Sequential Reliability Tests

Mindert H. Eiting
University of Amsterdam

Sequential tests for a stepped-up reliability estimator and coefficient alpha are developed. In a series of monte carlo experiments, the efficiency of the tests relative to each other and to fixed-sample tests is established, as well as the robustness of the alpha test. Both tests proved to be efficient, and the alpha test proved to be reasonably robust to deviations from normality and deviations from equal item error variances. On average, 47% of the sample size can be saved if a sequential test is applied instead of a fixed-sample test. *Index terms: monte carlo simulation, reliability hypothesis tests, sampling theory for reliability estimators, sequential probability ratio test.*

Making an acceptable decision about a reliability estimator is an important psychometric problem. Early contributions to this problem were made by Kristof (1963), Feldt (1965), and Lord and Novick (1968, pp. 206–208); Feldt, Woodruff, and Salih (1987) provide a recent summary concerning coefficient alpha (Cronbach, 1951). The problem is generally approached by applying analytical methods based on distributional assumptions. An alternative approach is the bootstrapping method (de Gruijter, 1988). In both methods, data are sampled in fixed quantities and decisions are made on the basis of a fixed sample.

A method of sequential evaluation of the reliability of psychometric instruments is proposed here. In this method, sample size is not fixed. Thus a test statistic is computed after each person is sampled, and a decision is made in each stage of the sampling process. This method is an example of the sequential probability ratio test (SPRT; Wald, 1947; Wetherill & Glazebrook, 1986) of a composite hypothesis in which both a hypothesis and an alternative are specified, as well as a Type I and Type II error rate. Compared with other methods, a considerable reduction of the sample size can be achieved by applying an SPRT. The reduction is due to the sample inspection procedure, which makes it possible to reach early decisions for samples that are favorable to an hypothesis or its alternative (Hoel, 1966, p. 358).

The sequential method may be useful in several psychometric applications. If an existing instrument has to be applied in a new population, it may be important to know whether the instrument is sufficiently reliable in that population, rather than to estimate precisely the item characteristics. An SPRT can detect this in small samples, especially if the actual reliability is low. This economy may be important if the costs of testing are considerable or if the population itself is small.

The sequential tests applied below were devised for the evaluation of language ability tests for preschool children from small ethnic minority groups in the Netherlands. It was estimated that the application of a fixed-sample method would sometimes require testing the complete population. A similar problem is encountered if a rating procedure has to be used with a limited set of products: If the complete set is rated and rater agreement proves to be insufficient, an improved attempt with the same raters requires a new set of products. This problem can be avoided if an early decision is made on the basis of a small subset of products. Another attractive feature of a sequential test is

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 2, June 1991, pp. 193-205

0146-6216/91/020193-13$1.90

that it can be inserted easily in any data entry program and indicate immediately whether a certain quality criterion is satisfied.

If a stepped-up reliability estimator (Kristof, 1963) is derived from the product-moment correlation between split-part scores of tests or between the ratings of two raters, and parallelism is assumed, a sequential test for reliability reduces to a sequential test for the product-moment correlation coefficient. Such a sequential test is developed below. If essentially tau-equivalent items in a test or essentially tau-equivalent raters are assumed, alpha is the reliability of the test or the ratings. A sequential test for alpha is also derived here, which can easily be adapted to Raju's (1977) coefficient beta.

It is worth noting that the stepped-up reliability can also be estimated with alpha or beta. Therefore, a correlation-based test should be preferred to the alpha test only if the former proves to be more efficient or robust than the latter. The following two sections compare sequential tests with each other and with fixed-size tests using monte carlo data.

