The Effects of Correlated Errors on Generalizability and Dependability Coefficients

James E. Bost
University of Pittsburgh

This study investigated the effects of correlated errors on the person × occasion design in which the confounding effect of equal time intervals results in correlated error terms in the linear model. Two specific error correlation structures were examined: the first-order stationary autoregressive (SAR1), and the first-order nonstationary autoregressive (NAR1) with increasing variance parameters. The effects of correlated errors on the existing generalizability and dependability coefficients were assessed by simulating data with known variances (six different combinations of person, occasion, and error variances), occasion sizes, person sizes, correlation parameters, and increasing variance parameters. Estimates derived from the simulated data were compared to their true values. The traditional estimates were acceptable when the error terms were not correlated and the error variances were equal. The coefficients were underestimated when the errors were uncorrelated with increasing error variances. However, when the errors were correlated with equal variances the traditional formulas overestimated both coefficients. When the errors were correlated with increasing variances, the traditional formulas both overestimated and underestimated the coefficients. Finally, increasing the number of occasions sampled resulted in more improved generalizability coefficient estimates than dependability coefficient estimates. *Index terms: changing error variances, computer simulation, correlated errors, dependability coefficients, generalizability coefficients.*

In the past 20 years, generalizability theory has emerged as an important method of analyzing the reliability or generalizability of test scores or observational data when multiple sources of variation occur simultaneously. Traditionally, the reliability of test scores was measured by classical test theory statistics such as the Pearson product-moment correlation coefficient and coefficient alpha, which often could not adequately account for all sources of variation. Ebel (1951), Horst (1949), Hoyt (1941), and Medley & Mitzel (1958) showed that classical test theory estimates of reliability could be written in terms of the ratio of mean squares derived using analysis of variance (ANOVA) techniques. Cronbach, Gleser, Nanda, & Rajaratnam (1972) demonstrated that by using ANOVA mean squares for models with several effects, reliability estimates could be derived for multiple sources of variation (e.g., occasions, items, and raters) in a testing situation. The study of such reliability estimates is known as generalizability theory.

Random or mixed effects ANOVA methods are used to estimate the variance components in generalizability theory. The ANOVA mean squares are derived for each effect in the model and set equal to their expectations. The variance components then are determined by algebraic manipulation of these equations. Finally, generalizability coefficients (ρ^2), which reflect the ratios of various combinations of these variance components, are estimated. Consequently, some of the ANOVA assumptions must also apply to the ρ^2s. One of these assumptions is that the error components of the underlying ANOVA linear model are mutually uncorrelated.

Researchers have shown that correlated errors in random or mixed effects ANOVA models lead to variances that are misestimated and to mean squares that are biased estimates of their expectations (Adke, 1986; Andersen, Jensen, & Schoul, 1981; Browne, 1977; Smith & Luecht, 1992). Consequently, if errors are correlated, the reliability estimates calculated using generalizability theory techniques may be overes-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 2, June 1995, pp. 191–203
© Copyright 1995 Applied Psychological Measurement Inc.
0146-6216/95/020191-13$1.90

timating or underestimating the true generalizability of the test scores.

In generalizability theory, the first step is to collect the data and derive initial variance component estimates. This part of the study is called the generalizability study (G-study). The specifications of the G-study define the universe of admissible observations. After the G-study, estimates of ρ^2 and the dependability coefficient (ϕ) are derived under different specifications. These estimates are the outcomes of the decision studies (D-studies) in which the specifications of each D-study define the universe of generalization. The universe of generalization can be the same as the universe of admissible observations or it can be different under certain restrictions. However, the structure and assumptions of the linear model—and consequently the underlying structure of the linear model's error terms—must be the same in both the universe of generalization and the universe of admissible observations, or the variance component estimates are no longer valid.

In terms of test scores or observational data, Cronbach & Furby (1970) stated the correlated errors problem as follows:

> Sometimes X and Y observations are "linked" as when the two scores are obtained from a single test or battery administered at one sitting, or when observations on different occasions are made by the same observer. The correlation between linked observations will ordinarily be higher than that between independent observations. (p. 69)

Purpose

The purpose of this study was to evaluate the effects of correlated errors and changing error variances on ρ^2 and ϕ. Both coefficients measure the degree to which a D-study's results would generalize to similar universes. ρ^2 is used when relative decisions are to be made. For example, in testing situations, ρ^2 is appropriate when relative scores (e.g., percentiles) are used. ϕ is used when absolute decisions are to be made; that is, when the magnitude of the individual's score is of interest regardless of its relative position. When correlated errors exist, the variance components used for estimating ρ^2 and ϕ are biased. Also, the formulas for calculating ρ^2 and ϕ do not take into account the correlated error structure. This study examined the robustness of the existing coefficient estimation method when error variances change over time, when errors are correlated, or both.

Correlated Errors in ANOVA, Reliability, and Generalizability Theory

Box (1954) derived the expected mean squares for a balanced row-column fixed effects design in which the observations across rows for a particular person were assumed to be correlated. The results showed that column variances were overestimated and row variances underestimated. Using a single-facet design, Maxwell (1968) investigated the general effect of correlated errors on Kuder-Richardson Formula 20 (KR20). First, the equality between KR20 and the single-facet ρ^2 was demonstrated. Then, KR20 was derived when the correlation between measurements of the facet within-person was not 0. Maxwell demonstrated that correlated errors (when unknown) change the expected mean squares and that, for this design, positively correlated errors result in KR20 being overestimated.

The Conners' Teacher Rating Scale was assessed in a generalizability study by Conger, Conger, Wallander, Ward, & Dygdon (1983). In one of their designs, teachers filled out the form on a sample of children on three occasions. They concluded that

> ... ratings on the second occasion involve the teacher's first occasion impression, modified, if at all, by behaviors occurring in the two-week interim period. Similarly, third occasion ratings are based on impressions gathered through the second occasion and are modified by behaviors during the second interim period. (p. 1026)

This indicates that errors in responses across occasions may not be independent and may follow a serially correlated structure. Suen, Lee, & Owen (1990) studied the effects of autocorrelation on single-subject single-facet crossed-design ρ^2s. Using 28 people and 144 facet levels, they concluded that the error in estimating ρ^2 was negligible when the data were transformed to contain autoregressive, moving average, and both autoregressive and moving average dependencies across facet levels.

