empirical sampling distribution, and the results are presented in Table 4-8 (frequencies were used for chi-square calculations, but proportions are presented for interpretation).

As can be derived from in Table 4-8, the empirical proportion in each region for the ε = 1.0 condition was very close to the corresponding theoretical value, and the goodness of fit test ($\chi^2_{(11)}$ = 9.2133, p = .6022) also indicates a high degree of agreement between theoretical and empirical sampling

distributions. However, significant goodness of fit tests for the $\varepsilon < 1.0$ conditions suggests that the empirical sampling distribution of $\hat{G}_1$ was not quite in agreement with the theoretical distribution when the circularity assumption failed, especially for severe noncircularity (i.e., $\varepsilon = .50$).

Table 4-8

Empirical sampling distribution of $\hat{G}_1$ and a goodness of fit test ($k = 5$, $n = 15$, $G_1 = .75$, and 6000 replications in each condition with continuous data only)

Theoretical percentile	Empirical proportions $\varepsilon = 1.0$	.70	.50
99.0	98.7	98.6	98.0
97.5	97.2	96.6	95.5
95.0	94.5	94.0	92.5
90.0	89.2	88.6	86.8
75.0	74.5	73.6	72.0
50.0	49.6	49.4	48.8
25.0	25.4	25.7	25.9
10.0	10.2	10.3	10.7
5.0	5.4	5.3	5.6
2.5	2.8	3.1	2.9
1.0	1.2	1.1	1.3
Mean of $\hat{G}_1$	.7077	.7080	.7074
Empirical SD	.1449	.1471	.1563
Theoretical SD	.1437	for all three conditions	
χ^2 (11)	9.2133	36.0073	124.4687
p	.6022	<.01	<.001

Examination of the empirical proportions for the $\varepsilon < 1.0$ conditions suggests that the sampling distribution of $\hat{G}_1$ was more spread out, thus leaving a larger proportion in both tails of the distribution. This is also evident by a larger empirical

standard deviation of $\hat{G}_1$ for the $\varepsilon < 1.0$ conditions. For example, the empirical proportion above the 95th percentile was 5.5%, 6.0%, and 7.5% for $\varepsilon = 1.0$, .70, and .50, respectively. Although a goodness of fit test for the $\varepsilon = .70$ condition was significant ($\chi^2_{(11)} = 36.0073$, $p < .01$), the departure of empirical proportion of $\hat{G}_1$ from the corresponding theoretical value in each region does not seem to be too serious for practical purposes -- with a large sample size (i.e., 6000), a chi-square test may detect even a minuscule departure from the theoretical distribution, and thus is almost certain to lead to a significant result. Nevertheless, these results indicate that the sampling theory of G coefficient is very satisfactory when the circularity assumption is met, but quite sensitive to severe noncircularity.

We now examine empirical sampling distributions of G coefficients in details across the simulated conditions. For each simulation condition 2000 replications were performed. Table 4-9 presents the empirical percentage of $\hat{G}_1$ falling beyond the theoretical limits of $100(1-\alpha)\%$ tolerance interval, averaged over the levels of G, n, and k (for $\alpha = .10$ and .05, two-tailed). Thus, the results in this table represents a general pattern of the effect of heterogeneity of covariance on the sampling distribution of $\hat{G}_1$ across the six categorical scales. To assess the extent to which the other simulated conditions affect the empirical proportion of $\hat{G}_1$, Table 4-9 was further broken down by the levels of G, n, or k. The respective results, presented in Tables 4-9 through 4-11 separately, were

used to investigate interaction effects on the empirical proportion across the categorical scales.

Table 4-9

Empirical percentage of $\hat{G}_1$ falling beyond the limits of the 100(1-α)% theoretical tolerance interval, averaged over the levels of k, n, and G

α	ε:	C	N5	U5	N3	U3	U2
Upper 5%	H:	5.2	2.5	2.4	.9	1.1	.5
	M:	5.9	2.9	2.9	1.3	1.4	.7
	L:	7.2	3.7	3.9	2.0	2.2	1.1
Lower 5%	H:	5.0	9.4	10.2	20.1	20.1	40.7
	M:	5.1	9.3	9.8	19.6	18.8	38.2
	L:	5.7	9.8	9.7	18.7	17.8	33.5
Upper 2.5%	H:	2.6	1.2	1.2	.4	.5	.3
	M:	3.3	1.4	1.5	.6	.7	.4
	L:	4.2	2.0	2.2	1.0	1.1	.6
Lower 2.5%	H:	2.5	5.1	5.5	12.6	12.4	30.4
	M:	2.6	5.1	5.3	12.3	11.7	28.5
	L:	2.8	5.4	5.2	11.5	10.8	24.4

As can be seen in Table 4-9, the empirical proportions of $\hat{G}_1$ falling beyond the limits of 100(1-α)% tolerance interval for the continuous data were very close to the nominal levels for the ε = 1.0 condition, indicating that the sampling theory of G_1 works well for continuous data. However, the results also indicate that there was some positive effects of heterogeneity of covariance on the sampling distribution of $\hat{G}_1$. These results are somewhat expected when we consider the results of the goodness of fit test reported above and the effect of ε conditions on the sampling variability of $\hat{G}_1$ described in the

previous section. Note also that the positive inflation was more apparent with the empirical percentage beyond the upper limit of the tolerance interval. The reason for this may be due to the skewness of F distribution. The pattern of the empirical percentage for the continuous data was very consistent across the levels of G, n, and k, as is shown later in Tables 4-9 through 4-11.

Although the sampling theory of the G coefficient seems to work well for the continuous data across all the simulated conditions, it is not adequate for the categorical scales, especially for a 3-point or less scale. As can be seen in Table 4-9 the empirical percentage beyond the lower limit increased considerably as the scale approached the U2 scale, whereas a reverse pattern was evident in the upper limit direction. Furthermore, the comparison among the three ε conditions reveals that this trend was more apparent with the higher ε condition.

The effect of the population G values on the empirical percentage of the tolerance interval of $\hat{G}_1$ is shown in Table 4-10. As discussed in the previous section, the $\hat{G}_1$ is a negatively biased estimator for its G_{cp}. Additionally, categorization resulted in a lower value of G_{cp} for the $\varepsilon = 1.0$ condition under categorical scales, which was considerably smaller than for the corresponding G_1 value. On the other hand, the theoretical limits of the tolerance interval of $\hat{G}_1$ are constructed using G_1, thus producing a narrower width of the theoretical limits for a larger G_1. Therefore, as can be seen in Table 4-10, it is not surprising that the $\varepsilon = 1.0$ and G = .90

conditions yielded a relatively larger empirical percentage beyond the lower limit under the categorical scales. For example, the mean (and standard deviation) of $\hat{G}_1$ for k = 5, n = 15, and G = .90 under the U2 scale is .7809 (.1088) (see Table 4-3), whereas the corresponding 90% theoretical lower and upper limits from Equation [4-4] are .7770 and .9488. Since the mean is already close to the lower limit, a large number of $\hat{G}_1$ would be falling below its lower limit, but little beyond its upper limit.

Table 4-10

Empirical percentage of $\hat{G}_1$ falling beyond the limits of the 90% theoretical tolerance interval, averaged over the levels of k and n

α	ε	G :	C	N5	U5	N3	U3	U2
Upper 5%	H	.90:	5.2	1.4	1.6	.3	.5	.3
		.75:	5.2	2.8	2.5	.9	1.2	.5
		.60:	5.2	3.4	3.1	1.6	1.8	.8
	M	.90:	6.1	1.6	2.0	.4	.6	.4
		.75:	5.8	3.3	3.0	1.4	1.5	.6
		.60:	5.8	3.9	3.7	2.0	2.1	1.1
	L	.90:	7.4	2.2	2.6	.6	.9	.4
		.75:	7.3	4.1	4.0	1.8	2.1	.8
		.60:	7.0	5.0	5.1	3.6	3.5	2.1
Lower 5%	H	.90:	4.9	12.7	13.5	32.6	31.3	64.6
		.75:	5.0	8.3	9.4	16.3	16.7	35.7
		.60:	5.1	7.3	7.9	11.3	12.3	21.8
	M	.90:	5.0	12.6	13.1	32.0	30.0	62.9
		.75:	5.3	8.4	9.1	15.9	15.7	33.3
		.60:	4.9	6.9	7.1	10.9	10.8	18.5
	L	.90:	5.5	13.2	13.3	30.5	29.1	59.4
		.75:	5.8	8.9	9.0	15.3	14.6	27.5
		.60:	5.8	7.3	7.0	10.1	9.5	13.5

Table 4-11

Empirical percentage of $\hat{G}_1$ falling beyond the limits of the 90% theoretical tolerance interval, averaged over the levels of k and G

α	ε	n:	C	N5	U5	N3	U3	U2
Upper 5%	H	15:	5.1	3.1	3.3	1.4	1.8	1.1
		30:	5.1	2.3	2.3	.9	1.0	.3
		45:	5.3	2.1	1.7	.6	.6	.1
	M	15:	5.7	3.7	3.8	1.9	2.2	1.3
		30:	5.9	2.6	2.6	1.1	1.2	.5
		45:	6.2	2.5	2.2	.9	.9	.2
	L	15:	7.3	4.6	5.0	2.8	3.2	1.8
		30:	7.1	3.4	3.6	1.7	1.8	.8
		45:	7.3	3.3	3.1	1.4	1.5	.6
Lower 5%	H	15:	5.1	7.4	7.8	12.2	12.3	24.6
		30:	5.3	9.8	10.5	20.6	20.6	43.0
		45:	4.5	11.1	12.4	27.5	27.5	54.5
	M	15:	4.9	7.2	7.6	12.3	11.9	23.3
		30:	5.4	9.7	9.7	19.8	19.4	40.3
		45:	4.9	11.0	11.9	26.7	25.2	51.1
	L	15:	5.5	7.7	7.6	11.8	11.7	20.6
		30:	5.8	10.0	10.1	19.1	18.2	35.1
		45:	5.8	11.7	11.6	25.0	23.4	44.6

With respect to the effect of the sample size on the empirical results, it is apparent from Equation [4-4] that the width of the theoretical tolerance limits of $\hat{G}_1$ becomes smaller as the sample size increases. Given that the G_{cp} for the categorical scales is already a lot smaller than its corresponding population G value, a smaller width of the limits due to a larger sample size would yield even greater proportion of $\hat{G}_1$ beyond the lower limit. Table 4-11 clearly shows the

aforementioned effect of the sample sizes in which the empirical proportion beyond the lower limit increased considerably with increasing sample sizes within the same value of ε.

It should be noted that if we had constructed the theoretical limits of the tolerance interval using the G_{cp} value, instead of using G_1, the empirical results would have shown a completely different pattern. For example, the G_{cp} under the U2 scale for the condition with G = .75, k = 5, and ε = H was equal to .6186. Replacing this value for the G_1 in Equation [4-4] (and using n = 30 for the critical F values) yields the lower and upper limits of 90% tolerance interval: .3516 and .7563, respectively (as compared to the corresponding theoretical limits of .5750 and .8403, which were calculated using the G_1 = .75). The mean of $\hat{G}_1$ for this condition was .5956 with the standard deviation of .1245, as shown in Table 4-3. Therefore, if the limits calculated using G_{cp}, instead of G_1, were used, most of 2000 sample estimates for this condition would have fallen within these limits. This example also demonstrates that the categorization was the main factor that resulted in larger empirical percentage beyond the theoretical tolerance limits of $\hat{G}_1$. We could also look at this problem from statistical power point of view. For example, we know that power increases as the sample size increases. If we presume that the difference between the G_1 and G_{cp} values is indeed a true difference, the empirical percentage beyond both limits would be interpreted as power for detecting that difference, instead of Type I error, and similarly, the empirical proportion

of $\hat{G}_1$ falling within the two limits would be considered as Type II error. From these two examples, it is clear that the sampling theory of G_1 is not adequate for the categorical data, at least, under the conditions simulated in the present study.

Finally, as can be seen in Table 4-12, the effect of the number of repeated measures (k) on the sampling distribution of $\hat{G}_1$ was relatively small. The absence of an effect of the levels of k on the sampling characteristics of $\hat{G}_1$ is also evident from Table 4-13. For example, the size of the theoretical tolerance interval was practically identical among the levels of k (see Table 4-13). Although the larger value of k tends to slightly reduce the empirical percentage beyond the lower limit within the same ε, the difference in the proportions among the three levels of k was relatively small. Therefore, unless a larger number of measurements are used in a design, the use of categorical scales, especially for a 3-point or less scale, could result in a serious downward bias in estimating the population G coefficient. The dimension of k used in the present study were certainly not large enough to produce substantial differences in the empirical results across the simulated conditions.

In summary, the empirical results for continuous data were very close to the corresponding theoretical values across the simulated conditions, which suggest the adequacy of the sampling theory of the G coefficient. It was also evident that the heterogeneity of covariance had some positive effects on the sampling distribution of $\hat{G}_1$. Although the effect was not

large, it was large enough to result in a moderate inflation in the empirical percentage beyond the upper limit of the theoretical tolerance interval for $\hat{G}_1$. However, the sampling theory was not adequate for categorical data, especially for a 3-point or less scale. This inadequacy was due to the effect of categorization on the G coefficient in terms of population characteristics, which led to a serious inflation in empirical percentage beyond the lower limit of the theoretical tolerance interval. These results led us to question the practical utility of sampling theory of G_1 with categorical data.

Table 4-12

Empirical percentage of $\hat{G}_1$ falling beyond the limits of the 90% tolerance interval, averaged over the levels of n and G

α	ε	k:	C	N5	U5	N3	U3	U2
Upper	H	3:	5.2	2.5	2.6	1.1	1.4	.8
5%		5:	5.4	2.5	2.4	.9	1.1	.5
		7:	5.0	2.5	2.3	.8	.9	.3
	M	3:	6.4	3.2	3.3	1.5	1.7	1.1
		5:	5.7	2.8	2.7	1.2	1.4	.6
		7:	5.6	2.9	2.6	1.1	1.1	.4
	L	3:	7.9	3.8	4.3	2.3	2.5	1.7
		5:	7.2	3.8	3.8	1.9	2.1	.8
		7:	6.6	3.6	3.6	1.8	1.9	.8
Lower	H	3:	4.9	10.4	10.9	22.7	21.8	42.3
5%		5:	5.1	9.2	10.2	19.6	19.7	40.5
		7:	4.9	8.7	9.7	17.9	18.8	39.2
	M	3:	5.0	10.0	9.9	21.7	20.3	39.0
		5:	5.2	9.1	9.7	19.4	18.6	38.3
		7:	5.1	8.8	9.7	17.7	17.6	37.4
	L	3:	5.7	10.5	9.5	19.8	18.5	32.5
		5:	5.8	9.5	9.9	18.7	17.8	34.6
		7:	5.6	9.4	9.9	17.5	17.0	33.3

Table 4-13

The lower and upper limits of the 90% theoretical tolerance interval for $\hat{G}_1$

		G =	.90		.75		0.60	
k	n	:	LL	UL	LL	UL	LL	UL
3	15	:	.7680	.9515	.4200	.8788	.0719	.8062
	30	:	.8243	.9399	.5607	.8497	.2971	.7594
	45	:	.8429	.9340	.6074	.8350	.3718	.7359
5	15	:	.7770	.9488	.4426	.8665	.1081	.7864
	30	:	.8300	.9361	.5750	.8403	.3200	.7445
	45	:	.8474	.9308	.6185	.8270	.3896	.7233
7	15	:	.7803	.9448	.4507	.8620	.1211	.7792
	30	:	.8320	.9348	.5800	.8369	.3281	.7390
	45	:	.8490	.9297	.6225	.8242	.3960	.7187

D. Sample estimates of epsilon and Type I error rates

As noted in the introduction to this chapter, a primary purpose of examining Type I error rates in this study was to use these results as a partial validation of the simulation procedure. Inflations in error rates similar to those reported in the literature, given the same epsilon values would suggest validity and accuracy in the simulation and subsequent calculations. A second purpose was to allow for a systematic examination of the sampling characteristics of epsilon for both continuous and categorical data, as there appears to be very little published on this topic. The results of this study have, in fact, provided the validation of the simulation procedure and added some detailed information on the sampling characteristics

of epsilon. However, these results have also revealed some interesting and unexpected findings with respect to epsilon estimates and Type I error rates -- findings which cause one to question the appropriateness of current practice in applying a "correction factor" to the conventional F test in the presence of low epsilon estimate. Thus, this section on the sampling characteristics of epsilon and Type I error rates, a section which had been expected to be very brief, has been expanded considerably to present and discuss these new findings.

A number of empirical studies have been conducted to investigate the effect of heterogeneity of covariance on the degree of inflation in Type I error rates. The results from these studies were generally in close agreement, and have been well documented in related literature (e.g., Collier, et al., 1967; Hertzog & Rovine, 1985; Huynh & Feldt, 1976, 1978; Stoloff, 1970; Wilson, 1975). Empirical studies with categorical data have also been conducted to investigate the degree of positive bias in the F tests. For example, Lunney (1970) investigated the Type I error rates of the F tests in various repeated measures designs with dichotomous data and demonstrated that the actual error rates were quite close to their nominal levels. Also, Hsu and Feldt (1969), and Gregoire and Driver (1987) examined the degrees of positive bias in the F tests with categorical data. However, in these studies, the conditions of heterogeneity of covariance were not part of their simulations. Furthermore, Myers, et al. (1982) found that heterogeneous covariance in dichotomous data had a positive

inflation in the empirical Type I error rates for the usual F tests. However, they did not examine the sampling characteristics of the epsilon estimates, nor did they incorporate the corresponding conditions of continuous data in their simulations. Thus, they failed to provide information regarding the degree of relative bias in Type I error rates for dichotomous data, in comparison to those for its parent continuous conditions.