A Sequential Test for the Product-Moment Correlation Coefficient

The method of sequential testing concerns a distributional parameter θ and a test of a hypothesis $\theta = \theta_0$ against the alternative $\theta = \theta_1$ $(\theta_0 < \theta_1)$. The likelihood ratio, P_m, is computed in the sample as

$$P_m = \frac{P_{1m}}{P_{0m}} = \frac{\prod_{i=1}^{m} \mathrm{f}(x_i \mid \theta_1)}{\prod_{i=1}^{m} \mathrm{f}(x_i \mid \theta_0)} \quad , \tag{1}$$

in which m is the number of examinees randomly sampled at a certain stage in the procedure, P_{1m} is the likelihood if the alternative is true, and P_{0m} is the likelihood if the hypothesis is true. The decision rule is:

If $P_m \leq \beta/(1 - \alpha)$, accept the hypothesis;
If $P_m \geq (1 - \beta)/\alpha$, accept the alternative;
If neither alternative is true, continue sampling.

With this rule, in which α and β are the nominal error rates, the sequential test will have approximately minimum strength (α,β) (Ghosh, 1967).

Several sequential tests for the product-moment correlation coefficient were proposed (Pradhan & Sathe, 1975). A number of these tests compare unfavorably with their fixed-sample analogue (Pradhan & Sathe, 1975, Table 1). The test of Pradhan and Sathe, for example, only uses the correlational information contained in two paired observations each time, which leads to a great loss of efficiency.

The SPRTs used here are tests of a composite hypothesis concerning some parameter of interest in the presence of nuisance parameters. An approach to this type of SPRT was advanced by Cox (1952; see also Wetherill & Glazebrook, 1986, pp. 57–62, 66), who proved that under certain conditions, the sampling distribution of a sequence of statistics obtained in the course of sequential sampling can be factorized in a distribution, depending only on the parameter of interest and a distribution that does not involve any of the original parameters. Cox's theorem was later improved by Hall, Wijsman, and Ghosh (1965) and fit in the general principle of invariance. Consider

$$\mathrm{f}(z_4, z_5, \ldots, z_m \mid \rho) = g(z_m \mid \rho)\, h(z_4, z_5, \ldots, z_m) \quad , \tag{2}$$

in which z_m is Fisher's transformation of the correlation coefficient r_m, which is asymptotically normally distributed in random samples from a bivariate normal parent distribution, and

$$z_m = \frac{1}{2}\log\,[(1 + r_m)/(1 - r_m)] \quad , \tag{3}$$

$$\mathrm{E}(z_m) = \frac{1}{2}\log\,[(1 + \rho)/(1 - \rho)] + \rho/(2m - 2) \quad , \tag{4}$$

and

$$\mathrm{Var}(z_m) \cong (m - 3)^{-1} \quad . \tag{5}$$

Note that the expected value is often presented in the literature (e.g., Cox, 1952) in its asymptotic form without the bias component for finite samples. Furthermore, the asymptotic variance may depend on ρ, but not if the parent distribution is bivariate normal or of certain related types (Hawkins, 1989). This means that the robustness to deviations from normality may generally be questionable.

Although the z statistics of Equation 2 are dependent with an unknown likelihood, their likelihood ratio becomes

$$P_m = \frac{g(z_m \mid \rho_1)}{g(z_m \mid \rho_0)} \quad , \quad \rho_0 < \rho_1 \quad . \tag{6}$$

Substituting the normal density for g and taking the logarithm of the ratio, the statistic t_m obtained is approximately

$$t_m = \frac{1}{2}(m - 3)\{[z_m - \mathrm{E}_0(z_m)]^2 - [z_m - \mathrm{E}_1(z_m)]^2\} \quad , \quad m > 3 \quad , \tag{7}$$

with E_0 and E_1 as the expected value of Equation 4 if either $(\rho = \rho_0)$ or $(\rho = \rho_1)$ is true. The test using this statistic (i.e., for $m > 3$) is termed the Fisher transform sequential test (FTST). Because the statistic is a log-likelihood ratio, the hypothesis boundary B_0 and the alternative boundary B_1 are

$$B_0 = \log\,[\beta/(1 - \alpha)] \tag{8}$$

and

$$B_1 = \log\,[(1 - \beta)/\alpha] \quad . \tag{9}$$

The decision rule for FTST is:

If $t_m \leq B_0$, accept the hypothesis;
If $t_m \geq B_1$, accept the alternative;
If neither alternative is true, continue sampling.

