Correlated Effects in Generalizability Studies

Philip L. Smith, University of Wisconsin, Milwaukee
Richard M. Luecht, The American College Testing Program

The analytical model typically used to perform generalizability analysis assumes that design effects are uncorrelated. Often, the assessment of behavioral data involves designs that employ multiple occasions or repeated trials (as in many observational and rating studies). In these cases, design effects may be serially correlated. The implications of serially correlated effects on the results of generalizability analyses are discussed. Simulated data are provided that demonstrate the biases that serially correlated effects introduce into the results. *Index terms: correlated effects, estimation of variance components, generalizability theory, observational studies, repeated trials, serial correlation.*

Generalizability analysis (Cronbach & Gleser, 1964; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach & Rajaratnam, 1965) has become an important and commonly used method for assessing the dependability of behavioral measures. Generalizability theory is flexible and able to handle different types of data. These capabilities have made the theory more appealing than classical test theory (Lord & Novick, 1968), which has no formal provision for assessing the dependability of data. Given the widespread use of generalizability analyses, information related to the behavior of the statistics associated with the theory is of interest to practitioners.

The estimation properties and applied uses of variance components, which form the central core of generalizability analysis, have been the focus of much previous research. Issues involving the estimation of variance components under many of the typical linear models applicable to generalizability analysis have received fairly extensive coverage in the literature (e.g., Brennan, 1983; Cronbach et al., 1972; Crump, 1946; Eisenhart, 1947; Henderson, 1953; Searle, 1971). Additional research has centered on the estimation of sampling distributions and the stability of estimates of variance components (e.g., Boardman, 1974; Brennan, 1983; Searle, 1970, 1971; Smith, 1978, 1982).

Other characteristics of the data that may influence the statistics resulting from a generalizability analysis are less well-researched. For example, generalizability theory employs a score model that makes certain assumptions, by design or for convenience, about the underlying distribution of the manifest variable. One of these assumptions is that all effects in the underlying design model are uncorrelated. This assumption is weaker than the assumption of independence used in typical ANOVA applications. In generalizability theory, most effects will be uncorrelated because of the random sampling assumptions and the definition of the linear model used (Brennan, 1983). Other effects simply are assumed to be uncorrelated.

For example, in the single-facet crossed design (often referred to as a $p \times i$ or "person by item" design), it is assumed that facet (i) and person (p) effects are uncorrelated, and that the residual effect is uncorrelated across both persons and levels of facet. If correlated effects exist in the data, the practitioner should be aware of how this might affect the resulting estimates of variance components and their standard errors.

Correlated Effects and Estimates of Variance Components

The impact of correlated effects on mean

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 3, September 1992, pp. 229-235

0146-6216/92/030229-07$1.60

square (MS) and estimates of variance components has attracted only limited attention in the past. None of this work has been done in the context of generalizability theory. Box (1954) showed that for a two-way ANOVA design with error (residual) effects correlated across columns (facet levels), the residual MS is underestimated by a fixed amount, and the row (person) MS is overestimated by a similar amount. The greater the correlations between residuals, the greater the bias. Although Box's formulation was based on quadratic forms rather than variance components, the results are the same. Maxwell (1968) provided a discussion of the effects of correlated errors on classical reliability coefficients by applying Box's results to a typical measurement problem. This type of correlation also would result in an inflated generalizability coefficient.

The effects predicted by Box's work can be anticipated for measurement situations if the covariance structure of measurement data is considered. Following Brennan (1983), it can be shown that in the single-facet design,

$$\mathop{\mathrm{E}}_{i}[\mathrm{Cov}(X_{pi},X_{pi})] = \sigma^2_p + \sigma^2_{pi,e} \quad (1)$$

and that

$$\mathop{\mathrm{E}}_{i\neq i'}[\mathrm{Cov}(X_{pi},X_{pi'})] = \sigma^2_p \quad . \quad (2)$$

Thus, the expected covariances in the $p \times i$ design, over persons and for common levels of the facet, will be the sum of the person and residual variance components. Correspondingly, the expected covariance across persons, but for different facet levels, is the person variance component (Brennan, 1983). These components can be estimated by averaging the relevant covariances over persons for scores on facet levels derived from a set of $p \times i$ data.