Since the early work by Box (1954) epsilon has been used as a correction factor in repeated measures ANOVA designs in order to control for probable inflation in the Type I error rates brought about by heterogeneity of covariance. However, despite its wide use, it appears that the properties of the sampling distribution of epsilon are unavailable. For any covariance matrix (either a sample or population matrix) the numerical maximum of ε is unity. If there is any deviation from the homogeneity of covariance (or from the circularity condition), ε will be less than unity. This means that a sample matrix can always be expected to exhibit some degrees of violation of the condition, even though the population matrix does not. Although we know that the estimator of ε is biased, an unbiased estimator of ε is not known [Huynh & Feldt (1976) reduced this bias by introducing a correction factor]. In addition, because of the theoretical upper and lower limits on ε [i.e., 1.0 and 1/(k-1)], $\hat{\varepsilon}$ is, in general, negatively skewed for high values of ε and positively skewed for low values of ε. Thus, we can also expect that the variability of $\hat{\varepsilon}$ for both large and small population

epsilons to be smaller than that for an ε in the middle range.

In addition to the aforementioned characteristics of epsilon, as discussed in relation to G_{cp}, we also know that categorization not only produces a smaller mean value of the covariances, but also affects (reduces) the degree of heterogeneity of covariance in the covariance matrix. Since the condition of ε is directly related to the Type I error rates in the F test, we first investigate the effect of categorization on the performance of epsilon. Second, the behavior of sample estimates of epsilon across simulated conditions is examined. Following this, the empirical Type I error rates of the conventional F test for the 'rater effect' (MS_r/MS_e) under the simulated conditions are observed and compared to previous findings in related literature.

Table 4-14 summarizes the effect of categorization of continuous data on ε across the levels of G, k, and ε. The values in each line in Table 4-14 were based on unique simulated population data sets of N = 90000. The epsilon calculated on the simulated population data was named the calculated population ε (ε_{cp}), in order to distinguish it from the population ε value. The results in Table 4-14 show that ε_{cp} values under the continuous data were virtually identical with the corresponding population ε values defined in the simulations. (Note that the difference in ε_{cp} among the three G values under the continuous scale is due to the initial population covariance matrices defined in the simulations). The results also indicate that when the population epsilon equals

unity (i.e., a perfect circularity condition), the categorization did not affect the structure of the covariance matrix, and the homogeneity of covariance condition remains the same for all categorical scales, regardless of the levels of k. However, for the $\varepsilon < 1.0$ conditions the ε_{cp} increased considerably as the scale approached U2, and this was more apparent for the higher G values (see the last column in Table 4-14). This trend in ε_{cp} was consistently shown in the three levels of k, within the same ε.

With the aforementioned characteristics of epsilon in mind, we now examine the empirical means and variabilities of $\hat{\varepsilon}$ across the categorical scales, which are summarized in Tables 4-14, 4-15, and 4-16 for some selected conditions. It is clear from Table 4-15 that the $\hat{\varepsilon}$ for $\varepsilon = 1.0$ was a (downward) biased estimator, and this bias became considerably larger with increasing k. However, for a fixed k, the magnitude of $\hat{\varepsilon}$ and standard deviations are very consistent across the categorical scales, though there was a slight decrease in $\hat{\varepsilon}$ and increase in its standard deviation at the U2 scale (see Table 4-15). As was the case in ε_{cp}, these results indicate that the categorization did not have any effect on the sample estimates of epsilon for the homogeneity of covariance condition.

Table 4-14

The effect of categorization of continuous data on epsilon (ε_{cp}) with simulated population data (N=90000)

k	ε	G :	C	N5	U5	N3	U3	U2	(U2-C)
3	H	.90:							
		.75:			All equal 1.00				
		.60:							
	M	.90:	.7052	.7918	.7697	.8542	.8301	.8815	.1763
		.75:	.7052	.7570	.7565	.8170	.8062	.8744	.1692
		.60:	.7082	.7500	.7501	.7979	.7919	.8522	.1440
	L	.90:	.5269	.6305	.6018	.7056	.6760	.7414	.2145
		.75:	.5397	.6020	.5940	.6721	.6535	.7356	.1959
		.60:	.5387	.5825	.5702	.6201	.6088	.6564	.1177
5	H	.90:							
		.75:			All equal 1.00				
		.60:							
	M	.90:	.6935	.7651	.7569	.8304	.8158	.8887	.1952
		.75:	.7061	.7549	.7599	.8083	.8074	.8752	.1691
		.60:	.7068	.7463	.7528	.7939	.7945	.8600	.1532
	L	.90:	.5010	.5883	.5767	.6815	.6566	.7640	.2630
		.75:	.5188	.5722	.5679	.6347	.6235	.7062	.1874
		.60:	.5050	.5469	.5419	.5974	.5851	.6549	.1499
7	H	.90:							
		.75:			All equal 1.00				
		.60:							
	M	.90:	.7009	.7623	.7591	.8233	.8147	.8852	.1843
		.75:	.7034	.7465	.7530	.7990	.7977	.8630	.1596
		.60:	.7001	.7370	.7407	.7830	.7816	.8440	.1439
	L	.90:	.5052	.5792	.5741	.6660	.6463	.7518	.2466
		.75:	.5072	.5530	.5468	.6080	.5932	.6627	.1555
		.60:	.5134	.5531	.5472	.5994	.5879	.6536	.1402

Note: Exact population ε values for the three levels of G and k were given in Table 3-1 in chapter III.

Table 4-15

The mean (standard deviation) of ^ε for the three levels of k (n = 15, averaged over the levels of G)

ε	Scale:	k = 3	5	7
High	C :	.8948 (.0817)	.7720 (.0769)	.6905 (.0641)
	N5:	.8932 (.0830)	.7735 (.0767)	.6945 (.0639)
	U5:	.8899 (.0846)	.7703 (.0776)	.6909 (.0637)
	N3:	.8916 (.0844)	.7729 (.0773)	.6920 (.0637)
	U3:	.8862 (.0871)	.7684 (.0776)	.6904 (.0642)
	U2:	.8702 (.1032)	.7637 (.0824)	.6895 (.0680)
Med.	C :	.6969 (.0889)	.6136 (.0921)	.5536 (.0772)
	N5:	.7437 (.0979)	.6468 (.0932)	.5808 (.0771)
	U5:	.7367 (.0992)	.6473 (.0933)	.5816 (.0759)
	N3:	.7831 (.1070)	.6785 (.0915)	.6065 (.0747)
	U3:	.7693 (.1052)	.6701 (.0916)	.6020 (.0753)
	U2:	.7919 (.1241)	.7006 (.0932)	.6324 (.0755)
Low	C :	.5369 (.0212)	.4744 (.0679)	.4329 (.0601)
	N5:	.6051 (.0557)	.5215 (.0765)	.4698 (.0633)
	U5:	.5883 (.0476)	.5169 (.0753)	.4668 (.0625)
	N3:	.6549 (.0845)	.5675 (.0852)	.5083 (.0694)
	U3:	.6364 (.0726)	.5557 (.0827)	.4978 (.0658)
	U2:	.6785 (.1095)	.6063 (.0954)	.5415 (.0748)

For the ε < 1.0 conditions, the bias in ^ε was almost null for k = 3 under the continuous scale, but decreased with increasing k. Note also that the variability of ^ε was greatest for the ε = M condition, within the same k. For the categorical scales, the nature of the bias in ^ε was shifted from negative to positive, but at which scale this change occurs varied depending on ε, k, and types of scale. This trend seemed to be a result of some interactive effects between the categorization, the theoretical limits on ε, and the nature of the downward bias in ^ε. In general, for the ε < 1.0 conditions the magnitude of

$\hat{\varepsilon}$ became smaller for larger k values, regardless of the conditions of ε and types of scale, but increased with increasing n (see Table 4-16).

Table 4-16

The mean (standard deviation) of $\hat{\varepsilon}$ for the three sample sizes (k = 5, averaged over the levels of G)

ε	Scale:	n = 15	30	45
High	C :	.7720 (.0769)	.8717 (.0499)	.9097 (.0369)
	N5:	.7735 (.0767)	.8749 (.0482)	.9127 (.0360)
	U5:	.7703 (.0776)	.8719 (.0496)	.9102 (.0374)
	N3:	.7729 (.0773)	.8744 (.0498)	.9121 (.0368)
	U3:	.7684 (.0776)	.8712 (.0497)	.9106 (.0371)
	U2:	.7637 (.0824)	.8682 (.0537)	.9094 (.0394)
Med.	C :	.6136 (.0921)	.6571 (.0755)	.6713 (.0646)
	N5:	.6468 (.0932)	.7004 (.0761)	.7180 (.0660)
	U5:	.6473 (.0933)	.6998 (.0754)	.7182 (.0661)
	N3:	.6785 (.0915)	.7426 (.0748)	.7635 (.0656)
	U3:	.6701 (.0916)	.7357 (.0761)	.7578 (.0657)
	U2:	.7006 (.0932)	.7814 (.0736)	.8116 (.0626)
Low	C :	.4744 (.0679)	.4924 (.0537)	.4971 (.0449)
	N5:	.5215 (.0765)	.5472 (.0608)	.5538 (.0517)
	U5:	.5169 (.0753)	.5403 (.0591)	.5474 (.0507)
	N3:	.5675 (.0852)	.6047 (.0691)	.6148 (.0592)
	U3:	.5557 (.0827)	.5900 (.0673)	.6004 (.0574)
	U2:	.6063 (.0954)	.6550 (.0792)	.6734 (.0691)

With respect to the effect of the population G conditions (see Table 4-17), within the same G, the size of $\hat{\varepsilon}$ was fairly consistent across the categorical scales for $\varepsilon = 1.0$, but increased as the scale approached U2 for the $\varepsilon < 1.0$ conditions. A comparison of the magnitude of $\hat{\varepsilon}$ among the three G values reveals that the $\hat{\varepsilon}$ was slightly larger for the G = .60

condition under ε = 1.0, but a reverse pattern was shown under the ε = L condition. However, the magnitude of these differences among the three G conditions was relatively small. Therefore, the magnitude of the mean of the covariances in the population matrix did not seem to have appreciable influence on the estimates of epsilon.

Table 4-17

The mean(standard deviation) of the epsilon estimates for the levels of G (k = 5, n = 15, 2000 replications)

ε	Scale:	G_1 = .90	.75	.60
High	C :	.7722 (.0775)	.7720 (.0768)	.7718 (.0764)
	N5:	.7662 (.0789)	.7739 (.0759)	.7803 (.0752)
	U5:	.7525 (.0815)	.7746 (.0772)	.7837 (.0742)
	N3:	.7609 (.0809)	.7755 (.0768)	.7822 (.0741)
	U3:	.7443 (.0817)	.7727 (.0768)	.7881 (.0742)
	U2:	.7230 (.0946)	.7754 (.0793)	.7928 (.0732)
Med.	C :	.6136 (.1016)	.6171 (.0933)	.6100 (.0814)
	N5:	.6512 (.1024)	.6478 (.0946)	.6415 (.0826)
	U5:	.6402 (.1032)	.6525 (.0938)	.6491 (.0829)
	N3:	.6833 (.0980)	.6799 (.0914)	.6722 (.0851)
	U3:	.6615 (.0970)	.6749 (.0930)	.6739 (.0848)
	U2:	.6730 (.1006)	.7117 (.0924)	.7170 (.0866)
Low	C :	.4764 (.0784)	.4826 (.0691)	.4643 (.0563)
	N5:	.5387 (.0899)	.5246 (.0776)	.5011 (.0621)
	U5:	.5290 (.0894)	.5225 (.0766)	.4993 (.0600)
	N3:	.5960 (.0994)	.5671 (.0840)	.5393 (.0721)
	U3:	.5759 (.0952)	.5589 (.0839)	.5324 (.0690)
	U2:	.6156 (.1016)	.6150 (.0968)	.5883 (.0877)

In summary, the results reported above indicate that the magnitude of $\hat{\varepsilon}$ decreased with increasing k and increased with increasing n. The results also showed that the $\hat{\varepsilon}$ was biased,

but constant across the categorical scale for $\varepsilon = 1.0$. However, for the $\varepsilon < 1.0$ conditions, the magnitude of $\hat{\varepsilon}$ increased considerably as the scale approached U2, regardless of the conditions of k, n, and G.

We now investigate the empirical Type I error rates of the conventional F test for the 'rater effect' (MS_r/MS_e) under the simulated conditions. The results of the Type I error rates are examined in relation to the behavior of the sample estimates of epsilon reported above, and also compared to previous findings in related literature.

Table 4-18 shows the general pattern of Type I error rates (α = .01, .05, and .10) for the three ε conditions across the categorical scales, averaged over the levels of k, n, and G. Note that the empirical percentage for $\varepsilon = 1.0$ was very close to the corresponding nominal levels, and consistent across the categorical scales. However, as expected the Type I error rates were noticeably inflated by the heterogeneous covariance, and the magnitudes of inflation were similar to those reported in the literature. For the $\varepsilon < 1.0$ conditions, the size of the error rates decreased as the scale approached U2. As a result, the Type I error rates with dichotomous data did not appear to be too serious, especially under the $\varepsilon = M$ condition. Note also that the error rates for the uniform distributions were slightly higher than those for the normal distributions, but the magnitude appears to be negligible. These results are generally in close agreement with previous empirical findings (e.g., Lunny, 1970; Myers, et al., 1982).

Table 4-18

Empirical percentage of the Type I error rates, averaged over the levels of k, n, and G

α	ε:	C	N5	U5	N3	U3	U2
.01	H:	0.9	0.9	0.9	0.9	1.0	0.9
	M:	2.2	2.0	2.0	1.8	1.8	1.4
	L:	3.5	3.1	3.2	2.6	2.8	2.3
.05	H:	4.7	4.6	4.9	4.9	4.9	4.6
	M:	7.0	6.6	6.8	6.3	6.4	5.7
	L:	8.9	8.0	8.4	7.5	7.8	7.0
.10	H:	9.4	9.4	9.9	10.0	10.0	9.7
	M:	11.8	11.3	11.6	11.2	11.1	10.4
	L:	13.6	12.7	13.1	12.2	12.4	11.5

The results in Table 4-18 were further broken down by the levels of k, n, or G for some selected conditions, and are presented in Table 4-19 for α = .05 only. The selected conditions for the Type I error rates presented in Table 4-19 are in accordance with those used to construct Tables 4-14 through 4-16 for the epsilon estimates. Thus, the effect of the magnitude of $\hat{\varepsilon}$ on the error rates can be easily examined by comparing appropriate entries in the corresponding tables.

Table 4-19

Empirical percentage of the Type I error rates for some selected conditions

			C	N5	U5	N3	U3	U2
α	ε	k:	n = 15, averaged over the levels of G					
.05	H	3:	5.6	4.9	5.2	5.1	4.9	4.8
		5:	4.6	4.3	4.8	5.0	4.8	4.6
		7:	4.7	4.6	4.8	5.0	5.1	4.4
	M	3:	7.1	6.0	6.4	6.0	5.8	5.4
		5:	7.3	6.5	7.2	6.3	6.2	5.9
		7:	7.8	7.5	7.4	7.0	6.8	5.9
	L	3:	8.5	7.3	8.1	7.3	7.2	6.7
		5:	9.2	8.0	8.6	7.4	8.0	7.6
		7:	9.8	8.8	9.1	7.9	8.2	7.2
α	ε	n	k = 5, averaged over the levels of G					
.05	H	15:	4.6	4.3	4.8	5.0	4.8	4.6
		30:	4.6	4.9	5.0	5.2	5.1	4.4
		45:	4.7	4.8	5.1	5.2	5.3	4.9
	M	15:	7.3	6.5	7.2	6.3	6.2	5.9
		30:	6.6	6.6	6.3	6.2	6.8	6.0
		45:	6.8	6.7	7.1	6.2	7.0	5.5
	L	15:	9.2	8.0	8.6	7.4	8.0	7.6
		30:	8.8	7.8	8.5	7.2	8.0	7.1
		45:	9.0	8.5	8.7	7.9	8.3	6.9
α	ε	G:	n = 15 and k = 5					
.05	H	.90:	4.8	4.2	4.9	4.9	4.7	4.9
		.75:	4.5	4.4	4.9	4.9	4.8	4.3
		.60:	4.6	4.3	4.7	5.1	4.9	4.5
	M	.90:	7.5	6.6	7.3	6.8	6.2	5.6
		.75:	7.0	6.1	7.2	6.2	6.0	5.8
		.60:	7.3	6.9	7.2	6.0	6.5	6.4
	L	.90:	9.7	7.9	8.7	7.2	7.9	6.8
		.75:	9.1	7.6	8.0	7.6	8.2	7.3
		.60:	8.9	8.4	9.1	7.3	7.9	8.6

As can be seen in Table 4-19, the general pattern of the Type I error rates across the categorical scales was repetitively demonstrated. That is, the Type I error rates for $\varepsilon = 1.0$ were all close to the nominal level, regardless of the conditions of k, n, and G. The results also showed that the heterogeneity of covariance conditions inflated the error rates, to some extent, but the magnitude decreased as the scale approached U2. This trend was consistent across the levels of k, n, and G. Additionally, for the $\varepsilon < 1.0$ conditions, the error rates increased somewhat with increasing k within the same ε, and this trend was consistent across the categorical scales. Note also in Table 4-19 that the error rates were slightly smaller for the two larger sample sizes, but there was no appreciable difference in the error rates among the three levels of G. From these results, it is evident that the present study replicated, in general, previous findings in related literature and demonstrated that the heterogeneity of covariance increased the Type I error rates for the usual F statistic.

