DOCUMENT RESUME

ED 423 304 TM 029 112

AUTHOR	Betebenner, Damian W.
TITLE	Improved Confidence Interval Estimation for Variance Components and Error Variances in Generalizability Theory.
PUB DATE	1998-00-00
NOTE	11p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Accountability; *Educational Change; Error of Measurement; *Generalizability Theory; *Performance Based Assessment; Research Design; *Sampling; Simulation
IDENTIFIERS	*Confidence Intervals (Statistics)

ABSTRACT

The zeitgeist for reform in education precipitated a number of changes in assessment. Among these are performance assessments, sometimes linked to "high stakes" accountability decisions. In some instances, the trustworthiness of these decisions is based on variance components and error variances derived through generalizability theory. Often overlooked is the fact that these statistics are subject to sampling error. This paper introduces techniques used to determine the accuracy of such statistics. It addresses the shortcomings of overlooking sampling error by presenting a number of results with respect to confidence intervals about linear combinations of expected mean squares appropriate for generalizability theory. Simulation results indicate that these intervals, particularly the two-sided and one-sided lower intervals, are accurate or conservative both in simple and complex designs with varying amounts of difference in degrees of freedom across effects. (Contains 1 figure and 12 references.) (Author/SLD)

Improved Confidence Interval Estimation for Variance Components and Error Variances in Generalizability Theory

Damian W. Betebenner
School of Education
University of Colorado

Abstract

The zeitgeist for reform in education precipitated a number of changes in assessment. Among these are performance assessments, sometimes linked to "high stakes" accountability decisions. In some instances the trustworthiness of these decisions are based upon variance components and error variances derived through generalizability theory. Often overlooked is the fact that these statistics are subject to sampling error. This paper introduces techniques used to determine the accuracy of such statistics.

Introduction

The climate today in educational assessment is vastly different than it was twenty years ago. Where standardized achievement tests once dominated the landscape, numerous competitor assessments regularly appear, assessments that incorporate previously unheard-of formats. The nontraditional nature of these assessments is an outgrowth of the search for tests that provide a measure of validity that typical multiple choice tests can't. Validity, however, is only a partial measure of the soundness of any measurement instrument. If a test's results aren't reliable, its superior validity is of little consequence. Older notions of reliability, like with antiquated notions of validity, don't possess the sophistication necessary to determine the accuracy of such instruments. Recent initiatives linking test results to rewards and penalties at the state, local and individual level are not uncommon. Providing a measure of accuracy to such results is critical. The statistical techniques developed to address these issues are typically referred to under the comprehensive heading generalizability theory.

Introduced nearly 30 years ago, today generalizability theory is one of the major statistical techniques used in assessment. Its adaptability makes it the model of choice in a number of assessment situations. Particularly with performance assessments, where IRT based procedures are inappropriate due to the small number of items, generalizability theory is sound. Depending upon the desired use of the assessment (e.g., to measure achievement at the individual level, the school level, or perhaps at the district level), generalizability theory allows for the computation of multifaceted error estimates that provide the most complete measure of reliability available from any procedure today. This subdivision of error is particularly valuable in pinpointing exactly where to improve an assessment.

Beginning with a model that includes all relevant facets and their interactions, the initial G study provides variance component estimates for both the main effects and for interactions associated with the model.[1] Using these variance components, the D study provides variance components associated with the means for sets of sampled conditions. These in turn yield the the absolute and relative error variances, $\hat{\sigma}^2(\Delta)$ and $\hat{\sigma}^2(\delta)$, respectively. The variance components and error variances provide both specific and overall information about how accurate the results from a test are and where sources of inaccuracy arise. More specifically, variance components provide a measure of the variability for each effect were it possible to collect numerous scores for the same student or aggregate of interest over all conditions in the universe of admissible observations. Similarly, $\hat{\sigma}^2(\Delta)$ and $\hat{\sigma}^2(\delta)$ provide a composite measure of variability for the same hypothetical collection of scores.