Therefore, correlations within persons will reduce the average covariance in the diagonal of a variance-covariance matrix over persons, and also will inflate the off-diagonal covariances. Again, these results suggest an overestimation of the variance component for persons and an underestimation of the variance component for the residual.

It should be noted that if the correlated effects are uniform across levels of the facet, the covariance matrix over persons will not be affected because the addition of a constant does not influence the covariance structure of the data. This will be true if facet effects, rather than residual effects, are correlated across levels of the facet. However, correlated facet effects will deflate the diagonal covariances in a covariance matrix formed over facet levels, resulting in an underestimation of the variance component for the facet. Although this variance component does not appear in the equation for "relative" error variance [$\sigma^2(\delta)$], it is part of the equation for "absolute" error variance [$\sigma^2(\delta)$] (Cronbach et al., 1972, pp. 24–25) and for a variety of other generalizability coefficients, such as those outlined by Brennan and Kane (1977) for use with criterion-referenced tests.

If correlated effects exist in the universe, such that the correlation between all pairs of effects are fairly homogeneous or vary in some random fashion, it can be argued that they are inherent to the measurement context or domain and are, therefore, correctly accommodated in the generalizability (G) study estimates of variance components. Subsequent decision (D) study applications will correctly reflect these correlations as well. However, this will not be true if the pattern of correlated effects is neither homogeneous nor random. A nonhomogeneous pattern will occur if the facet or residual effects (or some combination) are serially correlated (i.e., that adjacent or near adjacent pairs of effects are correlated more highly than nonadjacent effects). For example, a covariance matrix of serially correlated residuals would have higher values near the diagonal (adjacent and close pairs), and the values would decrease away from the diagonal.

Because the bias associated with correlated effects is a function of the average correlation between effects, the nature of the serial correlation and the number of levels of the correlated facet in the design are critical. If a lag – 1 serial correlation between effects is present (i.e., the direct correlation occurs only for adjacent levels of the

facet under consideration), the bias introduced into the estimates of the variance components will be inversely proportional to the number of levels of the facet used in the G study design. This will be true regardless of whether the correlations originate within the facet effects or within the residuals. Subsequent use of these biased components in D study estimation may be misleading.

In many measurement settings, some form of serial correlation might occur among facet effects in the data. This is especially likely among time-segment or occasion facets. Examples include situations in which ratings are made on repeated trials of some fixed task, or situations in which observational data are obtained ordinally over relatively brief time intervals. Correlated residuals across levels of any facet also might be present [in this context, correlated residuals imply $\sigma(\mu_{pi}, \mu_{pi'}) \neq 0$ rather than $\sigma(\mu_p, \mu_{pi}) \neq 0$]. For example, a serial correlation could be present for each person across levels of a facet, but its value may not be homogeneous across persons; this would result in a correlated interaction term (residual in the $p \times i$ design) in the ANOVA model employed.

Recently, Rowley (1989) raised a similar concern related to generalizability coefficients derived from correlated data in "single subject" designs. Rowley argued that the presence of serial correlations in "single subject" data also biases the resulting generalizability coefficient by underestimating the residual variance component. Suen and Ary (1989) reached similar conclusions, but reported that preliminary work with correlated data showed only minor variations in the expected values of the derived generalizability coefficients. Suen and Lee (1987) suggested using a time series analysis to remove the correlation from the data prior to analysis.

A Simulation Study

Method

To illustrate the bias introduced into estimates of variance components due to serially correlated effects, a monte carlo study was conducted using the single-facet crossed $p \times i$ design. The simulation assumed that correlated effects are not accommodated in typical applications of generalizability theory. Rather, these effects are represented in the modeled parametric form of the score effects (i.e., independent components) in the population of interest, which allows the appropriate application of the ANOVA sums of squares and the variance component quadratic models to the data. The simulations implemented in this study represent only two of an infinite number of potential violations of the essentially independent distributions of the data.

The study was designed to demonstrate that under this characterization in the typical single-facet model,

$$\sigma_x^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2 \qquad (3)$$

and, therefore, higher-order models may not hold. In that case, a new model may be warranted, or methods must be developed to remove the correlation from the data prior to applying the model.