The effects on the variance components of the person × item (p × i) design with lag-1 serially correlated errors (correlation with previous time period only) or facets was the focus of Smith & Luecht (1992). They simulated data with equal person item and error variances, person and facet sizes of 10, 25, and 50, and correlation parameters of .2, .4, .6, and .8. They found that with serially correlated errors, the residual component was underestimated, and the person component was overestimated by a nearly equal amount. These biases would result in overestimated D-study ρ^2s or ϕs. The bias became negligible as the number of facet levels increased.

There is general agreement that correlated errors statistically and substantively affect the reliability or generalizability of scores derived from a testing situation or observational study. Multivariate generalizability theory has been discussed as a method of linking correlated observations (Brennan, 1983) because current univariate methods do not allow such linking.

Method

This study evaluated the robustness of the existing ordinary least squares (OLS) random effects ANOVA method of estimating the ϕ and ρ^2s for the person × occasion (p × o) design. Using multivariate normal data, with and without correlated errors and changing error variances, simulations were conducted to examine the effects of correlated errors on the coefficients. Previous research suggested that correlated errors would lead to overestimation of these coefficients.

Design

The p × o design was selected for analysis because it is the most basic design when correlated errors may be present. For the p × o design, scores from the same battery or observations from the same observer are collected on different occasions; the sample of persons is the same at each occasion and the time between occasions is the same. For the p × o design, the error terms are often serially correlated; that is, they follow a simplex pattern (Edwards, 1991).

Because occasions are time dependent, the correlation pattern can be modeled using a time series design. Based on the literature (Chi & Reinsel, 1989; Edwards, 1991; Mansour, Nordheim, & Rutledge, 1985), two serially correlated error structures were examined here: the first-order stationary autoregressive error structure (SAR1) and the first-order nonstationary autoregressive model (NAR1). Under SAR1, the error term correlation from one occasion to another decreases as the time between occasions increases, and the occasion error variances remain the same. This is often the case when observations are made close in time.

When error terms follow an SAR1 correlation structure, the ρ^2 estimates may not be correct. Consider the following examples. Medley & Mitzel (1958) conducted two studies of teacher behavior to demonstrate the use of ANOVA techniques to estimate reliability when more than two scores per person were obtained. For one of the studies, one person observed the performance of a sample of teachers every week for four weeks. Under that design, correlated errors would be expected although they were not corrected for in this design. The reliability coefficient used was a modified intraclass correlation coefficient proposed by Ebel (1951). It is equivalent to ρ^2 when the number of D-study observations (future teacher visits in this case) equals 1. The coefficients ranged from .25 to .50. The reliability estimates would certainly have changed if they had been adjusted to reflect the correlational nature of the error terms because the variance estimates were biased.

In a study by Shavelson & Webb (1981), a group of junior high schoolers' communication skills were recorded in 1-minute intervals for 5 minutes to determine whether more similar behaviors occurred close in time. As noted by Shavelson and Webb, the matrix of intercorrelations exhibited a simplex pattern but their ρ^2s did not adjust for this pattern.

The second error structure examined was NAR1. It is similar to SAR1 but the variances can either increase or decrease over time. This often occurs when observations are made far apart in time or when an intervention (e.g., instruction) is introduced. For example, as part of a larger study, Egeland, Pianta, & O'Brien (1990) collected total score data from the Peabody Individual Achievement Test. Scores from the same set of children were collected at Grades 1, 2, and 3. Edwards (1991) fit the data to an SAR1 error model, but the observed change in the variances across time indicated that the NAR1 model might have fit the data better. The instructional intervention from year to year as well as the length of time between testings showed that the variances changed as the time between occasions changed.

Another situation in which an NAR1 error correlation structure may occur is in the grading of students' written compositions (Werts, Breland, Grandy, & Rock, 1980). It is not uncommon for students to submit writing samples on the same topic throughout the term, with the hope that the skills and techniques taught during the semester will improve their writing performance. Depending on the initial ability levels of the students and the quality of the instruction between submissions, the error variances could either increase or decrease over time. If the initial ability levels of the students is varied and instruction is exceptional, the error variances may decrease over time. If the initial ability levels are similar but the teaching method seems to help only certain groups of students, the error variances may increase over time.

This study required that data for a p × o design contain a specific error covariance structure. Simulating the data was preferred over collecting data from p × o studies for several reasons: (1) balanced data (i.e., equal sample sizes), which simplifies the calculations of ρ^2 and ϕ, can be generated easily in simulation; (2) simulating data facilitates the ability to vary the number of person and occasions sampled; (3) data can be simulated to contain a desired error correlation structure; (4) the person, occasion, and error variances can be fixed using simulated data so that the "true" values for ρ^2 and ϕ are known; (5) it is important to replicate each situation to assess precision of estimation, and anomalies in the data generated for a particular simulation can be accounted for by using averages; and (6) using several replications insured that the simulated data had, on average, the appropriate underlying correlation structure.

The robustness of the OLS ANOVA method in estimating ρ^2 and ϕ was studied when the error terms had a specific correlation structure with either constant or increasing error variances. Traditional estimates of ρ^2 and ϕ calculated from simulated data with known person and occasion variances, known and either constant or increasing error variances, and correlated or uncorrelated errors were compared to their true values. For the p × o design, the coefficients have the following form:

$$\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{po}^2}{n'_o}}, \tag{1}$$

and

$$\phi = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_o^2}{n'_o} + \frac{\sigma_{po}^2}{n'_o}}, \tag{2}$$

where

σ_p^2 is the person variance,
σ_o^2 is the occasion variance,
σ_{po}^2 is the error variance (confounded with the interaction variance), and
n'_o is the number of D-study occasions sampled.

The error structures, parameter values, person sample sizes, and number of occasions were varied in the simulation process. σ^2_p, σ^2_o, and σ^2_{po} were fixed so that the true values for ρ^2 and ϕ would be known. The estimated coefficients calculated from the simulated data were compared to the true values. The true values should be unaffected by the correlation pattern.