Any prudent user of a technique like generalizability theory must recognize that variance components and error variances are statistics and, as such, are subject to sampling error. The variance of variance components and error variances represents the fluctuation one might expect in those statistics were it possible to perform multiple G studies using students, schools, tasks, and raters from the same universe of admissible observations. Whereas the estimated variance components and error variances gauge the accuracy of the instrument with respect to a single administration, the standard error associated with variance components and error variances provides a measure of fidelity for the instrument across multiple administrations. Particularly with respect to decisions involving significant consequences, the extent of sampling error must be determined and accounted for. Determining the amount of sampling error is crucial to placing any faith in the use of a statistic and, in turn, any decision based upon that statistic (Cronbach, Linn, Brennan, & Haertel, 1995).

Brennan (1992) cites two methods for estimating standard errors of variance components. Unfortunately, because the exact distributions of variance components and error variances are very complex, even unbiased estimators of the variance of variance components become difficult to interpret. A more reliable method which bypasses this distributional difficulty is to compute confidence intervals. Previously, researchers relied on two procedures, the Satterthwaite and the Welch, to derive confidence intervals for variance components.[2] Though appropriate in a number of situations (Smith, 1936; Satterthwaite, 1941, 1946; Welch, 1956), a number of inadequacies make these two procedures less than ideal for use with generalizability theory applications.

The Satterthwaite procedure assumes large values of degrees of freedom for each source of variance and, in addition, assumes the difference in degrees of freedom across the sources of variation to be small. Both the Welch and Satterthwaite procedures assume that the estimated variance component about which the confidence interval is constructed be a linear combination of expected mean squares with only positive coefficients (Burdick & Graybill, 1992, pp. 30-31). That is, if

[1] Point estimates of variance components referred to in this paper are computed using linear combinations of mean squares. The results presented in this paper apply only to variance component estimates of this type.

[2] More specifically, the confidence intervals are for linear combinations of expected mean squares with positive scalar coefficients.

$$\sigma^2(\gamma) = \sum_{k=1}^{n} c_k E(S_k^2) \qquad c_k > 0, \tag{1}$$

then the Welch and Satterthwaite procedures are appropriate for confidence interval estimation. This last restriction proves fatal to any intended application one might put these two procedures to in generalizability theory.

As an example, consider the simple one-facet $p \times i$ design. If n_i and n_p denote the sample sizes associated with each facet, then the following equations provide the three variance components as functions of expected mean squares:

$$\begin{aligned} \sigma^2(p) &= [E(S_p^2) - E(S_{pi}^2)]/n_i \\ \sigma^2(i) &= [E(S_i^2) - E(S_{pi}^2)]/n_p \\ \sigma^2(pi) &= E(S_{pi}^2) \end{aligned}$$

Clearly, the scalar coefficients of the expected mean squares are not all positive, invalidating the use of both the Welch and Satterthwaite procedures. Worse yet, in typical implementations of this design, where p and i are main effects associated with persons and items, respectively, it is not at all uncommon for n_p to be significantly larger than n_i, undermining one of the Satterthwaite assumptions.

This simple design illustrates the need for alternate confidence interval estimation procedures for variance components and error variances in generalizability theory. The purpose of this article is to present new results concerning confidence intervals of variance components to statistics encountered using fully random balanced designs in generalizability theory. These results are tested for accuracy across numerous simple and complex designs found in generalizability theory applications. Appearing to overcome many of the shortcomings of the Welch and Satterthwaite procedures with respect to variance components, this research also yields the ability to give accurate confidence intervals about the most important statistic computed in generalizability theory, $\hat{\sigma}^2(\Delta)$.

Details of recent research

The variance components and error variances encountered using various designs in generalizability theory require alternate confidence interval construction procedures – procedures allowing both signed coefficients in Equation 1 and highly variable degrees of freedom. With respect to variance components, even the simplest designs produce variance components that are linear combinations of expected mean squares with both positive and negative coefficients. With respect to error variances, linear combinations of expected mean squares with both positive and negative coefficients *and* with only positive coefficients are not uncommon. Two recently derived procedures accommodate these scenarios and, in general, provide superior confidence coefficients to those produced using the Satterthwaite and Welch procedures.

