British Journal of Mathematical and Statistical Psychology (2009), 62, 385–400

The British Psychological Society
www.bpsjournals.co.uk

An effect size index for comparing two independent alpha coefficients

Hsin-Yun Liu and Li-Jen Weng*
Department of Psychology, National Taiwan University, Taipei, Taiwan

Since Cronbach proposed the α coefficient in 1951, researchers have contributed to the derivation of its sampling distribution and the testing of related statistical hypotheses. Yet, there has been no research on effect size index relevant to coefficient α to our knowledge. Considering the importance of effect size in understanding quantitative research findings, we therefore developed an effect size index Δ for the comparison of two independent αs with equal test length based on the asymptotic distribution of $(1/2)\ln(1-\hat{\alpha})$ under the assumptions of normality and compound symmetry. Simulations indicated that the index was applicable when the sample size was at least 100. The robustness of the derived index when the required assumptions were violated was also explored. It is suggested that the index should be applicable in most cases of unequal test lengths and could be extended to non-normally distributed component scores. Moreover, a small simulation was conducted to explore the behaviour of Δ with correlated errors, a frequently studied situation violating the assumption of compound symmetry. The proposed index was found to be robust unless a large number of highly correlated errors were present in the data.

1. Introduction

Coefficient α has been the most popular measure for assessment of the reliability of test scores. Hogan, Benjamin, and Brezinski (2000) reviewed the reliability information provided for 696 tests in the Directory of Unpublished Experimental Mental Measures, Volume 7. This sample was drawn from 2,078 tests published in 37 professional journals between 1991 and 1995. Out of the 801 reliabilities reported, coefficient α enjoyed the highest frequency of use at 66.5%, followed by test–retest reliability at 19%.

Since Cronbach introduced coefficient α in 1951, many of his successors have contributed to the development of its sampling distribution and the related hypothesis testing procedures (Alsawalmeh & Feldt, 1994a, 1994b, 1999; Feldt, 1965,

*Correspondence should be addressed to Professor Li-Jen Weng, Department of Psychology, National Taiwan University, Taipei 106, Taiwan (e-mail: ljweng@ntu.edu.tw).

DOI:10.1348/000711008X315518

1969, 1980; Feldt, Woodruff, & Salih, 1987; Hakstian & Whalen, 1976; Kristof, 1963; van Zyl, Neudecker, & Nel, 2000; Woodruff & Feldt, 1986). Yet, there has been no research on effect size indices relevant to coefficient α to our knowledge. Effect size is an important measure for understanding practical significance of the results from quantitative research (APA, 1994, 2001; Wilkinson & the APA Task Force on Statistical Inference, 1999). In this study, we proposed an index to define the effect size for the comparison of two independent αs from tests of equal length. The proposed effect size index would be useful in various research scenarios. One useful application of the effect size index developed is to compare the α coefficients obtained from tests of the same components but with different formats based on two independent samples. The effect size index could be used to select a scale format that yields test scores of higher internal consistency reliability for further research. Although a test statistic has been developed to test the equality of two independent αs (Charter & Feldt, 1996; Feldt, 1969), the conclusions drawn from the testing are affected by sample sizes.

Feldt (1969) derived the test statistic W for testing the equality of two independent αs. Let $\hat{\alpha}_1$ and $\hat{\alpha}_2$ be the sample estimates of α from two independent samples ($\hat{\alpha}_1 < \hat{\alpha}_2$) with sample size of N_1 and N_2, respectively. The W statistic, defined as $W = (1 - \hat{\alpha}_1)/(1 - \hat{\alpha}_2)$, is approximately distributed as a central F distribution with $N_2 - 1$ and $N_1 - 1$ degrees of freedom. Charter and Feldt (1996) presented an example to illustrate this test procedure. In the example, $\hat{\alpha}_1 = .71$ and $\hat{\alpha}_2 = .78$ with $N_1 = 151$ and $N_2 = 41$. The resulting W statistic of 1.318 yielded a p value of .242. Accordingly, the null hypothesis $H_0 : \alpha_1 = \alpha_2$ could not be rejected. However, if the sample size was raised to 250 for both samples, the p value would become .029 and would alter the conclusion of the test with the level of significance set at .05. The obtained p value associated with the testing decreased as the sample size increased. The example demonstrates the need for the development of an effect size index for comparing two independent αs.

Cohen (1988, 1992) defined the effect size to be the parameter discrepancy between the null hypothesis and the alternative hypothesis. Effect size is a critical tool in quantitative research for supplementing the limitations of null hypothesis significance testing (Cohen, 1990, 1994; Hubbard & Ryan, 2000; Kirk, 1996). As shown in the previous illustration, one of the frequently discussed limitations in null hypothesis significance testing is its dichotomous decision based on a single p value. The p value, however, decreases as the sample size increases as illustrated in the above example. In contrast to p values, effect size is independent of sample sizes and it enables researchers to evaluate the stability of results across studies. Accordingly, effect size has gained increasing attention from researchers in recent years (APA, 1994, 2001; Wilkinson & the APA Task Force on Statistical Inference, 1999).

The definition of the effect size index plays an important role in applications of effect size. Cohen is the major contributor to the development of effect size indices (Huberty, 2002). Under various statistical hypotheses, Cohen (1988, 1992) defined the effect size index and its corresponding values for large, medium, and small effect sizes. Take the comparison of means from two normal distributions with means μ_1 and μ_2 and common variance σ^2 as an example. The most frequently used effect size index under this condition is Cohen's d, $d = (\mu_1 - \mu_2)/\sigma$ for $\mu_1 > \mu_2$. Cohen's d reflects the difference of the mean values between two independent populations in a standard unit. Furthermore, Cohen proposed a d value of .8, .5, and .2 to represent large, medium, and small sizes of effect, respectively.

The present study applied the concept of Cohen's *d* to define the effect size index for the comparison of two independent αs based on large sample theory. The appropriateness of the derived effect size index under small samples was further investigated by a simulation study. The applicability of the derived index when the required assumptions were violated was also discussed.