For the first simulation, a simple data-generation algorithm was used to model one particular form of dependence in the distribution of the scores x_{pi} that might result in correlated residuals across facet levels within each person. In this model, x_{pi} can be characterized as

$$x_{pi} = z_p + z_i + z_{pi} \quad , \qquad (4)$$

where z_α denotes a normally distributed standard score, distributed [0,1] on the interval $\{-\infty, +\infty\}$, for the model term corresponding to the subscript α. If no other transformation is applied to z_α, then, given Equation 3,

$$\sigma_x^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2 = 3.0 \quad . \qquad (5)$$

For purposes of the first simulation condition, a lag - 1 dependence among facet effects within persons was introduced as a form of serial correlation within the residual component $\sigma_{pi,e}^2$ of the model in Equation 3. To produce this dependence, the data generation algorithm in Equation 4 was adapted so that

$$x^*_{pi} = z_p + z_i + z^*_{pi} \quad , \tag{6}$$

where

$$z^*_1 = z_{p1} \tag{7}$$

and

$$z^*_{pi} = \rho_{i-1,i}(z^*_{i-1}) + (1 - \rho^2_{i-1,i})^{1/2} z_{pi} \quad , \tag{8}$$
$$i = 2, \ldots, I.$$

In Equation 8, ρ defines the correlation of adjacent error terms ($i - 1$ and i) in the rightmost component of Equation 6. Note that despite the introduction of the lag - 1 dependence for $\mu_{pi-1,e}$ and $\mu_{pi,e}$, the independent variance of the third component score $\sigma^2_{pi,e}$ remains equal to 1 by definition.

Under this characterization, serial correlations might be introduced at any level of Equation 4 and with respect to any subscript. The choice of placing the serial correlation on the $z_{pi,e}$ component, for adjacent levels of i, merely restricts the serial correlation within the levels of p.

For the second simulation, which examined serially correlated facet effects, the same variance structure of Equation 4 was retained, but the focus was shifted to the second score component subscripted exclusively by "i." The process of introducing lag - 1 dependence among facet effects becomes

$$x^*_{pi} = z_p + z^*_i + z_{pi} \tag{9}$$

given the lag - 1 dependence on z^*_1 as

$$z^*_1 = z_1 \tag{10}$$

and

$$z^*_i = \rho_{i-1,i}(z^*_{i-1}) + (1 - \rho^2_{i-1,i})^{1/2} z_i \quad , \tag{11}$$
$$i = 2, \ldots, I.$$

For purposes of this study, a very restrictive set of conditions was studied to illustrate the potential bias that serially correlated data contribute to estimates of variance components and D study inferences. Specifically, the design selected was studied under a variety of conditions—all combinations of both n_i (the number of levels of the facet) and n_p (the number of persons) were set at 10, 25, or 50. Therefore, the smallest design studied was a 10 × 10 design, and the largest was a 50 × 50 design. These values were selected because they represent sample sizes that previously were found to produce stable estimates of variance components (Smith, 1978). A limited number of design configurations in which $n_i < 10$ also were simulated to illustrate the potential influence of sample size on the estimation bias introduced by correlated effects. For all design conditions, simulated data were replicated for lag - 1 serial correlations of .2, .4, .6, and .8.

In the major portion of the simulation, all effects were drawn from a unit normal distribution; therefore, the population values of the variance components (σ^2_p, σ^2_i, and $\sigma^2_{pi,e}$) were each constrained to 1.0 as indicated above. To provide some indication of how the relative magnitudes of the population variance components moderate the estimation bias associated with serially correlated effects, a limited set of design configurations was simulated such that the population value of the residual component was 4.0.

The various combinations of n_p, n_i, and ρ resulted in 36 different design configurations in the major portion of the study. Additional design configurations were simulated to more clearly establish the effects of varying n_i and the relative magnitude of the residual variance component. Each design was simulated 1,000 times.

