ACT Research Report Series 87-7

The Bootstrap and Other Procedures for Examining the Variability of Estimated Variance Components in Testing Contexts

Robert L. Brennan
Deborah J. Harris
Bradley A. Hanson

September 1987

ACT

THE BOOTSTRAP AND OTHER PROCEDURES FOR EXAMINING THE VARIABILITY OF ESTIMATED VARIANCE COMPONENTS IN TESTING CONTEXTS

Robert L. Brennan
Deborah J. Harris
Bradley A. Hanson

TABLE OF CONTENTS

Page

ABSTRACT.....iii
THE p x i RANDOM EFFECTS DESIGN AND
ASSOCIATED VARIANCE COMPONENTS.....2

STANDARD ERRORS AND CONFIDENCE INTERVALS
FOR VARIANCE COMPONENTS.....5
Traditional Approach.....6
Bootstrap--General Issues.....7
Bootstrap with the p x i Random Effects Design.....9
Jackknife--General Issues.....11

SIMULATION RESULTS FOR NORMALLY DISTRIBUTED DATA.....12
Data Generation.....13
Standard Errors.....14
Confidence Intervals.....16
Discussion.....18

SIMULATION RESULTS FOR BINARY DATA.....19
Bootstrap Sampling Procedures.....19
Standard Errors and Confidence Intervals.....21

SUMMARY AND CONCLUSIONS.....24

REFERENCES.....28

APPENDIX A--Estimated Standard Errors and Satterthwaite's
Confidence Intervals for Variance Components.....30

APPENDIX B--Jackknife Estimates, Their Standard Errors and
Confidence Intervals for the p x i Random Effects Design.....32

APPENDIX C--Tables Illustrating Results of Different
Procedures for Estimating Variance Components and
Their Standard Errors for Normally Distributed Data.....35

FOOTNOTES.....46

TABLES.....47

ABSTRACT

This paper examines the applicability of traditional, bootstrap, and jackknife methodologies for estimating standard errors and obtaining confidence intervals for the variance components for persons, items, and residuals in a random effects G study p x i design. Principal consideration is given to simulation results with binary data, although some simulation results for normally distributed data are also reported. The simulations suggest that the traditional approach produces accurate results with normally distributed data but poor results with binary data, at least for the variance component for residuals. The jackknife provides quite accurate results for both types of data and for all three variance components. The bootstrap can be "made to work" reasonably well but doing so seems to require several ad hoc procedures for defining bootstrap samples, which renders the bootstrap somewhat less satisfactory than the jackknife for the application considered here.

THE BOOTSTRAP AND OTHER PROCEDURES FOR EXAMINING THE VARIABILITY OF ESTIMATED VARIANCE COMPONENTS IN TESTING CONTEXTS

The Standards for Educational and Psychological Testing (APA, 1985) state that

> . . . the estimation of clearly labeled components of observed and error score variance is a particularly useful outcome of a reliability study, both for the test developer who wishes to improve the reliability of an instrument and for the user who wants to interpret test scores in particular circumstances with maximum understanding. Reporting standard errors, confidence intervals, or other measures of imprecision of estimates is also helpful. (p. 19)

The principal purpose of this paper is to examine the applicability of several methodologies for estimating standard errors and obtaining confidence intervals for variance components in testing contexts. Using the terminology of generalizability theory, the specific context considered here can be characterized as a G study with a random effects p x i design, in which all examinees respond to the same set of undifferentiated items.[1] This design can be used to estimate three basic variance components--one for persons, one for items, and one for residuals.

The variability of estimates of these three variance components is examined using traditional, bootstrap, and jackknife methodologies, with principal focus on various bootstrap approaches. Each of these methodologies is described in a subsequent section, which is followed by a discussion of simulation results for normally distributed data and binary data. Binary data simulations are considered more extensively, because binary data are more common in testing contexts.

The p x i Random Effects Design and Associated Variance Components

In generalizability theory (Cronbach, Gleser, Nanda, & Rajartnam, 1972, and Brennan, 1983) one begins by specifying a universe of conditions of measurement and a population of objects of measurement. Actually at least two universes can be specified, a universe of admissible observations and a universe of generalization. Here the universe under consideration is a universe of admissible observations consisting of $\underline{K}$ items, and the population consists of $\underline{N}$ persons. Usually, it is assumed that $K \to \infty$ and $N \to \infty$. This assumption is made below, unless otherwise noted.

Let X_{pi} denote the observed score for any person in the population on any item in the universe. The expected value over items of a person's observed score is

$$\mu_p = \underset{i}{E}\, X_{pi} \; . \tag{1}$$

The score μ_p can be conceptualized as the examinee's "mean" score over the universe of items. Similarly, the population "mean" for item $\underline{i}$ is

$$\mu_i = \underset{p}{E}\, X_{pi} \; , \tag{2}$$

and the "mean" over both the population and universe is

$$\mu = \underset{p}{E}\, \underset{i}{E}\, X_{pi} . \tag{3}$$

Although these mean scores are not themselves observable, an observable score can be expressed in terms of them using the following tautology:

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (X_{pi} - \mu_p - \mu_i + \mu) \tag{4}$$

or, in abbreviated form,

$$X_{pi} = \mu + \pi_p + \beta_i + \pi\beta_{pi}. \tag{5}$$

In Equation 5, $\pi_p = \mu_p - \mu$ represents the effect for person p, $\beta_i = \mu_i - \mu$ represents the effect for item i, and $\pi\beta_{pi} = X_{pi} - \mu_p - \mu_i + \mu$ represents the residual effect for person p and item i. Since there is only one observation for each person-item combination, interaction effects and other sources of random error are confounded in the residual effects. The manner in which the effects in Equation 5 have been defined implies that

$$\underset{p}{E}\, \pi_p = \underset{i}{E}\, \beta_i = \underset{p}{E}\, \pi\beta_{pi} = \underset{i}{E}\, \pi\beta_{pi} = 0 . \tag{6}$$

Also, most of the effects are necessarily uncorrelated by the manner in which they have been defined; others are assumed to be uncorrelated. (See Brennan, 1983, pp. 9-10 for more detail.)

For each effect, there is an associated variance, called a variance component. For example, the variance component for persons is

$$\sigma^2(p) = \underset{p}{E}(\mu_p - \mu)^2 = \underset{p}{E}\, \pi_p^2 . \tag{7}$$

Similarly,

$$\sigma^2(i) = \underset{i}{E}(\mu_i - \mu)^2 = \underset{i}{E}\, \beta_i^2 \quad \text{and} \tag{8}$$

$$\sigma^2(pi) = \underset{p}{E}\,\underset{i}{E}(X_{pi} - \mu_p - \mu_i + \mu)^2 = \underset{p}{E}\,\underset{i}{E}\,\pi\beta^2_{pi} \ . \tag{9}$$

These variance components can be estimated using the data that result from administering a random sample of k items from the universe to an independent random sample of n persons from the population. This design is denoted p x i and called a (G study) random effects design because it is associated with a random sampling process for both persons and items. Using this design, the usual estimators of the variance components (see Brennan, 1983) are:

$$\hat{\sigma}^2(p) = [MS(p) - (1-k/K)MS(pi)]/k \ , \tag{10}$$

$$\hat{\sigma}^2(i) = [MS(i) - (1-n/N)MS(pi)]/n \ , \quad \text{and} \tag{11}$$

$$\hat{\sigma}^2(pi) = MS(pi) \ , \tag{12}$$

where "MS" means "mean square."

Equations 10, 11, and 12 include the finite population and universe correction factors $1 - n/N$ and $1 - k/K$, respectively. When $N \to \infty$ and $K \to \infty$ they are both 1, which is the usual assumption. However, for the simulation studies considered later that involve binary data, $N < \infty$ and $K < \infty$ which necessitates using the finite correction factors. Also, in this case, all expectations in the above development of the p x i design should be replaced by the analogous summation notation. For example, μ_p in Equation 1 becomes the mean score over the finite universe of items for person p:

$$\mu_p = \sum_{i=1}^{K} X_{pi}/K \ . \tag{13}$$

Also, the grand mean in Equation 3 becomes

$$\mu = \sum_{p=1}^{N} \sum_{i=1}^{K} X_{pi}/NK , \tag{14}$$

and using Equations 13 and 14, the variance component for persons in Equation 7 becomes

$$\sigma^2(p) = \sum_{p=1}^{N} (\mu_p - \mu)^2/(N - 1). \tag{15}$$

Similarly, letting μ_i be the mean score over the finite population of persons for item i, the variance component for items is:

$$\sigma^2(i) = \sum_{i=1}^{K} (\mu_i - \mu)^2/(K - 1) \tag{16}$$

Finally, for the residuals,

$$\sigma^2(pi) = \sum_{p}^{N} \sum_{i}^{K} \frac{(X_{pi} - \mu_p - \mu_i + \mu)^2}{(N - 1)(K - 1)} . \tag{17}$$

The use of the divisors N-1 and/or K-1 rather than N and/or K in Equations 15-17 is based on the Cornfield and Tukey (1956) definitions of variance components, which are the usual definitions in generalizability theory (see Brennan, 1983, pp. 48-50).