A Sequential Test for Cronbach's Alpha

The FTST is a test for the reliability (ζ) on the assumption of two parallel parts. When the k items of a test or k raters are assumed to be essentially tau-equivalent, alpha is the reliability of the linear combination of the items or the ratings. Let ζ denote the population value, $\hat{\zeta}_m$ denote the estimator, and m denote the number of persons at any stage in an SPRT. Kristof (1963) and Feldt (1965) proved that the statistic $(1 - \zeta)/(1 - \hat{\zeta}_m)$ is distributed as central F, with $\nu_1 = (m - 1)$ and $\nu_2 = (k - 1)(m - 1)$, with the additional assumptions of (1) random sampling of both persons and items and (2) independently and normally distributed residual errors with homogeneous variance. Note that the requirement of equal error variances of the k items is more demanding than the requirement of essential tau-equivalence, which permits differing error variances.

Again applying the principle of invariance, the same type of factorization used in Equation 2 can be considered. The statistic contains the parameter of interest (ζ) this time, although the central F density does not. The statistic in Equation 3 does not contain the parameter (ρ) that appears in the expected value of the normal density. Ghosh (1967) proved that the principle of invariance applies when a related statistic is a function of the sample data only and has a density related to the central F density, yet contains the parameter of interest. Following the approach of Ghosh (1967, procedure II for the randomized block design; see also Feldt, Woodruff, & Salih, 1987), the factorization is applied to the statistic

$$z_m = 1 / [(1 - \hat{\zeta}_m)\ (k - 1)] \quad , \tag{10}$$

where

$$\hat{\zeta}_m = [k/(k - 1)]\left(1 - \sum_{j=1}^{k} \hat{\sigma}^2_{mj} / \hat{\sigma}^2_m\right) \quad , \tag{11}$$

in which $\hat{\sigma}^2_{mj}$ is the estimator of the variance of scores on the jth item after m persons were sampled, and $\hat{\sigma}^2_m$ is the variance estimator of the linear combination of the item scores. It does not matter whether biased or unbiased estimators are employed.

The density function of z_m (Ghosh, 1967, Equation 5) in the present notation is

$$g(z_m | \zeta) = \frac{(1 - \zeta)^{-\frac{1}{2}(m-1)(k-1)} z_m^{\frac{1}{2}(m-3)}}{B(\nu_1/2, \nu_2/2)[1/(1 - \zeta) + z_m]^{\frac{1}{2}k(m-1)}} \quad , \quad z_m \geq 0, \quad m,k > 1 \quad . \tag{12}$$

Following Equations 2 through 6, the following likelihood ratio is derived:

$$P_m = g(z_m | \zeta_1) / g(z_m | \zeta_0) \quad , \quad \zeta_0 < \zeta_1 \quad . \tag{13}$$

Equations 12 and 13 yield:

$$P_m = \left[\frac{1 - \zeta_1}{1 - \zeta_0}\right]^{\frac{1}{2}(m-1)} \left[\frac{1 + z_m(1 - \zeta_0)}{1 + z_m(1 - \zeta_1)}\right]^{\frac{1}{2}k(m-1)} \quad . \tag{14}$$

Taking the logarithm and substituting Equation 10 for the statistic z_m in Equation 14, the log-likelihood ratio for the sequential test is

$$t_m = \frac{1}{2}(m - 1) \log\left[\frac{1 - \zeta_1}{1 - \zeta_0}\right] + \frac{1}{2}k(m - 1) \log\left[\frac{k \sum_{j=1}^{k} \hat{\sigma}^2_{mj} - \zeta_0 \hat{\sigma}^2_m}{k \sum_{j=1}^{k} \hat{\sigma}^2_{mj} - \zeta_1 \hat{\sigma}^2_m}\right] , \quad m,k > 1 \quad . \tag{15}$$

The SPRT with the statistic of Equation 15 is termed the coefficient alpha sequential test (CAST) here (for $m, k > 1$), with boundaries of Equations 8 and 9. In addition, Ghosh proved that this type of test terminates with a probability of 1.0, and that it tests for $\zeta \leq \zeta_0$ against $\zeta \geq \zeta_1$ with minimum strength (α, β).