Before presenting the specific results, a brief survey on confidence intervals warrants presentation. Let α designate a prescribed significance level, then $1 - 2\alpha$ denotes the confidence coefficient of the two-sided interval $\{\sigma^2(\gamma) : L \leq \sigma^2(\gamma) \leq U\}$ and $1 - \alpha$ denotes the confidence coefficient of the two one-side intervals $\{\sigma^2(\gamma) : L \leq \sigma^2(\gamma) < \infty\}$ and $\{\sigma^2(\gamma) : 0 \leq \sigma^2(\gamma) \leq U\}$. More formally, $1 - 2\alpha = P[L \leq \sigma^2(\gamma) \leq U]$ and $1 - \alpha = P[0 \leq \sigma^2(\gamma) \leq U] = P[L \leq \sigma^2(\gamma) < \infty]$. In relatively few cases do the preceding equalities hold. When they do such intervals are called exact. In cases when equality fails to hold, intervals are called approximate. Moreover, if the probability exceeds the designated confidence coefficient, then the approximate interval is called conservative; otherwise the approximate interval is called liberal. Three types of approximate intervals are hereafter considered: Upper intervals of the form $[L, \infty)$, lower intervals of the form $[0, U]$, and two-sided intervals of the form $[L, U]$.

Graybill and Wang (1980) and Ting, Burdick, Graybill and Gui (1989) began by considering confidence intervals on positive sums of expected mean squares. Building on these results, Lu, Graybill and Burdick (1988) and Ting, Burdick, Graybill, Jeyaratnam, and Lu (1990) developed confidence interval estimation procedures on linear combinations of expected mean squares with both signs. With respect to the former, the researchers employed a modified large-sample procedure similar to that suggested by Welch (1956). Let $\sigma^2(\gamma)$ be defined as in Equation 1 and let $\hat{\sigma}^2(\gamma)$ be defined as follows:

$$\hat{\sigma}^2(\gamma) = \sum_{k=1}^{n} c_k S_k^2 \qquad c_k > 0.$$

Graybill and Wang defined the confidence interval containing $\sigma^2(\gamma)$ as follows:

$$\hat{\sigma}^2(\gamma) - \sqrt{\sum_{k=1}^{n} G_k^2 c_k^2 S_k^4} \leq \sigma^2(\gamma) \leq \hat{\sigma}^2(\gamma) + \sqrt{\sum_{k=1}^{n} H_k^2 c_k^2 S_k^4}, \tag{2}$$

where

$$G_k = 1 - \frac{1}{F_{\alpha:n_k,\infty}} \quad \text{and} \quad H_k = \frac{1}{F_{1-\alpha:n_k,\infty}} - 1$$

Using two methods, the authors tested the accuracy of these intervals against those available from the Satterthwaite and Welch procedures: (a) when $k = 2$ the authors employed numerical integration via an elegant result due to Fleiss (1971), and (b) when $k > 2$, the authors conducted simulation studies based upon 10,000 replications. In their study of two-sided intervals containing $\sigma^2(\gamma)$, Graybill and Wang demonstrated that their procedure was superior to those provided by Satterthwaite and Welch in cases where $k = 2$. Specifically, their confidence coefficients maintained at the stated α-level or were conservative across varying degrees of freedom. Later tests with $k > 2$ on one-sided intervals gave less conclusive results (Ting et al., 1989). On lower intervals the Graybill-Wang interval was superior to its Satterthwaite and Welch counterparts. In contrast, the Graybill-Wang interval was somewhat liberal with respect to upper intervals.

In cases where the variance component is a sum of expected mean squares with coefficients of both signs, analogous equations result from the modified large-sample approach just considered. Consider $\sigma^2(\gamma)$ defined by the following equation:

$$\sigma^2(\gamma) = \sum_{q=1}^{j} c_q E(S_q^2) - \sum_{r=j+1}^{k} c_r E(S_r^2), \tag{3}$$

where c_i, $1 \leq i \leq k$, are positive. If $\hat{\sigma}^2(\gamma)$ is defined by

$$\hat{\sigma}^2(\gamma) = \sum_{q=1}^{j} c_q S_q^2 - \sum_{r=j+1}^{k} c_r S_r^2,$$

where S_i represents the mean square associated with effect i, then the upper bound for a lower $1 - \alpha$ confidence interval is given by (Burdick & Graybill, 1992)