2. An effect size index for two independent αs with equal test length

The sample coefficient α, $\hat{\alpha}$, for a test of K components (Cronbach, 1951; Cronbach & Shavelson, 2004) is given by

$$\hat{\alpha} = \frac{K}{K-1}\left[1 - \frac{trS}{\mathbf{J}'S\mathbf{J}}\right], \tag{1}$$

where S is the sample covariance matrix of the K components, trS refers to the trace of matrix S, and $\mathbf{J}$ is a $K \times 1$ vector of ones. Under the assumption of essential tau equivalence, $\hat{\alpha}$ is an estimate of the reliability of test scores (Novick & Lewis, 1967).

After the introduction of α, researchers have worked on the sampling distribution of $\hat{\alpha}$. Assuming the K components to be normally distributed with the population covariance matrix Σ being of compound symmetry, Feldt (1965) showed that the sampling distribution of $\hat{\alpha}$ from a sample of N individuals was as follows:

$$\frac{1-\hat{\alpha}}{1-\alpha} \sim F_{(N-1)\times(K-1),(N-1)}, \tag{2}$$

with α being the population value of the coefficient. The mean and the variance of $\hat{\alpha}$ were

$$E(\hat{\alpha}) = \frac{-2}{N-3} + \frac{N-1}{N-3}\alpha, \text{ and} \tag{3}$$

$$\mathrm{Var}(\hat{\alpha}) = \frac{2(1-\alpha)^2(N-1)\{[(N-1)K]-2\}}{(K-1)(N-3)^2(N-5)}, \text{ respectively.} \tag{4}$$

Furthermore, van Zyl *et al.* (2000) obtained the asymptotic distribution of $\hat{\alpha}$ as

$$\sqrt{N-1}(\hat{\alpha}-\alpha) \xrightarrow{L} N\left(0, \frac{2(1-\alpha)^2K}{K-1}\right), \tag{5}$$

where '*L*' represented converging in distribution (e.g. Rao, 1973).

From the sampling distribution and the asymptotic distribution of $\hat{\alpha}$, it is noted that the variance of $\hat{\alpha}$ depends on the population value of α. The difference between two $\hat{\alpha}$s is thus not an optimal measure of effect size due to unequal scale unit.

van Zyl *et al.* (2000) further derived the asymptotic distribution of $(1/2)\ln(1-\hat{\alpha})$ under the assumptions of normality and compound symmetry to be as follows:

$$\sqrt{N-1}[(1/2)\ln(1-\hat{\alpha}) - (1/2)\ln(1-\alpha)] \xrightarrow{L} N\left(0, \frac{K}{2(K-1)}\right). \tag{6}$$

From equation (6), $(1/2)\ln(1-\hat{\alpha})$ is asymptotically normally distributed with a mean of $(1/2)\ln(1-\alpha)$ and its variance is independent of the population α value. Therefore, for independent α_1 and α_2 of equal test length of K components, the difference between independent $(1/2)\ln(1-\alpha_1)$ and $(1/2)\ln(1-\alpha_2)$ can be a measure of the effect size

for comparing two αs. Therefore, we propose the effect size index Δ for comparing two independent αs ($\alpha_1 < \alpha_2$ for Δ to be positive) with equal test length as

$$\Delta = \frac{(1/2)\ln(1-\alpha_1) - (1/2)\ln(1-\alpha_2)}{\sqrt{K/(2(K-1))}}. \quad (7)$$

The derivation of this effect size index Δ is parallel to the formulation of Cohen's *d*. For normally distributed *x*, $\sqrt{N}(\bar{x} - \mu) \sim N(0, \sigma^2)$. With common variance for independent μ_1 and μ_2 ($\mu_1 > \mu_2$), Cohen's *d* of $(\mu_1 - \mu_2)/\sigma$ represents the distance between two mean values divided by the common population standard deviation. Cohen's *d* is still applicable for non-normally distributed *x* as long as the samples are sufficiently large because sample means are normally distributed asymptotically. Following the idea of *d*, the effect size index Δ for comparing two independent αs is therefore defined by the asymptotic distribution of $(1/2)\ln(1-\hat{\alpha})$. Accordingly, Cohen's suggestion for large–medium–small effect size values of *d* can also be applied to Δ. A Δ greater than .8, .5, and .2 would suggest a large, medium, and small effect size, respectively.

Equation (7) shows that Δ is a function of α_1 and α_2. Hence, for a given *K*, α_2 can be derived with known Δ and α_1. Table 1 illustrates the relationship between Δ, α_1, and α_2 for a test of 30 components. The results indicate that the effect size, a given difference between two αs, depends on the magnitude of the αs. The higher the α, the larger effect size the difference suggests. Take the difference around .1 between two αs as an example. This difference indicates a small effect size for $\alpha_1 = .60$, a medium effect size for $\alpha_1 = .80$, and a large effect size for $\alpha_1 = .85$. Similar tables to Table 1 for tests of 10 components and tests with *K* going to infinity were also obtained, though not

Table 1. α_2 as a function of α_1 and Δ for a test of 30 components ($\alpha_2 > \alpha_1$)

	Δ								
α_1	.1	.2	.3	.4	.5	.6	.7	.8	.9
.00	.13	.25	.35	.44	.51	.58	.63	.68	.73
.05	.18	.29	.38	.47	.54	.60	.65	.70	.74
.10	.22	.32	.42	.49	.56	.62	.67	.72	.75
.15	.26	.36	.45	.52	.59	.64	.69	.73	.77
.20	.31	.40	.48	.55	.61	.66	.71	.75	.78
.25	.35	.44	.51	.58	.63	.68	.73	.76	.79
.30	.39	.47	.55	.61	.66	.70	.74	.78	.81
.35	.44	.51	.58	.63	.68	.73	.76	.79	.82
.40	.48	.55	.61	.66	.71	.75	.78	.81	.84
.45	.52	.59	.64	.69	.73	.77	.80	.83	.85
.50	.57	.62	.68	.72	.76	.79	.82	.84	.86
.55	.61	.66	.71	.75	.78	.81	.84	.86	.88
.60	.65	.70	.74	.77	.81	.83	.85	.87	.89
.65	.70	.74	.77	.80	.83	.85	.87	.89	.90
.70	.74	.77	.81	.83	.85	.87	.89	.91	.92
.75	.78	.81	.84	.86	.88	.89	.91	.92	.93
.80	.83	.85	.87	.89	.90	.92	.93	.94	.95
.85	.87	.89	.90	.92	.93	.94	.95	.95	.96
.90	.91	.92	.94	.94	.95	.96	.96	.97	.97
.95	.96	.96	.97	.97	.98	.98	.98	.98	.99

presented. The resulting α_2 under the same α_1 and Δ differs from the corresponding entry in Table 1 by less than .01. Therefore, Table 1 can serve as a quick reference for the estimation of the effect size between two αs of equal test length.