If a test is split into parts of differing length, essential tau equivalence generally does not apply to the part scores. Alpha, which is computed from the part scores, will underestimate the reliability. However, Raju's (1977) beta corrects this bias if the original items are essentially tau equivalent. Because CAST applies to alpha rather than the actual reliability, Raju's correction should be applied to the hypothesized and alternative parameter in Equation 15 as

$$\zeta'_v = \frac{\zeta_v(k - k\sum_{j=1}^{k} h_j^2)}{k - 1} \quad , \quad v = 0, 1 \quad , \tag{16}$$

in which k is the number of test parts and h_j is the number of items in part j relative to the total number of items in the test. With the simple correction of Equation 16, CAST is also an SPRT for Raju's beta (aside from robustness considerations).

Sample Numbers

The average sample number (ASN) indicates the number of persons for which the statistic in an SPRT is expected to remain between the boundaries so that sampling should be continued. For SPRTs of simple hypotheses, this number is known analytically. For SPRTs of composite hypotheses, such as the FTST and CAST, no exact expressions exist for the ASN and monte carlo studies may be needed.

SPRTs have a fixed-sample analogue with two choices instead of three. The fixed-sample analogue of the FTST, which is the Fisher Transform Fixed Test (FTFT), may be constructed for the same hypothesis and alternative and the same Type I and II error rates as the FTST. The fixed-sample number (FSN) can then be obtained analytically. The fixed-sample analogue of the CAST is (Ghosh, 1967) the F test of Kristof (1963) and Feldt (1965) (e.g., the coefficient alpha fixed test, CAFT). The FSN for this test can also be obtained.

For some problems ($\rho_0 = 0$), the FSN of tests for the product-moment correlation coefficient can be found in Cohen (1988). The FSN for the FTFT can be obtained by standard methods for a normally distributed statistic under both the hypothesis and the alternative. The FSN for the FTFT in sufficiently large samples is

$$\text{FSN} = \left\{ \frac{2(a + b)}{\log\left[\frac{(1 + \rho_1)(1 - \rho_0)}{(1 - \rho_1)(1 + \rho_0)}\right]} \right\}^2 + 3 \quad , \quad \rho_0 < \rho_1 \quad , \tag{17}$$

in which a and b are the abscissas for the $(1 - \alpha)$ and $(1 - \beta)$ ordinates of the cumulative standard normal distribution. The correlation coefficient of Equation 17 is related to reliability through the Spearman-Brown equation $\rho = \zeta/(2 - \zeta)$.

The FSN of the CAFT is more difficult to find. Let

$$ab = (1 - \zeta_0)\,/\,(1 - \zeta_1) \quad , \quad \zeta_0 < \zeta_1 \tag{18}$$

(Ghosh, 1967, Equations 19 and 25 with symbols m and k interchanged). Then a is the abscissa for the $(1 - \alpha)$ ordinate of the cumulative central F distribution with $v_1 = (\text{FSN} - 1)$ and $v_2 = (k - 1)(\text{FSN} - 1)$, and b is the abscissa for the $(1 - \beta)$ ordinate of the cumulative central F distribution with $v_1 = (k - 1)(\text{FSN} - 1)$ and $v_2 = (\text{FSN} - 1)$. The value of FSN that satisfies Equation 18 may be found by iterative numerical integration or by screening a detailed F table.

Most textbooks on mental testing indicate that a reliability of .8 or more is acceptable. Thus monte carlo studies with population reliabilities of approximately .848 were performed. Hypothesized values that implied large and small sample problems were selected at shorter and larger distances. Each problem was repeated for four combinations of nominal error rates of .05 and .01.