With respect to the relationship between the magnitude of epsilon estimates and the Type I error rates, the results in Tables 4-17 and 4-18 showed that for the $\varepsilon < 1.0$ conditions, the pattern and magnitude of changes in Type I error rates across the simulated conditions were very closely related to those shown for the sample estimates of epsilon. For example, as shown in Tables 4-14 through 4-16, the magnitude of $\hat{\varepsilon}$ decreased with increasing k, but increased with increasing n for the $\varepsilon < 1.0$ conditions. Furthermore, the magnitude of $\hat{\varepsilon}$ increased

considerably as the scale approached U2, regardless of the conditions of k, n, and G. Comparatively, the Type I error rates under the $\varepsilon < 1.0$ conditions decreased as the scale approached U2. Within the same ε, they increased for larger k values, and slightly decreased for the two larger sample sizes. Furthermore, the small difference in Type I error rates among the three G values shown on the bottom of Table 4-19 was in accordance with a marginal difference in $\hat{\varepsilon}$ among them shown in Table 4-17.

The dependency between the magnitude of $\hat{\varepsilon}$ and Type I error rates across simulated conditions is illustrated in Table 4-20, which presents a general pattern of the correlation between the two variables for the three ε conditions across the six scales. The values in each line in Table 4-20 were based on the estimates across the levels of k, n, and G (i.e., 27 conditions), which add up to 81 conditions for the total (i.e., the 'All' on the last line in the table). Inspection of Table 19 for the $\varepsilon < 1.0$ conditions shows that the correlations were negative and reasonably high under the continuous data, and were reduced as the scale approached U2. This trend in the correlation supports the aforementioned relationship between $\hat{\varepsilon}$ and Type I error rates.

While the size of the Type I error rates varied depending on the magnitude of $\hat{\varepsilon}$ across the simulated conditions for the $\varepsilon < 1.0$ conditions, it was not the case under the $\varepsilon = 1.0$ condition. As can be seen in Tables 4-14 through 4-16, the estimated epsilons for $\varepsilon = 1.0$ considerably varied in size,

ranging from .69 to .91, depending on a particular combination of simulated conditions. However, the associated Type I error rates were remarkably similar with one another, and very close to the corresponding nominal levels (see Table 4-19). This relationship is also evident from Table 4-20 as the correlation between the $\hat{\varepsilon}$ and Type I error rates for the $\varepsilon = 1.0$ condition was almost zero, especially under the continuous and 5-point scales. In other words, these results indicate that the inflation in probability of Type I error rates for the usual F tests may not be an issue, regardless of the magnitude of $\hat{\varepsilon}$, as long as it is certain that the sample data at hand are from a population which possesses homogeneity of covariance. However, when the correlation was computed based upon all three ε conditions combined, the aforementioned relationship between the $\hat{\varepsilon}$ and Type I error rates for the $\varepsilon = 1.0$ condition was completely obscured (see the correlation for the 'All' in Table 4-20).

Table 4-20

Correlation between the Type I error rates and the epsilon estimates for alpha = .05 only

ε:	C	N5	U5	N3	U3	U2
H:	-.0543	.0863	-.0246	-.1395	-.1552	.1907
M:	-.6680	-.6662	-.6080	-.5927	-.3981	-.2439
L:	-.8145	-.7709	-.7509	-.5512	-.7232	-.6784
All:	-.9304	-.8948	-.9093	-.8634	-.8683	-.7696

This highly negative correlation between the magnitude of $\hat{\varepsilon}$ and Type I error rates would lead to a misleading interpretation of the relationship between them, that is, the Type I error rates increase with decreasing sample estimates of epsilon, regardless of the population condition of ε. In fact, it is a common practice in repeated measures ANOVA designs to use $\hat{\varepsilon}$-adjusted F tests (i.e., Greenhouse-Gessier or Huynh-Feldt correction) in order to protect against a probable inflation in the Type I error rates whenever an observed epsilon is less than unity. Interestingly, an estimated epsilon from a sample covariance matrix is always less than unity due to the nature of the downward bias in $\hat{\varepsilon}$, and the population epsilon is always unknown in practice. Therefore, under such circumstances the $\hat{\varepsilon}$-adjusted F test would be correct only if one presumes that the population covariance matrix from which a sample at hand being taken possesses the heterogeneity of covariance condition. Otherwise, the $\hat{\varepsilon}$-adjusted F test would result in an unduly conservative test, and thus increase the probability of Type II error rate if the estimated epsilon is indeed from a population matrix with homogeneous covariance. This leads us to query the common practice of utilizing the $\hat{\varepsilon}$-adjusted F test in the repeated measures ANOVA designs, and raises the question: Is it always justifiable to use an $\hat{\varepsilon}$-adjusted F test when an observed epsilon is less than unity?

To summarize the results obtained in this section in terms of the relationships among sample estimates of epsilon, heterogeneity of covariance, observed mean squares, and Type I

error rates, the sample estimates of these variables are presented in Table 4-21 for some selected conditions. Consider first the observed mean squares (i.e, MS_e and MS_r). The magnitude of the mean of the observed mean squares was virtually identical among the three ε conditions, and very close to their expected value (note that $EMS_e = EMS_r$ under $\sigma^2_r = 0$, as defined in the simulation), and this trend was consistent across the categorical scales. However, the variability for both mean squares was greater under the $\varepsilon < 1.0$ conditions, while it was very close to their theoretical value under $\varepsilon = 1.0$ (note that the variance expression for mean squares was given on the bottom of Table 4-6). This resulted in a more variable sampling distribution for the F ratio than indicated by the theoretical F distribution under the $\varepsilon < 1.0$ conditions. Since interest lies in the upper tail of the distribution, the cumulative proportion beyond the upper limit are attributed to the Type I error rates, which were larger under the $\varepsilon < 1.0$ conditions as shown in Table 4-21.

For the categorical scales, the $\hat{\varepsilon}$ for $\varepsilon < 1.0$ became larger as the scale approached U2, and consequently there was little difference in the magnitude of $\hat{\varepsilon}$ among the three ε conditions under the U2 scale. As the size of $\hat{\varepsilon}$ had a positive impact on the variability of mean squares, this lack of difference in variability among the three ε conditions was shown in the observed mean squares as well as in the F ratio. Therefore, the difference in the error rates among the three ε conditions under the U2 scale was relatively small.

Table 4-21

Type I error rates, correlation, and descriptive statistics for some selected conditions (k = 5, n = 15, G_1 = .90)

Estimates:	C	N5	U3	U2
ε_{cp} (and ^ε)				
H:	1.00 (.7722)	1.00 (.7662)	1.00 (.7443)	1.00 (.7230)
M:	.6935(.6136)	.7651(.6512)	.8158(.6615)	.8887(.6730)
L:	.5010(.4764)	.5883(.5387)	.6566(.5759)	.7640(.6156)
MS_r (sd)				
H:	35.6962(24.6115)	.5471(.3783)	.3110(.2134)	.1397(.1009)
M:	35.9424(29.8885)	.5474(.4396)	.3088(.2451)	.1367(.1017)
L:	36.1523(35.4509)	.5443(.4929)	.3075(.2780)	.1367(.1110)
MS_e (sd)				
H:	35.5015(6.6744)	.5445(.1041)	.3080(.0689)	.1380(.0334)
M:	35.4153(7.8132)	.5461(.1173)	.3061(.0705)	.1369(.0337)
L:	35.4021(9.1513)	.5424(.1275)	.3021(.0739)	.1352(.0338)
EMS_e	35.71 (6.7486)	.5458	.3055	.1370
F (sd)				
H:	1.0446(.7616)	1.0418(.7613)	1.0459(.7439)	1.0461(.7611)
M:	1.0677(.9422)	1.0488(.8859)	1.0514(.8588)	1.0371(.7922)
L:	1.0958(1.1488)	1.0669(1.0385)	1.0774(1.0169)	1.0573(.8832)

Type I error (α = .05, .10)

H:	4.8	9.8	4.2	9.4	4.7	9.1	4.9	9.8
M:	7.5	12.3	6.6	11.6	6.2	11.7	5.6	10.6
L:	9.7	14.6	7.9	13.4	7.9	13.2	6.8	11.6

Correlations (^ε, MS_e, MS_r)

		^ε	MS_e	^ε	MS_e	^ε	MS_e	^ε	MS_e
H	MS_e:	.0031		.0008		.1065		.4330	
	MS_r:	.0236	-.0259	-.0062	.0105	.0143	.0898	.1991	.1307
M	MS_e:	-.4567		-.3468		-.1339		.3101	
	MS_r:	-.0068	-.0217	-.0015	.0035	.0595	.0664	.1684	.1193
L	MS_e:	-.4801		-.4754		-.2877		.1730	
	MS_r:	-.0198	-.0264	.0101	-.0218	.0602	.0100	.1596	.0880

Finally, examination of the correlations among the sample estimates ($\hat{\varepsilon}$, MS_e, and MS_r) reveals a very interesting but unclear phenomenon, especially the correlation between $\hat{\varepsilon}$ and MS_e across the levels of ε. For example, a negative correlation between them under the $\varepsilon < 1.0$ conditions suggests that a smaller $\hat{\varepsilon}$ is associated with a larger MS_e. Given that the correlations between $\hat{\varepsilon}$ and MS_r, and between MS_e and MS_r are almost zero, this negative correlation between $\hat{\varepsilon}$ and MS_e indicates that the ratio of MS_r/MS_e (F) tends to be smaller for a smaller $\hat{\varepsilon}$ under the $\varepsilon < 1.0$ conditions. This is opposite to what is reported in the literature. It appears that, given moderate to severe noncircularity in the population, F tests conducted on samples with moderate to severe noncircularity are probably <u>not</u> associated with inflated Type I error rates. If this is so, then the current practice of applying an epsilon-based correction factor to the F test in repeated measures ANOVA designs could be inappropriate. Additional studies are being conducted to explore this phenomenon and the results will be reported elsewhere (Eom & Schutz, 1993).

Simulation II: Two-facet design

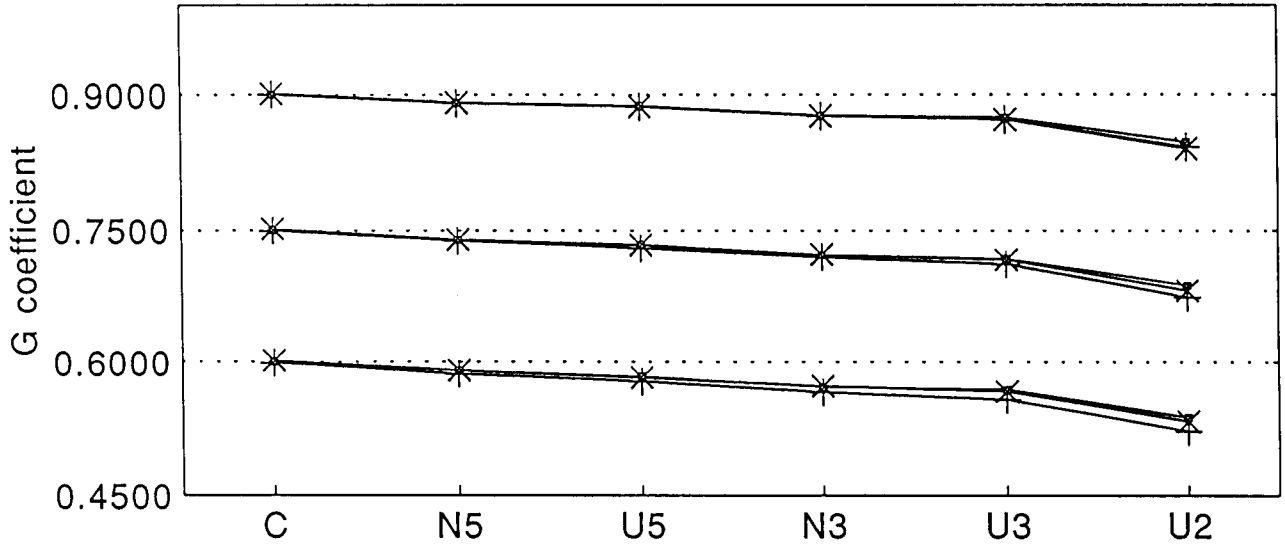
In this section, the simulation results for the two-facet (3 Occasions by 5 Raters), fully-crossed design are examined across the levels of G, n, ε, and types of scale. The results of the two-facet design closely paralleled those of the one-facet design in terms of the effect of categorization, population G value, and sample size, and thus they are discussed only briefly. More emphasis is placed on the effect of noncircularity on the sampling variability of $\hat{G}_2$, which resulted in some discrepancy between the two designs. To be consistent, the results are presented following a similar sequence as in the one-facet design. First, the effect of transformation of continuous data into categorical scales on the G coefficient is examined. Second, the characteristics of sample estimates of the G coefficient are compared across the simulated conditions. Third, the effect of noncircularity on the sampling variability of the G coefficient is investigated. Finally, empirical proportions of confidence intervals that failed to include a population G coefficient and Type I error rates of quasi F ratios for the "Occasion" and "Rater" main effects are examined and compared among the three local circularity conditions at specified alpha levels.

A. Calculated population G-coefficient (G_{cp})

Figure 4-3 illustrates the general pattern of the G coefficient (G_{cp}) across the six scales for a combination of ε and G. The values in each line shown in the bottom of Figure 4-3 were based on a unique simulated population of N = 90000 (i.e., 9 different simulated populations).

As can be seen in Figure 4-3, the G_{cp} under continuous data was identical to the corresponding population G value, and consistently so, regardless of ε conditions. However, as the scale approached U2, the G_{cp} gradually decreased. The pattern of changes in G_{cp} across categorical scales was virtually identical for the three G values, but the magnitude of decrease in G_{cp} was somewhat larger for G = .75 and .60. Note also that within the same G, the ε = O/OR condition produced a slightly smaller G_{cp} value, especially under the U2 scale, but the amount of difference among the three ε conditions seemed to be negligible. In comparison to the findings of the one-facet design, the results of the two-facet design showed a similar pattern of changes in G_{cp} across the simulated conditions. However, the magnitude of decrease in G_{cp} from C to U2 scales was somewhat smaller for the two-facet design, and also the interactive effects between categorization, G and ε conditions appear to be marginal. This difference between the two designs seemed to be due to a larger dimension involved in the two-facet design as well as due to a less variation in the covariance elements among the population covariance matrices defined in the two-facet design simulation.

Figure 4-3. Effect of categorization on the G coefficient with population data (two-facet design, N=90000)



	C	N5	U5	N3	U3	U2
G=.90, e=NONE	0.9004	0.8912	0.8877	0.8772	0.8751	0.8479
e=O/OR	0.9003	0.8903	0.8863	0.8765	0.8723	0.8426
e=R/OR	0.9011	0.8913	0.8868	0.8766	0.8724	0.8406
G=.75, e=NONE	0.7504	0.7394	0.7341	0.7225	0.7184	0.6872
e=O/OR	0.7495	0.7380	0.7295	0.7198	0.7116	0.6740
e=R/OR	0.7514	0.7397	0.7333	0.7230	0.7165	0.6815
G=.60, e=NONE	0.6001	0.5901	0.5840	0.5714	0.5691	0.5375
e=O/OR	0.5986	0.5860	0.5775	0.5661	0.5575	0.5221
e=R/OR	0.6014	0.5904	0.5820	0.5729	0.5666	0.5330

B. Estimated G coefficient ($\hat{G}_2$)

A sample estimate of the G coefficient ($\hat{G}_2$) for the two-facet, fully-crossed design is calculated based on the observed mean squares as:

[4-5]

$$\hat{G}_2 = 1 - \frac{MS_{po} + MS_{pr} - MS_e}{MS_p}$$

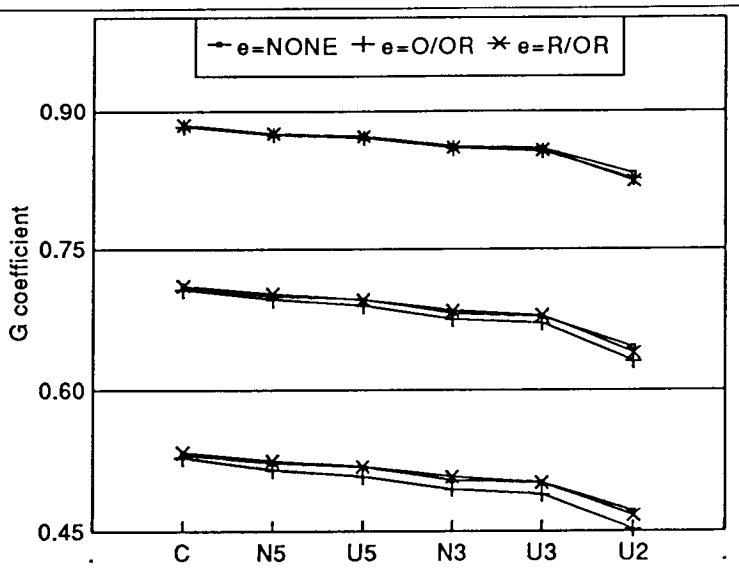
As described in chapter II, $\hat{G}_2$ is a negatively biased estimator for the population G_2 value, and the amount of bias is greater for the smaller G_2 value, but decreases with increasing sample sizes. Furthermore, it was also shown that the magnitude of bias in $\hat{G}_2$ is independent of the number of facets as well as the levels of a facet.

The aforementioned characteristics of $\hat{G}_2$ were well reflected in the empirical results, and they are graphically illustrated in Figure 4-4 separately for the three sample sizes. Note that the $\hat{G}_2$ values under the continuous data were almost identical among the three ε conditions within the same G and n. These results indicate that the violation of circularity condition does not have any effect on the magnitude of sample estimates of the G coefficient. However, under the categorical scales, the ε = O/OR condition yielded a slightly smaller $\hat{G}_2$ value, especially for a combination of $G < .90$ and $n = 15$ conditions, but the magnitude of difference in $\hat{G}_2$ among the three ε conditions was quickly reduced with increasing sample sizes.