Standard Errors and Confidence Intervals for Variance Components

Assuming that the residual effects, $\pi\beta_{pi}$, are normally and independently distributed, Searle (1971, pp. 408-419) shows that

$$\hat{\sigma}^2(pi) \sim \sigma^2(pi)\ \chi^2(df_{pi})/df_{pi} \quad \text{and}$$

$$\sigma[\hat{\sigma}^2(pi)] = \{2[\sigma^2(pi)]^2/df_{pi}\}^{\frac{1}{2}} , \tag{18}$$

where $df_{pi} = (n-1)(k-1)$ is the degrees of freedom for $MS(pi) = \hat{\sigma}^2(pi)$.

For $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(i)$ the distributions are unknown even under normality assumptions. However, under normality assumptions it can be shown that the standard errors are:

$$\sigma[\hat{\sigma}^2(p)] = \{2[\sigma^2(pi) + k\sigma^2(p)]^2/df_p + 2[\sigma^2(pi)]^2/df_{pi}\}^{\frac{1}{2}} , \text{ and} \tag{19}$$

$$\sigma[\hat{\sigma}^2(i)] = \{2[\sigma^2(pi) + n\sigma^2(i)]^2/df_i + 2[\sigma^2(pi)]^2/df_{pi}\}^{\frac{1}{2}} , \tag{20}$$

where $df_p = n - 1$ and $df_i = k - 1$ are the degrees of freedom for MS(p) and MS(i), respectively (see Searle, 1971).

Even under normality assumptions, of course, the standard errors formulas in Equations 18, 19, and 20 cannot be used directly unless the variance components themselves are known. With real data, variance components are unknown, the assumption of normally distributed score effects is often unreasonable and, therefore, the standard errors of estimated variance components are unknown.

Considered below, in general terms, are three approaches to estimating standard errors of estimated variance components and to obtaining confidence intervals for variance components.

Traditional Approach

Assuming mean squares are independent and score effects have a multivariate normal distribution, Appendix A provides a general formula for the estimated standard error of any estimated variance component. For the three estimated variance components of interest here, the resulting estimators of the standard errors are:

$$\hat{\sigma}[\hat{\sigma}^2(p)] = \sqrt{\frac{2[(MS(p)]^2}{k^2(n-1)} + \frac{2[(1-k/K)MS(pi)]^2}{k^2(n-1)(k-1)}} , \tag{21}$$

$$\hat{\sigma}[\hat{\sigma}^2(i)] = \sqrt{\frac{2[MS(i)]^2}{n^2(k-1)} + \frac{2[(1-n/N)MS(pi)]^2}{n^2(n-1)(k-1)}} , \text{ and} \tag{22}$$

$$\hat{\sigma}[\hat{\sigma}^2(pi)] = \sqrt{\frac{2[MS(pi)]^2}{(n-1)(k-1)}} , \tag{23}$$

As noted above, the distributions of $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(i)$ are unknown, even under normality assumptions. Under normality assumptions, however, Satterthwaite (1941, 1946) proposed a procedure for obtaining <u>approximate</u> confidence intervals. His procedure is described in Appendix A.

<u>Bootstrap--General Issues</u>

The bootstrap is a general methodology for assessing how accurate a particular $\hat{\theta}$ is as an estimate of θ . (See Efron, 1982, for a comprehensive theoretical treatment and Efron & Tibshirani, 1986, for a simpler and more applied treatment.) The bootstrap substitutes considerable amounts of computation for traditional, theoretical analysis. In doing so, often the bootstrap is able to deal with issues that are far too complicated for traditional statistical analyses. Furthermore, the bootstrap need not (and usually does not) involve any assumptions about distributional form. In this sense, it is (usually) a completely nonparametric procedure.

From a univariate sampling perspective, the bootstrap can be described in the following manner. Let $X_1, X_2, \ldots, X_s$ be independently and identically distributed as F, and let $\hat{\theta}(X_1, X_2, \ldots, X_s)$ be some statistic of interest. Also let the standard deviation of the sampling distribution of $\hat{\theta}$ be denoted $\sigma = \sigma(F; \hat{\theta}, s)$. Then, the bootstrap estimate of the standard error of $\hat{\theta}$ is

$$\hat{\sigma} = \sigma(\hat{F}; \hat{\theta}, s) , \tag{24}$$

where $\hat{F}$ is the empirical distribution putting equal probability mass on each of the s observed data points. That is, $\hat{F}$ is simply the observed set of s data points. Using a simple Monte Carlo algorithm, Equation 24 can be evaluated even without knowing the form of the expression for $\sigma(F; \hat{\theta}, s)$.

The algorithm is based on the results of multiple bootstrap samples, where each bootstrap sample consists of a random sample of size s drawn with replacement from the actual sample, $\hat{F}$. The three steps in the algorithm are: (i) using a random number generator, independently draw a large number of bootstrap samples, say B of them; (ii) for each sample evaluate the statistic of interest, say $\hat{\theta}_b$ (b = 1, 2, ..., B); and (iii) calculate the sample standard deviation of the $\hat{\theta}_b$:

$$\hat{\sigma}(\hat{\theta}_b) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b - \hat{\theta}_B)^2} \tag{25}$$

where

$$\hat{\theta}_B = \sum_{b=1}^{B} \hat{\theta}_b / B \tag{26}$$

will be called here the "bootstrap estimate" of $\hat{\theta}$. As $B \to \infty$, $\hat{\sigma}(\hat{\theta}_b)$ will approach $\hat{\sigma}$ in Equation 24, the bootstrap estimate of the standard error of $\hat{\theta}$. For estimating standard errors, B in the range of 50 to 200 is quite adequate according to Efron and Tibshirani (1986, p. 56).

The bootstrap can also be used to produce approximate confidence intervals. For example, an 80% approximate confidence interval for θ can be defined as the 10th and 90th percentile points of the distribution of the $\hat{\theta}_b$. For confidence intervals, however, the computational requirements are more substantial. Usually, one wants $B \geq 1000$ bootstrap samples (see Efron & Tibshirani, 1986, p. 67).

Bootstrap with the p x i Random Effects Design

Although the bootstrap is conceptually simple in univariate situations, it will become evident subsequently that it is unclear how to extend it to estimates of variance components generated from the random effects p x i design. The crux of the matter is to specify how to draw a bootstrap sample from the n x k matrix of observed scores. It might seem that the obvious way to do so is: (i) draw a random sample of n persons with replacement from the sampled persons; (ii) draw an independent random sample of k items with replacement from the sampled items; and (iii) let the bootstrap sample consist of the responses of the sampled persons to the sampled items. This double sampling procedure will be denoted "boot-p,i."

The boot-p,i procedure seems obvious in that it is similar to the random sampling process that generates the observed n x k data matrix. However, it is important to note that the boot-p,i procedure involves sampling with replacement from the observed data. This means that, except when the bootstrap sample is the observed sample, the bootstrap sample matrix will contain some repeated persons and some repeated items. Now, it is possible to show that $\hat{\sigma}^2(p)$ in Equation 10 is a function of the item covariances, and $\hat{\sigma}^2(i)$ in Equation 11 is a function of the person covariances. For example, for $K \to \infty$ an expression equivalent to Equation 10 for $\hat{\sigma}^2(p)$ is:

$$\hat{\sigma}^2(p) = \frac{1}{k(k-1)} \left\{ \sum_{i \neq i'} \sum \left[\sum_{p} \frac{(X_{pi} - \bar{X}_i)(X_{pi'} - \bar{X}_{i'})}{n-1} \right] \right\}, \tag{27}$$

which is the average of the unbiased estimates of the item covariances. When items are repeated, Equation 27 suggests that $\hat{\sigma}^2(p)$ is likely to be an inflated estimate of $\sigma^2(p)$, especially when k is relatively small. A similar statement holds for $\hat{\sigma}^2(i)$. The consequences of the boot-p,i procedure for $\hat{\sigma}^2(pi)$ and the standard errors of each of the estimated variance

components are not immediately obvious, but they will be illustrated in the simulation studies.

Because the boot-p,i procedure seems suspect, three other procedures for obtaining bootstrap samples are considered in the simulation studies discussed later.

The "boot-p,i,r" procedure involves random sampling with replacement for both persons and items as in the boot-p,i procedure plus random sampling with replacement from the nk residuals of the form

$$e_\ell = X_{pi} - \bar{X}_p - \bar{X}_i + \bar{X} \quad (\ell = 1, 2, \ldots, nk).$$

Specifically, suppose the independently sampled person, item, and residual are denoted p*, i*, and ℓ*, respectively. Then, the data element for person p* and item i* in the bootstrap sample matrix is

$$X_{p*i*} = \bar{X} + (\bar{X}_{p*} - \bar{X}) + (\bar{X}_{i*} - \bar{X}) + e_{\ell*} \quad .$$

The boot-p,i,r procedure was considered because it would appear to circumvent the kind of problem indicated above for the boot-p,i procedure.

The other two procedures considered for obtaining a bootstrap sample involve sampling only one dimension. The "boot-p" procedure involves sampling n persons with replacement, but not items. The "boot-i" procedure involves sampling k items with replacement, but not persons. The boot-p and boot-i procedures keep items and persons fixed, respectively, in obtaining bootstrap samples. Since results are wanted for the situation in which both persons and items are random, one would expect that neither of these procedures would be completely satisfactory for all variance components. However, it was

hypothesized that these procedures might provide some useful results or insights.