Specification of Error Structures

First-order stationary autoregressive. The population variance for a particular score (Y_{ij}) on the *i*th occasion for the *j*th person has the following form:

$$\sigma^2_{Y_{ij}} = \sigma^2_p + \sigma^2_o + \frac{\sigma^2_e}{1-\xi^2}, \tag{3}$$

where $\sigma^2_e = \sigma^2_{po}$ is the error term variance, and ξ is the correlation parameter.

If each person is observed four times, the population covariance matrix for the error term across the four occasions for a particular person is

$$\sigma^2_e\mathbf{V} = \left[\sigma^2_e/\left(1-\xi^2\right)\right]\begin{bmatrix} 1 & \xi & \xi^2 & \xi^3 \\ \xi & 1 & \xi & \xi^2 \\ \xi^2 & \xi & 1 & \xi \\ \xi^3 & \xi^2 & \xi & 1 \end{bmatrix}. \tag{4}$$

In general, each cell of the matrix takes the form $v_{lk} = \xi^{|l-k|}/\left(1-\xi^2\right)$.

First-order nonstationary autoregressive. A NAR1 structure assumes that, for each person, the correlation between scores decreases as the time between occasions increases; the variance at each occasion also may change. The error covariance structure can take on different forms, depending on the degree of variability built into the model. The model proposed by Edwards (1991) allows for increasing or decreasing variances over time. The population error covariance matrix for a particular person on four occasions would have the form

$$\sigma^2_e\mathbf{V} = \left[\sigma^2_e/\left(1-\xi^2\right)\right]\begin{bmatrix} \eta_1^2 & \eta_1\eta_2\xi & \eta_1\eta_3\xi^2 & \eta_1\eta_4\xi^3 \\ \eta_1\eta_2\xi & \eta_2^2 & \eta_2\eta_3\xi & \eta_2\eta_4\xi^2 \\ \eta_1\eta_3\xi^2 & \eta_2\eta_3\xi & \eta_3^2 & \eta_3\eta_4\xi \\ \eta_1\eta_4\xi^3 & \eta_2\eta_4\xi^2 & \eta_3\eta_4\xi & \eta_4^2 \end{bmatrix}, \tag{5}$$

where η_i is the increasing or decreasing variance parameter. The general form of each cell is

$$v_{lk} = \eta_l\eta_k\xi^{|l-k|}/\left(1-\xi^2\right). \tag{6}$$

The population error correlation matrix is the same as the SAR1 matrix. In fact, if $\eta_1 = \eta_2 = \eta_3 = \eta_4 = 1$, the above covariance matrix would equal its SAR1 counterpart. Note that it is the error variance that increases across occasions and not the occasion variance (which remains constant because occasion is a random facet). The increasing error variances can be considered a confounding time effect not accounted for in the model.

For purposes of notation and the simulation program, it was assumed that the design matrix entered the occasions first, followed by persons, and then the error term. If the design matrix is so ordered, the population error covariance matrix for the entire set of observations has the form $\mathbf{W} = \sigma^2_e(\mathbf{V} \otimes \mathbf{I}_{n_p})$, where $\sigma^2_e\mathbf{V}$ is either the SAR1 or NAR1 covariance matrix and n_p is the number of persons. The symbol $\otimes$ is the Kronecker product of a matrix.

Simulating p × o Data

The p × o design is time dependent and often has a within-person covariance structure that is not the identity matrix. The model has the following form:

$$Y_{ij} = \alpha + \tau_i + \beta_j + z_{ij}\,, \tag{7}$$

where

α is the grand mean,
$\tau_i \sim \mathrm{E}(\tau_i) = 0$ and variance $= \sigma_o^2$, $i = 1, \ldots, n_o$ occasions;
$\beta_j \sim \mathrm{E}(\beta_j) = 0$ and variance $= \sigma_p^2$, $j = 1, \ldots, n_p$ persons;
$z_{ij} \sim \mathrm{E}(z_{ij}) = 0$ and variance-covariance matrix $\sigma_e^2(\mathbf{V} \otimes \mathbf{I}_{n_p})$; and
$\mathbf{V} = \{v_{lk}\}$ $(l = 1, \ldots, n_o;\ k = 1, \ldots, n_o)$, where v_{ll} represents the diagonal element at occasion l, and v_{lk} represents the off-diagonal element for occasions l and k.

Because two error correlation conditions were simulated (SAR1 and NAR1), the simulation method had to incorporate the above covariance structures into the model and allow the number of persons, the number of occasions, and the covariance matrix parameters to vary. It also had to allow the user to specify σ_p^2, σ_o^2, and σ_e^2 in the data. Finally, for convenience it had to transform the generated data values to insure positive responses.

The set of data values generated (**Y**) had to have a variance-covariance matrix equal to $(\sigma_o^2\mathbf{I}_{n_o} + \sigma_e^2\mathbf{V}) \otimes \sigma_p^2\mathbf{I}_{n_p}$. The method used to generate multivariate scores with this specific variance-covariance structure combined the Cholesky decomposition method (Kennedy & Gentle, 1980) used by Edwards (1991) and the additive effects method (Smith, 1978); it included the following steps:

1. Generate an $n_p \times n_o$ matrix of standard normal deviates Z;
2. Generate **V**, the desired within-person error covariance matrix (SAR1 or NAR1) before incorporating the error variances;
3. Let c be the selected σ_e^2;
4. Let $\mathbf{D} = c\mathbf{V}$ in order to obtain the desired intermediate covariance matrix $\sigma_e^2\mathbf{V}$;
5. Derive the Cholesky decomposition matrix **A** of the covariance matrix **D**;
6. Calculate $\mathbf{Y} = \mathbf{ZA}$ so that each row of **Y** is $\mathrm{N}(\mathbf{0}, \mathbf{D})$;
7. Let f be the selected σ_o^2;
8. Generate **K**, a $1 \times n_o$ vector of normal random deviates with variance-covariance matrix $\mathbf{G} = f{*}\,\mathbf{I}$, where **I** is the $n_o \times n_o$ (occasion) identity matrix;
9. Add **K** to each row of **Y** so that each row of **Y** is now $\mathrm{N}_o(\mathbf{0}, \mathbf{D} + \mathbf{G})$;
10. Generate **M**, an $n_p \times 1$ vector of normal random deviates with mean 0 and variance-covariance matrix $\mathbf{N} = t{*}\mathbf{I}$ where $t = [1 - (c + f)]$ is σ_p^2 and **I** is the $n_p \times n_p$ (person) identity matrix (the specific values selected for c, f, and t are explained below); and
11. Add **M** to each column of **Y** so that **Y** is now $[\mathbf{0},(\mathbf{D} + \mathbf{G})\otimes\mathbf{N}]$.