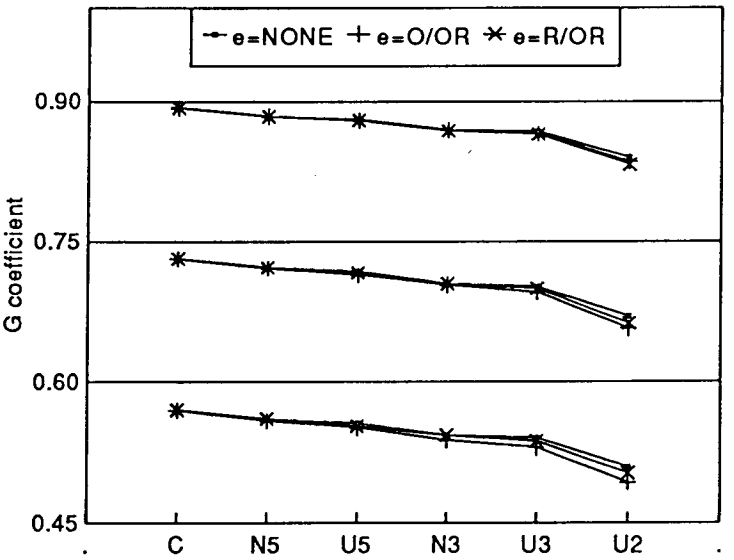
Figure 4-4

The mean of $\hat{G}_2$ for the three sample sizes (2000 replications)

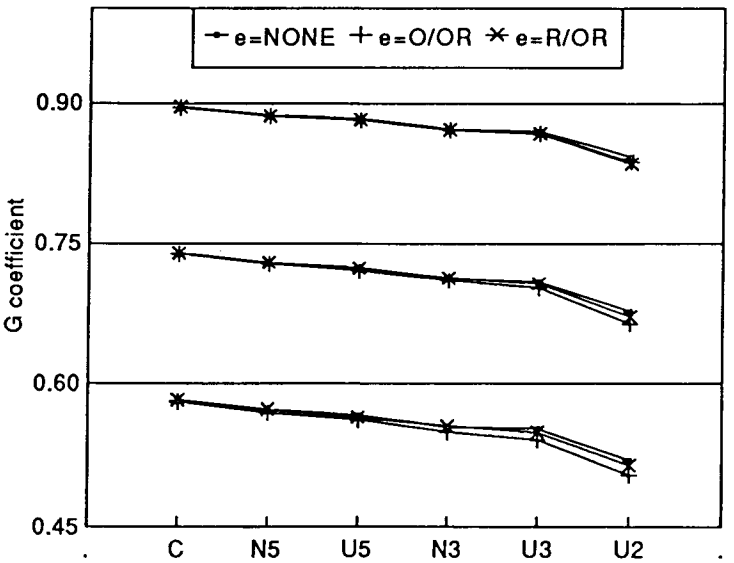
(a) n = 15



(b) n = 30



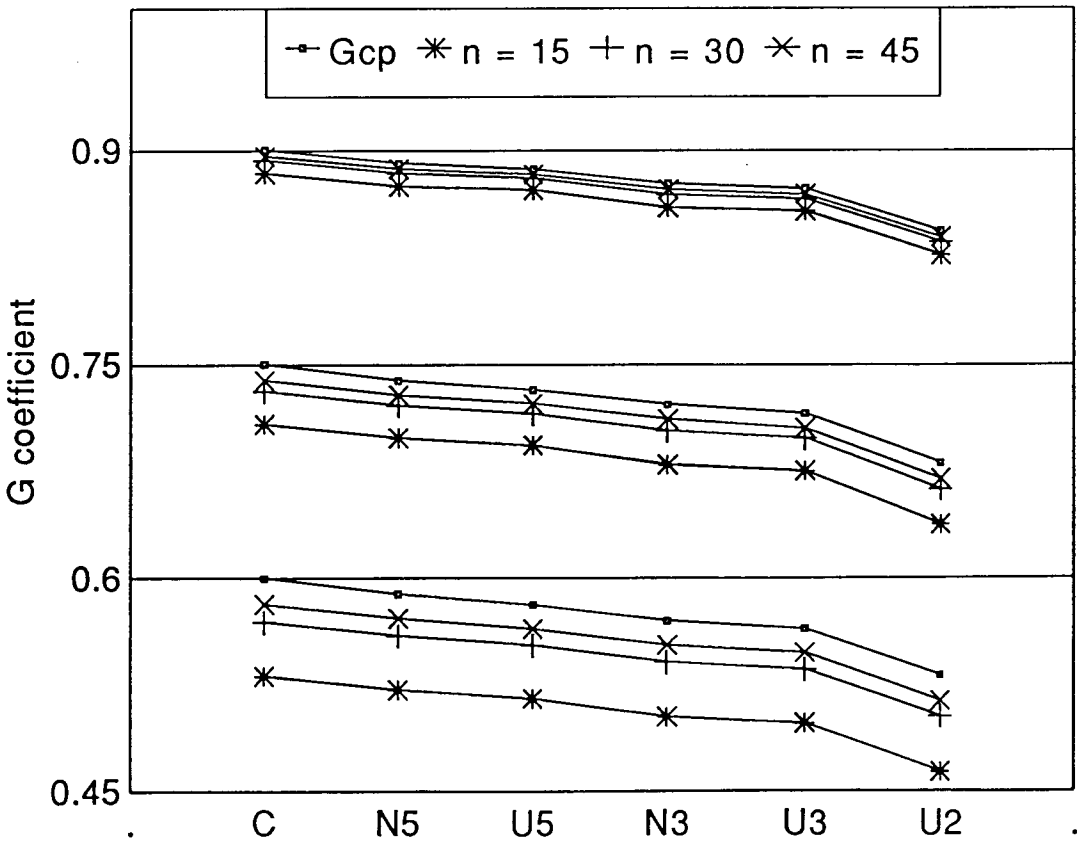
(c) n = 45



To examine the nature of the bias in the sample estimator of G_2 under categorical scales, the values of $\hat{G}_2$ were compared to the corresponding G_{cp} values and depicted in Figure 4-5 for the three sample sizes, averaged over the three ε conditions. Figure 4-5 clearly illustrates the effect of population G value and sample size on $\hat{G}_2$, that is, the amount of bias in $\hat{G}_2$ increased for the smaller G_2 values, but decreased with increasing sample sizes. The parallel trends between G_{cp} and $\hat{G}_2$ across the categorical scales also indicate that the $\hat{G}_2$ was biased, but the amount of bias was consistent across all six scales.

Figure 4-5

Comparison of G_{cp} and $\hat{G}_2$, averaged over the levels of epsilon



In summary, the pattern of changes in $\hat{G}_2$ across the simulated conditions resembled that of G_{cp} shown in Figure 4-3, except for the magnitude of bias in $\hat{G}_2$. Furthermore, the effects of the simulated conditions (i.e., categorization, G, and n) on the sample estimates were similar to those of the one-facet design, but to somewhat less extent. Although there was some discrepancy in the degree of noncircularity between the two designs, it was apparent from the results of both designs that the violation of circularity condition did not have any effect on the magnitude of both $\hat{G}_1$ and $\hat{G}_2$ under continuous data.

With respect to the sampling variability of $\hat{G}_2$, it is apparent from Equation [4-6] that the variability of $\hat{G}_2$ increases with decreasing G_2 and decreases with increasing sample sizes. Note also that the variance expression involves the Satterthwaite's adjusted degrees of freedom, f_a.

[4-6]

$$\text{var}(\hat{G}_2) = (1 - G_2)^2 \frac{2(n_p-1)^2 (f_a + n_p - 3)}{f_a(n_p-3)^2 (n_p - 5)}$$

The theoretical variabilities of $\hat{G}_2$ across the simulated conditions are presented in Table 4-22, and these values were used to examine whether the sampling theory of G coefficient is robust to noncircularity by comparing them with the corresponding empirical values among the ε conditions.

Table 4-22

Expected mean squares (EMS), Satterthwaite's degrees of freedom (f_a), and theoretical standard deviation (SD) of $\hat{G}_2$

G :	EMS_p	EMS_{po}	EMS_{pr}	EMS_e	n	f_a	SD($\hat{G}_2$)
.90 :	900	66	55	31	15	37.13	.0600
					30	76.91	.0353
					45	116.69	.0273
.75 :	680	116	85	31	15	46.75	.1462
					30	96.84	.0859
					45	146.93	.0664
.60 :	550	151	100	31	15	48.32	.2332
					30	100.11	.1369
					45	151.89	.1059

Table 4-23 presents the mean and standard deviation of $\hat{G}_2$ based on 2000 replications across the three ε conditions for some selected conditions. Note that the standard deviation for the ε = O/OR condition with n = 15 was somewhat larger than the other two conditions, but this difference vanished as the sample size increased. Examination of the range of $\hat{G}_2$ for this condition showed that the large standard deviation was mainly due to some outliers (e.g., the smallest value of $\hat{G}_2$ for ε = O/OR was .1087 with a corresponding z value of -11.9 under the continuous data, as compared to .4461 and .5060 for the other two conditions). In general, the empirical standard deviations under the continuous data were virtually identical among the three ε conditions for a fixed value of G and were close to the corresponding theoretical values reported in Table 4-22. However, it was not the case for the categorical scales. As can

be seen in Table 4-23, the variability of $\hat{G}_2$ increased considerably as the scale approached U2, and was substantially larger than its theoretical counterpart. As was the case in the one-facet design, the large sampling variability of $\hat{G}_2$ under the categorical scales appeared to be a result of the initial difference between G_2 and G_{cp} brought about by the effect of categorization. When considering the fact that the standard deviations were similar among the three ε conditions under the same categorical scale, it appears that the violation of circularity assumption had a minimal effect on the sampling variability of $\hat{G}_2$ for the categorical scales as well. Therefore, taken together, these results suggest that the sampling theory of G_2 is quite robust to the violation of circularity condition.

With respect to the effect of the ε conditions on the sampling variability of $\hat{G}_2$, the results from the two-facet design were somewhat different from those of the one-facet design. For example, as shown in Tables 4-3, under the continuous data the standard deviation of $\hat{G}_1$ was larger for the $\varepsilon < 1.0$ conditions which indicate some positive effects of heterogeneity of covariance on the sampling variability of $\hat{G}_1$. However, in the two-facet design although the standard deviation of $\hat{G}_2$ in Table 4-23 tends to be slightly larger for the ε = O/OR and R/OR conditions, the magnitude of difference among the three ε conditions seemed to be very small. Even though some discrepancy may be expected between the two designs because of the unmatched degree of ε defined in each design as well as the

design itself, the reason for why noncircularity did not have any appreciable effect on the sampling variability of $\hat{G}_2$ needs further clarification.

Table 4-23

The mean (standard deviation) of $\hat{G}_2$ for the two-facet design (n = 30 only, 2000 replications)

n	G	ε :	C	N5	U3	U2
15	.90	NONE:	.8838(.0598)	.8752(.0649)	.8595(.0751)	.8312(.0898)
		O/OR:	.8830(.0649)	.8741(.0699)	.8560(.0798)	.8256(.0964)
		R/OR:	.8851(.0587)	.8758(.0636)	.8571(.0726)	.8232(.0934)
30	.90	NONE:	.8926(.0352)	.8839(.0378)	.8678(.0437)	.8404(.0542)
		O/OR:	.8931(.0357)	.8841(.0387)	.8655(.0459)	.8352(.0579)
		R/OR:	.8932(.0359)	.8839(.0388)	.8650(.0452)	.8323(.0572)
	.75	NONE:	.7308(.0854)	.7210(.0894)	.7009(.0961)	.6697(.1050)
		O/OR:	.7315(.0855)	.7209(.0891)	.6948(.1000)	.6562(.1136)
		R/OR:	.7316(.0870)	.7215(.0903)	.6985(.0973)	.6617(.1111)
	.60	NONE:	.5688(.1365)	.5597(.1415)	.5402(.1476)	.5099(.1565)
		O/OR:	.5689(.1383)	.5580(.1428)	.5298(.1533)	.4933(.1662)
		R/OR:	.5695(.1397)	.5602(.1422)	.5366(.1507)	.5029(.1597)
45	.90	NONE:	.8958(.0265)	.8871(.0289)	.8711(.0340)	.8437(.0418)
		O/OR:	.8959(.0272)	.8869(.0297)	.8684(.0358)	.8379(.0458)
		R/OR:	.8965(.0267)	.8872(.0291)	.8684(.0346)	.8359(.0447)

To investigate the cause for this discrepancy, we examined the sampling characteristics of the observed mean squares. Table 4-24 presents the means and standard deviations of mean square estimates and pair-wise correlations between them across the three ε conditions for some selected conditions. As expected, the mean of the observed mean squares (MS_p, MS_{po}, MS_{pr}, MS_e) was very similar among the three ε conditions within

the same G, and close to their corresponding population value (to be more precise, they were almost identical to those mean squares calculated on the simulated population data set).

Table 4-24

The mean (standard deviation) of observed mean squares and their correlations (continuous data only, n = 30, 2000 replications)

		G coefficient		
MS	ε :	.90	.75	.60
MS_p	NONE:	906.13(241.03)	683.22(181.48)	551.96(146.17)
	O/OR:	903.66(240.70)	681.07(182.05)	549.78(146.57)
	R/OR:	905.34(241.44)	683.04(181.53)	551.63(146.58)
MS_{po}	NONE:	66.58(12.60)	117.09(21.99)	152.46(28.53)
	O/OR:	66.31(15.13)	116.93(27.05)	152.46(35.04)
	R/OR:	65.87(12.15)	116.55(21.90)	151.84(28.42)
MS_{pr}	NONE:	55.06(7.28)	85.15(11.33)	100.21(13.36)
	O/OR:	54.83(7.04)	84.82(10.97)	99.66(12.81)
	R/OR:	55.05(8.88)	85.08(13.61)	100.12(16.29)
MS_e	NONE:	30.98(2.84)	30.97(2.83)	30.97(2.82)
	O/OR:	30.98(3.51)	30.93(3.46)	30.91(3.50)
	R/OR:	31.08(4.30)	31.07(3.79)	31.06(3.96)

Correlation:

		MS_e	MS_{po}	MS_e	MS_{po}	MS_e	MS_{po}
NONE	MS_{po}:	-.02		-.02		-.02	
	MS_{pr}:	-.05	.00	-.05	-.01	-.05	-.01
O/OR	MS_{po}:	-.02		-.01		-.01	
	MS_{pr}:	.14	-.01	.25	-.02	.44	-.01
R/OR	MS_{po}:	.13		.06		.02	
	MS_{pr}:	-.01	-.01	-.02	-.02	.00	-.02

Note: correlations between MS_p and the other MS's were all less than |.05|.

With respect to the variability of mean squares, Table 4-24 shows that the standard deviation of MS_p was virtually identical for the three ε conditions, but that of MS_{po}, MS_{pr}, and MS_e was positively inflated under the violation of respective local circularity. For example, the standard deviation (15.13) of MS_{po} under the ε = O/OR condition was considerably larger than its theoretical counterpart [12.26 = $(2(EMS_{po})^2/df)^{1/2}$ = $(2(66)^2/58)^{1/2}$]. A similar result was shown for MS_{pr} under the ε = R/OR condition. Note also that the standard deviation of MS_e was positively inflated under both ε = O/OR and R/OR conditions, and it was slightly larger under the ε = R/OR condition -- this was due to a smaller ε value defined for the O by R interaction term in the simulation (see Table 3-4). Taken together, these results indicate that the mean square estimates were unbiased, but more variable under the violation of local circularity. Given that the variability of MS_p was fairly consistent across the three ε conditions, the variance of a linear combination of the observed mean squares in the numerator of Equation [4-5] appears to be the main source that determines the degree of sampling variability of $\hat{}G_2$. Since the mean squares in the numerator of Equation [4-5] were more variable when circularity fails, we might expect that the variability of $\hat{}G_2$ would also be somewhat larger under noncircularity condition, as was the case in the one-facet design. However, the results in Table 4-23 showed that this did not happen, and there was only a minimal effect of noncircularity on the sampling variability of $\hat{}G_2$ for the two-facet design. The

reason for this is perhaps due to the fact that the computational formula for $\hat{G}_2$ involves a combination of mean squares.

Inspection of Table 4-24 in relation to Equation [4-5] suggests a possible explanation for why the sampling variability of $\hat{G}_2$ was not sensitive to noncircularity. For example, the ε = O/OR condition produced more variable MS_{po} and MS_e, whereas the ε = R/OR condition yielded a larger variability for MS_{pr} and MS_e. Therefore, in order for $\hat{G}_2$ to be more variable under noncircularity conditions, either MS_{po} and MS_e, or MS_{pr} and MS_e must have at least some degree of negative correlation, given that the correlation between MS_{po} and MS_{pr} was essentially zero under both ε = O/OR and R/OR conditions. However, as shown on the bottom of Table 4-24, it was not the case. All correlations were nearly zero, except for those between MS_{pr} and MS_e under the ε = O/OR, and between MS_{po} and MS_e under the ε = R/OR condition. A positive correlation between MS_{pr} and MS_e, for example, may be due to the fact, in part, that MS_{pr} is a composite of MS_e and $n_o\hat{\sigma}^2_{pr}$. Thus, under the ε = O/OR condition, MS_{pr} would tend to fluctuate with MS_e, given that the quantity $n_o\hat{\sigma}^2_{pr}$ is almost independent of MS_e. The cause of the varied size of the correlation coefficients for these pairs across the levels of G is unclear, but it could be a result of the difference in the magnitude of the mean squares among the levels of G. Nonetheless, these results suggest that more-variable individual mean squares are not necessarily associated with a particularly small or large value of $\hat{G}_2$. Since the

characteristics of the sampling variability of $\hat{G}_2$ are directly related to the properties of sampling distribution of $\hat{G}_2$, we examine this problem further in the following section.