Table 1 provides a summary of ASNs of the sequential test (PSST) of Pradhan and Sathe (1975; see also Hoel, 1966, p. 361) and the FSNs of the FTFT and CAFT for all problems in the next sections. The ASNs of the PSST exceed the FSNs of the alternative fixed tests for all problems. The sample

Table 1
Sample Numbers for Varying Nominal Error Rates (α, β), Number of Items, ζ = .848, and Combinations of ζ_0 and ζ_1 for PSST, FTFT, and CAFT

Number of Items	Test	ζ_0 = .848 ζ_1 = .956	ζ_0 = .474 ζ_1 = .848	ζ_0 = .848 ζ_1 = .900	ζ_0 = .769 ζ_1 = .848
α = .05, β = .05					
2	PSST	83	47	587	479
2	FTFT	31	31	250	250
2	CAFT	30	30	250	250
15	CAFT	17	17	134	134
30	CAFT	16	16	130	130
α = .05, β = .01					
2	PSST	130	52	924	526
2	FTFT	44	44	363	363
2	CAFT	43	43	363	363
15	CAFT	25	25	199	199
30	CAFT	24	24	193	193
α = .01, β = .05					
2	PSST	91	73	645	754
2	FTFT	44	44	363	363
2	CAFT	43	43	363	363
15	CAFT	23	23	191	191
30	CAFT	22	22	184	184
α = .01, β = .01					
2	PSST	140	79	996	813
2	FTFT	59	59	497	497
2	CAFT	58	58	498	498
15	CAFT	32	32	267	267
30	CAFT	31	31	258	258

numbers of FTFT and CAFT are remarkably similar. Table 1 also demonstrates that an increment of the number of items (especially from 2 to 15) corresponds with a considerable increment of efficiency.

Comparing the FTST and CAST for Two Items

Method

In the first monte carlo experiment, the FTST and CAST were compared with each other and with their fixed analogues with respect to estimated error rates and sample numbers. The experiment concerned two bivariate standard normally distributed variables (k = 2) with a correlation of .7361 and a reliability of .848. Therefore, the marginal distributions had equal means and variances. Numbered pairs were sampled randomly from the population using the KR algorithm of Kinderman and Ramage (1976).[1] Hypothesized and alternative parameters were selected as displayed in Table 1. If no decision could be made during a trial in favor of the hypothesis or the alternative, the sample size was increased by 1. Error rates and sample numbers were recorded. Because the sample numbers have a skewed distribution in sequential tests, the proportion of trials (PLF) in which the sample number is less than the FSN is as useful as the ASN. This proportion was also recorded. All evaluations were based on 2,000 independent trials per problem.

[1] All programs for the monte carlo experiments were written in Microsoft QuickBASIC operating with 32 bits precision. The programs were run on a fast IBM-compatible PC under MS-DOS.

Table 2
Estimated Error Rate (EER), Average Sample Number (ASN), and Probability Sample Number Less Than Fixed (PLF) for Varying Nominal Error Rates (α, β) and Combinations of ζ_0 and ζ_1, in 2,000 Trials per Problem Sampled from Bivariate Standard Normally Distributed Numbers ($k = 2$), for FTST and CAST

	FTST ($\zeta = .848$)				CAST ($\zeta = .848$)			
Statistic	$\zeta_0 = .848$ $\zeta_1 = .956$	$\zeta_0 = .474$ $\zeta_1 = .848$	$\zeta_0 = .848$ $\zeta_1 = .900$	$\zeta_0 = .769$ $\zeta_1 = .848$	$\zeta_0 = .848$ $\zeta_1 = .956$	$\zeta_0 = .474$ $\zeta_1 = .848$	$\zeta_0 = .848$ $\zeta_1 = .900$	$\zeta_0 = .769$ $\zeta_1 = .848$
$\alpha = .05, \beta = .05$								
EER	.0380	.0410	.0480	.0460	.0415	.0315	.0485	.0355
ASN	19	19	130	127	19	19	135	134
PLF	.8620	.8655	.9025	.9080	.8585	.8535	.8820	.8930
$\alpha = .05, \beta = .01$								
EER	.0320	.0075	.0385	.0095	.0420	.0050	.0425	.0110
ASN	28	20	200	140	28	20	204	146
PLF	.8620	.9420	.9130	.9540	.8670	.9350	.8960	.9440
$\alpha = .01, \beta = .05$								
EER	.0065	.0355	.0085	.0540	.0060	.0350	.0100	.0380
ASN	21	27	143	198	20	27	141	202
PLF	.9380	.8720	.9570	.9055	.9370	.8800	.9625	.9015
$\alpha = .01, \beta = .01$								
EER	.0075	.0060	.0085	.0115	.0075	.0050	.0065	.0095
ASN	28	29	214	216	29	29	213	217
PLF	.9540	.9490	.9570	.9490	.9405	.9385	.9650	.9580