Results

Table 1, which summarizes the results of the simulations for the serially correlated facet effect and residual effect models, reveals predictable trends. In the case of serially correlated errors, the residual component was underestimated, and the person component was overestimated by a nearly equal amount. Together or separately, these biases will result in an overestimation of the computed generalizability coefficient for any D study. When facet effects were serially correlated, the residual and person components were unaffected; however, σ^2_i was underestimated as a

Table 1
Mean Variance Component Estimates for Various Combinations of n_p, n_i, and ρ Under Two Conditions of Correlated Effects

	Condition					
	Correlated Residual			Correlated Facet		
n_p, n_i, and ρ	Person	Facet	Residual	Person	Facet	Residual
10, 10, .2	1.048	.995	.939	1.041	.934	.991
10, 10, .4	1.151	1.019	.868	.998	.876	1.003
10, 10, .6	1.238	1.018	.757	.989	.831	.999
10, 10, .8	1.496	1.008	.505	.999	.540	.998
10, 25, .2	1.002	1.017	.982	1.019	.978	1.008
10, 25, .4	1.095	.987	.945	1.006	.957	.997
10, 25, .6	1.097	.986	.899	1.008	.896	1.001
10, 25, .8	1.303	1.012	.726	1.030	.697	1.000
10, 50, .2	.959	1.000	.990	.996	.975	1.001
10, 50, .4	1.016	1.003	.973	1.006	.976	1.000
10, 50, .6	1.034	1.008	.938	1.010	.929	.999
10, 50, .8	1.153	.999	.853	.989	.833	.995
25, 10, .2	1.031	1.008	.952	1.024	.965	1.006
25, 10, .4	1.138	1.019	.878	1.001	.841	1.006
25, 10, .6	1.265	.989	.753	1.004	.732	.999
25, 10, .8	1.470	1.010	.509	1.004	.474	.999
25, 25, .2	1.025	.997	.981	.979	.989	1.000
25, 25, .4	1.017	1.031	.957	1.012	.946	.998
25, 25, .6	1.104	1.014	.897	.996	.888	.999
25, 25, .8	1.289	1.006	.738	.988	.711	.995
25, 50, .2	1.011	.995	.989	.990	1.003	.999
25, 50, .4	1.036	.988	.976	.981	.970	1.000
25, 50, .6	1.052	.988	.946	1.010	.945	1.000
25, 50, .8	1.130	1.004	.853	1.016	.855	.997
50, 10, .2	1.052	1.026	.946	.987	.974	1.000
50, 10, .4	1.127	.966	.877	1.006	.896	1.000
50, 10, .6	1.258	.991	.756	.995	.772	.997
50, 10, .8	1.491	.981	.508	.982	.519	1.002
50, 25, .2	1.029	.993	.981	.988	1.005	.998
50, 25, .4	1.063	1.006	.950	.987	.949	1.001
50, 25, .6	1.102	1.015	.894	1.006	.886	1.000
50, 25, .8	1.277	.996	.736	.992	.726	.998
50, 50, .2	1.009	.996	.991	.987	.997	1.000
50, 50, .4	1.035	.990	.977	.995	.967	1.000
50, 50, .6	1.045	1.012	.947	1.003	.956	.999
50, 50, .8	1.134	.984	.860	.995	.845	1.000

function of both the serial correlation and n_i. Although this underestimation does not directly affect the typical generalizability coefficient and "relative" error term, it does affect the estimation of the "absolute" error term, universe score confidence intervals, and other statistics that make use of this component (e.g., index of dependability for mastery tests).

For designs that employ only a few levels of a facet, the bias introduced by serially correlated effects can be significant even for modest levels of correlation (ρ). Table 2 shows this bias for some selected conditions in which $n_i < 10$ (the information reported in Table 2 also is based on 1,000 replications of each design configuration). Note that in the typical case in which only a few observations are made on each person, the bias in the estimates of variance components can be

Table 2
Mean Variance Component Estimates for the Correlated Residuals Model for Values of $n_i < 10$, Where $\rho = .2$ and $n_p = .25$

	Variance Component		
n_i	Person	Facet	Residual
5	1.088	1.020	.907
4	1.107	1.013	.886
3	1.162	.895	.856
2	1.190	1.034	.780

quite substantial even for small values of serial correlation.

Note that the amount of bias introduced in the estimates of variance components for each condition is constrained by the relative size of the population values of the variance components represented. In the simulation study, the residual component was modeled to be comparable in magnitude to the person and facet component. In most applications of generalizability theory, the residual component is usually far larger than that for persons. Consequently, the practical effect of serial correlation can be expected to be even greater for real data. That is, the relative amount of variance redistributed to the person component from the residual component is proportional to the population value of the residual component for fixed values of n_p, n_i, and the serial correlation.