The above algorithm was modified to simulate an NAR1 error structure. The vector of variance parameters (**g**) was multiplied by its transpose, resulting in an $n_p \times n_p$ matrix. Element-wise multiplication of this matrix by **V** created the NAR1 error term matrix. The above simulation algorithm was programmed using PROC IML of SAS (SAS, 1989). Because PROC IML was used to calculate the variance components and coefficients, it was advantageous to write the simulation program in SAS.

Specifying the Variances

A critical step in the above simulation is selecting values for $f = \sigma_o^2$, $c = \sigma_e^2$, and $t = \sigma_p^2$. Like other generalizability simulation studies (e.g., DiStefano, 1979), the variances in this study were set to specific values to represent possible true variance values for the p × o design. Seven studies (e.g., Edwards, 1991; Lee & Kim, 1990; Medley & Mitzel, 1958; Webb & Shavelson, 1981;) were consulted to evaluate the percent of total variance usually attributed to σ_o^2, σ_e^2, and σ_p^2. From these studies, 20 occasion and error variances were evaluated. Between 0% and 30% of the total variance was attributed to the occasion component. The percent of total variance attributed to the error term was between 10% and 50%. To select the variances, the total variance $(\sigma_o^2 + \sigma_p^2 + \sigma_e^2)$ was fixed at 1. Based on the above percentages, .1 and .3 were selected for σ_o^2; .1, .3, and .5 were selected for σ_e^2. In each case, $\sigma_p^2 = 1 - (\sigma_o^2 + \sigma_e^2)$. Table 1 provides a list of the six variance component sets for the simulation study. These six combinations were used to generate the simulated data from which the estimated ρ^2 and ϕ values were calculated.

Table 1
Six Variance Combination Sets Used in the Simulation

Variance Combination	$\sigma_e^2 = \sigma_{po}^2$	σ_o^2	σ_p^2
1	.1	.1	.8
2	.1	.3	.6
3	.3	.1	.6
4	.3	.3	.4
5	.5	.1	.4
6	.5	.3	.2

Simulated Parameter Values and Sample Sizes

For the SAR1 and NAR1 correlation structure, a total of 3 (occasion sample size) × 3 (person sample size) × 4 (ξ values) = 36 different simulations were run for each of the six variance combinations. Each simulation was replicated 50 times and a new seed value was used for each replication; this mimicked the process of selecting a different randomly parallel set of persons and occasions each time. 50 replications were selected to adequately account for the variability inherent in the simulation process.

Three sample sizes were simulated for persons (n_p = 15, 25, and 50). These sample sizes were selected to reflect class sizes and are consistent with person sample sizes from published studies. Three sizes were simulated for occasions ($n_o = n'_o$ = 3, 5, and 7, where n'_o is the number of occasions in the D-study and n_o is the number of occasions in the G-study). These number of occasions were selected to be consistent with studies having similar designs (e.g., Egeland et al., 1990; Medley & Mitzel, 1958; Shavelson & Webb, 1981) and to investigate the possibility of trends in ρ^2 and ϕ as a function of the number of occasions. ξ was simulated at four values—0.0, .3, .5, and .7—to represent a range from 0 to high correlation. The η values selected for the NAR1 structure had an increasing variance pattern. The specific values for η were 1.0, 1.2, and 1.5 for $n_o = 3$; 1.0, 1.2, 1.5, 2.0, and 2.5 for $n_o = 5$; and 1.0, 1.2, 1.5, 2.0, 2.5, 3.0, 3.5 for $n_o = 7$. These values were selected to simulate a relatively steady increase in variability and are identical to the values used in Edwards (1991).

When $\sigma_o^2 = .1$, the true values for ρ^2 and ϕ ranged from .66 to .98 (depending on the number of occasions) with a difference in the two coefficients of approximately .04. When $\sigma_o^2 = .3$, the true values for ρ^2 and ϕ ranged from .43 to .98 (depending on the number of occasions), and the two coefficients were much further apart. Consequently, several possible scenarios were examined while staying approximately within the bounds of previous studies that have used this design. For each simulation, ρ^2 and ϕ (Equations 1 and 2) were calculated assuming the number of occasions in the D-study (n'_o) was the same as the number of occasions in the G-study (n_o).

Statistics Used to Assess Robustness

To determine the robustness of the OLS ANOVA method of estimating ρ^2 and ϕ when the error terms have a specific correlation structure with either constant or increasing error variances, the following statistics were calculated: $\bar{\hat{\rho}}^2$ and $\bar{\hat{\phi}}$ (the means of the 50 estimates of ρ^2 and ϕ calculated from the replicated simulated data), $\bar{\hat{\sigma}}_o^2$, $\bar{\hat{\sigma}}_p^2$, and $\bar{\hat{\sigma}}_{po}^2$ (the means of the estimates of the individual variance components calculated from the replicated data). In addition to these estimates, the difference between the estimated and true values ($\bar{\hat{\rho}}^2 - \rho^2$ and $\bar{\hat{\phi}} - \phi$) also were calculated along with the percent difference calculated as

$$\Delta\rho^2 = 1 - \frac{\bar{\hat{\rho}}^2}{\rho^2}, \tag{8}$$

and

$$\Delta\phi = 1 - \frac{\bar{\hat{\phi}}}{\phi}. \tag{9}$$

For each variance combination, the average of the above statistics was calculated based on the 3 occasion sizes × 3 person sizes × 4 ξ values. Finally, the above summaries were done separately for each set of person sizes, occasion sizes, and ξs.