$$U = \hat{\sigma}^2(\gamma) + \sqrt{V_U}, \tag{4}$$

where

$$\begin{aligned}
V_U &= \sum_{q=1}^{j} H_q^2 c_q^2 S_q^4 + \sum_{r=j+1}^{k} G_r^2 c_r^2 S_r^4 \\
&\quad + \sum_{q=1}^{j} \sum_{r=j+1}^{k} H_{qr} c_q c_r S_q^2 S_r^2 + \sum_{r=j+1}^{k-1} \sum_{t>r}^{k} H_{rt}^* c_r c_t S_r^2 S_t^2 \\
H_q &= \frac{1}{F_{1-\alpha:n_q,\infty}} - 1 \qquad q = 1, \ldots, j \\
G_r &= 1 - \frac{1}{F_{\alpha:n_r,\infty}} \qquad r = j+1, \ldots, k \\
H_{qr} &= \frac{(1 - F_{1-\alpha:n_q,n_r})^2 - H_q^2 F_{1-\alpha:n_q,n_r}^2 - G_r^2}{F_{1-\alpha:n_q,n_r}} \quad \text{and} \\
H_{rt}^* &= \left[\left(1 - \frac{1}{F_{\alpha:n_r+n_t,\infty}} \right)^2 \frac{(n_r + n_t)^2}{n_r n_t} - \frac{G_r^2 n_r}{n_t} - \frac{G_t^2 n_t}{n_r} \right] \Big/ (k - j - 1) \\
& \qquad\qquad t = r+1, \ldots, k
\end{aligned}$$

A similarly formidable set of equations gives the lower bound for an upper $1 - \alpha$ confidence interval.

$$L = \hat{\sigma}^2(\gamma) - \sqrt{V_L}, \tag{5}$$

where

$$\begin{aligned}
V_L &= \sum_{q=1}^{j} G_q^2 c_q^2 S_q^4 + \sum_{r=j+1}^{k} H_r^2 c_r^2 S_r^4 \\
&\quad + \sum_{q=1}^{j} \sum_{r=j+1}^{k} G_{qr} c_q c_r S_q^2 S_r^2 + \sum_{q=1}^{j-1} \sum_{u>q}^{j} G_{qu}^{*} c_q c_u S_q^2 S_u^2 \\
G_q &= 1 - \frac{1}{F_{\alpha:n_q,\infty}} \qquad q = 1, \ldots, j \\
H_r &= \frac{1}{F_{1-\alpha:n_r,\infty}} - 1 \qquad r = j+1, \ldots, k \\
G_{qr} &= \frac{(F_{\alpha:n_q,n_r} - 1)^2 - G_q^2 F_{\alpha:n_q,n_r}^2 - H_r^2}{F_{\alpha:n_q,n_r}} \qquad \text{and} \\
G_{qu}^{*} &= \left[\left(1 - \frac{1}{F_{\alpha:n_q+n_u,\infty}}\right)^2 \frac{(n_q+n_u)^2}{n_q n_u} - \frac{G_q^2 n_q}{n_u} - \frac{G_u^2 n_u}{n_q} \right] \Big/ (j-1) \\
& \qquad u = q+1, \ldots, j
\end{aligned}$$

In a similar fashion as Graybill and Wang (1980) and Ting et al. (1989), Ting et al. (1990) applied two methods to test the prescribed intervals accuracy depending upon the number of terms in Equation 3. When $\sigma^2(\gamma) = c_1 E(S_1^2) - c_2 E(S_2^2)$, the authors utilized numeric integration and the corrolary due to Fleiss (1971) to determine the accuracy of the confidence coefficients. When $k > 2$ the authors conducted a simulation study based upon 10,000 replications for all possible signed combinations of the expected mean squares with variable degrees of freedom across the different signed combinations. Put in the context of the variance components encountered in generalizability theory, their results cover all variance components encountered in one and two facet designs as well as most encountered in three facet designs. Results (Burdick & Graybill, 1992, Table 3.3.1, p.42) are superior across the three types of confidence intervals. Particularly impressive were the results for the lower and two-sided intervals – these results either held at their stated α-level or were conservative. Like with the Graybill-Wang interval, the upper one-sided interval, under certain conditions, proved too liberal.

Methodology and results

Building on the results of Ting et al. (1989) and Ting et al. (1990), this paper further examines their confidence interval estimation methods and tests their appropriateness for $\hat{\sigma}^2(\Delta)$ and $\hat{\sigma}^2(\delta)$ in all possible one and two facet designs, and select three facet designs. Consider the following inequalities implied from Equations 4 and 5:

$$\hat{\sigma}^2(\gamma) - \sqrt{V_L} \le \sigma^2(\gamma) \le \hat{\sigma}^2(\gamma) + \sqrt{V_L} \qquad (6)$$

Table 1: Simulated Ranges of 95% Confidence Coefficients of Lower and Upper Intervals and 90% Two-Sided Intervals on $\hat{\sigma}^2(\Delta)$ and $\hat{\sigma}^2(\delta)$.