3. Simulations on effect size index Δ against small samples

The proposed effect size index Δ was defined by the asymptotic distribution of $(1/2)\ln(1-\hat{\alpha})$. However, the behaviour of Δ and $(1/2)\ln(1-\hat{\alpha})$ under small samples was unknown and warranted further investigations. Prior to the examination of Δ against small samples, the behaviour of $(1/2)\ln(1-\hat{\alpha})$ with small samples was first studied by a simulation.

3.1. Small sample behaviour of $(1/2)\ln(1-\hat{\alpha})$

The sampling distribution of $\hat{\alpha}$ for a test of parallel components was affected by test length, component reliability, and sample size (Feldt, 1965, 1969). Hence, these three factors were manipulated in the simulation to examine the sampling distribution of $(1/2)\ln(1-\hat{\alpha})$ under small samples. The length of a test included the conditions of 10, 15, 20, and 25 components. Component reliability (ρ) was defined as the ratio of the true-score variance to the observed score variance for each component. With component reliability so defined, coefficient α for a test of K components equalled $K\rho/(1+(K-1)\rho)$. Component reliability in this simulation was chosen to be .15, .25, and .35. Accordingly, the resulting α coefficients for the entire test ranged from .64 to .98, with values widely applicable throughout much research. In order to understand the behaviour of $(1/2)\ln(1-\hat{\alpha})$ under small samples, sample size varied across seven conditions ranging from 30 to 500 (30, 60, 100, 200, 300, 400, and 500). One thousand replication samples were generated for each of the 84 conditions (4 test length × 7 sample size × 3 component reliability) to examine whether the sampling distribution of $(1/2)\ln(1-\hat{\alpha})$ was normal in small samples.

The test scores were generated according to the classical test theory and were simulated by a SAS program. Let $\mathbf{x_j}$ be a $K \times 1$ vector consisting of the scores of the K components for observation j, $\mathbf{x_j} = [x_{1j}, x_{2j}, \ldots, x_{Kj}]'$, $j = 1, \ldots, N$, where $x_{ij} = T_{ij} + E_{ij}$, with T_{ij} and E_{ij} being the true and the error scores on component i for observation j. The components were assumed to be parallel such that they had the same true score (T_j for observation j) and equal error variance. In other words, every component score x_{ij} was the sum of T_j and E_{ij} and the covariance matrix of $\mathbf{x_j}$ met the requirement of compound symmetry for $(1/2)\ln(1-\hat{\alpha})$ to have the asymptotic distribution derived by van Zyl *et al.* (2000). In order to yield the desired test reliability, T_j and E_{ij} were decomposed as $T_j = \sqrt{\rho}t_j$ and $E_{ij} = \sqrt{1-\rho}e_{ij}$, with ρ being the component reliability and t_j and e_{ij} being distributed as $N(0,1)$, generated by the RANNOR function in SAS.

The Wilk–Shapiro W test suggested by Thode (2002) was used to test whether the distribution of $(1/2)\ln(1-\hat{\alpha})$ over 1,000 replications followed a normal distribution under small samples. The t test for single mean and the chi-squared test for variance were further employed to test whether the mean of $(1/2)\ln(1-\hat{\alpha})$ equalled $(1/2)\ln(1-\alpha)$ and the variance of $(1/2)\ln(1-\hat{\alpha})$ equalled $K/(2(N-1)(K-1))$. Table 2 presents the p values associated with the test of normality (p_1) and the test of mean (p_2) under each condition. As shown in Table 2, the hypothesis that $(1/2)\ln(1-\hat{\alpha})$ followed a normal distribution for samples size greater than 100 could

Table 2. *p* values for tests of normality and expected value of (1/2)ln(1 − $\hat{\alpha}$)

		Component reliability					
		.15		.25		.35	
K	*N*	p_1	p_2	p_1	p_2	p_1	p_2
10	30	.31	.00	.00	.08	.00	.00
	60	.32	.05	.00	.04	.00	.01
	100	.05	.26	.01	.02	.34	.39
	200	.81	.76	.25	.96	.24	.03
	300	.27	.03	.21	.04	.81	.27
	400	.30	.26	.79	.86	.47	.14
	500	.86	.06	.88	.99	.56	.53
15	30	.29	.00	.00	.00	.24	.00
	60	.97	.00	.00	.00	.55	.00
	100	.25	.44	.54	.00	.02	.51
	200	.42	.31	.18	.27	.86	.63
	300	.52	.88	.31	.02	.95	.19
	400	.73	.31	.14	.49	.54	.60
	500	.02	.03	.80	.13	.30	.27
20	30	.03	.00	.00	.01	.00	.00
	60	.01	.01	.15	.05	.01	.02
	100	.07	.14	.62	.99	.18	.07
	200	.04	.20	.54	.35	.82	.28
	300	.52	.97	.33	.29	.26	.44
	400	.64	.12	.53	.04	.48	.31
	500	.21	.42	.79	.13	.83	.80
25	30	.03	.00	.00	.00	.00	.06
	60	.09	.01	.06	.00	.01	.00
	100	.15	.13	.01	.05	.57	.03
	200	.24	.69	.12	.24	.21	.74
	300	.12	.65	.86	.03	.54	.03
	400	.50	.36	.05	.13	.32	.08
	500	.71	.16	.56	.49	.17	.74

Note. $p_1 = p$ values for tests of normality of (1/2)ln(1 − $\hat{\alpha}$); $p_2 = p$ values for the test of expected value of (1/2)ln(1 − $\hat{\alpha}$); *K*= test length; *N*= sample size.

not be rejected with all the *p* values being greater than .01. When the sample size was less than 100, there might be significant discrepancies between $(1/2)\ln(1-\hat{\alpha})$ and its expected value $(1/2)\ln(1-\alpha)$.[1] The results of the tests of the variance being equal to $K/(2(N-1)(K-1))$ were all non-significant ($p > .01$) and therefore not presented.