An illustration of this effect is shown in Table 3. The data in Table 3 are based on simulations for the correlated residual model in which the residual variance component was modeled to be 4.0 instead of 1.0 (again, the values in Table 3 are based on 1,000 replications of each design configuration). Note that the proportion of residual variance transferred to the person component estimate is nearly identical to that shown in Table 1; therefore, the bias introduced into the person component estimate is much greater because the residual component was initially larger.

Discussion

The results reported here indicate that serial correlation in measurement effects can have a marked influence on the impression of the dependability of the measurement data. Perhaps the most troubling aspect of these findings is that the bias introduced by the correlation is a function of the number of levels of the facet represented in the G study design and of the relative magnitude of the population values of the variance components. The applied implication of the former finding is not only that the estimates of variance components will be biased, but that the bias will decrease as the number of observations on each person increases. Consequently, when n_i is large, the effects will be minimal. Because of these dependencies, the normal D study estimates (in which n_i is varied and typically smaller than that used in the G study) will not be accurate because the variance components resulting from the G study are based on a different n_i than that employed in the D study. This, of course, provides still another strong case for large sample G study designs in which the bias is minimal.

The effects noted above are equally likely to be present in designs employing more than a single facet. Because economy of observations becomes an important concern in more complex designs, the biases present could be quite large. Consider, for example, a simple nested design with ratings on repeated trials nested within

Table 3
Mean Variance Component Estimates for Various Combinations of n_i and ρ, Where $\sigma^2_p/\sigma^2_{pi,e}$ and $\sigma^2_i/\sigma^2_{pi,e}$ Equal 4.0 ($n_p = 50$)

		Variance Component		
n_i	ρ	Person	Facet	Residual
10	.2	1.178	1.017	3.795
10	.4	1.484	.960	3.492
10	.6	1.999	1.032	3.002
10	.8	2.956	.987	2.045
25	.2	1.076	.987	3.925
25	.4	1.214	.998	3.803
25	.6	1.458	1.008	3.561
25	.8	2.053	.992	2.962
50	.2	1.044	1.010	3.964
50	.4	1.120	.995	3.904
50	.6	1.243	1.001	3.785
50	.8	1.609	.992	3.434

occasions for a particular observation schedule. In such studies, it would not be uncommon to include only a few occasions in the G study (certainly fewer than 10). In the presence of a serially correlated occasion effect, the variance component for occasions might be seriously underestimated and therefore leave the impression that the occasion of the observations has a minimal effect on the generalizability of the data.

References

Boardman, T. J. (1974). Confidence intervals for variance components: A comparative Monte Carlo study. *Biometrics, 30,* 251–262.

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 of correlation between errors in the two way classification. *Annals of Mathematical Statistics, 29,* 885–891.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City IA: 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.

Cronbach, L. J., & Gleser, G. C. (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement, 24,* 467–480.

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

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

Crump, S. L. (1946). The estimation of variance components in analysis of variance. *Biometrics, 2,* 7–11.

Eisenhart, C. (1947). The assumptions underlying analysis of variance. *Biometrics, 3,* 1–21.

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

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics, 9,* 226–253.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

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

Rowley, G. L. (1989). Assessing error in behavioral data: Problems of sequencing. *Journal of Educational Measurement, 26,* 273–284.

Searle, S. R. (1970). Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrics, 26,* 505–524.

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

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

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.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale NJ: Erlbaum.

Suen, H. K., & Lee, P. S. C. (1987, April). *Generalizability assessment of autocorrelated data via Box-Jenkins back forecasting.* Paper presented at the annual meeting of the American Educational Research Association, Chicago IL.

Acknowledgments

The authors are indebted to Robert Brennan for his many useful and helpful comments on earlier drafts of this manuscript. Likewise, the authors wish to thank the reviewers for their comments and suggestions during the final revisions to the manuscript.

Author's Address

Send requests for reprints or further information to Philip L. Smith, Department of Educational Psychology, Enderis Hall 777, P.O. Box 413, University of Wisconsin, Milwaukee WI 53217, U.S.A.