C. Empirical sampling distribution of $\hat{G}_2$

As presented in chapter II, the ratio $(1 - G_2)/(1 - \hat{G}_2)$ is approximately distributed as an F-variate with degrees of freedom (n_p-1) for the numerator and f_a for the denominator. It was also shown that from this a $100(1-\alpha)\%$ confidence interval for a population G_2 value can be constructed as:

[4-7]

$$\text{Lower limit} < G_2 < \text{Upper limit}$$

$$= 1 - (1 - \hat{G}_2)F_U < G_2 < 1 - (1 - \hat{G}_2)F_L.$$

The terms F_L and F_U are the critical values corresponding to the lower $\alpha/2$ and upper $(1-\alpha/2)$ percentage points, respectively, of the F distribution with degrees of freedom (n_p-1) for the numerator and $\hat{f}_a$ for the denominator. The quantify $\hat{f}_a$ is estimated using observed mean squares, and thus varies over replications. In the simulation, the lower and upper limits of a $100(1-\alpha)\%$ confidence interval for a specified population G_2 value were obtained for each replication. The mean and standard deviation of the limits of 2000 confidence intervals are presented in Table 4-25 for some selected conditions.

Table 4-25

The mean (standard deviation) of the limits of the 90% confidence intervals for a G_2 in the two-facet design for some selected conditions (2000 replications)

			C		U2	
n	G	ε :	LL	UL	LL	UL
15	.90	NONE:	.7703(.1161)	.9490(.0264)	.6486(.1770)	.9274(.0395)
		O/OR:	.7685(.1260)	.9487(.0287)	.6411(.1888)	.9246(.0425)
		R/OR:	.7723(.1140)	.9496(.0260)	.6344(.1831)	.9237(.0411)
30	.90	NONE:	.8263(.0562)	.9379(.0204)	.7344(.0874)	.9092(.0314)
		O/OR:	.8270(.0573)	.9382(.0208)	.7276(.0932)	.9059(.0336)
		R/OR:	.8272(.0573)	.9383(.0209)	.7220(.0917)	.9044(.0332)
	.75	NONE:	.5731(.1351)	.8428(.0499)	.4644(.1682)	.8094(.0610)
		O/OR:	.5740(.1358)	.8432(.0499)	.4452(.1821)	.8012(.0660)
		R/OR:	.5743(.1375)	.8433(.0509)	.4525(.1774)	.8047(.0646)
	.60	NONE:	.3178(.2160)	.7478(.0798)	.2101(.2501)	.7163(.0910)
		O/OR:	.3179(.2197)	.7480(.0807)	.1877(.2656)	.7059(.0966)
		R/OR:	.3188(.2208)	.7483(.0818)	.1995(.2552)	.7122(.0930)
45	.90	NONE:	.8456(.0389)	.9328(.0172)	.7632(.0617)	.9004(.0271)
		O/OR:	.8457(.0401)	.9328(.0177)	.7558(.0675)	.8964(.0296)
		R/OR:	.8466(.0392)	.9332(.0174)	.7522(.0658)	.8953(.0289)

With respect to the relationship between mean squares and $\hat{G}_2$ under noncircularity conditions, it can be anticipated from Equation [4-7] that if the more-variable mean squares obtained under the ε = O/OR or R/OR conditions are associated with a particularly small or large $\hat{G}_2$, it would result in even greater fluctuation in the limits of the confidence interval because the critical value of F_U and F_L in Equation [4-7] varies depending on its denominator degrees of freedom, $\hat{f}_a$. As can be seen in Table 4-25, the results, however, do not suggest the aforementioned relationship between mean squares and $\hat{G}_2$. There

was little difference among the three ε conditions in the mean and standard deviation of both limits of 90% confidence intervals under the continuous data within the same G and n. Therefore, these results also support the robustness of sampling theory of G_2 to the violation of circularity. One further note from Table 4-25 is that the width of the confidence interval limits for the U2 scale became narrower with increasing G and n, and thus its upper limit is already close to its population G_2. Therefore, as is shown later, a larger empirical proportion of the confidence intervals would fail to include its population G_2.

To assess the adequacy and robustness of the sampling theory of the G coefficient in the two-facet design, the empirical proportion of 2000 confidence intervals that failed to include a specified population G_2 value in either the lower or upper direction was obtained at three significance levels (α = .10, .05, .01, two-tailed). Table 4-26 presents the empirical proportion for the three ε conditions, averaged over the levels of G and n. Thus, the results in this table represent a general pattern of the effect of noncircularity on the sampling distribution of $\hat{G}_2$ across the six scales. Table 4-26 was further broken down by the levels of G or n, and the results are presented in Table 4-27 for α = .10 only. Note that the values for the upper 5%, for example, in the tables are the empirical proportion of the confidence intervals whose lower limit was greater than a specified population G_2 value. Thus, the proportion can be interpreted as a Type I error rate.

From Tables 4-25 and 4-26 it is apparent that the sampling theory of G_2 is robust to the violation of circularity condition under the continuous data as the empirical proportion was almost identical among the three ε conditions. Although the proportion was slightly larger for the ε = O/OR and R/OR conditions, the magnitude of difference seemed to be negligible. Furthermore, both upper and lower empirical proportions were all close to the corresponding nominal levels, regardless of the levels of G and n.

Table 4-26

Empirical proportion of confidence intervals that failed to include a population G_2 value (averaged over the levels of G and n)

α	ε :	C	N5	U5	N3	U3	U2
Upper	NONE:	4.7	3.7	3.5	2.5	2.3	1.3
5%	O/OR:	5.2	4.0	3.7	2.3	2.2	1.0
	R/OR:	5.2	3.7	3.5	2.4	2.2	1.1
Lower	NONE:	5.0	6.6	7.3	10.2	10.7	19.5
5%	O/OR:	5.2	7.1	8.1	10.8	12.5	22.3
	R/OR:	5.2	6.8	7.7	10.1	11.7	22.1
Upper	NONE:	2.2	1.7	1.7	1.2	1.2	.6
2.5%	O/OR:	2.7	1.8	1.8	1.1	1.0	.4
	R/OR:	2.5	1.8	1.7	1.1	1.1	.5
Lower	NONE:	2.5	3.4	3.6	5.5	5.9	12.3
2.5%	O/OR:	2.6	3.5	4.2	5.8	6.9	15.0
	R/OR:	2.6	3.4	4.0	5.6	6.6	14.4
Upper	NONE:	.5	.4	.4	.2	.2	.1
.5%	O/OR:	.5	.4	.3	.2	.2	.1
	R/OR:	.5	.4	.4	.2	.2	.1
Lower	NONE:	.4	.7	.7	1.2	1.3	3.7
.5%	O/OR:	.4	.7	.9	1.4	1.7	5.1
	R/OR:	.5	.6	.7	1.2	1.4	4.9

Table 4-27

Empirical percentage of 90% confidence intervals that failed to include a specified population G_2 value

α	ε	G :	C	N5	U5	N3	U3	U2
			averaged over the levels of n					
Upper 5%	NONE	.90:	4.6	2.9	2.8	1.3	1.2	.4
		.75:	5.0	3.9	3.6	2.7	2.5	1.5
		.60:	4.6	4.4	4.0	3.5	3.1	1.9
	O/OR	.90:	5.3	3.2	3.1	1.2	1.2	.4
		.75:	5.2	4.2	4.0	2.5	2.4	1.0
		.60:	5.2	4.5	4.0	3.3	3.0	1.6
	R/OR	.90:	4.9	2.9	2.8	1.2	1.2	.3
		.75:	5.2	3.7	3.7	2.6	2.3	1.2
		.60:	5.4	4.5	4.1	3.5	3.1	1.8
lower 5%	NONE	.90:	5.1	7.9	8.8	14.1	15.4	32.7
		.75:	4.9	6.2	6.9	8.8	9.3	15.3
		.60:	5.0	5.8	6.3	7.5	7.5	10.5
	O/OR	.90:	5.6	8.3	10.2	15.0	17.6	35.5
		.75:	5.0	6.6	7.5	9.3	11.2	18.7
		.60:	5.0	6.3	6.7	7.9	8.6	12.7
	R/OR	.90:	4.8	7.6	9.4	14.4	17.2	37.8
		.75:	5.4	6.6	7.1	8.8	9.8	17.1
		.60:	5.3	6.1	6.5	7.3	8.1	11.5
α	ε	n :	averaged over the levels of G					
Upper 5%	NONE	15:	5.0	4.7	4.5	3.4	3.2	2.1
		30:	4.6	3.5	3.3	2.2	2.0	.8
		45:	4.7	3.0	2.6	1.9	1.6	.8
	O/OR	15:	5.5	4.8	5.0	3.3	3.5	1.7
		30:	5.2	3.8	3.3	2.0	1.8	.7
		45:	5.0	3.4	2.9	1.7	1.3	.5
	R/OR	15:	5.4	4.1	4.5	3.1	3.0	1.9
		30:	5.0	3.6	3.4	2.4	2.0	.9
		45:	5.1	3.3	2.7	1.8	1.6	.4
Lower 5%	NONE	15:	5.3	6.4	6.7	8.1	8.8	13.0
		30:	5.5	7.0	7.7	10.3	11.1	19.9
		45:	4.2	6.5	7.6	12.0	12.3	25.5
	O/OR	15:	5.7	6.7	6.8	8.8	9.5	14.3
		30:	5.2	7.1	8.5	10.9	12.6	22.7
		45:	4.8	7.5	9.0	12.6	15.3	29.9
	R/OR	15:	5.4	6.4	6.7	8.2	8.7	13.9
		30:	5.6	7.2	8.0	10.8	12.2	22.8
		45:	4.5	6.8	8.3	11.4	14.2	29.8

The results in Tables 4-25 and 4-26 also indicate that the sampling theory of G_2 is not adequate for the categorical scales, especially for a 3-point or less scale. As expected, the empirical proportion decreased in the upper direction and increased considerably to an unacceptable level in the lower direction as the scale approached U2, and this trend was more apparent for larger G and n values (see Table 4-27). In general, these results reflected the characteristics of the sampling variability of $\hat{G}_2$ reported above. That is, a condition with a larger variability of $\hat{G}_2$ is associated with a larger empirical proportion in the sampling distribution of $\hat{G}_2$. Therefore, it can be concluded that the sampling theory of G_2 for the two-facet design works well under continuous data, and is quite robust to the violation of circularity condition. However, it was not acceptable for categorical scales, especially for a 3-point or less scale. As discussed in relation to the one-facet design, the cause for this inadequacy was mainly due to the effect of categorization in terms of population characteristics.

D. Type I error rates in quasi F tests

In this section, we present some results of empirical Type I error rates for quasi F ratios for the test of Rater and Occasion main effects in the context of a three-way (Subjects by Raters by Occasions) random effects ANOVA model. In general, for any given design requiring the quasi F there are several different ways to form a test statistic (Winer, 1971). The two

quasi F ratios used for our two-facet design are presented at the bottom of Table 4-28. Under the null hypothesis that σ^2_r = 0 (i.e., no Raters effect), for example, the numerator and denominator of both QFR_1 and QFR_2 have the same structure of expected values of mean squares. Thus, the test statistic can be set up in the usual way, but the degrees of freedom for those terms associated with a combination of mean squares are obtained from the Satterthwaite's procedure. In general, Satterthwaite's adjusted degrees of freedom is fractional, and thus an exact critical F value using a fractional degrees of freedom was obtained by referring to F-inverse function in IMSL subroutine in the simulation.

The reason for including the first form of the quasi F ratio (Form 1 in Table 4-28) was that the structure of first quasi F ratio is similar to that involved in the sampling distribution of $\hat{}G_2$ for the two-facet design. As reported in the one-facet design, the effect of noncircularity on the empirical proportion of the sampling distribution of $\hat{}G_1$ was not as large as that on the Type I error rates in the F test. The main cause for this is that the F test involves both MS_r and MS_e which are more variable when the circularity assumption fails, whereas the sampling distribution of $\hat{}G_1$ include MS_e and MS_p, but the variability of MS_p is not sensitive to noncircularity conditions. Therefore, for similar reasons, it may be expected that noncircularity would have a larger effect on the quasi F test than it would have on the sampling distribution of $\hat{}G_2$, and thus produce at least some positive effects on the quasi F test

for either the Rater or Occasion main effect because the quasi F ratio includes more-variable mean squares in both numerator and denominator. As can be seen in Table 4-28, the results do show that the Type I error rates for the test of both Occasion and Rater main effects using the first form of quasi F ratio were somewhat positively inflated under the violation of respective local circularity. For example, the test for the Occasion effect resulted in slightly larger Type I error rates under the ε = O/OR condition, whereas the Type I error rates for the same test were somewhat conservative or close to the nominal level under the other two ε conditions. Similar but slightly larger inflation in the Type I error rates were produced for the test of Rater effect under the ε = R/OR condition.

With respect to the second form of the quasi F ratio, Maxwell and Bray (1986) used this form in a simulation study to investigate the effect of violating sphericity (circularity) on the quasi F ratio in a three-way ANOVA design with one nested factor. They concluded that the quasi F ratio was in general quite robust to noncircularity, though it produced conservative results for some of the conditions simulated. Although the design, and thus the expected value of the mean squares, is not the same in their study and the present study, we expect that the results of this section would also show the robustness of quasi F tests to the violation of circularity. These results would then serve as a partial validation of the simulation procedure and subsequent calculation implemented in the simulation program. As can be seen in Table 4-28, the Type I

error rates for the test of the Occasion effect were very close to the nominal level, indicating the robustness of the quasi F, whereas the same test under the other two conditions was somewhat conservative. A slightly larger inflation was shown for the test of Rater effect under the ε = R/OR condition. In general, the Type I error rates for the second form of the quasi F test were smaller than those for the first form. This may be due to the fact that the second form involves the Satterthwaite's degrees of freedom in both numerator and denominator, which could be adjusted in such a way that the critical F value is larger or smaller, keeping the actual probability of a Type I error rate near the nominal level. One further note from these results is that the Type I error rates were virtually identical between the normal and uniform distributions within the same number of response scale, and they decreased as the scale approached U2, regardless of the ε conditions and the types of quasi F ratio.

In summary, the results of the two-facet design closely paralleled those of the one-facet design in terms of the effect of categorization, sample size, and population G value. However, some discrepancy was observed in terms of the effect of noncircularity on the sampling variability of the G coefficient between the two designs. That is, the sampling distribution of $\hat{G}_2$ was quite robust to the violation of circularity condition, whereas the one-facet was not. Finally, as was the case in the one-facet design, the noncircularity had more effect on the F test (and quasi F) than it had on the sampling distribution of

$\hat{G}_2$. The results also indicate that the quasi F tests were relatively robust to noncircularity, and their Type I error rates were generally in close agreement with previous findings in the literature.

Table 4-28

Empirical percentage of the Type I error rates for quasi F tests in the three-way random effects ANOVA model (averaged over the levels of G and n, each condition having 2000 replications, alpha = .05)

Quasi F	ε :	C	N5	U5	N3	U3	U2
QFO_1	NONE:	4.5	4.2	4.2	4.3	4.2	3.9
	O/OR:	6.1	6.1	5.7	5.6	5.5	5.1
	R/OR:	4.3	4.5	4.0	4.1	4.2	3.5
QFR_1	NONE:	5.3	5.3	5.4	5.0	5.2	4.6
	O/OR:	3.1	3.3	3.2	3.4	3.5	3.2
	R/OR:	7.6	7.1	7.1	6.3	6.6	5.7
QFO_2	NONE:	3.4	3.2	3.2	3.3	3.1	2.9
	O/OR:	5.0	5.0	4.6	4.4	4.4	4.0
	R/OR:	3.1	3.3	3.0	3.0	3.1	2.5
QFR_2	NONE:	4.0	4.2	4.1	4.0	4.1	3.7
	O/OR:	2.2	2.3	2.3	2.5	2.5	2.6
	R/OR:	6.4	5.8	6.0	5.3	5.4	5.0

	Occasion effect	Rater effect
Form 1:		
	$QFO_1 = \frac{MS_o}{MS_{po} + MS_{or} - MS_e}$,	$QFR_1 = \frac{MS_r}{MS_{pr} + MS_{or} - MS_e}$
Form 2:		
	$QFO_2 = \frac{MS_o + MS_e}{MS_{po} + MS_{or}}$	$QFR_2 = \frac{MS_r + MS_e}{MS_{pr} + MS_{or}}$.

CHAPTER FIVE: SUMMARY AND CONCLUSIONS

This chapter presents a brief summary and the findings of the present study, followed by the implications of the empirical results. It concludes with limitations of the present study and suggestions for future research.

The present study employed Monte Carlo procedures to investigate the interactive effect of data categorization and heterogeneity of covariance on the generalizability coefficient for the one-facet and two-facet designs as well as on the Type I error rates for the F tests in repeated measures ANOVA designs. The primary focus was to examine and compare the sampling characteristics of the G coefficients obtained on both categorical scales and their parent continuous data under the violation of the circularity assumption. Computer programs were developed to construct the population covariance matrices of interest with desired G and ε values, and to conduct a series of simulations under various sampling conditions.

One-facet design

An overview of the results with respect to the G coefficient and to the Type I error rates is illustrated in a tree diagram and presented in Table 5-1 and Table 5-2, respectively, for the one-facet design.