Jackknife--General Issues

Quenouille (1949) invented a nonparametric estimator of bias, subsequently called the jackknife, although the term "jackknife" is usually associated with Tukey, probably because of his extension of Quenouille's idea to a nonparametric estimator of the standard error of a statistic (Tukey, 1958). An often-referenced overview of the jackknife is given by Mosteller and Tukey (1968) who also discuss how to use the jackknife in obtaining confidence intervals.

Suppose θ is some parameter of interest and one obtains a set of s data points to estimate θ. In general terms, the steps involved in using the jackknife are as follows: (i) obtain $\hat{\theta}$ for all s data points; (ii) obtain the s estimates of θ that result from eliminating each one of the data points, and let each such estimate be designated $\hat{\theta}_{-j}$; (iii) obtain the s "pseudo-values" $\hat{\theta}_{*j} = \hat{\theta} + (s-1)(\hat{\theta} - \hat{\theta}_{-j})$; (iv) obtain the jackknife estimator of θ which is the mean of the pseudo-values, $\hat{\theta}_J$; (v) obtain the estimate of the standard error of the jackknife estimate of θ :

$$\hat{\sigma}(\hat{\theta}_J) = \{\sum_{j=1}^{s} (\hat{\theta}_{*j} - \hat{\theta}_J)^2/[s(s-1)]\}^{\frac{1}{2}} ,$$

which is simply the sample standard deviation of the pseudo-values divided by $\sqrt{s}$; and (vi), if desired, obtain the jackknife 100(1 - α) percent confidence interval for θ :

$$\hat{\theta}_J - t\ \hat{\sigma}(\hat{\theta}_J) \leq \theta \leq \hat{\theta}_J + t\ \hat{\sigma}(\hat{\theta}_J) ,$$

where t is the (1 - α/2) percentage point of Student's t distribution with

s - 1 degrees of freedom. The extension of the jackknife to estimated variance components resulting from data for a random effects p x i design is discussed in Appendix B. The basic steps are those outlined above, but several of the steps are somewhat more complicated conceptually and computationally.

Efron (1982) considers in some detail similarities and dissimilarities between the bootstrap and the jackknife. Both are based on resampling models and are primarily nonparametric procedures. As such, they are quite flexible and have considerable appeal in complicated contexts such as estimating the variability of estimated variance components with a crossed design. In this context, however, an apparently important difference between the bootstrap and the jackknife is that the bootstrap involves sampling with replacement while the jackknife does not. (Also, to establish confidence intervals with the jackknife requires an assumption about the distributional form of jackknife estimates, whereas approximate bootstrap intervals require no comparable assumption.)

Simulation Results for Normally Distributed Data

The principal, intended focus of this paper is on bootstrap procedures with binary data. However, since binary data often introduce added complexities in psychometric analyses, and since it is not immediately obvious how to extend the bootstrap to a random effects p x i design, it seemed prudent to consider normal data, first, in conjunction with traditional, bootstrap, and jackknife approaches. The advantage of considering normally distributed data is that at least some of the properties of the distributions of estimated variance components are known. The principal disadvantage is that the assumption of normally distributed data is often unrealistic for the p x i design in generalizability theory applications. However, if any one of

the approaches considered here does not work with normal data, then it seems highly unlikely that such an approach would have much general utility.

Data Generation

Each element in an n x k data matrix was generated using the following formula:

$$X_{pi} = \mu + \sigma(p)z_p + \sigma(i)z_i + \sigma(pi)z_{pi} , \tag{28}$$

where μ, $\sigma(p)$, $\sigma(i)$, and $\sigma(pi)$ are prespecified parameter values and z_p, z_i, and z_{pi} are randomly and independently sampled values from a unit normal distribution. The z-scores were obtained using the IMSL (1984) subroutine GGNML. Since the X_{pi} are the sum of normally distributed variables (plus a constant, μ), it follows that the X_{pi} themselves are normally distributed with mean μ and variance $\sigma^2(p) + \sigma^2(i) + \sigma^2(pi)$. Implicit in this data generation procedure is the assumption that $N \to \infty$ and $K \to \infty$.

The simulation results in Table 1 are for $\sigma^2(p) = 4$, $\sigma^2(i) = 16$, and $\sigma^2(pi) = 64$ with n = 200 persons and k = 20 items. (The parameter μ was set to 50, but since the focus here is on variance components the value of μ is irrelevant.) There is always some subjectivity involved in choosing parameter values and sample sizes for a simulation study. The rationale for the choices made here was as follows. First, the magnitudes of the variance components were chosen to be relatively large in order to highlight relatively small differences between comparable statistics. Second, $\sigma^2(pi)$ was chosen to be considerably larger than either of the other two variance components because, in generalizability theory, this is almost always the case with real data. Third, $\sigma^2(p)$ and $\sigma^2(i)$ were chosen to be substantially different because this too is a common occurrence in generalizability theory. Fourth, n was chosen to be quite large because, with most testing programs that the authors encounter, there are a large number of person records available for

analysis. Finally, k was chosen to be rather small because a principal focus of this research was to examine the potential applicability of the bootstrap when relatively small numbers of items are associated with content categories in a table of specifications.

Insert Table 1 about here

Standard Errors

With normally distributed data and known parameter values, Equations 18, 19, and 20 can be used to obtain exact standard errors for $\hat{\sigma}^2(pi)$, $\hat{\sigma}^2(p)$, and $\hat{\sigma}^2(i)$, respectively. With $\sigma^2(p) = 4$, $\sigma^2(i) = 16$, $\sigma^2(pi) = 64$, $n = 200$, and $k = 20$, these standard errors are reported in the first row of Table 1. They are the target values for evaluating the results of the various approaches to estimating standard errors. The rest of the table provides results for one sample (trial) of size $n = 200$ persons and $k = 20$ items.

The row labeled "traditional" provides the estimated variance components (using Equations 10-12) and the estimated standard errors (using Equations 21-23) for the sample. The estimates are all quite close to the corresponding parameters for this particular sample. The largest differences are for items (which is to be expected since k is only 20), but even these differences are relatively small. For example, $\hat{\sigma}^2(i) = 13.66$ with $\sigma^2(i) = 16$, but since the exact standard error is 5.29, this difference does not seem very dramatic. Also, the estimated standard error of $\hat{\sigma}^2(i)$, namely 4.53, is lower than the parameter value 5.29 primarily because $\hat{\sigma}^2(i)$ is less than $\sigma^2(i)$. It is not unexpected that the traditional results are reasonably good approximations of the parameters because, under the normality conditions that generated the data, the estimated variance components are unbiased estimates and the squares

of the estimated standard errors are nearly unbiased estimates (see Appendix A).

The next four lines in Table 1 provide bootstrap results for the four ways of obtaining a bootstrap sample that were outlined previously. For all results, B = 1000 bootstrap samples were employed. This number is larger than required for estimating standard errors. However, as discussed later, these bootstrap samples were used to get approximate confidence intervals, too.

The boot-p,i results are not very accurate. In particular, the bootstrap estimate of the variance component for persons, $\hat{\sigma}^2_B(p)$, is much too large, as are the estimated standard errors of the estimated variance components for persons and residuals, $\hat{\sigma}[\hat{\sigma}^2_b(p)]$ and $\hat{\sigma}[\hat{\sigma}^2_b(pi)]$, respectively. Also, $\hat{\sigma}^2_B(pi)$ is too small. The results for boot-p,i,r are only marginally better, primarily because $\hat{\sigma}\ [\hat{\sigma}^2_b(pi)]$ is quite close to its parameter value.

For the boot-p procedure, the results in boldface for persons and residuals are quite accurate, but the estimated standard error for items, $\hat{\sigma}[\hat{\sigma}^2_b(i)]$, is much too low. For the boot-i procedure, the results in boldface for items are quite accurate [although $\hat{\sigma}^2_B(i)$ is a little low], but the results for persons and residuals are less accurate.

In short, the boot-p,i and boot-p,i,r procedures do not produce very accurate results for these data, the boot-p procedure produces accurate results for persons and residuals but not items, and the boot-i procedure produces accurate results for items but not persons or residuals. This summary is something of an oversimplification, primarily because boot-i, boot-p,i, and boot-p,i,r provide reasonably comparable results for items. However, the results in Table 1 provide no compelling reason for preferring the more complicated boot-p,i or boot-p,i,r procedures over the simpler and computationally quicker boot-p or boot-i procedures--whether one considers persons, items, or residuals.

It is notable that the boldfaced bootstrap estimates of variance components (3.90, 13.08, and 63.45) are all less than the corresponding unbiased estimates in the traditional row (3.93, 13.66, and 63.74). From these results (and numerous others not reported here) it appears that, as $B \to \infty$: (i) for boot-p, $n/(n-1)\hat{\sigma}^2_B(p)$ and $n/(n-1)\ \hat{\sigma}^2_B(pi)$ are nearly unbiased estimates; and (ii) for boot-i, $k/(k-1)\ \hat{\sigma}^2_B(i)$ and $k/(k-1)\ \hat{\sigma}^2_B(pi)$ are nearly unbiased estimates. With $n = 200$ and $k = 20$, $n/(n-1) = 1.00503$ and $k/(k-1) = 1.05263$, and when these correction factors are applied to the boldfaced estimated variance components and standard errors in Table 1, the results are those in the row labeled "Boot^a." The results in this row are taken as <u>the</u> bootstrap results, in the sense of the "best" results that were obtained using the bootstrap in this situation. They appear quite good relative to the traditional results and the parameter values.