[1] *We would like to thank one anonymous reviewer for pointing out the erratic pattern of the* p *values associated with* N *in Table 2. It was suspected that the erratic pattern was because only one set of data, as representing one simulated sampling distribution of* $(1/2)\ln(1-\hat{\alpha})$*, was generated for each test. If several sets of data were generated for each* N*, we would expect the mean* p *values to increase with* N*. We took the case of 10 components and component reliability being .15 with sample sizes ranging from 30 to 500 (30, 60, 100, 200, 300, 400, and 500) as an example to test our conjecture. Ten sampling distributions of* $(1/2)\ln(1-\hat{\alpha})$ *for each* N *were simulated. The results indicated that for any given sample size the* p *values across the 10 simulated sampling distributions varied substantially and hence yielded large standard deviations. Yet, the mean* p *values as supporting our conjecture appeared to increase with* N *as expected.*

Feldt (1969) concluded that with the sample size of at least 100, the derived F distribution with $N_2 - 1$ and $N_1 - 1$ degrees of freedom could be used to approximate the sampling distribution of W statistic for testing the difference between two independent αs with no need for adjustment in degrees of freedom. Similar results seemed to apply to the behaviour of $(1/2)\ln(1 - \hat{\alpha})$ in small samples. According to Table 2, the behaviour of $(1/2)\ln(1 - \hat{\alpha})$ appears robust against small sample sizes. More specifically, $(1/2)\ln(1 - \hat{\alpha})$ is approximately normally distributed with the desired expected value and variance when the sample size is at least 100.

3.2 Small sample behaviour of Δ

Cohen (1988) provided the measures of non-overlap U_1, U_2, and U_3 for interpreting large, medium, and small effects for d. These measures of percent non-overlap were used to evaluate the appropriateness of the effect size index Δ under small samples. The measures were developed on the basis of two normally distributed populations with equal variability. Let the mean of population B be greater than the mean of population A. Measure U_1 is the percentage of non-overlapping under both populations combined (Cohen, 1988, p. 21). Measure U_2 is 'the percentage in the B population that exceeds the same percentage in the A population' (Cohen, 1988, p. 21). Measure U_3 is 'the percentage of the A population which the upper half of the cases of the B population exceeds' (Cohen, 1988, p. 21). The larger the effect, the greater the U measures.

The U measures associated with different values of Cohen's d can be computed by the cumulative standard normal distribution function $P_x : U_3 = P_d$, $U_2 = P_{d/2}$, and $U_1 = (2U_2 - 1)/U_2$ (Cohen, 1988, p. 23). For example, for a small effect size of $d = .2$, two normally distributed populations with equal variability have only 14.7% (U_1) of their combined area non-overlapped. And the highest 54% in population B exceeds the lowest 54% in population A with measure U_2 therefore being equal to 54%. Also, the upper half of the B population exceeds 57.9% of the population A, implying that $U_3 = 57.9\%$. Similarly, when $d = .5$ as the operational definition of a medium effect size, $U_1 = 33\%$, $U_2 = 59.9\%$, and $U_3 = 69.1\%$. When $d = .8$ as representing a large effect size, $U_1 = 47.4\%$, $U_2 = 65.5\%$, and $U_3 = 78.8\%$.

Because Δ is defined by the asymptotic distribution of $(1/2)\ln(1 - \hat{\alpha})$, values of U_1, U_2, and U_3 of Δ are expected to approach the values of U_1, U_2, and U_3 associated with d in large samples. More specifically, a Δ equal to .8, .5, or .2 is expected to yield the U measures given above. Therefore, the discrepancies between the U measures of Δ and the corresponding entities of d under large, medium, and small effects can be used to assess the appropriateness of the effect size index Δ under small samples. In order to evaluate the discrepancies, the data generated in the previous simulations were taken as the sampling distribution of $(1/2)\ln(1 - \hat{\alpha}_1)$, and an additional set of data were simulated to represent the sampling distribution of $(1/2)\ln(1 - \hat{\alpha}_2)$. The same data generation procedures of 1,000 replications were employed except with the component reliability ρ_2 carefully chosen. The component reliability ρ_2 was chosen so that the associated Δ value satisfied the required magnitude of .8, .5, or .2 in the population. For example, for $\rho_1 = .15$ and $K = 10$, to simulate a medium effect size, ρ_2 was set at .325 so that the resulting Δ yielded a value of .5.

A histogram with 50 classes based on the simulated values of $(1/2)\ln(1 - \hat{\alpha}_1)$ and $(1/2)\ln(1 - \hat{\alpha}_2)$ was constructed to estimate U_1. The histogram was created from 2,000 values simulated under a large effect size. The average class interval of the histogram was around 0.11. U_1 therefore equalled the percentage of non-overlapping

area over the total area under the entire histogram. The estimation of U_2 and U_3 for different values of Δ was straightforward and could be directly obtained from the sampling distributions of $(1/2)\ln(1-\hat{\alpha}_1)$ and $(1/2)\ln(1-\hat{\alpha}_2)$ based on the 1,000 replications.