Table 5-1

An overview of the results regarding G_{cp}, $\hat{G}_1$, and empirical proportions beyond the theoretical limits of the tolerance interval of $\hat{G}_1$ in the one-facet design

G_1 Scale	ε	G_{cp}	k	G_{cp}	n	$\hat{G}_1$	sd	L%	U%
G_1=.90	C	1.0	.90	.90	45	.8958	.0242	4.1	5.0
					15	.8841	.0550	5.3	5.3
			7	.90	45	.8960	.0274	4.5	5.2
			3	.90	15	.8845	.0610	4.9	4.9
		.50	.90	.90	45	.8956	.0258	5.2	6.7
					15	.8844	.0580	5.2	6.5
			7	.90	45	.8959	.0312	5.7	8.3
			3	.90	15	.8847	.0660	5.6	7.9
	U2	1.0	.79	.81	45	.8066	.0457	82.8	0.0
					15	.7903	.0991	36.8	0.2
			7	.78	45	.7742	.0666	83.7	0.0
			3	.78	15	.7589	.1333	44.3	1.6
		.50	.81	.82	45	.8122	.0451	78.8	0.0
					15	.7972	.0975	33.4	0.2
			7	.80	45	.7944	.0627	76.5	0.3
			3	.80	15	.7807	.1205	37.3	1.9
G_1=.60	C	1.0	.60	.60	45	.5810	.0992	4.7	4.9
					15	.5363	.2190	5.1	5.1
			7	.60	45	.5818	.1119	4.5	5.1
			3	.60	15	.5323	.2531	5.4	5.0
		.50	.60	.60	45	.5789	.1069	6.2	6.2
					15	.5373	.2383	5.8	7.2
			7	.60	45	.5829	.1197	5.7	7.1
			3	.60	15	.5344	.2661	5.2	7.2
	U2	1.0	.46	.47	45	.4501	.1264	29.8	0.2
					15	.3979	.2772	12.9	1.2
			7	.46	45	.4356	.1508	29.0	0.4
			3	.46	15	.3737	.3492	15.2	1.9
		.50	.52	.52	45	.4940	.1252	20.0	1.0
					15	.4454	.2760	10.3	2.2
			7	.54	45	.5190	.1287	12.5	2.3
			3	.54	15	.4678	.2854	8.7	3.8

Table 5-2

An overview of the results regarding ε_{cp}, $\hat{\varepsilon}$, and Type I error rates in the one-facet design

ε	G_1	Scale	ε_{cp}	k	ε_{cp}	n	$\hat{\varepsilon}$	sd	α=.05
1.0	.90	C	1.0	7	1.0	45	.8714	.0343	4.9
						15	.6906	.0639	4.7
				3	1.0	45	.9604	.0360	5.1
						15	.8952	.0817	5.7
		U2	1.0	7	1.0	45	.8540	.0396	4.9
						15	.6541	.0762	4.3
				3	1.0	45	.9389	.0561	4.5
						15	.8295	.1292	4.8
	.60	C	1.0	7	1.0	45	.8710	.0347	4.4
						15	.6905	.0643	4.8
				3	1.0	45	.9602	.0358	4.8
						15	.8943	.0817	5.6
		U2	1.0	7	1.0	45	.8875	.0315	4.5
						15	.7152	.0630	4.4
				3	1.0	45	.9634	.0352	4.6
						15	.8990	.0842	4.8
.52	.90	C	.52	7	.51	45	.4850	.0577	9.1
						15	.4429	.0779	9.8
				3	.53	45	.5274	.0084	7.8
						15	.5285	.0167	8.9
		U2	.75	7	.75	45	.6781	.0673	6.0
						15	.5557	.0835	6.9
				3	.74	45	.7244	.0742	5.7
						15	.6839	.1161	6.7
	.60	C	.53	7	.51	45	.4809	.0301	9.9
						15	.4284	.0498	9.7
				3	.54	45	.5394	.0119	8.2
						15	.5404	.0231	8.2
		U2	.66	7	.65	45	.6107	.0462	8.2
						15	.5352	.0687	7.2
				3	.66	45	.6553	.0713	8.0
						15	.6498	.1117	7.2

Cost of data categorization. Categorization of continuous data had a marked influence on the G coefficient, resulting in a considerably smaller G_{cp} than for the parent continuous data, especially for a 3-point or less scale. Although the magnitude of reduction in G_{cp} from C to U2 scales varied in a rather complicated manner, depending on a particular combination of G, k, and ε of the simulated conditions, it was the largest with a small number of measures (i.e., k = 3). In practice, researchers rarely have control of the population parameters (G and ε), but these results can be used as a guide for planning a G study. The findings of the one-facet design suggest that in situations where a categorical response is inevitable (e.g., a G study in observational research), the practitioner should consider using a 5-point or more scale and try to avoid using a combination of a 3-point or less response category and a small number of raters (i.e., k = 3) in a G study. Otherwise, he or she may estimate a G coefficient which is already about 20% lower than its population G coefficient.

Sample estimates. The sample estimate of the G coefficient ($\hat{G}_1$) is a downward biased estimator, as shown in the mathematical derivation where $E(\hat{G}_1) < G_1$, and the amount of bias varies as a function of the size of G_1 and n, but independently of the number of measures (k). The empirical results in the present study suggest that this bias became substantially larger when $G_1 \leq .75$ and $n < 30$. In such circumstances, the use of the unbiased estimator of the G

coefficient is strongly recommended.

Although $\hat{G}_1$ is a biased estimator of G_1, the mean of $\hat{G}_1$ over the replications was very close to its expected value [i.e., $E(\hat{G}_1)$] and was consistently so across all simulated conditions (k, ε, and types of scale) for a given G_1 (or G_{cp}). These findings suggest that the degree of heterogeneity of covariance did not introduce any additional bias to the magnitude of $\hat{G}_1$, nor did nonnormality, nor a moderate departure from homogeneity of variance (i.e., a ratio of .6 to 1.4 among the variances). However, heterogeneity of covariance, especially ε = .5, did result in more variable estimates of MS_e (but not for MS_p). This in turn produced a larger sampling variability of $\hat{G}_1$. Especially, the magnitude of $\hat{G}_1$ with k = 3, G = .60, ε = .5, and n = 15 varied markedly, ranging anywhere from .93 to -2.0.

The variability of $\hat{G}_1$ across the categorical scales showed a rather complicated trend. The magnitude of empirical standard deviations for the categorical data was considerably larger than for its parent continuous data, but this result was deemed to be a result of the initial difference between G_{cp} and G_1 brought about by the interactive effects among population conditions (G, ε, types of scale). When the effect of categorization was partialled out, the variability of $\hat{G}_1$ appeared to be very close to its corresponding theoretical value, although it was still considerably larger than that for the continuous data. Furthermore, the comparison of empirical standard deviations among the three ε conditions for the categorical data showed

that the variability of $\hat{G}_1$ for the ε = H condition was larger, especially under the U2 scale. This result was deemed to be due to the sampling characteristics of epsilon as well as to the effect of categorization -- the categorization resulted in a smaller G_{cp} for the ε = H condition and produced a larger epsilon estimate, especially under the U2 scale.

Therefore, these results suggest that the violation of the circularity assumption did not add any bias to the estimate, but yielded more variable estimates of the G coefficient for continuous data. Thus, it is likely to produce too many large estimates of the G coefficient (as well as too many small ones). However, the sampling variability of $\hat{G}_1$ for categorical data was less sensitive to the heterogeneity of covariance, especially for a 3-point or less scale.

Sampling distribution of $\hat{G}_1$. Heterogeneity of covariance had some positive effects, though not large, on the sampling distribution of $\hat{G}_1$ as evident by the inflated empirical proportions beyond the upper limit of the theoretical tolerance interval of $\hat{G}_1$, especially under the ε = .5 condition. The empirical proportions were, in general, about 6% for ε = .7 and about 7.2% for ε = .5 at $\alpha/2$ = .05. When considering the criteria of robustness suggested by Bradley (1978, p.146) -- a stringent criterion being $0.9\alpha <$ actual value $< 1.1\alpha$ and the most liberal one being $0.5\alpha <$ actual value $< 1.5\alpha$, these results indicate that the sampling theory of the G coefficient is fairly robust to a moderate departure from circularity (i.e., under the

ε = .7 condition), but somewhat sensitive to severe noncircularity. Therefore, when the circularity assumption is not seriously violated, the sampling theory of G coefficient can be adequately applied to an inferential test for an estimated G coefficient.

The sampling theory of the G coefficient was not adequate for categorical scales. The empirical proportion beyond the theoretical upper limit was reduced to close to zero, and that beyond the lower limit increased considerably to an unacceptable level -- for the U2 scale it was as large as about 80% for n = 45 and G_1 = .90 for all three levels of k. These results were deemed to be mainly due to the initial difference between G_1 and G_{cp} brought about by the effect of categorization. An interesting yet rather contradictory interpretation of these results would be that the empirical proportion of $\hat{G}_1$ falling within the two limits of the tolerance interval could be interpreted as a Type II error, if one could presume that the difference between the G_1 and G_{cp} is indeed a true difference (but the sampling theory and statistical model assume that observed variables have an underlying continuous metric). Nevertheless, these findings indicate that the sampling theory of the G coefficient and its inferential procedure are not adequate for categorical data, especially for a 3-point or less scale, unless a large number of measures (k) (large enough to bring up the G_{cp} close to its parent continuous G_1 value) are involved in a design.

Type I error rates in the F test. As expected, the results showed that the Type I error rates of the F test for the Rater main effect (MS_r/MS_e) were inflated when circularity failed. For categorical data, the degree of this inflation in the error rates decreased as the scale approached U2. This appeared to be due to the sampling characteristics of epsilon -- epsilon estimates were larger under categorical scales than under continuous data. As a result, the error rates for categorical scales were not too serious for a moderate departure from circularity, especially for a 3-point or less scale. The results also suggest that the effect of noncircularity on the F test was larger than that on the inferential procedure of the G coefficient (i.e., about 7% vs. 9% under $\varepsilon = .50$). In general, the empirical results in the present study were in close agreement with previous findings in the literature, and thus provided the validation of the simulation procedure and accuracy of subsequent calculations implemented in the simulation programs.

Examination of the relationships among the population epsilon, the sample estimate, and the Type I error rates revealed an interesting phenomenon. There was a strong negative relationship between the magnitudes of epsilon estimates and the Type I error rates across simulated conditions when covariances were heterogeneous, thus supporting current theory. However, for the $\varepsilon = 1.0$ condition, although the magnitude of the sample estimates varied widely, the associated Type I error rates were all close to the nominal level, thus yielding a near zero

correlation between them. Further investigation of the correlations among the sample estimates ($\hat{\varepsilon}$, MS_e, and MS_r) showed the presence of a negative correlation between $\hat{\varepsilon}$ and MS_e for the low epsilon conditions. This negative correlation indicates that the ratio of MS_r/MS_e tends to be smaller for a smaller $\hat{\varepsilon}$ under the violation of circularity assumption, which is a contradictory relationship of what is reported in the literature.

Two-facet design

The results of the two-facet design closely paralleled those of the one-facet design in terms of the effects of categorization, sample size, and population G value. However, a primary difference in the findings between the two designs was that the violation of the circularity assumption did not have any appreciable effect on the sampling characteristics of the G coefficient for the two-facet design. The results of the sampling variability and empirical distribution of $\hat{G}_2$ suggest that the sampling theory of the G coefficient for the two-facet design, which is based on an approximated F distribution using the Satterthwaite's procedure, was very satisfactory and quite robust to the violation of the circularity assumption for continuous data and for a 5-point scale.

With respect to quasi F ratios, the magnitude of Type I error rates for the Rater or Occasion effect test varied somewhat depending on the particular form of the F ratios. It was found that the Type I error rates for the quasi F ratio,

which include a combination of mean squares only in the denominator of the F ratio, were somewhat positively inflated under noncircularity. As in the conventional F test, the degree of inflation in the error rates was reduced to close to the nominal level as the scale approached U2. However, the second form of the quasi F test, which includes a combination of mean squares in both numerator and denominator of the test, resulted in tests which were quite robust to noncircularity. On the other hand, the same test was somewhat conservative when the circularity assumption was met. These results for the second form of the quasi F test were in close agreement with those in the related literature.

Implications of the present study

The findings of the present empirical study have implications for the use of G theory. First, the results showed that categorization of continuous data into a 3-point or less scale resulted in a considerable loss of measurement information. For example, a G coefficient for a 5-point scale was only about 5% smaller than that for the parent continuous data. However, it was about 10% and about 20% smaller for a 3-point scale and for a dichotomous scale, respectively. Therefore, dichotomization of any Likert-scale variables or the use of dichotomous scale for the simplicity of ratings should be avoided, whenever possible.

Second, it was shown that $\hat{G}_1$ is a downward biased estimator of G_1, and the amount of bias increases with

decreasing G and n. Furthermore, sample size and the magnitude of the G value have relatively larger influence on the variability of $\hat{G}_1$, in comparison to the number of measures. Therefore, if possible, the use of a large sample size should be considered in conducting a G or D study in order to reduce the amount of bias as well as the sampling variability of the G coefficient. However, when a small scale G study is inevitable, researchers should consider using an unbiased estimator of G_1. For example, consider a measurement design under which a researcher obtained a G coefficient of .68 with n = 15. The corresponding unbiased value for the same data would be .73 = [(.68)(15-3) + 2]/(15-1). The two values may lead him/her to reach quite a different conclusion in a decision making, one being classified as an unacceptable value and the other being considered as a reasonable value.

Third, the sampling theory for the one-facet design was quite robust to a moderate violation of the circularity assumption (i.e., ε = .70). Researchers applying an inferential test for a G coefficient may not need to worry too much about noncircularity unless a severe noncircularity is observed (i.e., ε = .50). For the two-facet design, noncircularity did not have any effect on the sampling distribution of G coefficient. However, for both designs the use of inferential procedures for a G coefficient with categorical data, especially with a 3-point or less scale, should be given extra consideration. Since a G coefficient for a 3-point or less scale is already considerably lower than that for the parent continuous data, the limits of a

confidence interval for an unknown population G coefficient would be shifted downward, and thus give a misleading range for a true population G coefficient.

Fourth, Type I error rates for continuous data were inflated when circularity failed, but the error rates for categorical data were not too serious for a moderate departure from circularity, especially for a 3-point or less scale. However, the Type I error rates were close to the nominal level for the $\varepsilon = 1.0$ condition, regardless of the size of epsilon estimates. These findings suggest that an $\hat{\varepsilon}$- or $\tilde{\varepsilon}$-adjusted F test would be correct only if one can presume that the population covariance matrix from which a sample being taken exhibits noncircularity. Otherwise, the adjusted F test would result in a conservative test, and thus increase the probability of a Type II error if the estimated epsilon is indeed from the population matrix with homogeneous covariances. Furthermore, empirical results revealed the presence of negative correlations between MS_e and $\hat{\varepsilon}$ for the $\varepsilon < 1.0$ conditions. Given that correlations between MS_r and $\hat{\varepsilon}$ and between MS_r and MS_e were near zero, the negative relationship between MS_e and $\hat{\varepsilon}$ suggests that a smaller $\hat{\varepsilon}$ is associated with a larger MS_e, which in turn tends to produce a smaller F ratio (MS_r/MS_e). These results led us to question the validity of the common practice of utilizing the $\hat{\varepsilon}$- or $\tilde{\varepsilon}$-adjusted F test in the repeated measures ANOVA designs.

Suggestions for future research

As with any simulation study, the results obtained in the present study must be interpreted with a certain degree of caution. There are a number of limitations imposed by the conditions simulated, and these limitations suggest possible directions for future research in this area. One such limitation is that the present study investigated the sampling behavior of only one form of the G coefficient, that is, the G coefficient for a relative decision. There are a number of reliability-like indices frequently used in practice, such as intraclass correlation coefficients and the G coefficient for absolute decisions, and it is uncertain as to the extent which the results obtained for one specific index can be generalized to other indices. Further research is also needed to investigate the extent to which the results obtained for the specific measurement designs used in the present study can be generalized to other measurement designs, such as fixed and nested designs, and to a design having a large number of facets and levels within a facet.

Another restriction on the generalizability of the present results is that the simulated population data were generated from a particular form of population covariance matrix, and the data were transformed to the categorical scales having the normal and uniform distribution. Although it seems clear that the sampling characteristics of the G coefficient would be insensitive to a certain level of variation in nonnormality represented by the uniform distribution and in heterogeneity of

variances / covariances, it is less clear whether similar conclusions can be made about the performance of the G coefficient under other distributional forms (e.g., exponential) or under a radically different form of covariance matrices with severe noncircularity for all facets and their interactions.

Finally, it would be worthwhile to conduct further investigations on the relationships between the population epsilon, the sample estimates, and the Type I error rates for the F tests (perhaps for quasi F ratios as well). An interesting yet contradictory preliminary finding regarding the correlations among the sample estimates raises a question about the current practice of utilizing an $\hat{\varepsilon}$-adjusted F test in repeated measures ANOVA designs, and this question requires further confirmation with extensive empirical work.

REFERENCES

Algina, J. (1978). Comment on Bartko's "On various intraclass correlation reliability coefficients". Psychological Bulletin, 85, 135-138.

Alsawalmeh, Y.M, & Feldt, L.S. (1992). Test of the hypothesis that the intraclass reliability coefficient is the same for two measurement procedures. Applied Psychological Measurement, 16, 195-205.

Andersen, A.H., Jensen, E.B., & Schou, G. (1981). Two-way analysis of variance with correlated errors. International Statistical Review, 49, 153-167.

Bartko, J.J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83, 762-765.

Bay, K.S. (1973). The effect of non-normality on the sampling distribution and standard error of reliability coefficient estimates under an analysis of variance model. British Journal of Mathematical Psychology, 26, 45-57.

Bell, J.F. (1986). Simultaneous confidence intervals for the linear functions of expected mean squares used in generalizability theory. Journal of Educational Statistics, 11, 197-205.

Bert, R.A. (1978). An analysis of variance model for assessing reliability of naturalistic observations. Perceptual and Motor Skills, 47, 271-278.

Bert, R.A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. American Journal of Mental Deficiency, 83, 460-472.

Birch, N.J., Burdick, R.K., & Ting, N. (1990). Confidence intervals and bounds for a ratio of summed expected mean squares. Technometrics, 32, 437-444.