Error Variance	Interval	Design				
		$p \times i$	$i : p$	$p \times i \times h$	$p \times (i : h)$	$(i : p) \times h$
$\hat{\sigma}^2(\Delta)$	$[L, \infty)$	.942-.951	.941-.953	.935-.959	.938-.950	.937-.955
	$[0, U]$	.948-.952	.949-.953	.946-.955	.947-.954	.948-.953
	[L,U]	.901-.905	.900-.905	.896-.905	.899-.903	.900-.906
$\hat{\sigma}^2(\delta)$	$[L, \infty)$	.944-.950	.941-.951	.943-.954	.944-.952	.939-.952
	$[0, U]$	.947-.952	.949-.953	.946-.960	.947-.957	.951-.958
	[L,U]	.902-.908	.899-.908	.898-.908	.897-.902	.902-.905

Error Variance	Interval	Design				
		$i : (p \times h)$	$(i \times h) : p$	$i : h : p$	$p \times i \times h \times o$	$(p \times i \times h) : o$
$\hat{\sigma}^2(\Delta)$	$[L, \infty)$	.944-.954	.945-.950	.939-.947	.928-.960	.935-.956
	$[0, U]$	.946-.957	.950-.959	.946-.954	.945-.964	.948-.962
	[L,U]	.899-.909	.904-.915	.897-.902	.889-.928	.900-.916
$\hat{\sigma}^2(\delta)$	$[L, \infty)$	.944-.960	.935-.957	.943-.959	.924-.972	.934-.964
	$[0, U]$	.947-.961	.945-.959	.948-.958	.943-.977	.941-.967
	[L,U]	.896-.918	.894-.908	.897-.905	.890-.942	.893-.925

Dividing all terms by $\sum_{i=1}^{k} c_i\theta_i$ and utilizing the fact that $n_i S_i^2/\theta_i$, $i = 1, \ldots, k$, are independently distributed chi-square distributions for balanced, random, normal probability models, the inequality becomes a function of $c_i E(S_i^2)/\sum_{i=1}^{k} c_i E(S_i^2)$. Notice that if all the c_is are equal, as is the case with variance components, then the inequality reduces to a function of $E(S_i^2)/\sum_{i=1}^{k} E(S_i^2)$. This is an explicit assumption in Ting et al. (1989) and an implicit assumption in Ting et al. (1990). Though the assumption doesn't impinge on the confidence interval tests associated with variance components, it does affect tests involving $\hat{\sigma}^2(\Delta)$ and $\hat{\sigma}^2(\delta)$, since all c_is are not equal in those cases.

The study examined several combinations of n_i, $i = 1, \ldots, k$, which yield different c_i, and $\rho_i = c_i E(S_i^2)/\sum_{i=1}^{k} c_i E(S_i^2) > 0$. Because not all the c_is are equal, the simulation study placed particular emphasis on non-standard values of ρ_i. Dependent upon whether the linear combination of expected mean squares for $\hat{\sigma}^2(\Delta)$ and $\hat{\sigma}^2(\delta)$ contained all positive or both negative and positive signs, a SAS program produced 10,000 values associated with Expressions 2 and 6 using the random number generator for the gamma distribution, RANGAM, with $\alpha = \nu/2$ degrees of freedom and $\beta = 2$. These were used to determine the number of lower, upper and two-sided intervals containing $\sum_{q=1}^{j} \rho_q - \sum_{r=j+1}^{k} \rho_r$. Based on the simulation study involving 10,000 replications and the normal approximation to the binomial distribution, the chance that the simulated values differ from the actual values in magnitude by more than 0.004 is less than 5 percent. Table 1 provides results of the simulation study. The range of values for each error variance and design represents the variation across different ρ_i and n_i combinations.