Tables 3–5 present the deviations of the U measures between Δ and d under large, medium, and small size of effect, respectively. The deviations associated with U_2 and U_3 were all small unless the sample size was as little as 30. In other words, when Δ was set at .8, .5, and .2, the estimated values of U_2 and U_3 approached those of d at .8, .5, and .2. The deviations of the U_1 for large and medium effect sizes seemed small as well. However, the estimated U_1 for a small effect size of Δ might differ from the expected magnitude of U_1 of d at .2. A small effect size implied that the two sampling distributions were close. Accordingly, the majority of the two sampling distributions overlapped. As a result, the U_1 measure, defined as the percent area non-overlapping over the entire area covered by the two sampling distributions, might be easily affected by the distribution of

Table 3. Deviation between U measures of Δ and d for a large effect size (U_Δ–U_d)

		Component reliability								
		.15			.25			.35		
K	*N*	U_1	U_2	U_3	U_1	U_2	U_3	U_1	U_2	U_3
10	30	.06	.03	.03	−.01	−.01	−.02	.02	.01	.03
	60	.02	.01	.01	.01	.00	−.02	−.01	−.01	−.01
	100	.02	.00	.01	.05	.02	.02	−.02	−.01	.00
	200	−.02	−.01	−.02	.01	.01	.03	.02	.01	.00
	300	.03	.01	.02	.02	.00	.02	.04	.02	.00
	400	.03	.02	.03	.01	−.01	−.02	.03	.01	.02
	500	.01	.00	.00	.00	.00	−.04	.02	.01	.00
15	30	.02	.00	.01	.01	−.01	−.01	.03	.01	−.01
	60	.04	.01	.00	−.02	−.01	−.01	.03	.01	.00
	100	−.01	.00	−.02	.05	.02	.03	−.03	−.01	−.03
	200	.02	.00	.02	.04	.02	.02	.01	.00	.00
	300	.01	.01	−.01	.02	.01	.00	.00	−.01	−.02
	400	.06	.02	.02	−.01	.00	.00	.01	.00	.01
	500	.01	.00	−.02	−.01	.00	.00	.00	.00	−.01
20	30	.02	.01	.02	−.01	−.01	−.03	.01	.00	.01
	60	.01	.00	−.02	−.01	.00	−.02	.02	.01	.02
	100	−.02	−.01	−.02	−.02	−.02	−.04	.01	.01	.02
	200	.00	.00	−.01	−.01	.00	.00	.01	.00	.00
	300	−.02	−.01	.00	.01	−.01	.01	.00	−.01	.02
	400	.08	.03	.05	.00	.00	.00	.00	.00	−.02
	500	−.01	−.01	.00	−.03	−.01	−.04	−.01	−.01	−.01
25	30	.03	.01	−.01	.00	−.01	−.01	−.06	−.02	−.05
	60	−.01	−.01	−.01	−.02	−.01	.00	.03	.01	.02
	100	−.01	−.01	−.02	−.01	−.01	−.01	.00	.00	−.03
	200	.00	−.01	.01	.01	.00	−.02	.00	.00	.00
	300	−.04	−.02	−.03	.04	.02	.01	.03	.01	.01
	400	.01	−.01	.01	.02	.01	.01	.00	.00	−.01
	500	.03	.02	.01	.00	.00	−.04	−.03	−.02	−.03

Note. $U_1 = .474$, $U_2 = .655$, and $U_3 = .788$ for a large effect size; *K*= test length; *N*= sample size.

Table 4. Deviation between *U* measures of Δ and *d* for a medium effect size (U_Δ–U_d)

		Component reliability								
		.15			.25			.35		
K	*N*	U_1	U_2	U_3	U_1	U_2	U_3	U_1	U_2	U_3
10	30	.07	.02	.05	−.04	−.01	−.02	.04	.01	.03
	60	.03	.00	.00	.02	.01	−.02	−.01	−.01	−.02
	100	.02	.00	.00	.04	.00	.00	−.02	−.01	−.02
	200	.00	.00	−.02	.04	.00	.02	.01	.00	.00
	300	.04	.01	.02	.03	.01	.02	.07	.02	.02
	400	.06	.02	.04	−.03	−.01	−.02	.04	.01	.03
	500	.01	.00	.00	.02	−.01	−.02	.03	.01	.00
15	30	.02	.01	.00	.00	−.01	.00	.04	.01	.00
	60	.03	.01	.01	.00	−.01	−.01	.04	.01	.02
	100	−.01	−.01	−.01	.07	.02	.04	−.02	−.02	−.02
	200	.04	.00	.02	.05	.01	.03	.01	.00	.01
	300	.02	.01	.00	.02	.01	−.01	−.02	−.01	−.01
	400	.07	.02	.04	.00	−.01	.00	.04	.01	.02
	500	.03	−.01	−.02	.01	.00	.00	.00	.00	.00
20	30	.05	.01	.02	.02	−.01	−.01	.02	.00	.01
	60	.01	−.04	−.01	.01	.00	−.01	.03	.01	.01
	100	−.01	−.01	−.03	−.03	−.02	−.03	.01	.00	.01
	200	.02	.00	−.01	.00	.01	.00	.03	.00	.00
	300	−.01	−.01	−.01	.03	.00	.02	.02	−.01	.00
	400	.09	.03	.05	.04	−.01	.00	.01	.00	−.01
	500	.02	−.02	.00	−.04	−.02	−.03	−.03	−.02	−.02
25	30	.03	.01	.00	.02	.00	.00	−.05	−.02	−.06
	60	.03	.00	.00	−.01	−.01	−.01	.04	.01	.02
	100	.00	−.01	.00	.00	−.01	−.02	.01	−.01	−.03
	200	.02	.00	.00	.02	.01	−.01	.01	.00	.00
	300	−.03	−.02	−.03	.05	.02	.03	.04	.01	.03
	400	.02	.00	.02	.02	.00	.01	.00	.00	−.01
	500	.05	.02	.03	.00	.00	−.02	−.02	−.02	−.04

Note. $U_1 = .330$, $U_2 = .599$, and $U_3 = .691$ for a medium effect size; *K*= test length; *N*= sample size.

the simulated data and deviated from the expected magnitude. Yet, considering the small sample behaviours of $(1/2)\ln(1-\hat{\alpha})$ and Δ simultaneously, the proposed Δ appeared robust as an effect size index for comparing two independent αs of equal test length in small samples, if the sample size was at least 100.