Boardman, T.J. (1974). Confidence intervals for variance components - A comparative Monte Carlo study. Biometrics, 30, 251-269.

Boik, R.J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. Psychometrika, 46, 241-255.

Booth, C.L., Mitchell, S.K., & Solin, F.K. (1979). The generalizability study as a method of assessing intra-and interobserver reliability in observational research. Behavior Research Methods & Instrumentation, 11, 491-494.

References

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effect of inequality of variance and correlation between errors in the two-way classification. Annals of Mathematical Statistics, 25, 484-498.

Bradley, J.V. (1978). Robustness? British Journal of Statistical Psychology, 31, 144-152.

Bradley, J.V. (1980). Nonrobustness in one-sample Z and t tests: A large-scale sampling study. Bulletin of the Psychonomic Society, 15, 29-32.

Bradley, J.V. (1980). Nonrobustness in classical tests on means and variances: A large-scale sampling study. Bulletin of the Psychonomic Society, 15, 275-278.

Bradley, J.V. (1980). Nonrobustness in Z, t, and F tests at large sample sizes. Bulletin of the Psychonomic Society, 16, 333-336.

Brennan, R.L. (1983). Elements of generalizability theory. Iowa: The American College Testing Program.

Brennan, R.L., & Kane, M.T. (1977). An index of dependability for mastery tests. Journal of Educational Measurement, 14, 277-289.

Brennan, R.L., & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Burdick, R.K., & Graybill, F.A. (1988). The present status of confidence interval estimation on variance components in balanced and unbalanced random models. Communications in Statistics A: Theory and method, 17, 1165-1195.

Burt, C. (1955). Test reliability estimated by analysis of variance. British Journal of Statistical Psychology, 8, 103-118.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 13, 119-135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. Journal of Educational Measurement, 18, 183-204.

Carmines, E.G., & Zeller, R.A. (1979). Reliability and validity assessment. California: Sage Publications, Series 07-017.

References

Cicchetti, D.V., Showalter, D., & Tyrer, P.J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. Applied Psychological Measurement, 9, 31-36.

Cohen, J. (1983). The cost of Dichotomization. Applied Psychological Measurement, 7, 249-253.

Collier, Jr. R.O., Baker, F.B., Mandeville, G.K., & Hayes, T.F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrika, 32, 339-353.

Connor, R.J. (1972). Grouping for testing trends in categorical data. Journal of American Statistical Association, 67, 601-604.

Cornfield, J., & Tukey, J.W. (1956). Average values of mean squares in factorials. Annals of Mathematical Statistics, 27, 907-949.

Cox, D.R. (1957). Note on grouping. Journal of American Statistical Association, 52, 543-547.

Craig, A.T. (1938). On the dependence of certain estimates of variance. Annals of Mathematical Statistics, 9, 48-55.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Crocker, L., Llabre, M., & Miller, M.D. (1988). The generalizability of content validity ratings. Journal of Educational Measurement, 25, 287-299.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N., (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). The theory of generalizability: A liberalization of reliability theory. The British Journal of Statistical Psychology, 16, 137-163.

Davenport, J.M., & Webster, J.T. (1973). A comparison of some approximate F-tests. Technometrics, 15, 779-789.

Doverspike, D., Carlisi, A.M., Barrett, G.V., & Alexander, R.B. (1983). Generalizability analysis of a point-method job evaluation instrument. Journal of Applied Psychology, 68, 476-483.

References

Ebel, R.L. (1951). Estimation of the reliability of ratings. Psychometrika, 16, 407-424.

Eom, H.J, & Schutz, R. W. (1993, August). Data categorization, noncircularity, and Type I error rates in repeated measures ANOVA designs. Paper presented at the meeting of the American Statistical Association, San Francisco, CA.

Erickson, R.S. (1978). Analyzing one variable-three wave panel data: A comparison of two methods. Political Methodology, 5, 151-166.

Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Rechardson reliability coefficient twenty. Psychometrika, 30, 357-370.

Feldt, L.S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34, 363-373.

Feldt, L.S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Psychometrika, 45, 99-105.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), Educational measurement (3rd ed.) (pp.105-146). New York: American Council on Education.

Finsturen, K., & Campbell, M.E. (1979). Further comments on Bartko's "On various intraclass correlation reliability coefficients". Psychological Reports, 45, 375-380.

Fleiss, J.L. (1971). On the distribution of a linear combination of independent chi squares. Journal of American Statistical Association, 66, 142-144.

Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. Biometrics, 31, 651-658.

Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley & Sons.

Fleiss, J.L., & Shrout, P.E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika, 43, 259-262.

Gaylor, D.W., & Hopper, F.N. (1969). Estimating the degrees of freedom for linear combinations of mean squares by Satterthwaite's formula. Technometrics, 11, 691-706.

Geisser, S., & Greenhouse, S.W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. Annals of Mathematical Statistics, 29, 885-891.

References

Gessaroli, M.E, & Schutz, R.W. (1983). Variable error: Variance-covariance heterogeneity, block size and Type I error rates. Journal of Motor Behavior, 15, 74-95.

Ghiselli, E.E., Campbell, J.P., & Zedeck, S. (1981). Measurement theory for the behavioral sciences. New York: W.H. Freeman and Company.

Gibbons, J.D. (1985). Nonparametric methods for quantitative analysis (second edition). Ohio: American Sciences Press, Inc..

Gillmore, G.M. (1983). Generalizability theory: Applications to program evaluation. In L.J.Fyans,Jr. (Ed.). Generalizability theory: Inferences and practical applications. San Francisco: Jossey-Bass.

Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Gleser, G.C., Cronbach, L.J., & Rajaratnam, N., (1965). Generalizability of scores influenced by multiple sources of variance. Psychometrika, 30, 395-418.

Godbout, P., & Schutz, R.W. (1983). Generalizability of ratings of motor performances with reference to various observational designs. Research Quarterly for Exercise and Sport, 54, 20-27.

Green, S.B., Lissitz, R.W., & Mulaik, S.A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 37, 827-839.

Greenhouse, S.W., & Geisser, S. (1959). On methods in the analysis of profile data. Psychometrika, 24, 95-112.

Gregoire, T.G., & Driver, B.L. (1987). Analysis of ordinal data to detect population differences. Psychological Bulletin, 101, 159-165.

Grieve, A.P. (1984). Tests of sphericity of normal distributions and the analysis of repeated measures designs. Psychometrika, 49, 257-267.

Hakstian, A.R., & Whalen, T.E. (1976). A k-sample significance test for independent alpha coefficient. Psychometrika, 41, 219-231.

Heise, D.R. (1969). Separating reliability and stability in test-retest correlation. American Sociological Review, 34, 93-101.

References

Horst, P. (1949). A generalized expression for the reliability of measures. Psychometrika, 14, 21-31.

House, A.E., House, B.J., & Campbell, M.B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. Journal of Behavioral Assessment, 3, 37-57.

Hoyt, C.J. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Huck, S.W. (1978). A modification of Hoyt's analysis of variance reliability estimation procedure. Educational and Psychological Measurement, 38, 725-736.

Hudson, J.D., & Krutchkoff, R.G. (1968). A Monte Carlo investigation of the size and power of tests employing Satterthwaite's synthetic mean squares. Biometrika, 55, 431-433.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. Psychometrika, 43, 161-175.

Huynh, H., & Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F distributions. Journal of American Statistical Association, 65, 1582-1589.

Huynh, H., & Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational Statistics, 1, 69-82.

Huynh, H., & Mandeville, G.K. (1979). Validity conditions in repeated measures designs. Psychological Bulletin, 86, 964-973.

Huysamen, G.K. (1990). The application of generalizability theory to the reliability of ratings. South African Journal of Psychology, 20, 200-205.

IMSL. (1991). International Mathematical and Statistical Libraries (10th ed.). Houston, TX.

Jenkins, Jr, G.D., Taber, T.D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. Journal of Applied Psychology, 62, 392-398.

Johnson, S., & Bell, J.F. (1985). Evaluating and predicting survey efficiency using generalizability theory. Journal of Educational Measurement, 22, 107-119.

Joreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. Psychometrika, 36, 109-133.

Joreskog, K.G., & Sorbom, D. (1989). LISREL 7: User's reference guide. Mooresville: Scientific Software, Inc..

Kane, M.T. (1986). The role of reliability in criterion-referenced tests. Journal of Educational Measurement, 23, 221-224.

Kane, M.T., & Brennan, R.L. (1977). The generalizability of class means. Review of Educational Research, 47, 267-292.

Kane, M.T., & Gilmore, G.M., & Crooks, T.J. (1976). Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement, 13, 171-183.

Kenny, D.A., & Judd, C.M. (1986). Consequences of violating the independence assumption in analysis of variance. Psychological Bulletin, 99, 422-431.

Khuri.A.I. (1981). Simultaneous Confidence intervals for functions of variance components in random models. Journal of American Statistical Association, 76, 878-885.

Khuri, A.I., & Sahai, H. (1985). Variance components analysis: A selective literature survey. International Statistical Review, 53, 279-300.

Kogan, L.S. (1948). Analysis of variance - repeated measurements. Psychological Bulletin, 45, 131-143.

Kraemer, H.C. (1981). Extension of Feldt's approach to testing homogeneity of coefficients of reliability. Psychometrika, 46, 41-45.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. Psychometrika, 28, 221-238.

Kristof, W. (1970). On the sampling theory of reliability estimation. Journal of Mathematical Psychology, 7, 371-377.

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Lahey, M.A., Downey, R.G., & Saal, F.E. (1983). Intraclass correlations: There's more there than meets the eye. Psychological Bulletin, 93, 586-595.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. Applied Measurement in Education, 2, 195-205.

Lissitz, R.W., & Green, S.B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. Journal of Applied Psychology, 60, 10-13.

Lomax, R.G. (1982). An application of generalizability theory to observational research. Journal of Experimental Education, 51, 22-30.

Looney, M.A., Heimerdinger, B.M. (1991). Validity and generalizability of social dance performance ratings. Research Quarterly for Exercise and Sport, 62, 399-405.

Lunney, G.H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. Journal of Educational Measurement, 7, 263-269.

Macready, G.B. (1983). The use of generalizability theory for assessing relations among items within domains in diagnostic testing. Applied Psychological Measurement, 7, 149-157.

Marcoulides, G.A. (1989). The estimation of variance components in generalizability studies: A resampling approach. Psychological Reports, 65, 883-889.

Marcoulides, G.A. (1990). An alternative method for estimating variance components in generalizability theory. Psychological Reports, 66, 379-386.

Masters, J.R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. Journal of Educational Measurement, 11, 49-53.

Matell, M.S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items: Study I: Reliability and validity. Educational and Psychological Measurement, 31, 657-674.

Maxwell, A.E. (1968). The effect of correlated errors on estimates of reliability coefficients. Educational and Psychological Measurement, 28, 803-811.

Maxwell, S.E., & Bray, J.H. (1986). Robustness of the Quasi F statistic to violations of sphericity. Psychological Bulletin, 99, 416-421.

McCall, R.B., & Appelbaum, M.I. (1973). Bias in the analysis of repeated-measures designs: Some alternative approaches. Child Development, 44, 401-415.

References

McHugh, R.B., Sivanich, G., & Geisser, S. (1961). On the evaluation of changes by psychometric test profiles. Psychological Reports, 7, 335-344.

Mendoza, J.L., Toothaker, L.E., & Crain, B.R. (1976). Necessary and sufficient conditions for F ratios in the L x J x K factorial design with two repeated factors. Journal of American Statistical Association, 71, 992-993.

Mitchell, S.K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376-390.

Mitzel, H.C., & Games, P.A. (1981). Circularity and multiple comparisons in repeated measure designs. British Journal of Mathematical and Statistical Psychology, 34, 253-259.

Morgan, S. (1988). Diagnostic assessment of autism: A review of objective scales. Journal of Psychoeducational Assessment, 6, 139-151.

Morrow, J.R,Jr., et al. (1986). Generalizability of the AAHPERD health related skinfold test. Research Quarterly for Exercise and Sport, 57, 187-95.

Morrow, J.R,Jr. (1989). Generalizability theory. In M.J. Safrit & T.M. Wood (Eds.), Measurement concepts in physical education and exercise science (pp. 73-96). Illinois: Human Kinetics.

Myers, J., DiCecco, J.V., White, B.J., & Borden, V.M. (1982). Repeated measurements on dichotomous variables: Q and F tests. Psychological Bulletin, 92, 517-525.

Novick, M.R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32, 1-13.

Paulson, E. (1942). An approximate normalization of the analysis of variance distribution. Annals of Mathematical Statistics, 13, 233-235.

Rao, C.R. (1973). Linear statistical inference and its applications (2nd ed.). New York: John Wiley & Sons.

Rasmussen, J.L. (1989). Analysis of Likert-scale data: A reinterpretation of Greoire and Driver (1987). Psychological Bulletin, 105, 167-170.

Rasmussen, J.L., Heumann, K.A., Heumann, M.T., & Botzum, M. (1989). Univariate and multivariate groups by trails analysis under violation of variance-covariance and normality assumptions. Multivariate Behavioral Research, 24, 93-105.

References

Rogan, J.C., & Keselman, H.J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. American Educational Research Journal, 14, 493-498.

Rogan, J.C., Keselman, H.J., & Mendoza, J.L. (1979). Analysis of repeated measurements. British Journal of Mathematical and Statistical Psychology, 32, 269-286.

Rouanet, H., & Lepine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. The British Journal of Mathematical and Statistical Psychology, 23, 147-163.

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 9, 99-103.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Sahai, H. (1979). A bibliography on variance components. International Statistical Review, 47, 177-222.

Sahai, H., Khuri, A.I., & Kapadia, C.H. (1985). A second bibliography on variance components. Communications in Statistics A: Theory and Method, 14, 63-115.

Santa, J.L., Miller, J.J., & Shaw, M.L. (1979). Using Quasi F to prevent alpha inflation due to stimulus variation. Psychological Bulletin, 86, 37-46.

Satterthwaite, F.E. (1941). Synthesis of variance. Psychometrika, 6, 309-316.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. Biometrics, 2, 110-114.

Scheffe, H. (1959). The analysis of variance. New York: Wiley.

Schroeder, M.S., & Hakstian, R. (1990). Inferential procedures for multifaceted coefficients of generalizability. Psychometrika, 55, 429-447.

Searle, S.R. (1971). Linear models. New York: John Wiley.

Searle, S.R., Casella, G, & McCulloch, C.E. (1992). Variance components. New York: John Wiley.

Sedere, M.U., & Feldt, L.S. (1976). The sampling distributions of the Kristof reliability coefficient, the Feldt coefficient, and Guttman's lambda-2. Journal of Educational Measurement, 14, 53-62.

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166.

Shavelson, R.J., & Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park: Sage.

Shavelson, R.J., Rowley, G.L., & Webb, N.M. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-93.

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Smith, P.L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. Journal of Educational Statistics, 3, 319-346.

Smith, P.L. (1981). Gaining accuracy in generalizability theory: Using multiple designs. Journal of Educational Measurement, 18, 147-154.

Smith, P.L. (1982). A confidence interval approach for variance component estimates in the context of generalizability theory. Educational and Psychological Measurement, 42, 459-466.

Smith, P.L., & Luecht, R.M. (1992). Correlated effects in generalizability studies. Applied Psychological Measurement, 16, 229-235.

Stayrook, N., & Corno, L. (1979). An application of generalizability theory in disattenuating a path model of teaching and learning. Journal of Educational Measurement, 16, 227-237.

Stoloff, P.H. (1970). Correcting for heterogeneity of covariance for repeated measures designs of the analysis of variance. Educational and Psychological Measurement, 30, 909-924.

Tukey, J.W. (1949). One degree of freedom for non-additivity. Biometrics, 5, 232-242.

Ulrich, D., Ulrich, B.D., & Branta, C.F. (1988). Developmental gross motor skill ratings: A generalizability study. Research Quarterly for Exercise and Sport, 59, 203-209.

Verdooren, L.R. (1982). How large is the probability for the estimate of a variance component to be negative. Biometrical Journal, 24, 339-360.

Violato, C., & Travis, L.D. (1988). An application of generalizability theory to the consistency-specificity problem: The transituational consistency of behavioral persistence. The Journal of Psychology, 122, 389-407.

Webb, N.M., Rowley, G.L., & Shavelson, R.J. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Wike, E.L., & Church, J.D. (1980). Nonrobustness in F tests: 1. A replication and extension of Bradley's study. Bulletin of the Psychonomic Society, 20, 165-167.

Wike, E.L., & Church, J.D. (1982). Nonrobustness in F tests: 2. Further extensions of Bradley's study. Bulletin of the Psychonomic Society, 20, 168-170.

Wilcox, R.R. (1987). New designs in analysis of variance. Annual Review of Psychology, 38, 29-60.

Wilson, K. (1975). The sampling distribution of conventional, conservative and corrected F-ratios in repeated measurements designs with heterogeneity of covariance. Journal of Statistical Computation and Simulation, 3, 201-215.

Winer, B.J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Woodruff, D.J., & Feldt, L.S. (1988). Tests for equality of several alpha coefficients when their sample estimates are dependent. Psychometrika, 51, 393-413.

Zimmerman, D.W. (1980). Is classical test theory 'robust' under violation of the assumption of uncorrelated errors? Canadian Journal of Psychology, 34, 227-237.

Appendix A

Circularity assumptions in repeated measures ANOVA

Repeated measures analysis of variance (ANOVA) procedures are extensively used in educational and psychological research. When the repeated measures are obtained from the same individuals, it is naturally believed that the successive measures or responses will tend to positively correlated. In this case, besides the usual ANOVA assumptions of normality of distribution and homogeneity of variances, there is an additional assumption regarding the pattern of these correlated measures. A number of empirical studies on the effect of violating ANOVA assumptions of normality of distribution and homogeneity of variances on Type I error rates have shown that ANOVA F statistic is generally robust with regard to moderate departures from these assumptions, especially if sample sizes are equal (e.g., Glass, Peckham, & Sanders, 1972), but see Bradley (1978). However, ANOVA loses its robustness when the covariance matrix underlying the repeated measures deviates from a certain pattern, referred to as compound symmetry or circularity.