Results

Table 2 shows that the FTST and the CAST have correct estimated error rates for bivariate standard normally distributed variables. If either the hypothesis or its alternative was true, the ASN was approximately 51% of the FSN. Both tests were almost equally efficient for $k = 2$. The tests were approximately four times as efficient as the PSST (see Table 1). The efficiency increased with decreasing nominal error rate. The ASN was approximately 46% of the FSN for the $\alpha = .01$, $\beta = .01$ conditions, and the sample number did not exceed the FSN in approximately 95% of the trials. The ASN was approximately 56% of the FSN for the (.05,.05) conditions, and the sample number did not exceed the FSN in approximately 88% of the trials. It should be noted that the ASN was not the same as the FSN for the (.05,.01) and (.01,.05) conditions. When the hypothesis was true, the (.01,.05) condition was more efficient, whereas the reverse was true for the true alternative.

The results of Table 2 apply at a fixed population reliability of .848. For example, if $\zeta_0 = .474$, $\zeta_1 = .848$, and $\alpha = .05 = \beta$, and varying population values are selected, the operation characteristic (OC) function (i.e., the probability of accepting the hypothesis), as well as the ASN and the PLF functions of the test are obtained. This was done for the CAST in 2,000 trials for each of 13 selected values. The estimated functions are shown in Figure 1. It can be seen that the OC was strictly decreasing in ζ, which implies that the CAST is a test for $\zeta \leq \zeta_0$ against $\zeta \geq \zeta_1$ with minimum strength (α,β). Outside the interval $\zeta_0 - \zeta_1$, the ASN was the same or less than that reported in Table 2. Inside the interval $\zeta_0 - \zeta_1$, the ASN was larger, and it exceeded the FSN slightly in the worst case. The PLF function demonstrated that even in the worst case, the probability that the sample number was less than FSN was still approximately .6.

Robustness Experiments for the CAST

Method

Because the CAST proved to be almost as efficient as the FTST for the two-variable case, and the robustness of the FTST to deviations from normality was already questionable, only continuing experiments with the CAST were performed. Seeger and Gabrielsson (1968; see also Feldt, 1965; Feldt et al., 1987) demonstrated that the F test is reasonably robust to deviations from normality in a variance analysis of dichotomous variables. It remained to be seen whether this was also true for the sequential test when dichotomous items with unequal error variances were employed (i.e., essentially tau-equivalent dichotomous items).

Study 1. The first analysis used 15 artificial dichotomous items ($k = 15$) that were essentially tau-equivalent. The sampling procedure was as follows: Let the item scores, X_{ij} ($X_{ij} = 0,1$; $i = 1, 2, \ldots, m$; $j = 1,2, \ldots, k$) have the same covariance c with all other item scores, but possibly different means, p_j, and variances, $p_j q_j$. The items are then essentially tau-equivalent, and