Results

Robustness Assessment With Increasing Error Variances

For the NAR1 error structure, when $\xi = 0$, the simulated data did not have correlated errors but the error term variances increased over time. Table 2 contains the summarized results for ρ^2. For all six variance sets, ρ^2 was underestimated. The difference between the estimated and true values ranged from −.031 to −.447. As the proportion of variance attributed to error increased, the underestimation dramatically increased. For example, when $\sigma_{po}^2 = .5$, the true coefficients indicated moderate to high generalizability whereas the estimates suggested only low to moderate generalizability. These results imply that unless the error term variances are very small, ρ^2 is substantially underestimated using the traditional OLS random effects ANOVA method. By looking at the variance estimates, it is clear that the underestimation is the result of σ_{po}^2 being substantially overestimated. For example, when $\sigma_{po}^2 = .3$ (variance combinations 3 and 4), $\bar{\hat{\sigma}}_{po}^2$ ranged from .471 to 1.560. The results for ϕ (not shown here) were similar to those for ρ^2, except that the values for $\Delta\phi$ (Equation 9) were slightly smaller due to the fact that $\Delta\phi$ has the additional occasion variance component in ϕ, which is independent of the error term variance σ_{po}^2.

For each of the six variance combinations, as the number of occasions increased, the estimates of ρ^2 and ϕ became increasingly less than the true values. This is due to the increasing heterogeneity of variances as a result of the large η values. Figure 1, a plot of the difference between the true relative error variance [$\sigma^2(\delta) = \sigma_{po}^2/n_o$] and its estimate by occasion size for the three different error term variances, shows that the number of occasions (and consequently the number and magnitude of the η values) affected the size of $\bar{\hat{\sigma}}_{po}^2$ as well as the accuracy of $\bar{\hat{\rho}}^2$. Similar results were demonstrated for ϕ and the absolute error variance [$\sigma^2(\Delta)$]. Finally, increasing the number of persons sampled had a negligible effect on the estimates of ρ^2 and ϕ.

Robustness Analysis With the SAR1 Error Structure

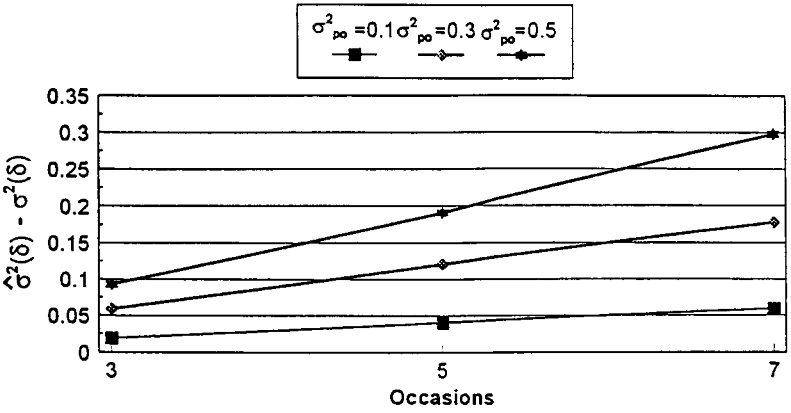
The SAR1 error structure assumes that, for a given individual, the error term correlation decreases as the time between occasions increases, and the error variances remain constant. Table 2 contains the summarized results for ρ^2 and ϕ. Table 2 shows that as σ_{po}^2/σ_p^2 increased, ρ^2 became increasingly overestimated. For example, when $\rho^2 = .55$ ($\sigma_{po}^2/\sigma_p^2 = .5/.2$) and $n_o = 3$, $\hat{\rho}^2 = .701$. When $\sigma_{po}^2 = .5$, 64% of the ρ^2 values were

Table 2
Robustness Statistics Averaged Over 50 Replications of ρ^2 and Over 3 Person Values for NAR1 With $\xi = 0$, and Over 3 Person and 4 ξ Values for SAR1

Variance Combination,	NAR1, $\xi = 0$ (No Correlation and Increasing Error Variance)						SAR1					
ρ^2, and n_o	$\bar{\hat{\rho}}^2$	$\bar{\hat{\rho}}^2 - \rho^2$	$\Delta\rho^2$	$\bar{\hat{\sigma}}^2_{po}$	$\bar{\hat{\sigma}}^2_o$	$\bar{\hat{\sigma}}^2_p$	$\bar{\hat{\rho}}^2$	$\bar{\hat{\rho}}^2 - \rho^2$	$\Delta\rho^2$	$\bar{\hat{\sigma}}^2_{po}$	$\bar{\hat{\sigma}}^2_o$	$\bar{\hat{\sigma}}^2_p$
Variance Combination 1												
$\rho^2 = .96, n_o = 3$	.929	−.031	.033	.156	.098	.776	.965	.005	−.006	.084	.105	.860
$\rho^2 = .98, n_o = 5$	.920	−.056	.057	.304	.103	.766	.976	0.000	0.000	.094	.098	.845
$\rho^2 = .98, n_o = 7$	.907	−.075	.076	.511	.096	.804	.981	−.002	.002	.103	.101	.835
Variance Combination 2												
$\rho^2 = .95, n_o = 3$	.911	−.036	.038	.158	.318	.609	.955	.007	−.008	.083	.308	.650
$\rho^2 = .97, n_o = 5$	.899	−.069	.071	.300	.263	.598	.968	0.000	0.000	.095	.299	.632
$\rho^2 = .98, n_o = 7$	.882	−.095	.097	.521	.301	.612	.975	−.002	.002	.102	.301	.617
Variance Combination 3												
$\rho^2 = .86, n_o = 3$	.783	−.074	.086	.471	.093	.636	.888	.030	−.035	.254	.092	.766
$\rho^2 = .91, n_o = 5$	.731	−.178	.195	.928	.099	.571	.916	.007	−.007	.285	.101	.717
$\rho^2 = .93, n_o = 7$	.693	−.237	.258	1.528	.099	.571	.935	.002	−.002	.308	.108	.708
Variance Combination 4												
$\rho^2 = .80, n_o = 3$	.687	−.113	.141	.483	.319	.405	.838	.038	−.048	.257	.298	.533
$\rho^2 = .87, n_o = 5$	.664	−.206	.237	.882	.280	.400	.886	.017	−.019	.285	.287	.511
$\rho^2 = .90, n_o = 7$	.614	−.290	.321	1.560	.297	.423	.910	.007	−.008	.306	.288	.497
Variance Combination 5												
$\rho^2 = .80, n_o = 3$	.574	−.132	.186	.791	.097	.411	.780	.074	−.105	.426	.104	.643
$\rho^2 = .87, n_o = 5$	.550	−.250	.313	1.448	.091	.414	.839	.039	−.049	.475	.098	.604
$\rho^2 = .90, n_o = 7$	.482	−.367	.432	2.574	.109	.406	.868	.020	−.023	.514	.104	.571
Variance Combination 6												
$\rho^2 = .55, n_o = 3$	.437	−.109	.199	.769	.326	.227	.701	.156	−.287	.420	.282	.459
$\rho^2 = .67, n_o = 5$	.398	−.268	.402	1.462	.303	.225	.755	.088	−.132	.478	.298	.407
$\rho^2 = .74, n_o = 7$	.290	−.447	.607	2.596	.299	.173	.790	.054	−.073	.513	.286	.365