4. Discussion

Coefficient α is one of the most widely used measures for assessing the reliability of test scores. The present research proposes an effect size index Δ for comparing two independent αs based on the asymptotic distribution of $(1/2)\ln(1-\hat{\alpha})$. The overall findings from the simulations indicate that the proposed effect size index Δ is applicable for samples of at least 100 observations. In this study, Δ was derived on the basis of three assumptions: two tests of equal number of components, normally

Table 5. Deviation between *U* measures of Δ and *d* for a small effect size (U_Δ–U_d)

		Component reliability								
		.15			.25			.35		
K	*N*	U_1	U_2	U_3	U_1	U_2	U_3	U_1	U_2	U_3
10	30	.16	.02	.07	.06	−.01	−.02	.10	.01	.03
	60	.08	−.01	−.01	.05	.00	−.01	.06	.00	−.01
	100	.09	−.01	−.01	.07	.01	.01	.06	−.01	−.03
	200	.07	−.01	.00	.09	.01	.02	.02	.00	.00
	300	.09	.02	.03	.08	.01	.03	.09	.02	.03
	400	.09	.01	.03	.04	−.01	−.01	.09	.01	.04
	500	.09	−.01	−.01	.06	−.01	−.02	.06	.00	.01
15	30	.08	.01	.01	.05	.00	.00	.09	.00	.00
	60	.08	.01	.00	.06	−.02	−.03	.08	.01	.03
	100	.05	−.01	−.02	.12	.02	.04	.03	−.01	−.03
	200	.06	.01	.03	.06	.01	.03	.07	.00	−.01
	300	.07	.01	.01	.04	.00	.00	.04	−.01	−.02
	400	.12	.02	.04	.02	−.01	−.01	.12	.01	.02
	500	.12	.00	−.01	.08	.01	.00	.10	.00	−.01
20	30	.11	.02	.03	.04	.00	−.01	.06	.00	.01
	60	.04	.00	−.01	.09	−.01	−.02	.10	.01	.02
	100	.03	−.01	−.01	.05	−.01	−.03	.08	.01	.00
	200	.06	.01	.00	.09	.00	.01	.05	.01	.01
	300	.06	−.01	−.01	.07	.01	.01	.06	.00	.00
	400	.15	.03	.05	.05	−.01	−.02	.07	.00	.02
	500	.12	−.02	−.02	.03	−.02	−.03	.07	−.01	−.02
25	30	.06	−.01	−.01	.09	.03	.05	.05	−.03	−.06
	60	.09	.00	.01	.06	.00	−.01	.08	.01	.02
	100	.06	.00	−.01	.03	−.01	−.02	.10	−.02	−.03
	200	.09	.01	.00	.06	.01	.01	.05	−.01	−.01
	300	.06	−.02	−.05	.10	.01	.03	.10	.01	.01
	400	.08	.01	.01	.07	.00	.00	.04	−.01	.00
	500	.10	.01	.03	.05	.00	−.02	.05	−.02	−.05

Note. $U_1 = .147$, $U_2 = .54$, and $U_3 = .579$ for a small effect size; *K*= test length; *N*= sample size.

distributed component scores, and compound symmetry of the covariance matrix of the components. The applicability of Δ when violation of these assumptions is present was discussed and explored.

When two tests consist of equal number of components, $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha}_1)$ and $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha}_2)$ share the same population asymptotic variance of $K/(2(K-1))$. As a result, the effect size index Δ represents the degree of discrepancy between two independent αs on a common basis. When αs from tests of different lengths are to be compared, $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha}_1)$ and $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha}_2)$ no longer have the same asymptotic variance. Can Δ still be applicable when $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha}_1)$ and $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha}_2)$ have unequal variances? Based on the asymptotic variances of $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha})$ with varying test lengths as illustrated below, we tend to suggest that Δ is highly likely to be applicable with unequal test lengths as long as none of the tests contains fewer than 5 components.

Figure 1 showed the asymptotic variance of $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha})$ with test length ranging from 1 to 65. The asymptotic variance varied substantially when the test contained fewer than 5 components and approached .5 when test length was greater than 10. Specifically, the asymptotic variance of $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha})$ with test length of 5, 10, 15, 20, and 25 was .63, .56, .54, .53, and .52, respectively. As a result, the difference between the asymptotic variances of $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha}_1)$ and $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha}_2)$ from tests of different lengths approached zero unless one of the tests contained fewer than 5 components. Therefore, the effect size index Δ should be applicable to most cases of unequal test lengths. In such cases, the average of the two variances could be used as an estimate of the common variance in calculating Δ. If, instead, the researcher prefers a conservative estimate of the effect size between two independent αs with unequal test lengths, the variance of $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha})$ obtained from the shorter test could be used as the denominator in calculation of Δ.

Non-normally distributed data are frequently encountered in educational and psychological research (Micceri, 1989). The sampling distribution of $\hat{\alpha}$ with non-normally distributed components has been examined in the past. Zimmerman, Zumbo, and Lalonde (1993) compared the distribution of $\hat{\alpha}$ when test components followed normal, uniform, exponential, and mixed-normal distributions, and found the central tendency and the variability of $\hat{\alpha}$ to be similar regardless of the distributions of the component scores. Moreover, Yuan and Bentler (2002) noted that the results of van Zyl *et al.* (2000) held for component scores with heterogeneous skewnesses and kurtoses. These studies seem to suggest that the distribution of component scores exerts very little influence on the sampling distributions of $\hat{\alpha}$ and the asymptotic distribution of $(1/2)\ln(1-\hat{\alpha})$. Therefore, in view of the previous results on the robustness of $\hat{\alpha}$ and $(1/2)\ln(1-\hat{\alpha})$ against non-normality, Δ should also be robust to the violation of normality assumption and could be extended to non-normally distributed component scores.