The results parallel those found by Ting et al. (1989) and Ting et al. (1990). Overall, the lower one-sided and two-sided confidence intervals either held at the designated α-level or were conservative. The performance of the upper one-sided intervals is not as good, sometimes yielding liberal confidence coefficients. Because of the excellent results, both for error variances and variance components, the intervals provided by Inequalities 2 and 6 are excellent candidates for use with generalizability theory applications. Indeed, because of the excellent performance of the lower one-sided intervals and because upper bounds for variance components and error variances provide more crucial information than do lower bounds, the lower one-sided intervals provided by (2) and (6) are particularly appropriate for generalizability theory applications. The following section provides an application of these confidence intervals to a situation involving cut-scores.

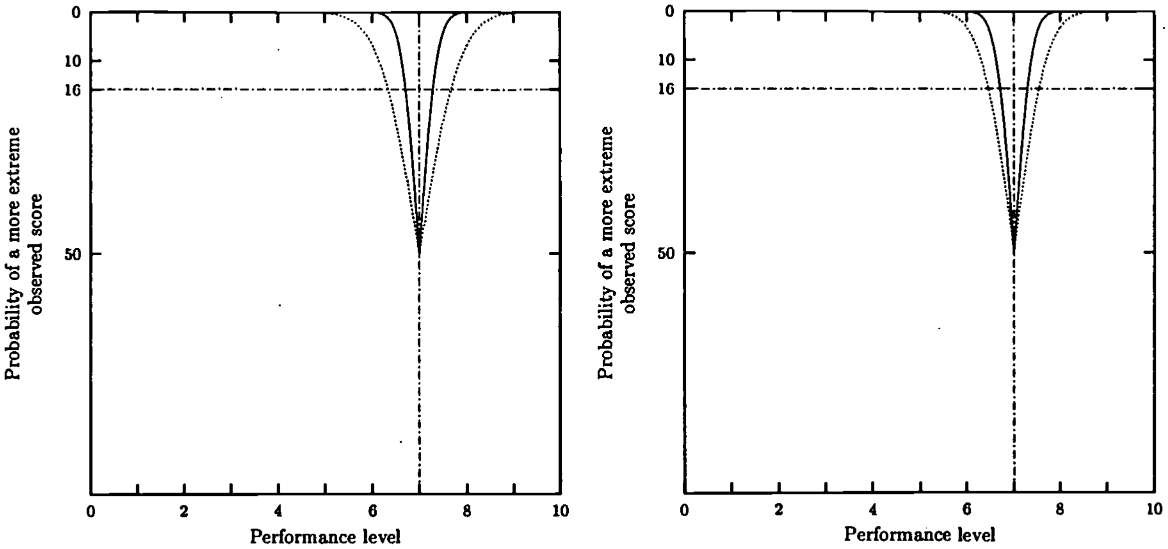
An application

In a number of circumstances, the numerical score received by a student on a test has a set of standards applied to it. In some instances, these standards provide a cut-score, above which the students pass and below which the students don't. Classification errors within this pass-or-fail scenario are directly proportional to the standard error of the measurement instrument, $\hat{\sigma}^2(\Delta)$. Clearly, as sampling error increases the variability of $\hat{\sigma}^2(\Delta)$, the potential for classification errors increases. The following is an investigation of the extent to which misclassification increases using the above tested confidence intervals.

Data collected from 100 students responding to five different essay prompts were used to test misclassification rates with respect to the variability of error variances. Three raters assessed each prompt from each student and issued scores ranging from one to nine. Using mean squares available from the GENOVA output, confidence intervals about $\hat{\sigma}^2(\Delta)$ were derived using the same number of raters and tasks originally provided in the G study, $n_i' = 5 \text{and} n_r' = 3$, as well as with $n_i' = 10 \text{and} n_r' = 3$. A *Mathematica* notebook designed for computing said intervals performed the necessary computations. Borrowing an efficient graphical depiction from Cronbach et al. (1995), results are presented in Figure 1.