overestimated by more than 10%. The estimates of ρ^2 were exaggerated due to the overestimated person variance components ($\hat{\sigma}^2_p > \sigma^2_p$), which also became worse as σ^2_{po}/σ^2_p increased. σ^2_{po} also was underestimated in each case by a very small magnitude. Similar results were obtained for ϕ, which was expected

Figure 1
The Difference Between Estimated and True Relative Error Variance [$\hat{\sigma}^2(\delta) - \sigma^2(\delta)$] as a Function of the Number of Occasions Sampled for Each Error Variance Assessed



because the occasion variances were accurately estimated.

For each of the six variance combinations, as the number of occasions increased, the estimates of ρ^2 and ϕ improved. For variance combination 6, ρ^2 was overestimated by 28.7% when $n_o = 3$ but by only 7.3% when $n_o = 7$. This was primarily due to less measurement error when the number of observations increased and smaller ξs when occasions became farther apart in time. Increasing the number of persons simulated did not improve the estimates—it made them slightly worse. This may be due to the fact that errors were correlated within persons but not across persons. As expected, increasing the size of ξ increased the amount that ρ^2 and ϕ were overestimated (results for the 4 levels of ξ are not shown here). This is consistent with the results presented by Maxwell (1968) for KR20 and Williams & Zimmerman (1977) for the reliability of difference scores when the scores are correlated. Also, the larger the ξ, the less precise were the estimates. Across all six variance combination sets, the standard deviation of $\Delta\rho^2$ increased from .026 when $\xi = 0$ to .148 when $\xi = .7$, and the standard deviation of $\Delta\phi$ increased from .026 when $\xi = 0$ to .209 when $\xi = .7$.

Robustness Analysis With NAR1 Error Structures

Recall that the NAR1 error structure was the same as the SAR1 error structure except increasing variance parameters were added. Results in Table 2 for NAR1 with $\xi = 0$ indicate that increasing error term variances resulted in estimates of ρ^2 and ϕ that were too low and that correlated errors resulted in estimates of ρ^2 and ϕ that were too high. The effects of both of these anomalies occurred simultaneously in this study.

Table 3 contains the summarized results for ρ^2 and ϕ. At first glance, the estimates seem fairly accurate with the means of $\bar{\hat{\rho}}^2 - \rho^2$ never falling below $-.2$ and the average $\Delta\rho^2$ never exceeding 19%. However, the variance estimates and the ranges of $\Delta\rho^2$ provide additional information. σ^2_{po} became more overestimated as the true proportion of total variability attributed to error (σ^2_{po}) increased. For example, in variance combination 5 in which $\sigma^2_{po} = .5$ and $n_o = 7$, $\hat{\sigma}^2_{po} = 2.736$. The overestimations were of the same magnitude as the situations in which the error term variances were increasing and the error terms were uncorrelated (see Table 2). However, the overestimation of σ^2_p was much larger under NAR1 than SAR1 [e.g., for variance combination 5 and $n_o = 7$, $\hat{\sigma}^2_{po} = 2.736$ for NAR1 and $\hat{\sigma}^2_{po} = .514$ for SAR1 (see Table 2)]. σ^2_o was accurately estimated across all the variance sets. As the ratio of σ^2_{po}/σ^2_p increased, $\Delta\rho^2$ increased but then started to decrease. For $n_o = 5$, starting with variance combination 1, $\Delta\rho^2$ began to increase, peaked for variance combination 5 ($\Delta\rho^2 = .129$), and decreased for variance combination 6 to $\Delta\rho^2 = .074$. The underestimation of ρ^2 due to increasing error term variances was most pronounced when σ^2_{po}/σ^2_p was small. The correlated errors reversed this effect, especially when $\sigma^2_{po}/\sigma^2_p > 1$. Also, although the mean $\Delta\rho^2$ for each variance set was not large, the majority of $\Delta\rho^2$ values were either below $-.1$ or above .1 (a 10% difference). In variance combination 6 and $n_o = 5$, 50% of the $\Delta\rho^2$ values were greater than 10% due to σ^2_{po} being large. However, 25% of the $\Delta\rho^2$ values were less than -10% due to more weight given to the correlation matrix elements as σ^2_{po} increased. In many situations a discrepancy of 10% or more would be unacceptable. Finally, Table 3 shows that increasing σ^2_{po} had a more pronounced effect on the estimates of ρ^2 than the correlated errors due to more $\Delta\rho^2$ values being above 10% then below 10%.

Table 3 also includes summary statistics for ϕ to show that the estimates of dependability behaved differently than those of generalizability under an NAR1 error structure. For the sets of variances with the same occasion variance (.1 or .3), as the ratio of σ^2_{po}/σ^2_p increased, $\Delta\phi$ increased but then decreased. When $\sigma^2_o = .3$, the $\Delta\phi$ values were more often smaller than when $\sigma^2_o = .1$. This implies that the overestimation effects of correlated errors became stronger as the percentage of total variance attributed to occasions increased. In fact, for variance combination 6, the $\Delta\phi$ values indicate that ϕ was most often overestimated.