A covariance matrix is said to possess the property of circularity if the variances of all pair-wise differences between the repeated measures are equal. A special case of circularity is compound symmetry, a covariance matrix with equal variances and equal covariances (Huynh and Feldt, 1970; Rouanet and Lepine, 1970; Winer, 1971). For example, for a two-way ANOVA model, which is the equivalent model to a one-facet crossed design (i.e., subjects by raters) in G theory, the compound symmetry implies that a r by r covariance matrix exhibits equal variances in the diagonal and equal covariances in the off-diagonal. Furthermore, for a three-way ANOVA model, i.e., a two-facet (Persons x Raters x Occasions) fully crossed design in G theory, each of the covariance matrices Σ_r (n_r x n_r), Σ_o (n_o x n_o), and Σ_{ro} ($n_r n_o$ x $n_r n_o$) is required to possess a local circularity (Rouanet & Lepine, 1970) in order for the F or quasi F statistic to be valid (Huynh & Mandeville, 1979; Maxwell & Bray, 1986; Mendoza, Toothaker, & Crain, 1976).

Box (1954) has shown that violating this assumption yields more variable estimates of the mean squares, and thus results in more extreme large as well as small F ratios than indicated by the theoretical F distribution. Since interest is directed to the upper tail of this distribution, a cumulative proportion beyond the theoretical upper limit is attributed to Type I error rates. Consequently, he developed a measure of the degree of departure from compound symmetry, known as epsilon (ε). Epsilon is used to correct a positive bias in the usual F test by adjusting the degrees of freedom by an amount proportional to ε or of its estimate $\hat{\varepsilon}$. The epsilon is a function of the variances and covariances in the population matrix (Σ_x), and can be calculated as:

[A1]

$$\varepsilon = \frac{k^2 \, (\underline{\sigma}_{ii} - \underline{\sigma}..)^2}{(k-1)\,(\Sigma\Sigma\sigma^2_{ij} - 2k\Sigma\sigma^2_{i.} + k^2 \, \underline{\sigma}^2..)}$$

where;

k = the order of the covariance matrix,

$\underline{\sigma}_{ii}$ = the mean of the variances (diagonals),

$\underline{\sigma}..$ = the grand mean of the covariance matrix,

$\sigma_{i.}$ = the mean of the i^{th} row or column of the covariance matrix, and

σ_{ij} = an individual element in the matrix (where; i and j = 1,2,..k).

Huynh and Feldt (1970), and Rouanet and Lepine (1970) demonstrated independently that the compound symmetry condition of the covariance matrix is a sufficient condition for the ratio of mean squares to have an F distribution, but it is not a necessary condition. That is, a matrix Σ may have other patterns, but the ratio of the mean squares may still have an F distribution with $\varepsilon = 1.0$. If the difference scores between all pairs of measures are equally variable, this produces a covariance matrix which possesses circularity (Rouanet & Lepine, 1970) condition. This property in the covariance matrix indicates that when a k x k covariance matrix Σ_x is transformed orthonormally, using a k-1 by k orthonormal matrix **M**, then a resultant (k-1) by (k-1) matrix Σ_y contains a set of orthonormal variables. If the original matrix Σ_x has the circularity pattern, then the resultant matrix Σ_y has sphericity condition which results in $\Sigma_y = M\Sigma_x M' = cI$, where I is the identity matrix of order (k-1), and c is a constant. From this relationship, the epsilon ε can be alternatively defined in terms of orthonormally transformed matrix as:

[A2]

$$\varepsilon = \frac{(\Sigma \, c_i)^2}{(k-1) \; \Sigma \, c^2_i}$$

where;

c is the (k-1) eigenvalues of a (k-1) by (k-1) matrix Σ_y.

Under circularity or sphericity conditions, all eigenvalues are equal. Consequently, $(\Sigma \, c_i)^2 = (k-1) \; \Sigma \, c^2_i$, and $\varepsilon = 1.0$. Under maximum departure from sphericity, all eigenvalues, except one, are equal to zero (e.g., Boik, 1981; Grieve, 1984), thus $(\Sigma \, c_i)^2 = \Sigma \, c^2_i$ and $\varepsilon = 1/(k-1)$.

Appendix A

Box's work was extended to more complex designs by Geisser and Greenhouse (1958), Greenhouse and Geisser (1959), McHugh, Sivanich, and Geisser (1961), and Huynh (1978). In addition, many subsequent simulation studies (with continuous dependent variables) have shown that the degree of bias introduced by violating circularity assumption is quite substantial in a variety of specific cases (Collier, Baker, Mandeville, & Hayes, 1967; Greenhouse & Geisser, 1959; Huynh, 1978; Rasmussen, Heumann, Heumann, & Botzum, 1989; Wilson, 1975). For example, Collier et al. (1967) showed that computing $\hat{\varepsilon}$ from a sample covariance matrix and adjusting the degrees of freedom for the critical F by amount of $\hat{\varepsilon}$ produced an approximate F test that is relatively robust for reasonable samples of 15 or larger.

However, Stoloff (1970), and Huynh and Feldt (1976) have demonstrated, through Monte Carlo studies, that $\hat{\varepsilon}$-adjusted test is negatively biased (i.e., too conservative). This bias is greatest when a population ε is near or above .75, especially when the sample size is small. Consequently, Huynh and Feldt proposed an alternative estimator of ε. Their estimator $\tilde{\varepsilon}$ is a function of n (sample), g (group), k (level) and Box's $\hat{\varepsilon}$:

[A3]

$$\tilde{\varepsilon} = \frac{n(k-1)\,\hat{\varepsilon} - 2}{(k-1)\,[n-g-(k-1)\,\hat{\varepsilon}]}$$

Thus, for any value of n and k, $\tilde{\varepsilon}$ is equal to or greater than $\hat{\varepsilon}$, and the equality holds when $\hat{\varepsilon} = 1/(k-1)$. The upper bound of $\tilde{\varepsilon}$ was set to unity, though it theoretically can be greater than unity. Huynh and Feldt(1976, 1978) and Rogan, Keselman and Mendoza (1979) reported that $\tilde{\varepsilon}$-adjusted test produced a test size closer to a specified alpha level than did the $\hat{\varepsilon}$-adjusted test.

Most empirical studies mentioned above have focused more on examining the behavior of the ratio of the mean square estimates and the degree of positive bias in the F test. However, knowing that the departure from the circularity condition in the covariance matrix results in unstable estimates of mean squares, it is quite likely that the more-variable mean squares would add an additional variability to the estimates of the G coefficient since the G coefficient is a function of observed mean squares. This would be especially true with a small scale of a measurement design. There is, however, little research in the literature that has systematically examined the effects of noncircularity on the magnitude of the sample estimates of the G coefficient as well as on the sampling variability of the estimates.

Appendix B

Input population covariance matrices

Table B-1

Population covariance matrices for the one-facet design (k = 3)

	G = .90	.75	.60
ε =	1.0	1.0	1.0
	100	100	100
	75 100	50 100	33.33 100
	75 75 100	50 50 100	33.33 33.33 100
ε =	.7051	.7051	.7088
	100	100	100
	75 100	50 100	14 100
	61 89 100	22 78 100	10 76 100
ε =	.5268	.5395	.5389
	100	100	100
	76 100	50 100	3 100
	54 95 100	10 90 100	2 95 100

Table B-2

Population covariance matrices for the one-facet design (k = 5)

G = .90:

```
100                    ε = 1.0          100          ε = .6930
 64.29 100                               77 100
 64.29  64.29 100                        64  77 100
 64.29  64.29  64.29 100                 51  64  77 100
 64.29  64.29  64.29  64.29 100          41  51  64  77 100

100                    ε = .5001
 83 100
 65  83 100
 41  65  83 100
 34  41  65  83 100
```

G = .75:

```
100                    ε = 1.0          100          ε = .7061
 37.50 100                               60 100
 37.50  37.50 100                        33  60 100
 37.50  37.50  37.50 100                 12  33  60 100
 37.50  37.50  37.50  37.50 100          12  12  33  60 100

100                    ε = .5179
 80 100
 25  50 100
 10  25  50 100
 10  10  25  90 100
```

G = .60:

```
100                    ε = 1.0          100          ε = .7066
 23.08 100                               30 100
 23.08  23.08 100                         1  40 100
 23.08  23.08  23.08 100                  0   5  60 100
 23.08  23.08  23.08  23.08 100           0   0  25  70 100

100
 95 100                ε = .5047
  0  25 100
  0   0  26 100
  0   0   0  90 100
```

Table B-3

Population covariance matrices for the one-facet design (k = 7)

G = .90:

100						
56.25	100					
56.25	56.25	100				
56.25	56.25	56.25	100			
56.25	56.25	56.25	56.25	100		
56.25	56.25	56.25	56.25	56.25	100	
56.25	56.25	56.25	56.25	56.25	56.25	100

$\varepsilon = 1.0$

100						
71	100					
60	71	100				
54	60	71	100			
40	54	60	71	100		
40	40	54	60	71	100	
40	40	40	54	60	71	100

$\varepsilon = .7024$

100						
80	100					
60	80	100				
50	60	80	100			
39	50	60	80	100		
35	39	50	60	80	100	
15	35	39	50	60	80	100

$\varepsilon = .5069$

G = .75:

100						
30	100					
30	30	100				
30	30	30	100			
30	30	30	30	100		
30	30	30	30	30	100	
30	30	30	30	30	30	100

$\varepsilon = 1.0$

100						
65	100					
20	50	100				
20	20	50	100			
15	20	20	50	100		
10	15	20	30	50	100	
10	10	20	30	40	65	100

$\varepsilon = .7046$

100						
90	100					
25	40	100				
15	25	90	100			
10	15	25	40	100		
50	10	15	25	40	100	
5	10	10	15	25	90	100

$\varepsilon = .5080$

Table B-3 - continued

Population covariance matrices for the one-facet design (k = 7)

G = .60:

100 ε = 1.0

```
100
 17.65 100
 17.65  17.65 100
 17.65  17.65  17.65 100
 17.65  17.65  17.65  17.65 100
 17.65  17.65  17.65  17.65  17.65 100
 17.65  17.65  17.65  17.65  17.65  17.65 100
```

100 ε = .6993

```
100
 20 100
 10  50 100
 10  10  70 100
  0  10  10  20 100
  0   0  10  10  50 100
  0   0   0  10  10  70 100
```

100 ε = .5136

```
100
 90 100
  5  25 100
  0   5  25 100
  0   0   5  90 100
  0   0   0   5  25 100
  0   0   0   0   5  90 100
```

Table B-4

Population covariance matrices for the two-facet (3 Occasions by 5 Raters) design

---

G = .90, Matrix: NONE, Epsilon

100															
61	100														
61	61	100													
61	61	61	100												
61	61	61	61	100											
62	54	54	54	54	100										
54	62	54	54	54	61	100									
54	54	62	54	54	61	61	100								
54	54	54	62	54	61	61	61	100							
54	54	54	54	62	61	61	61	61	100						
62	54	54	54	54	62	54	54	54	54	100					
54	62	54	54	54	54	62	54	54	54	61	100				
54	54	62	54	54	54	54	62	54	54	61	61	100			
54	54	54	62	54	54	54	54	62	54	61	61	61	100		
54	54	54	54	62	54	54	54	54	62	61	61	61	61	100	

Epsilon	
O	= 1.0
R	= 1.0
OR	= 1.0

G = .90, Matrix: O/OR, Epsilon

100															
75	100														
75	75	100													
75	75	75	100												
75	75	75	75	100											
66	54	54	54	54	100										
54	66	54	54	54	60	100									
54	54	66	54	54	60	60	100								
54	54	54	66	54	60	60	60	100							
54	54	54	54	66	60	60	60	60	100						
40	54	54	54	54	80	54	54	54	54	100					
54	40	54	54	54	54	80	54	54	54	48	100				
54	54	40	54	54	54	54	80	54	54	48	48	100			
54	54	54	40	54	54	54	54	80	54	48	48	48	100		
54	54	54	54	40	54	54	54	54	80	48	48	48	48	100	

Epsilon	
O	= .6729
R	= 1.0
OR	= .6752

G = .90, Matrix: R/OR, Epsilon

100															
80	100														
60	80	100													
40	60	80	100												
30	40	60	80	100											
70	54	54	54	54	100										
54	60	54	54	54	80	100									
54	54	60	54	54	60	80	100								
54	54	54	60	54	40	60	80	100							
54	54	54	54	60	30	40	60	80	100						
70	54	54	54	54	70	54	54	54	54	100					
54	60	54	54	54	54	60	54	54	54	80	100				
54	54	60	54	54	54	54	60	54	54	60	80	100			
54	54	54	60	54	54	54	54	60	54	40	60	80	100		
54	54	54	54	60	54	54	54	54	60	30	40	60	80	100	

Epsilon	
O	= 1.0
R	= .6810
OR	= .4495

Table B-4 - continued

Population covariance matrices for the two-facet (3 Occasions by 5 Raters) design

---

G = .75, Matrix: NONE, Epsilon

100															
51	100														
51	51	100													
51	51	51	100												
51	51	51	51	100											
52	34	34	34	34	100										
34	52	34	34	34	51	100									
34	34	52	34	34	51	51	100								
34	34	34	52	34	51	51	51	100							
34	34	34	34	52	51	51	51	51	100						
52	34	34	34	34	52	34	34	34	34	100					
34	52	34	34	34	34	52	34	34	34	51	100				
34	34	52	34	34	34	34	52	34	34	51	51	100			
34	34	34	52	34	34	34	34	52	34	51	51	51	100		
34	34	34	34	52	34	34	34	34	52	51	51	51	51	100	

Epsilon	
O	= 1.0
R	= 1.0
OR	= 1.0

G = .75, Matrix: O/OR, Epsilon

100															
80	100														
80	80	100													
80	80	80	100												
80	80	80	80	100											
50	34	34	34	34	100										
34	50	34	34	34	50	100									
34	34	50	34	34	50	50	100								
34	34	34	50	34	50	50	50	100							
34	34	34	34	50	50	50	50	50	100						
30	34	34	34	34	76	34	34	34	34	100					
34	30	34	34	34	34	76	34	34	34	23	100				
34	34	30	34	34	34	34	76	34	34	23	23	100			
34	34	34	30	34	34	34	34	76	34	23	23	23	100		
34	34	34	34	30	34	34	34	34	76	23	23	23	23	100	

Epsilon	
O	= .6543
R	= 1.0
OR	= .6742

G = .75, Matrix: R/OR, Epsilon

100															
70	100														
50	70	100													
30	50	70	100												
20	30	50	70	100											
40	40	30	30	30	100										
40	50	40	30	30	70	100									
30	40	50	40	30	50	70	100								
30	30	40	50	40	30	50	70	100							
30	30	30	40	70	20	30	50	70	100						
40	40	30	30	30	40	40	30	30	30	100					
40	50	40	30	30	40	50	40	30	30	70	100				
30	40	50	40	30	30	40	50	40	30	50	70	100			
30	30	40	50	40	30	30	40	50	40	30	50	70	100		
30	30	30	40	70	30	30	30	40	70	20	30	50	70	100	

Epsilon	
O	= 1.0
R	= .6810
OR	= .5673

Table B-4 - continued

Population covariance matrices for the two-facet (3 Occasions by 5 Raters) design

---

```
100                                  G = .60,  Matrix: NONE,  Epsilon
 46 100                                                       ----------
 46  46 100                                                    O = 1.0
 46  46  46 100                                               R = 1.0
 46  46  46  46 100                                          OR = 1.0
 45  22  22  22  22  100
 22  45  22  22  22   46 100
 22  22  45  22  22   46  46 100
 22  22  22  45  22   46  46  46 100
 22  22  22  22  45   46  46  46  46 100
 45  22  22  22  22   45  22  22  22  22  100
 22  45  22  22  22   22  45  22  22  22   46 100
 22  22  45  22  22   22  22  45  22  22   46  46 100
 22  22  22  45  22   22  22  22  45  22   46  46  46 100
 22  22  22  22  45   22  22  22  22  45   46  46  46  46 100

100                                  G = .60,  Matrix: O/OR,  Epsilon
 82 100                                                       ----------
 82  82 100                                                   O =  .6552
 82  82  82 100                                               R = 1.0
 82  82  82  82 100                                          OR =  .6542
 33  22  22  22  22  100
 22  33  22  22  22   36 100
 22  22  33  22  22   36  36 100
 22  22  22  33  22   36  36  36 100
 22  22  22  22  33   36  36  36  36 100
 20  22  22  22  22   82  22  22  22  22  100
 22  20  22  22  22   22  82  22  22  22   20 100
 22  22  20  22  22   22  22  82  22  22   20  20 100
 22  22  22  20  22   22  22  22  82  22   20  20  20 100
 22  22  22  22  20   22  22  22  22  82   20  20  20  20 100

100                                  G = .60,  Matrix: R/OR,  Epsilon
 70 100                                                       ----------
 40  70 100                                                   O = 1.0
 20  40  70 100                                               R =  .6566
 20  20  40  70 100                                          OR =  .5149
 30  30  20  15  15  100
 30  45  30  15  15   70 100
 20  30  45  30  20   40  70 100
 15  15  30  45  30   20  40  70 100
 15  15  20  30  60   20  20  40  70 100
 30  30  20  15  15   30  30  20  15  15  100
 30  45  30  15  15   30  45  30  15  15   70 100
 20  30  45  30  20   20  30  45  30  20   40  70 100
 15  15  30  45  30   15  15  30  45  30   20  40  70 100
 15  15  20  30  60   15  15  20  30  60   20  20  40  70 100
```