The gullwings in Figures 1(a) and 1(b) are truncated normal ogives reflected about the cut-score of seven. The outer gullwing in each case represents the upper bound on the lower 95 percent confidence interval about $\hat{\sigma}^2(\Delta)$, represented by the inner gullwing. Lower confidence intervals provide the most relevant information in almost all generalizability applications for the simple reason that knowing how small $\hat{\sigma}^2(\Delta)$ is isn't nearly as important as knowing how large it is. In Figure 1(a), with $n_i' = 5$ and $n_r' = 3$, $\hat{\sigma}^2(\Delta) = .293$ whereas the outer gullwing was constructed from a normal distribution with standard deviation equal to .674. The main reason for such a large upper bound for the lower 95 percent confidence interval was a large mean square associated with the task effect. By doubling the number of tasks and leaving the number of raters fixed at three, the standard deviation associated with the outer gullwing is reduced to .544, as shown in Figure 1(b). The interaction between increasing and decreasing the number of D study sampled conditions and the upper and lower bounds for the confidence intervals is highly nonlinear and difficult to predict.

In either case, the results are troubling. If, for example, this test determined those passing a writing course versus those failing. Across repeated administrations of the test with respect to the intended universe of generalization, one should expect highly variable misclassification rates. Misclassification rates based upon $\hat{\sigma}^2(\Delta) = .3$ might be considered



(a) Depiction of $0 \leq \hat{\sigma}^2(\Delta) \leq U$ where $\hat{\sigma}^2(\Delta) = .293$ and $U = .674$

(b) Depiction of $0 \leq \hat{\sigma}^2(\Delta) \leq U$ where $\hat{\sigma}^2(\Delta) = .293$ and $U = .544$

Figure 1. Absolute standard error with respect to the 95 percent lower confidence interval in two D studies.

tenable under certain circumstances – if a score of 6.5 represents the highest failing score, then one should expect less than five percent of those students with a passing score of seven to fail. Yet, misclassification rates based upon $\hat{\sigma}^2(\Delta) = .6$ are almost always indefensible – nearly 20 percent of those student with a passing score of seven fail. The importance here being to recognize the amount of variability possible across administrations and its impact upon misclassification.

Conclusions

Sampling error in variance components and error variances used in generalizability theory is often overlooked when making determinations about the accuracy of a given test. This paper attempts to address this shortcoming by presenting a number of results with respect to confidence intervals about linear combinations of expected mean squares appropriate for generalizability theory. Simulation results indicate that these intervals, particularly the two-sided and one-sided lower intervals, are accurate or conservative both in simple and complex designs with varying amounts of difference in degrees of freedom across effects. In practice the lower intervals might prove to be most relevant since knowing how potentially large an error variance is is important, while knowing how small it is isn't. The potential for grossly underestimating standard errors seems a real possibility. More analysis using larger data sets is needed.

References

Brennan, R. L. (1992). *Elements of generalizability theory* (Revised ed.). Iowa City, Iowa: ACT Publications.

Burdick, R. K., & Graybill, F. A. (1992). *Confidence intervals on variance components*. New York: Marcel Dekker, Inc.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). *Generalizability analysis for educational assessments* (Evaluation Comment, Summer 1995). Los Angeles: The National Center for Research on Evaluation, Standards, and Student Testing.

Fleiss, J. L. (1971). On the distribution of a linear combination of independent chi squares. *Journal of the American Statistical Association*, *66*(333), 142–144.

Graybill, F. A., & Wang, C.-M. (1980). Confidence intervals on nonnegative linear combinations of variances. *Journal of the American Statistical Association*, *75*(372), 869–873.

Lu, T.-F. C., Graybill, F. A., & Burdick, R. K. (1988). Confidence intervals on a difference of expected mean squares. *Journal of Statistical Planning and Inference*, *18*, 375–380.

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.

Smith, H. F. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council of Scientific and Industrial Research*, *9*, 211–212.

Ting, N., Burdick, R. K., Graybill, F. A., & Gui, R. (1989). One-sided confidence intervals on nonnegative sums of variance components. *Statistics and Probability Letters*, *8*(2), 129–135.

Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnam, S., & Lu, T.-F. C. (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computation and Simulation*, *35*, 135-143.

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

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM029112

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Improved confidence interval estimation for variance components and error variances in generalizability theory	
Author(s): Damian W. Betebenner	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Damian Betebenner	Printed Name/Position/Title: Damian Betebenner / grad. student	
Organization/Address: University of Colorado at Boulder School of Education Boulder, CO 80309-0249	Telephone: 303-492-8976	FAX:
	E-Mail Address: betebenn@ucsu.colorado. edu	Date: 6-18-98

(over)

ERIC®

Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

March 20, 1998

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae.net.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (424)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1998/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1] If you are an AERA chair or discussant, please save this form for future use.