Both ρ^2 and ϕ became more underestimated as the number of occasions sampled increased. The $\Delta\rho^2$ and $\Delta\phi$ values were not as large as the corresponding values when the errors only had increasing variances, because of the competing effects of the correlated errors. Across all six variance combinations, when $n_o =$

Table 3
Robustness Statistics Averaged Over 50 Replications of ρ^2 and Over 3 Person Values for NAR1 With $\xi = 0$, and Over 3 Person and 4 ξ Values for NAR1, and Number of $\Delta\rho^2$ and $\Delta\phi$ Values Below −10% (% < −.1) and Above 10% (% > .1)

Variance Combination, ρ^2,				$\Delta\rho^2$									$\Delta\phi$	
ϕ, and n_o	$\bar{\hat{\rho}}^2$	$\bar{\hat{\rho}}^2 - \rho^2$	$\Delta\rho^2$	% < −.1	% > .1	$\bar{\hat{\sigma}}^2_{po}$	$\bar{\hat{\sigma}}^2_{o}$	$\bar{\hat{\sigma}}^2_{p}$	$\bar{\hat{\phi}}$	$\bar{\hat{\phi}} - \phi$	$\Delta\phi$	% < −.1	% > .1
Variance Combination 1														
$\rho^2 = .96, \phi = .92, n_o = 3$	.944	−.016	.017	0	0	.134	.102	.864	.908	−.015	.017	0	0
$\rho^2 = .98, \phi = .95, n_o = 5$	.932	−.044	.045	0	0	.299	.097	.913	.912	−.040	.042	0	0
$\rho^2 = .98, \phi = .97, n_o = 7$	.914	−.068	.070	0	0	.549	.096	.945	.901	−.065	.067	0	0
Variance Combination 2														
$\rho^2 = .95, \phi =. 82, n_o = 3$	.930	−.017	.018	0	0	.135	.289	.678	.825	.007	−.008	0	0
$\rho^2 = .97, \phi = .88, n_o = 5$	.913	−.055	.057	0	0	.297	.277	.700	.848	−.035	.039	0	0
$\rho^2 = .98, \phi = .91, n_o = 7$	.896	−.081	.083	0	16.7	.551	.295	.763	.850	−.063	.069	0	8.3
Variance Combination 3														
$\rho^2 = .86, \phi = .82, n_o = 3$	.843	−.014	.016	0	0	.405	.099	.850	.815	−.003	.004	0	0
$\rho^2 = .91, \phi = .88, n_o = 5$	.810	−.099	.109	0	50.0	.898	.104	.930	.794	−.089	.100	0	50.0
$\rho^2 = .93, \phi = .91, n_o = 7$	.782	−.152	.163	0	75.0	1.651	.102	1.058	.772	−.141	.155	0	75.0
Variance Combination 4														
$\rho^2 = .80, \phi = .67, n_o = 3$	.791	−.009	.011	16.7	25.0	.403	.311	.646	.702	.036	−.053	33.3	16.7
$\rho^2 = .87, \phi = .77, n_o = 5$	.761	−.108	.124	0	58.3	.902	.303	.726	.710	−.059	.077	0	33.3
$\rho^2 = .90, \phi = .82, n_o = 7$	.732	−.171	.189	0	75.0	1.634	.284	.844	.703	−.121	.147	0	66.7
Variance Combination 5														
$\rho^2 = .71, \phi = .67, n_o = 3$	.729	.023	−.032	50.0	25.0	.667	.094	.796	.706	.040	−.059	50.0	25.0
$\rho^2 = .80, \phi = .77, n_o = 5$	.696	−.104	.129	0	50.0	1.476	.101	.945	.684	−.086	.111	0	50.0
$\rho^2 = .85, \phi = .82, n_o = 7$	.665	−.183	.216	0	75.0	2.736	.101	1.129	.658	−.166	.202	0	75.0
Variance Combination 6														
$\rho^2 = .55, \phi = .67, n_o = 3$	.651	.105	−.193	58.3	25.0	.672	.289	.594	.587	.158	.369	75.0	16.7
$\rho^2 = .67, \phi = .77, n_o = 5$	.617	−.050	.074	25.0	50.0	1.468	.294	.732	.582	.026	−.047	50.0	33.3
$\rho^2 = .74, \phi = .82, n_o = 7$	.581	−.156	.212	16.7	58.3	2.781	.314	.937	.560	−.076	.120	25.0	50.0

3, ρ^2 and ϕ were most often overestimated. However, when $n_o = 7$ and $\sigma^2_{po} = .3$, 71% of the $\Delta\rho^2$ and 60% of the $\Delta\phi$ values were larger than 10%, indicating substantial parameter underestimation.

Although not displayed in Table 3, for each of the six variance sets, increasing n_p had a negligible effect on the estimates of ρ^2 and ϕ. As expected, the larger the ξ, the smaller the values of $\Delta\rho^2$ and $\Delta\phi$. The overestimation of ρ^2 and ϕ became more pronounced as the error terms within each person exhibited a stronger correlation. Also, the standard deviation of $\Delta\rho^2$ and $\Delta\phi$ increased as ξ increased, indicating that the estimates of ρ^2 and ϕ became less precise as ξ increased.

Discussion

The purpose of this study was to assess the effects of correlated errors on ρ^2 and ϕ for the p × o design. There are many testing situations and observational studies in which occasion is a factor. In these situations, occasion is confounded with time and the error terms from one time period to the next are most likely correlated. Furthermore, the correlation pattern from one time period to the next often follows a known pattern, such as the SAR1 or NAR1 structure. When the correlation pattern is known, determining the effects of such an error structure on ρ^2 and ϕ can and should be assessed.

When the assumptions of uncorrelated errors and equal variances are satisfied, the traditional method of using OLS random effects ANOVA provides accurate estimates of the variance components for ρ^2 and ϕ for the p × o design (Bost, 1993). It is expected that this method would perform similarly for more complex designs. When the assumption of uncorrelated errors was satisfied but the error term variances increased over time, the traditional ANOVA method underestimated ρ^2 and ϕ. This was primarily due to overestimated error variances (σ^2_{po}).

When the error terms within a person had an SAR1 structure (correlated with equal variance), the traditional ANOVA method overestimated ρ^2 and ϕ primarily due to underestimated person variance (σ_p^2). In fact, the results of this study suggest that when the ratio of error to person variance is greater than .5, the traditional method should not be used. The traditional method estimates improved as more occasions were sampled and ξ approached 0.