The covariance matrix among the component scores was assumed to have compound symmetry in the derivation of the asymptotic distribution of $(1/2)\ln(1-\hat{\alpha})$. When component scores are parallel, the assumption of compound symmetry is met. However, compound symmetry could be violated in many situations. Among the circumstances that compound symmetry could be violated, the condition of correlated errors of measurement was most frequently studied. Novick and Lewis (1967)

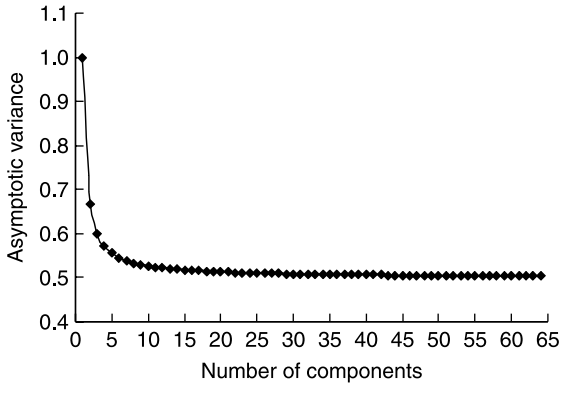

Figure 1. The asymptotic variances of $\sqrt{N-1}(1/2)\ln(1-\hat{\alpha})$ with varying numbers of components in a test.

termed the assumption of uncorrelated measurement errors the 'assumption of linear experimental independence', and made it one of the fundamental assumptions in their derivation of coefficient α. However, Allen and Yen (1979) pointed out that this assumption was not reasonable if the test scores had been affected by fatigue, practice effects, or environmental conditions. Zimmerman and Williams (1980) illustrated the deviation of test reliability from its actual value when the assumption of uncorrelated errors was violated. The discrepancy enlarged as the correlation between errors increased except when the test scores had fairly high reliabilities. The bias of $\hat{\alpha}$ was also being reported to increase systematically with larger number of correlated errors or higher degree of correlation between errors (e.g. Komaroff, 1997; Zimmerman *et al.*, 1993). Thus, with the demonstrated influence of correlated errors on the value of coefficient α, to what extent would correlated errors affect Δ? If both αs were affected by correlated errors, would the effects of correlated errors be cancelled out in the calculation of Δ? A small simulation was conducted to explore the issue.

Two sets of components under classical test theory with uncorrelated errors were first generated to yield a desired level of Δ. Then, correlations among errors were incorporated into the data and the value of Δ from components with correlated errors was estimated, symbolized as $\hat{\Delta}$. The differences between the resulting $\hat{\Delta}$ and Δ in population signalled the effect of correlated errors on the estimation of Δ. The component scores were simulated with varying test lengths (K), component reliabilities (ρ), degrees of correlation between errors (r), proportions of test components having correlated errors (P), and sample sizes (N) as detailed in the Appendix. It was found that the discrepancies between $\hat{\Delta}$ under $N = 500$ and $N = 100$ were less than .01, lending additional support to the earlier finding of the acceptable behaviour of Δ with sample size of 100 and above. Moreover, most of the resulting $\hat{\Delta}$ under N of 500 were closer to the expected magnitudes of .8, .5, and .2 than N of 100. Therefore, to save space, Table 6 represented the mean differences between $\hat{\Delta}$ and Δ with correlated errors only at the sample size of 100. Take the medium effect size .5 with $\rho_1 = .15$, $K = 10$, $r = .1$, and $P = .4$ as an example. According to the formulation of Δ, ρ_2 was set at .33 to yield the desired Δ of .50. Next, the two sets of test components with $\rho_1 = .15$ and $\rho_2 = .33$ having 40% of the error scores correlated at .1 were generated. The resulting mean $\hat{\Delta}$ of .48 over 1,000 replications gave the corresponding entry of $-.02$ in the table. Because of the positive correlations simulated among errors, all the mean differences were either negative or zero. The standard errors of $\hat{\Delta}$s, ranging within .13 and .15, were all close to the expected asymptotic value of .142 ($\sqrt{2/(N-1)} = .142$ with $N = 100$) and thus are not presented.

Parallel to previous findings on the effect of correlated errors on test score reliability (Komaroff, 1997; Zimmerman & Williams, 1980; Zimmerman *et al.*, 1993), when error scores were correlated, the absolute mean differences between $\hat{\Delta}$ and Δ increased with larger number of correlated errors (as reflected in higher proportion of errors correlated or larger number of test components), higher correlation between errors, and lower component reliability. In addition, the absolute mean difference between $\hat{\Delta}$ and Δ also increased with rising effect size. Yet, considering the magnitudes of the discrepancies between $\hat{\Delta}$ and Δ, the result seemed to suggest that $\hat{\Delta}$, though with correlated errors, could still closely approximate the size of Δ unless in cases of a large number of highly correlated errors.

This study has taken a step to define the effect size index Δ for comparing two independent αs. Following the discussion above, Δ is expected to be applicable to cases of unequal test lengths, non-normally distributed test components, and even certain

Table 6. Mean deviations between Δ and $\hat{\Delta}$ with correlated errors ($\hat{\Delta} - \Delta$, $N = 100$)