Finally, the last row in Table 1 provides the jackknife results. All things considered, they appear to be at least as accurate as the traditional or bootstrap results. For the jackknife, the estimated standard errors for persons and residuals are a little low, but the estimated standard error for items is considerably closer to the parameter value than is the traditional or bootstrap result.

Confidence Intervals

In principal, one can evaluate standard errors without reference to confidence intervals but, in practice, frequently when a standard error is calculated, it is used (explicitly or implicitly) to establish an interval of some kind. Therefore, it seems highly desirable that the traditional, bootstrap, and jackknife approaches be evaluated, in part, with respect to the approximate confidence intervals that can be obtained using them.

Based on the same data and parameter values used to generate the results in Table 1, Table 2 provides approximate 80% confidence intervals for the variance components.[2] Since the distributions of $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(i)$ are unknown, even under normality assumptions, a simulation study was conducted to obtain empirical sampling distributions for estimates of each of the variance components. This study involved 2000 random samples (or trials) of size $n = 200$ and $k = 20$. The observed data for each trial were generated using Equation 28 with $\sigma^2(p) = 4$, $\sigma^2(i) = 16$, and $\sigma^2(pi) = 64$. The 10th and 90th percentile points of the three distributions are reported in Table 2. They can be used as approximate target values for evaluating the confidence intervals from the Satterthwaite, bootstrap, and jackknife procedures.

Insert Table 2 about here

The Satterthwaite results are virtually identical to the simulation results, as might be expected since Satterthwaite's procedure assumes normally distributed data. The bootstrap and jackknife results are quite similar and reasonably close to the target values. The most discrepant results are for $\sigma^2(i)$, where the limits of the bootstrap and jackknife confidence intervals are a little too low. As discussed in Appendix B, the jackknife can be employed also using logarithms of estimated variance components. For this simulation, however, the last row in Table 2 suggests no advantage to using logs. Indeed, using logs, the upper limit of the interval for $\sigma^2(i)$ is too high.

Technically, the above manner of evaluating procedures for establishing confidence intervals is <u>ad hoc</u>. In principal, such procedures should be evaluated in terms of the proportion of intervals that cover the parameter of

interest. This was done with Satterthwaite's procedure for establishing 80% confidence intervals. To do so, 1000 random samples (trials) of size n = 200 and k = 20 were generated using Equation 28 with $\sigma^2(p) = 4$, $\sigma^2(i) = 16$, and $\sigma^2(pi) = 64$. The resulting proportions of intervals that covered $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$ were 81.7, 81.5, and 79.7, respectively. Apparently, for the conditions of this simulation, the Satterthwaite intervals for $\sigma^2(p)$ and $\sigma^2(i)$ tend to be a little bit too broad, but not by much.

Coverage simulations were not undertaken for the bootstrap and jackknife procedures with normal data because such simulations are quite costly. As discussed later, however, more extensive coverage simulations were conducted with binary data, which are of greater interest in this paper.

Discussion

The normal data results presented above are somewhat limited in their generalizability for two principal reasons.

First, except for Satterthwaite confidence intervals, the basic results are for one trial (i.e., one random sample of size n = 200 and k = 20 with prespecified values for variance components). Results for a small number of trials not reported here tend to confirm the results discussed above, but a "complete" simulation study would involve a systematic analysis from a large number of trials. (This is done with the binary data discussed later.)

Second, the results presented above all use $\sigma^2(p) = 4$, $\sigma^2(i) = 16$, and $\sigma^2(pi) = 64$ with n = 200 and k = 20. Conceivably, conclusions might differ for different patterns of variance components and/or sample sizes. To address this issue, at least in part, Appendix C provides one-trial results for standard errors (B = 100) for each of three sets of variance components and each of three pairs of sample sizes.

The results presented above and those in Appendix C suggest that, with normal data: (a) Satterthwaite's procedure produces very accurate results;

(b) jackknifing produces quite accurate results, but using logs seems unnecessary or even ill-advised; and (c) bootstrapping can be "made to work". Specifically, these results suggest using boot-p for persons, boot-i for items, and boot-p for residuals if $n \geq k$ or boot-i for residuals if $k \geq n$. With normal data, however, there appears to be no need to employ anything other than Satterthwaite's procedure, which is by far the simplest and quickest to use.

Simulation Results for Binary Data

To examine traditional, bootstrap, and jackknife approaches with binary data, a population of persons and a universe of items were used that consisted of N = 2000 persons and K = 200 items. (These persons are a subset of those in a data base for a large licensure testing program.) For this finite population and universe, $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$ were obtained using Equations 15-17. The results are reported in the last row of Table 3 labeled "Parameters," along with standard errors of estimated variance components for n = 200 and k = 20. These (approximate) standard errors are the standard deviations of the distributions of $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(pi)$ resulting from 2000 random samples of size n = 200 and k = 20 from the finite population and universe.

Insert Table 3 about here

Bootstrap Sampling Procedures

The previous results with normal data and $n > k$ suggest using boot-p with persons and residuals and boot-i with items. However, as discussed below, an alternative procedure works better with binary data.

To examine different procedures for creating bootstrap samples with binary data, a simulation study was conducted that employed the boot-p,

boot-i, and boot-p,i procedures in estimating variance components and their standard errors. This study involved taking 100 random samples (trials) of size n = 200 and k = 20 from the finite population and universe, with B = 100 bootstrap samples per trial. (Note that each trial involved independent random sampling <u>without</u> replacement.) The results are summarized in Table 3.

Consider, for example, the boot-p results for persons. Since there were 100 trials, there were 100 values of $\hat{\sigma}^2_B(p)$, with each value being the average $\hat{\sigma}^2_b(p)$ for 100 bootstrap samples. The mean, over the 100 trials, of the $\hat{\sigma}^2_B(p)$ was .0069; and the standard deviation, over the 100 trials, of the $\hat{\sigma}^2_B(p)$ was .0017. Similarly, there were 100 values of $\hat{\sigma}[\hat{\sigma}^2_b(p)]$ with each value being the standard deviation of the $\hat{\sigma}^2_b(p)$ for 100 bootstrap samples. The mean, over 100 trials, of the $\hat{\sigma}[\hat{\sigma}^2_b(p)]$ was .0016; and the standard deviation, over the 100 trials, was .0003. The other entries in Table 3 are interpretable in a similar manner.

Note that with boot-p the bootstrap estimates of $\sigma^2(p)$ and $\sigma^2(pi)$ that are reported in Table 3 are those resulting from multiplying the computed estimates by the correction factor n/(n-1) = 200/199 = 1.00503. Similarly, for boot-i the correction factor k/(k-1) = 20/19 = 1.05263 was applied to the computed bootstrap estimates of $\sigma^2(i)$ and $\sigma^2(pi)$. Finally, with boot-p,i the correction factor [n/(n-1)][(k/(k-1)] = 1.05792 was applied to the computed bootstrap estimates of $\sigma^2(pi)$.

All things considered Table 3 suggests that boot-p is preferable for the estimated variance component and standard error for persons, boot-i is preferable for the estimated variance component and standard error for items, and boot-p,i and boot-i provide quite accurate results for the residuals. It is particularly noticeable that the mean of the 100 trial values of $\hat{\sigma}[\hat{\sigma}^2_b(pi)]$ using boot-p is much too small, which is <u>not</u> the case with normal data. Apparently, the nature of the underlying data has considerable

influence on the "best" way to obtain bootstrap samples with the random effects p x i design.

Although Table 3 indicates that boot-p is preferable to boot-i and boot-p,i for estimating the standard error of $\hat{\sigma}^2(p)$, this standard error is still not well estimated. The parameter value is .0021, and the boot-p estimate is .0016 with a standard deviation of .0003. This result casts some doubt on the applicability of the bootstrap with binary data, even when using the "best" procedure for obtaining bootstrap samples.

Standard Errors and Confidence Intervals

Table 4 provides traditional, bootstrap, and jackknife estimates of standard errors for five random samples (trials) of size n = 200 and k = 20 that were drawn from the finite population and universe. As such, Table 4 for binary data is analogous to Table 1 for normal data.

Insert Table 4 about here

The traditional estimates of standard errors in Table 4 for $\hat{\sigma}^2(i)$ bracket the parameter value, but for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ the estimated standard errors are systematically too low. Furthermore, $\hat{\sigma}[\hat{\sigma}^2(pi)]$ is 2-3 times too low!

For each trial, bootstrap results were based on B = 1000 bootstrap samples, using boot-p for persons, boot-i for items, and boot-p,i for residuals. The bootstrap estimates of standard errors in Table 4 are generally closer to the parameter values than are the traditional estimates. However, for persons and residuals, the bootstrap estimates of standard errors are too low, even though they are better than the traditional results.

The jackknife estimates of standard errors are noticeably more variable than either the traditional or bootstrap estimates, but on average the

jackknife estimates seem to be at least as accurate or more accurate than the traditional or bootstrap results.

Table 5 provides approximate 80% confidence intervals using the Satterthwaite, bootstrap, and jackknife procedures for the five trials in Table 4. As was done for the normal data simulations, the target values for limits of the confidence intervals were defined as the 10th and 90th percentile points of the distributions of each of the estimated variance components, based on 2000 random samples of size n = 200 and k = 20 from the finite population and universe. These results are provided in the first row of Table 5.