When the error terms within a person had an NAR1 structure and increasing error variance parameters, the traditional ANOVA method both overestimated and underestimated ρ^2 and ϕ. When $.5 < \sigma^2_{po}/\sigma_p^2 < 1$, the underestimation effect of increasing variance had the most pronounced effect on ρ^2 and ϕ. When $\sigma^2_{po}/\sigma_p^2 > 1$, both the correlated errors and the unequal variances affected the estimates in opposite directions. The overestimation effect of correlated errors was also stronger for ϕ than for ρ^2. Increasing the number of occasions sampled exacerbated the underestimation problem; increasing the value of ξ increased the overestimation problem. The traditional formulas were acceptable only when most of the total variability in the model was attributed to persons.

The results of this study imply that if the errors were correlated in published generalizability studies, the reported ρ^2s would most likely be overestimates. For example, the study by Conger et al. (1983) reported $\rho^2 = .758$ for a p × o design with $n_o = n'_o = 3$. Unless the errors within persons had zero correlation, .758 is an overestimate of the study's generalizability. Also, the study outlined in Webb & Shavelson (1981), in which students were observed at five different time periods, showed that the data exhibited a simplex pattern in its correlation matrix across observations. Using the traditional ρ^2 or ϕ formula would not accurately quantify the generalizability of the study's outcomes to future studies.

The findings of this study are applicable to other generalizability theory designs as well as to classical reliability estimates. ρ^2 and ϕ derived from more complex designs with correlated errors would have similarly biased ρ^2 estimates. These include designs with more than an occasion factor (the person × item × occasion design, in which the errors within a person and item across occasions are correlated), designs in which more than one factor could affect the error correlation matrix [the person × occasion × rater design in which the error correlation structure for each person and rater is SAR1 (i.e., across occasions)], the case in which two factors are confounded (a person × occasion × rater design with one error correlation structure across raters and a different error correlation structure across persons), and nested designs (the person × item nested within domain design in which errors of items in the same domain are correlated).

This study has demonstrated that the traditional ρ^2s and ϕs are often not appropriate when there are within-person correlations; future work will concentrate on developing alternative formulas for ρ^2 and ϕ. These formulas must incorporate the error term covariance matrix in the formulas so that the universe of generalizability is developed under the same assumptions as the universe of admissible observations.

References

Adke, S. R. (1986). One way ANOVA for dependent observations. *Communications in Statistics—Theory and Methodology, 15,* 1515–1528.

Andersen, A. H., Jensen, E. B., & Schoul, G. (1981). Two-way analysis of variance with correlated errors. *International Statistical Review, 49,* 153–169.

Bost, J. E. (1993). The effects of correlated errors on generalizability coefficients (Doctoral dissertation, University of Pittsburgh, 1992). *Dissertation Abstracts International,* 1718.

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

Brennan, R. L. (1983). *Elements of generalizability theory.*

Iowa City IA: The American College Testing Program.

Browne, M. W. (1977). The analysis of patterned correlation matrices by generalized least squares. *British Journal of Mathematics, Statistics, and Psychology, 30,* 113–124.

Chi, E. M., & Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association, 84*(406), 452–459.

Conger, A. J., Conger, J. C., Wallander, J., Ward, D., & Dygdon, J. (1983). A generalizability study of the Conners' teacher rating scale—revised. *Educational and Psychological Measurement, 43,* 1019–1031.

Cronbach, L. J., & Furby, L. (1970). How should we "measure" change—or should we? *Psychological Bulletin, 74,* 68–80.

Cronbach, L. J., Gleser, G. C., Nanda, H. A. N., & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: Wiley.

DiStefano, J. A. (1979). *A Monte Carlo approach to the sampling error of estimates of the generalizability statistics from a two facet completely crossed generalizability study.* Unpublished doctoral dissertation, University of Maryland, College Park.

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

Edwards, L. K. (1991). Fitting a serial correlation pattern to repeated observations. *Journal of Educational Statistics, 16,* 53–76.

Egeland, B., Pianta, R., & O'Brien, M. (1990). *Maternal intrusiveness in infancy and child maladaptation in early school years.* Manuscript submitted for publication.

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

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

Kennedy, W. J., Jr., & Gentle, J. E. (1980). *Statistical computing.* New York: Marcel Dekker.

Lee, J. S., & Kim, Y. B. (1990, April). *An application of generalizability theory to a scientific thinking and research skills test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston MA.

Mansour H., Nordheim, E. V., & Rutledge, J. J. (1985). Maximum likelihood estimation of variance components in repeated measures designs assuming autoregressive errors. *Biometrics, 41,* 287–294.

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

Medley, D. M., & Mitzel, H. E. (1958). Applications of analysis of variance to the estimation of the reliability of observations of teachers' classroom behavior. *Journal of Experimental Education, 27,* 23–35.

SAS (1989). *SAS/IML software: Usage and reference* (Ver. 6, 1st ed.). Cary NC: SAS Institute Inc.

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

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., & Luecht, R. M. (1992). Correlated effects in generalizability studies. *Applied Psychological Measurement, 16,* 229–235.

Suen, H. K., Lee, P. S. C., & Owen, S. V. (1990). Effects of autocorrelation on single-subject single-facet crossed-design generalizability assessment. *Behavioral Assessment, 12,* 305–315.

Webb, N. M., & Shavelson, R. J. (1981). Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement, 18,* 13–22.

Werts, C. E., Breland, H. M., Grandy, J., & Rock, D. R. (1980). Using longitudinal data to estimate reliability in the presence of correlated measurement errors. *Educational and Psychological Measurement, 40,* 19–29.

Williams, R. H., & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement, 37,* 679–689.

Acknowledgments

Thanks to Michael Harwell, Clement Stone, Louis Pingel, Leon Gleser and especially Suzanne Lane for their contributions to the initial development of this manuscript, and to the reviewers whose comments and suggestions led to substantial improvements.

Author's Address

Send requests for reprints or further information to James E. Bost, American College of Cardiology, 9111 Old Georgetown Road, Bethesda MD 20814-1699, U.S.A. Internet: jbost@acc.org.