				Correlation between errors											
					.1				.3				.5			
					Proportion of errors correlated				Proportion of errors correlated				Proportion of errors correlated			
ρ_1	K	ρ_2	Δ	0	.2	.4	.6	.8	.2	.4	.6	.8	.2	.4	.6	.8
.15	10	.21	.2	.00	−.01	−.01	−.01	−.03	.00	−.03	−.05	−.07	.00	−.05	−.07	−.09
		.33	.5	.00	.00	−.02	−.04	−.06	−.01	−.04	−.09	−.14	−.02	−.06	−.13	−.20
		.45	.8	.00	−.01	−.02	−.04	−.09	.00	−.06	−.12	−.21	−.02	−.08	−.19	−.29
	25	.20	.2	.00	−.01	−.01	−.03	−.04	−.01	−.03	−.06	−.08	−.02	−.04	−.09	−.11
		.29	.5	.00	−.01	−.02	−.05	−.08	−.02	−.05	−.12	−.19	−.03	−.09	−.18	−.24
		.39	.8	.00	−.01	−.03	−.07	−.12	−.02	−.09	−.17	−.26	−.04	−.13	−.25	−.36
.35	10	.43	.2	.00	.00	−.01	−.01	−.02	−.01	−.01	−.02	−.04	.00	−.02	−.04	−.05
		.56	.5	.00	−.01	−.01	−.02	−.03	.00	−.02	−.05	−.08	−.01	−.03	−.07	−.12
		.67	.8	.00	.00	−.01	−.02	−.04	−.01	−.03	−.06	−.10	−.01	−.04	−.09	−.16
	25	.42	.2	.00	.01	−.01	−.01	−.02	−.01	−.01	−.02	−.04	−.01	−.02	−.04	−.06
		.53	.5	.01	.00	.00	−.02	−.03	−.01	−.03	−.06	−.09	−.01	−.04	−.07	−.14
		.64	.8	.00	.00	−.01	−.02	−.05	−.01	−.03	−.07	−.12	−.01	−.05	−.12	−.19

Note. ρ_1 = component reliability for Test 1; ρ_2 = component reliability for Test 2; K = test length; Δ = effect size level; 0 = no correlated errors.

conditions under correlated errors. Therefore, although the deviation of Δ was based on several assumptions, Δ could be extended to a wider range of research scenarios beyond the scope restricted by the assumptions.

One line of research worthy of further pursuit is the development of effect size index for dependent coefficient αs. For example, respondents may be asked to fill out the same test items under different instructions or response formats. Feldt and his colleague (Alsawalmeh & Feldt, 1994a, 1994b, 1999; Feldt, 1980) have worked on the sampling distribution of the difference between correlated α coefficients. Future research could possibly apply the derived sampling distribution to develop the appropriate effect size index for comparing dependent αs.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Books-Cole.

Alsawalmeh, Y. M., & Feldt, L. S. (1994a). A modification of Feldt's test of the equality of two dependent alpha coefficients. *Psychometrika*, *59*, 49–57.

Alsawalmeh, Y. M., & Feldt, L. S. (1994b). Testing the equality of two related intraclass reliability coefficients. *Applied Psychological Measurement*, *18*, 183–190.

Alsawalmeh, Y. M., & Feldt, L. S. (1999). Testing the equality of independent alpha coefficients adjusted for test length. *Educational and Psychological Measurement*, *59*, 373–383.

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

Charter, R. A., & Feldt, L. S. (1996). Testing the equality of two alpha coefficients. *Perceptual and Motor Skills*, *82*, 763–768.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

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

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391–418.

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

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

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

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*, 93–103.

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

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523–531.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology and its future prospects. *Educational and Psychological Measurement*, *60*, 661–681.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*, 227–240.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.
Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error on coefficient alpha. *Applied Psychological Measurement*, *21*, 337–348.
Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, *28*, 221–238.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.
Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1–13.
Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.
Thode, H. C. (2002). *Testing for normality*. New York: Marcel Dekker.
van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*, 271–280.
Wilkinson, L., & the APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.
Woodruff, D. J., & Feldt, L. S. (1986). Tests for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika*, *51*, 393–413.
Yuan, K.-H., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika*, *67*, 251–259.
Zimmerman, D. W., & Williams, R. H. (1980). Is classical test theory 'robust' under violation of the assumption of uncorrelated error. *Canadian Journal of Psychology*, *34*, 227–237.
Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, *53*, 33–49.

Received 27 March 2006; revised version received 21 April 2008

Appendix

The procedures for simulating $\hat{\Delta}$ with correlated errors

The simulation generated tests of 10 and 25 components with the component reliability ρ_1 being .15 or .35 for sample size (N) of 100 and 500. Based on the definition of Δ, ρ_2 was carefully chosen to satisfy the magnitude of .8, .5, or .2 in the population. Similar to previous studies (e.g. Komaroff, 1997; Zimmerman *et al.*, 1993), the degree of correlation between error scores was determined by the correlation between two errors ($r = .1, .3,$ or $.5$) and the proportion of test components having correlated errors ($P = .2, .4, .6,$ or $.8$). Correlated errors were generated by the procedures as in Zimmerman *et al.* One thousand replications were generated for each of the 288 conditions (2 sample size × 3 effect size level × 2 component reliability × 2 test length × 3 degrees of correlation between errors × 4 proportions of test components having correlated errors).

The component scores were again simulated based on the classical test theory. The component score for observation j on component i (x_{ij}) was represented as $x_{ij} = T_j + E_{ij}$, where T_j was the true score for observation j, and E_{ij} was the error score for observation j on component i. In order to yield the desired component reliability, T_j was further decomposed as $T_j = \sqrt{\rho/(1-\rho)}t_j$, with ρ being the component reliability and t_j being distributed as $N(0,1)$. Next, let E_{ij} and E_{kj} be the error scores for observation j on components i and k. In order for the two error scores to have a correlation of r, let

$E_{ij} = (W_i + bZ)/\sqrt{1 + b^2}$ and $E_{kj} = (W_k + bZ)/\sqrt{1 + b^2}$, where W_i, W_k, and Z were independent standard normal variables, and $b = \sqrt{r/(1 - r)}$. As a result, E_{ij} and E_{kj} had means of 0 and variances of 1, and their correlation equalled r. Details of the procedures could be found in Zimmerman *et al.* (1993). The W_i, W_k, Z, and t_j were generated by the RANNOR function in SAS.

The simulation results for $N = 100$ were presented in Table 6. A similar table with $N = 500$ was also obtained, though not presented. The discrepancies between $\hat{\Delta}$ under the two sample sizes were all less than .01, and most of the resulting $\hat{\Delta}$ under $N = 500$ were closer to the expected magnitudes of .8, .5, and .2 than the corresponding entries in Table 6.