Insert Table 5 about here

Because the magnitudes of the estimated standard errors are generally quite small, even quite large parameter-estimate discrepancies in standard errors are not likely to lead to confidence intervals that <u>appear</u> to be dramatically wrong. This is illustrated by many of the intervals in Table 5.

To gain some perspective on these results, for each trial the one or two "best" intervals for each variance component are identified with an asterisk in Table 5. These judgments about "best" are admittedly subjective, but they do suggest that: (a) Satterthwaite's procedure does not provide as accurate intervals as the other procedures; (b) the bootstrap and jackknife procedures provide comparable intervals; and (c) there is no advantage to be gained in jackknifing the logarithms of the variance components.

The results in Tables 4 and 5 are limited in that they are based on only five trials. Table 6 provides traditional and jackknife estimates of variance components and standard errors for 1000 trials, with each trial consisting of a random sample (without replacement) of size n = 200 and k = 20 from the

finite population and universe. These results can be compared with the bootstrap results in Table 3, although the bootstrap results are for only 100 trials.

Insert Table 6 about here

The results in Table 6 confirm the observations made about the traditional and jackknife procedures based on the five trials in Table 4. With the traditional approach, the standard error of $\hat{\sigma}^2(p)$ is somewhat underestimated and the standard error of $\hat{\sigma}^2(pi)$ is <u>dramatically</u> underestimated. In fact, under the traditional approach, the <u>maximum</u> value of $\hat{\sigma}[\hat{\sigma}^2(pi)]$ is .0052 which is about 50% smaller than the parameter value of .0118. Clearly, Equation 23 (for residuals) provides poor results with binary data. With the jackknife approach, for all three variance components the mean of the estimated standard errors is quite close to the parameter value, but there is much more variability in the estimated standard errors for persons and residuals than is the case for the traditional and bootstrap procedures. For example, the standard deviation of the 1000 jackknife estimates of $\sigma[\hat{\sigma}^2(p)]$ is .0005, which is 2-3 times larger than the comparable traditional and bootstrap results.

Using the same 1000 trials that resulted in the statistics reported in Table 6, 1000 Satterthwaite and jackknife confidence intervals were obtained for each variance component using nominal coverage coefficients of 50%, 80%, and 90%. The percents of intervals that actually covered the parameters are reported in Table 7. Comparable bootstrap results were not obtained because doing so would have required $B \geq 1000$ bootstrap samples for boot-p, boot-i, and boot-p,i for each of the 1000 trials. This was judged to be excessively expensive.

Insert Table 7 about here

For the Satterthwaite intervals, the actual coverages for $\sigma^2(p)$ are somewhat low, and the actual coverages for $\sigma^2(i)$ are somewhat high. In other words, the $\sigma^2(p)$ intervals are somewhat too narrow, and the $\sigma^2(i)$ intervals are somewhat too broad. Most importantly, however, the actual coverages for $\sigma^2(pi)$ are dramatically low, which means that the confidence intervals for $\sigma^2(pi)$ using Satterthwaite's procedure are <u>much</u> too narrow with binary data. This result is consistent with the excessively low value for $\hat{\sigma}[\hat{\sigma}^2(pi)]$ using the traditional approach. Clearly, the fact that binary data violate the normality assumptions in Satterthwaite's procedure causes this procedure to work poorly with the variance component for residuals.

Almost without exception, the actual coverages for the jackknife intervals are somewhat too low, implying that the intervals are a little bit too narrow. However, the results in Table 7 for the jackknife suggest that, all things considered, the jackknife provides confidence intervals that are quite accurate for practical use.

Summary and Conclusions

The simulation studies reported in this paper suggest the following conclusions with respect to examining the variability of estimated variance components for the p x i random effects design:

(a) The traditional approach (estimated standard errors using Equations 18-20 and Satterthwaite confidence intervals) provides accurate results for all three variance components with normal data; however, with binary data, the traditional approach provides only moderately accurate results for the person and item variance components and quite inaccurate results for the residual variance component;

(b) The jackknife approach provides quite accurate results with both normal and binary data for all three variance components--its primary limitation being more variability in estimated standard errors than was found with the other approaches;

(c) Computing jackknife pseudo-values based on the logarithms of variance components is not advisable--at least for situations similar to those that characterized these simulations; and

(d) As discussed below, the bootstrap results are mixed, largely because there seems to be no single "best" way to obtain bootstrap samples.

The bootstrap simulation results suggest that the "best" procedure with normal data is to use boot-p for persons, boot-i for items, boot-p for residuals if $n \geq k$ or boot-i for residuals if $k \geq n$. In particular, it does not appear that it is advisable to use boot-p,i with normal data for estimating standard errors or obtaining confidence intervals for any of the variance components. With binary data, however, all things considered the "best" procedure is to use boot-p for persons, boot-i for items, and boot-p,i for residuals. Furthermore, it seems advisable to adjust the boot-p, boot-i, and boot-p,i estimates of variance components and their standard errors by the correction factors $n/(n-1)$, $k/(k-1)$, and $[n/(n-1)][k/(k-1)]$, respectively.

In this sense, it might be stated that the bootstrap "works," but it is somewhat disconcerting that correction factors are needed, and that different bootstrap sampling procedures are required depending on the particular variance component under consideration and the nature of the underlying data. Furthermore, for the binary data simulations, the boot-p estimates of $\sigma[\hat{\sigma}^2(p)]$ are somewhat low, which implies that confidence intervals for $\sigma^2(p)$ are somewhat too narrow for the parameters and sample sizes in this study.

Another limitation of the bootstrap is that it is not clear how to extend its use to estimating standard errors and confidence intervals for functions of variance components. For example, the estimated error variance $\hat{\sigma}^2(\Delta) = [\hat{\sigma}^2(i) + \hat{\sigma}^2(pi)]/k$ is often reported in generalizability analyses. To estimate its standard error using the bootstrap, a particular bootstrap sampling procedure must be chosen. However, the simulation results reported above do not clearly indicate how best to obtain such bootstrap samples. Since generalizability analyses often involve several different functions of variance components, this problem is of some consequence.

In terms of complexity and computational requirements, the traditional approach is simple and quick, and the jackknife is conceptually complex but not too demanding computationally--at least for the sample sizes considered here. The application of the bootstrap considered here, however, is somewhat complex and requires considerably more computation than the jackknife. Recall that bootstrap samples need to be formed in two or three ways, and $B \geq 1000$ sample are necessary to establish confidence intervals. In effect, this means that applying the bootstrap with the p x i random effects design requires 2000-3000 analyses of variance with matrices of size n x k if one wants to establish confidence intervals for the variance components. These computational requirements may be excessive in some circumstances.

On balance, it would appear that the bootstrap has some limitations as a methodology for addressing questions about the variability of estimated variance components with the random effects p x i design. This is unfortunate since the nonparametric characteristics of the bootstrap appear to make it especially attractive in testing contexts where normality assumptions are known to be violated because the data are binary.

On a positive note, however, the results in this paper suggest that the jackknife produces quite accurate results in such contexts. Furthermore, future research may reveal that there is a different way to conceive of bootstrap samples that avoids some of the limitations identified in this paper.

REFERENCES

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC:. Author.

Arvesen, J. N., & Schmitz, T. H. (1970). Robust procedures for variance component problems using the jackknife. Biometrics, 26, 677-686.

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.

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

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

Collins, J. D. (1970). Jackknifing generalizability. Unpublished doctoral dissertation, University of Colorado, Boulder.

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

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

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, CBMS-NSF Monograph, 38, Philadelphia.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1, 54-77.

Graybill, F. A. (1961). An introduction to linear statistical models. New York: McGraw-Hill.

Graybill, F. A. (1976). Theory and application of the linear model. North Scituate, MA: Duxbury Press.

International Mathematical and Statistical Libraries. (1984). IMSL Libraries (9th ed.). Houston: Author.

Jarjoura, D., & Brennan, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. Applied Psychological Measurement, 6, 161-171.

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

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.). The handbook of social psychology, Vol. 2. Research methods (2nd ed.) (pp. 80-203). Reading, MA: Addison-Wesley.

Quenouille, M. (1949). Approximate tests of correlation in time series. Journal of the Royal Statistical Society--Series B, 11, 18-84.

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

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

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

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.

Tukey, J. (1958). Bias and confidence in not quite large samples, abstract. Annals of Mathematical Statistics, 29, 614.

Welch, B. L. (1956). On linear combinations of several variances. American Statistical Association Journal, 51, 132-148.

APPENDIX A

Estimated Standard Errors and Satterthwaite's Confidence Intervals for Variance Components

Assuming mean squares are independent and score effects have a multivariate normal distribution, Searle (1971, pp. 415-417) shows that an estimated standard error of any estimated variance component, $\hat{\sigma}^2(\gamma)$, is:

$$\hat{\sigma}[\hat{\sigma}^2(\gamma)] = [\sum_j 2(f_j MS_j)^2/df_j]^{\frac{1}{2}} \qquad \text{(A1)}$$

where j indexes the mean squares that enter $\hat{\sigma}^2(\gamma)$, f_j is the coefficient of MS_j in the linear combination of the MS_j that gives $\hat{\sigma}^2(\gamma)$, and df_j is the degrees of freedom for MS_j. Technically, $\hat{\sigma}^2[\hat{\sigma}^2(\gamma)]$ is a biased estimator of the variance of $\hat{\sigma}^2(\gamma)$. It can be transformed to an unbiased estimator by replacing df_j with $df_j + 2$. All estimated standard errors in this paper employ Equation A1 (rather than the companion equation involving $df_j + 2$) for two reasons. First, technically, Equation A1 is required for the confidence intervals discussed below. Second, because degrees of freedom are relatively large for all analyses reported in this paper, using $df_j + 2$ rather than df_j makes virtually no noticeable difference in the results.

The exact distribution of estimated variance components is generally unknown, even under normality assumptions. Based on such assumptions, however, Satterthwaite (1941, 1946) proposed the following <u>approximate</u> $100(1 - \alpha)$ percent confidence interval for estimated variance components:

$$\frac{\hat{\sigma}^2(\gamma)\nu}{\chi^2_{1-\alpha/2}(\nu)} \leq \sigma^2(\gamma) \leq \frac{\hat{\sigma}^2(\gamma)\nu}{\chi^2_{\alpha/2}(\nu)}, \qquad \text{(A2)}$$

where ν is called the "effective" degrees of freedom, and $\chi^2_{1-\alpha/2}(\nu)$ and $\chi^2_{\alpha/2}(\nu)$ are the $1-\alpha/2$ and $\alpha/2$ percentage points, respectively, of the chi-squared distribution with ν degrees of freedom.

Brennan (1983) shows that, for the types of estimators of variance components considered in this paper, $\nu = 2r^2$ where

$$r = \hat{\sigma}^2(\gamma)/\hat{\sigma}[\hat{\sigma}^2(\gamma)] \tag{A3}$$

is the ratio of the estimated variance component to its estimated standard error. Based on this simplification, Brennan (1983, pp. 137-140) provides a table for 67, 75, 80, 90, and 95 percent confidence intervals given any one of 150 values of r between 2 and 100. For most purposes Brennan's table provides accurate enough results. For purposes of this paper, however, more numerical accuracy was desired, which involved using the IMSL (1984) subroutine MDCHI in conjunction with Equation A2.

Welch (1956) proposed a relatively complicated correction factor for use with Satterthwaite's approximate confidence intervals. Welch derived his correction factor using assumptions different from those of Satterthwaite, and Welch himself expressed reservations about practical use of the correction factor (see Welch, 1956, p. 145). Even so, Cronbach et al. (1972, p. 53), Graybill (1961, pp. 368-382), and Smith (1982) seem to be favorably disposed to using the correction factor. However, a study by Boardman (1974) does not support its use, and more recently Graybill (1976, pp. 642-643) has chosen to neglect it completely in his textbook treatment of variance components. For these reasons, Welch's correction factor was not employed for the analyses reported in this paper.

APPENDIX B

Jackknife Estimates, Their Standard Errors and Confidence Intervals for the p x i Random Effects Design

Based on advice from John Tukey, Cronbach et al. (1972, pp. 54-57, 66, 70-72) outline a jackknife procedure for estimating the standard errors of estimated variance components and for establishing confidence intervals for variance components. The basic elements of this procedure were employed by Collins (1970) in a fairly large simulation study.[3] Provided below is an outline of how to employ the jackknife with the p x i design.

Consider the following notational conventions:

$\hat{\theta}$ = any estimated variance component for the p x i design based on analyzing the full n x k matrix [i.e., $\hat{\theta}$ could be $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(i)$, or $\hat{\sigma}^2(pi)$]

$\hat{\theta}_{-pi}$ = value of $\hat{\theta}$ for the (n-1) x (k-1) matrix that results from eliminating person p and item i;

$\hat{\theta}_{-p0}$ = value of $\hat{\theta}$ for (n-1) x k matrix that results from eliminating person p;

$\hat{\theta}_{-0i}$ = value of $\hat{\theta}$ for the n x (k-1) matrix that results from eliminating item i;

$\hat{\theta}_{-00}$ = value of $\hat{\theta}$ for original n x k matrix (i.e., $\hat{\theta} = \hat{\theta}_{-00}$) .

Now, the pseudo-value for person p and item i is:

$$\hat{\theta}_{*pi} = nk\hat{\theta}_{-00} - (n-1)k\hat{\theta}_{-p0} - n(k-1)\hat{\theta}_{-0i} + (n-1)(k-1)\hat{\theta}_{-pi} \ . \qquad \text{(B1)}$$

The mean of all nk pseudo-values is the jackknife estimator of θ :

$$\hat{\theta}_J = \sum_{p=1}^{n} \sum_{i=1}^{k} \hat{\theta}_{*pi}/nk \ . \qquad \text{(B2)}$$

For the types of estimators usually employed in generalizability theory $\hat{\theta}_J = \hat{\theta}$ which is an unbiased estimator of θ . For example, if $\hat{\theta} = \hat{\sigma}^2(p) = [MS(p) - MS(pi)]/k$, then $\hat{\theta}_J = \hat{\sigma}^2_J(p) = \hat{\sigma}^2(p) = \hat{\theta}$, which is an unbiased estimator of $\theta = \sigma^2(p)$.

To estimate the standard error of $\hat{\theta}_J = \hat{\theta}$ using the jackknife procedure, we employ the matrix of pseudo-values, which has n rows and k columns. For this matrix let $\hat{\sigma}^2(\text{rows})$, $\hat{\sigma}^2(\text{cols})$, and $\hat{\sigma}^2(\text{res})$ be the estimated variance components taking into account sampling from a finite population and/or universe if $N < \infty$ and/or $K < \infty$, respectively. (Use Equations 10, 11, and 12 with "rows," "cols," and "res" replacing p, i, and pi, respectively.) Then the estimated standard error of $\hat{\theta}_J$ is:

$$\hat{\sigma}(\hat{\theta}_J) = [c_n\hat{\sigma}^2(\text{rows})/n + c_k\hat{\sigma}^2(\text{cols})/k + c_n c_k\hat{\sigma}^2(\text{res})/nk]^{\frac{1}{2}} \ , \qquad \text{(B3)}$$

where $c_n = 1 - n/N$ and $c_k = 1 - k/K$ are the finite population and universe correction factors, respectively. [Note that, in their discussion of the jackknife, Cronbach et al. (1972, pp. 56, 71) incorrectly suggest that the result in Equation B3 be divided by the square root of nk.]

To this point the jackknife procedure is nonparametric in that none of the above results make any assumptions about distributional form. To establish confidence intervals, however, student's t distribution is usually employed (see Mosteller and Tukey, 1968, p. 135). Thus, a $100(1-\alpha)$ percent confidence interval for θ is

$$\hat{\theta}_J - t\ \hat{\sigma}(\hat{\theta}_J) \le \theta \le \hat{\theta}_J + t\ \hat{\sigma}(\hat{\theta}_J) \ , \qquad \text{(B4)}$$

where t is the $(1 - \alpha/2)$ percentage point of the t distribution with

nk - 1 degrees of freedom (see Collins, 1970, p. 29). In this paper, interest focuses on designs in which nk - 1 is quite large, and for such designs the unit normal distribution can be used in place of Student's t.

Actually, Cronbach et al. (1972) and, to an extent, Collins (1970) both suggest jackknifing the logarithm of estimated variance components rather than the estimates themselves. Specifically, for variance components Cronbach et al. (1972) use as pseudo-values

$$\hat{\theta}_{*pi} = nk \log \hat{\theta}_{-00} - (n-1)k \log \hat{\theta}_{-p0} - n(k-1) \log \hat{\theta}_{-0i}$$

$$+ (n-1)(k-1) \log \theta_{-pi} \; . \qquad \text{(B5)}$$

Then, when the limits of a confidence interval are obtained in the log metric, they are transformed back to the original metric using antilogs. Of course, the usual estimates of variance components can be negative, and one cannot take the log of a negative number. To avoid this, Tukey suggests the possibility of using fifth roots rather than logs in Equation B5 when negative estimates are likely. With samples as large as those of interest in this paper, negative estimates are not very likely, however. Usually, the principal effect of using logs is that the resulting confidence intervals are broader, especially for effects with relatively small numbers of observations.

APPENDIX C

Tables Illustrating Results of Different Procedures for Estimating Variance Components and Their Standard Errors for Normally Distributed Data

Subsequent pages provide illustrative results for the following nine combinations of three sets of variance components and three sets of sample sizes.

Table	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$	n	k
C1	4	16	64	200	20
C2	4	16	64	100	40
C3	4	16	64	50	50
C4	16	16	64	200	20
C5	16	16	64	100	40
C6	16	16	64	50	50
C7	16	16	16	200	20
C8	16	16	16	100	40
C9	16	16	16	50	50

Note that all bootstrap results are based on B = 100 bootstrap samples.

TABLE C1

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 200 and k = 20

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	4.00	.73	16.00	5.29	64.00	1.47
Estimates						
Traditional	4.36	.76	18.19	6.01	63.72	1.47
Boot p,i	7.22	1.58	17.38	5.27	60.37	2.73
Boot p,i,r	7.46	1.23	17.81	6.48	60.14	1.34
Boot p	4.01	.87	18.40	.96	63.39	1.42
Boot i	7.59	1.22	17.12	5.21	60.56	1.57
Boot[a]	4.03	.87	18.02	5.48	63.71	1.43
Jackknife[b]	4.36	.85	18.19	6.59	63.72	1.22

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor $n/(n-1) = 1.00503$, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor $k/(k-1) = 1.05263$.

[b]Based on full matrix of 200 x 20 = 4000 pseudo-values.

TABLE C2

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 100 and k = 40

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	4.00	.79	16.00	3.77	64.00	1.46
Estimates						
Traditional	3.75	.76	8.26	2.02	63.68	1.45
Boot p,i	5.25	1.46	8.31	2.11	61.74	2.30
Boot p,i,r	5.31	.96	8.51	1.75	61.41	1.36
Boot p	3.68	.84	8.83	.73	63.14	1.34
Boot i	5.35	.77	7.83	1.77	62.26	1.50
Boot[a]	3.72	.85	8.03	1.82	63.77	1.35
Jackknife[b]	3.75	.73	8.27	1.86	63.68	1.59

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor n/(n-1) = 1.01010, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor k/(k-1) = 1.02564.

[b]Based on full matrix of 100 x 40 = 4000 pseudo-values.

TABLE C3

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 50 and k = 50

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	4.00	1.07	16.00	3.49	64.00	1.85
Estimates						
Traditional	4.25	1.12	11.50	2.58	63.30	1.83
Boot p,i	5.21	1.53	12.50	2.86	60.97	3.28
Boot p,i,r	5.46	1.26	12.40	2.49	60.26	1.39
Boot p	3.98	.98	12.79	1.04	62.10	2.06
Boot i	5.47	.82	11.37	2.32	62.27	1.78
Boot[a]	4.06	1.00	13.05	2.37	63.36	2.10
Jackknife[b]	4.25	1.07	11.50	2.39	63.30	2.11

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor n/(n-1) = 1.02041, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor k/(k-1) = 1.02041.

[b]Based on full matrix of 50 x 50 = 2500 pseudo-values.

TABLE C4

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 200 and k = 20

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	16.00	1.93	16.00	5.29	64.00	1.47
Estimates						
Traditional	16.20	1.94	18.86	6.22	62.37	1.43
Boot p,i	19.12	3.02	18.38	7.53	58.95	2.47
Boot p,i,r	19.40	2.34	17.17	6.19	58.94	1.28
Boot p	15.84	2.37	19.21	1.11	62.29	1.27
Boot i	19.48	1.41	18.04	7.21	59.12	1.71
Boot[a]	15.92	2.38	18.99	7.59	62.60	1.28
Jackknife[b]	16.20	2.19	18.86	7.92	62.37	1.21

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor $n/(n-1) = 1.00503$, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor $k/(k-1) = 1.05263$.

[b]Based on full matrix of 200 x 20 = 4000 pseudo-values.

TABLE C5

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 100 and k = 40

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	16.00	2.50	16.00	3.77	64.00	1.47
Estimates						
Traditional	16.25	2.53	20.82	4.56	62.36	1.42
Boot p,i	17.44	2.77	21.31	4.64	60.43	2.58
Boot p,i,r	17.54	2.92	20.48	3.97	59.99	1.29
Boot p	15.96	2.37	21.46	1.13	61.88	1.51
Boot i	17.73	1.22	20.62	4.24	60.61	1.37
Boot[a]	16.13	2.39	21.15	4.35	62.50	1.53
Jackknife[b]	16.25	2.73	20.82	4.45	62.36	1.62

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor $n/(n-1) = 1.01010$, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor $k/(k-1) = 1.02564$.

[b]Based on full matrix of 100 x 40 = 4000 pseudo-values.

TABLE C6

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 50 and k = 50

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	16.00	3.49	16.00	3.49	64.00	1.85
Estimates						
Traditional	10.06	2.30	18.41	3.98	64.98	1.88
Boot p,i	10.41	3.10	19.66	4.09	62.25	3.00
Boot p,i,r	11.23	2.51	18.84	4.08	62.35	1.62
Boot p	9.21	2.37	19.66	1.57	63.63	1.69
Boot i	11.43	1.43	18.37	3.45	63.47	1.87
Boot[a]	9.40	2.42	18.74	3.52	64.93	1.72
Jackknife[b]	10.06	2.69	18.41	4.06	64.98	1.65

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor n/(n-1) = 1.02041, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor k/(k-1) = 1.02041.

[b]Based on full matrix of 50 x 50 = 250 pseudo-values.

TABLE C7

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 200 and k = 20

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	16.00	1.68	16.00	5.22	16.00	.37
Estimates						
Traditional	16.47	1.73	9.11	2.98	15.20	.35
Boot p,i	17.01	1.92	8.67	2.87	14.35	.75
Boot p,i,r	17.09	1.58	8.75	2.70	14.38	.28
Boot p	16.21	1.67	9.17	.37	15.18	.34
Boot i	17.25	.61	8.65	2.83	14.38	.51
Boot[a]	16.30	1.68	9.11	2.98	15.25	.34
Jackknife[b]	16.46	1.76	9.11	3.09	15.20	.42

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor $n/(n-1) = 1.00503$, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor $k/(k-1) = 1.05263$.

[b]Based on full matrix of 200 x 20 = 4000 pseudo-values.

TABLE C8

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 100 and k = 40

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	16.00	2.33	16.00	3.66	16.00	.36
Estimates						
Traditional	17.20	2.50	13.07	3.00	16.30	.37
Boot p,i	17.10	2.78	13.02	2.35	15.86	.62
Boot p,i,r	17.37	2.55	12.97	1.97	15.71	.34
Boot p	16.79	2.55	13.26	.54	16.21	.37
Boot i	17.59	.55	12.79	2.27	15.93	.36
Boot[a]	16.96	2.58	13.12	2.33	16.38	.37
Jackknife[b]	17.20	2.60	13.07	2.31	16.30	.39

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor $n/(n-1) = 1.01010$, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor $k/(k-1) = 1.02564$.

[b]Based on full matrix of 100 x 40 = 4000 pseudo-values.

TABLE C9

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 50 and k = 50

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	16.00	3.30	16.00	3.30	16.00	.46
Estimates						
Traditional	15.23	3.14	16.65	3.43	15.99	.46
Boot p,i	15.58	3.01	16.31	3.78	15.34	.74
Boot p,i,r	15.05	2.65	16.58	3.75	15.44	.49
Boot p	15.20	2.78	17.12	.67	15.67	.40
Boot i	15.59	.70	15.88	3.68	15.68	.40
Boot[a]	15.51	2.84	16.21	3.76	15.99	.41
Jackknife[b]	15.23	2.73	16.65	3.71	15.99	.39

Note. Bootstrap results are based on 100 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor n/(n-1) = 1.02041, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor k/(k-1) = 1.02041.

[b]Based on full matrix of 50 x 50 = 2500 pseudo-values.

Author Notes

The authors gratefully acknowledge the advice of Michael J. Kolen at various stages of this research.

Footnotes

[1]A more realistic design for most testing contexts would be p x (i:h) where each item is nested within a single content category, h, and categories are fixed (see, for example, Jarjoura & Brennan, 1982). This more realistic design was judged to be too complicated, at this stage of research, for a comparative treatment of traditional, bootstrap, and jackknife methodologies.

[2]Throughout this paper, the nominal coverage for confidence intervals (confidence coefficient) is for a single variance component. If one were equally interested in confidence intervals for all three variance components in some specific situation, then one might want to consider the joint coverage for all three variance components simultaneously (see Bell, 1986). This can be done using a method proposed by Khuri (1981) or Satterthwaite's procedure can be modified. Specifically, for q joint $100(1 - \alpha)$ percent confidence intervals, α in equation A2 should be replaced by $\alpha' = 1 - (1-\alpha)^{1/q}$. This is equivalent to obtaining Satterthwiate intervals with individual coverages if $100(1 - \alpha')$ percent. For example, for $q = 3$ joint 80% confidence intervals, one uses $\alpha' = 1 - (1 - .2)^{1/3} = .0717$, which is equivalent to obtaining intervals with individual coverages of about 93%.

[3]Arvesen and Schmitz (1970) also employ the jackknife in a variance components situation, but they treat a nested design, only, not a crossed design.

TABLE 1

Estimates of Variance Components and Their Standard Errors for Normally Distributed Data Based on a Sample of Size n = 200 and k = 20

	Persons		Items		Residuals	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	4.00	.73	16.00	5.29	64.00	1.47
Estimates						
Traditional	3.93	.72	13.66	4.53	63.74	1.47
Boot-p,i	7.09	1.55	13.41	4.96	60.26	2.55
Boot-p,i,r	7.11	1.02	13.44	4.88	60.24	1.43
Boot-p	**3.90**	**.70**	13.98	1.04	**63.45**	**1.42**
Boot-i	7.10	1.19	**13.08**	**4.77**	60.58	1.70
Boot[a]	3.92	.70	13.77	5.02	63.77	1.43
Jackknife[b]	3.93	.66	13.66	5.22	63.74	1.39

Note. Bootstrap results are based on 1000 bootstrap samples.

[a]Obtaining $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(pi)$ and their standards error from bootstrapping persons; and $\hat{\sigma}^2(i)$ and its standard error from bootstrapping items. Reported values for $\hat{\sigma}^2(p)$ and $\hat{\sigma}^2(pi)$ and their standard errors employ the correction factor $n/(n-1) = 1.00503$, and the reported value of $\hat{\sigma}^2(i)$ and its standard error employs the correction factor $k/(k-1) = 1.05263$.

[b]Based on the full matrix of 200 x 20 = 4000 pseudo-values.

TABLE 2

Approximate 80 Percent Confidence Intervals for Variance Components Based on Results in Table 1

	$\sigma^2(p) = 4$	$\sigma^2(i) = 16$	$\sigma^2(pi) = 64$
Simulation Results[a]	(3.1, 5.0)	(9.5, 22.6)	(62.0, 65.9)
Estimates			
Satterthwaite	(3.2, 5.1)	(9.5, 22.6)	(61.9, 65.7)
Bootstrap[b]	(3.0, 4.8)	(7.4, 20.6)	(61.9, 65.6)
Jackknife[c]	(3.1, 4.8)	(7.0, 20.3)	(62.0, 65.5)
Jackknife, log[d]	(3.2, 4.9)	(8.7, 25.4)	(62.0, 65.6)

[a]Reported values are the 10th and 90th percentile points of the distribution of each estimated variance component for 2000 random samples of size n = 200 and k = 20.

[b]Based on row labeled "Boot" in Table 1, with the limits of the intervals for $\sigma^2(p)$ and $\sigma^2(pi)$ multiplied by the correction factor $n/(n-1) = 1.00503$, and with the limits of the interval for $\sigma^2(i)$ multiplied by the correction factor $k/(k-1) = 1.05263$.

[c]Based on pseudo-values given by Equation B1.

[d]Based on pseudo-values given by Equation B5.

TABLE 3

Bootstrap Estimates of Variance Components and Their Standard Errors for Binary Data Based on 100 Random Samples (Trials) of n = 200 Persons and k = 20 Items From a Population and Universe of Sizes N = 2000 and K = 200, Respectively

		Means and SD's Over 100 Trials with B = 100 per Trial					
		Persons		Items		Residual	
Boot	Statistic	$\hat{\sigma}^2_B(p)$	$\hat{\sigma}[\hat{\sigma}^2_b(p)]$	$\hat{\sigma}^2_B(i)$	$\hat{\sigma}[\hat{\sigma}^2_b(i)]$	$\hat{\sigma}^2_B(pi)$	$\hat{\sigma}[\hat{\sigma}^2_b(pi)]$
p	Mean	**.0069**[a]	**.0016**[a]	.0360	.0025	.1898[a]	.0033[a]
	SD	.0017[a]	.0003[a]	.0103	.0004	.0127[a]	.0004[a]
i	Mean	.0164	.0043	**.0334**[b]	**.0098**[b]	.1901[b]	.0121[b]
	SD	.0022	.0005	.0090[b]	.0033[b]	.0128[b]	.0021[b]
p,i	Mean	.0161	.0047	.0339	.0103	**.1911**[c]	**.0123**[c]
	SD	.0020	.0005	.0092	.0030	.0117[c]	.0019[c]
Parameters		.0068	.0021	.0346	.0101	.1902	.0118

Note. An individual trial involves sampling without replacement from the population and universe, whereas a bootstrap sample involves sampling with replacement from the person and/or item vectors constituting a given trial.

[a]Computed estimates multiplied by the correction factor n/(n-1) = 1.00503.

[b]Computed estimates multiplied by the correction factor k/(k-1) = 1.05263.

[c]Computed estimates multiplied by the correction factor [n/(n-1)][k/(k-1)] = 1.05792.

TABLE 4

Estimates of Variance Components and Their Standard Errors for Binary Data Based on Five Random Samples (Trials) of n = 200 Persons and k = 20 Items From a Population and Universe of Sizes N = 2000 and K = 200, Respectively

	Persons		Items		Residual	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	.0068	.0021	.0346	.0101	.1902	.0118
Estimates						
Traditional	.0078	.0017	.0232	.0078	.1934	.0044
	.0066	.0017	.0148	.0051	.2180	.0050
	.0087	.0018	.0263	.0088	.2106	.0048
	.0094	.0018	.0368	.0122	.1971	.0045
	.0087	.0017	.0429	.0142	.1864	.0043
Bootstrap[a]	.0078	.0017	.0233	.0117	.1938	.0094
	.0066	.0017	.0149	.0039	.2177	.0094
	.0086	.0017	.0264	.0064	.2106	.0096
	.0094	.0021	.0369	.0093	.1972	.0090
	.0087	.0016	.0423	.0099	.1868	.0117
Jackknife[b]	.0078	.0021	.0232	.0118	.1934	.0085
	.0066	.0015	.0148	.0036	.2180	.0078
	.0087	.0020	.0263	.0063	.2106	.0082
	.0094	.0033	.0368	.0093	.1971	.0071
	.0087	.0021	.0429	.0105	.1864	.0107

[a]Obtaining $\hat{\sigma}^2(p)$ and $\hat{\sigma}[\hat{\sigma}^2(p)]$ from bootstrapping persons, $\hat{\sigma}^2(i)$ and $\hat{\sigma}[\hat{\sigma}(i)]$ from bootstrapping items, and $\hat{\sigma}^2(pi)$ and $\hat{\sigma}[\hat{\sigma}^2(pi)]$ from bootstrapping both persons and items. Reported values for $\hat{\sigma}^2(p)$, $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(pi)$ (and their standard errors) employ the correction factors $n/(n-1) = 1.00503$, $k/(k-1) = 1.05263$, and $[n/(n-1)][k/(k-1)] =$ 1.05792, respectively. (B = 1000)

[b]Based on the full matrix of 200 x 20 = 4000 pseudo-values.

TABLE 5

Approximate 80 Percent Confidence Intervals for Variance Components for Results of the Five Trials in Table 4

	$\sigma^2(p) = .0068$	$\sigma^2(i) = .0346$	$\sigma^2(pi) = .1902$
Simulation Results[a]	(.0043, .0096)	(.0219, .0479)	(.1744, .2052)
Estimates			
Satterthwaite	(.0061, .0105)	(.0160, .0387)*	(.1878, .1993)
Bootstrap	(.0057, .0101)*	(.0092, .0397)	(.1814, .2062)*
Jackknife	(.0051, .0105)*	(.0082, .0383)	(.1825, .2043)*
Jackknife, log	(.0058, .0112)	(.0120, .0685)	(.1830, .2048)
Satterthwaite	(.0050, .0095)*	(.0101, .0251)	(.2117, .2245)
Bootstrap	(.0043, .0088)*	(.0100, .0200)	(.2048, .2291)*
Jackknife	(.0047, .0085)*	(.0101, .0194)	(.2079, .2280)*
Jackknife, log	(.0053, .0085)	(.0111, .0211)	(.2084, .2283)
Satterthwaite	(.0068, .0118)	(.0182, .0438)*	(.2045, .2169)
Bootstrap	(.0066, .0109)*	(.0182, .0349)	(.1980, .2225)*
Jackknife	(.0062, .0113)*	(.0181, .0344)	(.2000, .2211)*
Jackknife, log	(.0069, .0117)	(.0197, .0375)	(.2005, .2215)
Satterthwaite	(.0075, .0124)	(.0255, .0608)	(.1914, .2031)
Bootstrap	(.0069, .0124)*	(.0253, .0486)*	(.1853, .2084)*
Jackknife	(.0051, .0136)*	(.0248, .0487)*	(.1880, .2062)
Jackknife, log	(.0065, .0153)	(.0271, .0539)	(.1883, .2066)
Satterthwaite	(.0069, .0115)	(.0298, .0707)	(.1810, .1920)
Bootstrap	(.0066, .0108)*	(.0288, .0551)*	(.1716, .2016)*
Jackknife	(.0060, .0113)*	(.0294, .0563)	(.1727, .2001)*
Jackknife, log	(.0066, .0210)	(.0321, .0613)	(.1736, .2009)

[a]Reported values are the 10th and 90th percentile points of the distribution of each estimated variance component for 2000 random samples (trials) of size n = 200 and k = 20 from the population and universe sizes N = 2000 and K = 200, respectively.

TABLE 6

Traditional and Jackknife Estimates of Variance Components and Their Standard Errors for Binary Data Based on 1000 Random Samples (Trials) of Size n = 200 and k = 20 From a Finite Population and Universe of Sizes N = 2000 and K = 200, Respectively

	Persons		Items		Interaction	
	$\sigma^2(p)$	$\sigma[\hat{\sigma}^2(p)]$	$\sigma^2(i)$	$\sigma[\hat{\sigma}^2(i)]$	$\sigma^2(pi)$	$\sigma[\hat{\sigma}^2(pi)]$
Parameters	.0068	.0021	.0346	.0101	.1902	.0118
Estimates						
Traditional						
Mean	.0067	.0015	.0350	.0116	.1903	.0044
SD	.0020	.0002	.0103	.0033	.0120	.0003
Jackknife						
Mean	.0068	.0021	.0344	.0101	.1905	.0116
SD	.0021	.0005	.0102	.0033	.0117	.0021

Note. Jackknife results are not based on taking the logarithms of estimated variance components.

TABLE 7

Confidence Interval Coverage Results for Satterthwaite and Jackknife Procedures for Binary Data and the 1000 Trials Summarized in Table 6

Normal Coverage (Percent)	Actual Coverage (Percent)					
	Satterthwaite			Jackknife		
	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$	$\sigma^2(p)$	$\sigma^2(i)$	$\sigma^2(pi)$
	50	49.5	54.0	19.2	51.3	49.748.1
	80	72.4	83.9	36.8	78.9	76.477.2
	90	85.3	93.2	46.0	87.4	84.286.9

Note. Jackknife results are not based on taking logarithms of estimated variance components.