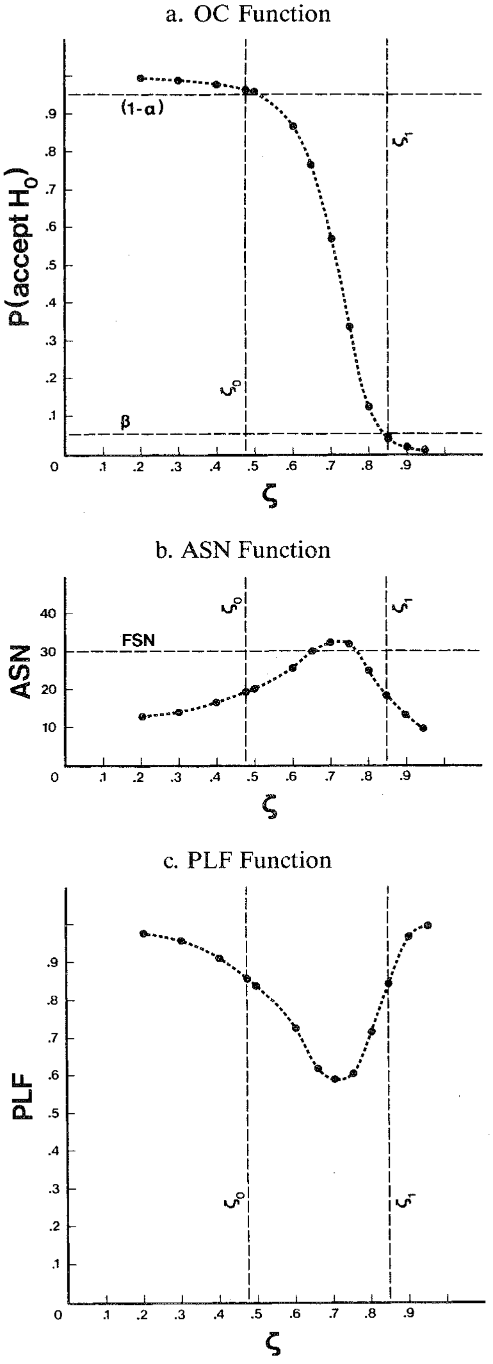
$$\sum_{j=1}^{k} p_j q_j = ck(k/\zeta - k + 1) \quad . \tag{19}$$

The item means must be selected to satisfy Equation 19, and to avoid out-of-range probabilities:

$$p_j p_h,\ q_j q_h < 1 - c, \quad p_j q_h > c, \quad c^{1/2} < p_j < 1 - c^{1/2}, \quad \text{for all } j,h\ (j \neq h) \quad . \tag{20}$$

For each case i, randomly sample a reference score, X_{i0}, which is kept outside the dataset, thus satisfying $P(X_{i0} = 1) = 1/2$. Then sample the k item scores according to:

Figure 1
Estimated CAST Functions for $\zeta_0 = .474$, $\zeta_1 = .848$, and $\alpha = \beta = .05$ in 2,000 Trials per Point
[Sampled From Bivariate Standard Normally Distributed Numbers ($k = 2$)]


a. OC Function

b. ASN Function

c. PLF Function

$$P(X_{ij} = 1 \mid X_{i0} = 1) = p_j + c^{1/2} \quad , \tag{21}$$

and

$$P(X_{ij} = 1 \mid X_{i0} = 0) = p_j - c^{1/2} \quad . \tag{22}$$

The set of item means in the experiment varied from .29 to .71 in increments of .03. For $\zeta = .848$, it follows that $c = .06322046$.

It should be noted that the linear combination of 15 item scores for this reliability will have a rather extreme bimodal distribution (i.e., the instrument is highly discriminating). Because discrete variables may have 0 variance in small samples, the decision rule for the CAST had to be liberalized (i.e., sampling continued if alpha could not be not computed). All problems in the first experiment are as shown in Table 1.

Study 2. In a second monte carlo analysis, the CAST was applied to real data. The source of randomly sampled data (with replacement) consisted of dichotomous scores of 500 Dutch children approximately 9-1/2 and 11-1/2 years old on 30 semantic ability items. The item means ranged from .036 to .932. The linear combination of item scores had a unimodal skewed distribution (mean 20.0, mode 25, variance 29.9, skewness –.85, kurtosis –.03). Alpha was .848—that is, $\zeta = .848$ in this population. The reliability was also computed in a linear model, assuming only congeneric items (Fleishman & Benson, 1987) that were analyzed with the FABIN-2 method (Hagglund, 1982). The reliability in this model was .850 with an adjusted goodness of fit of .97, which suggests that an assumption of essential tau-equivalence might be reasonable. However, the sum of the item variances was 5.39, which implies a covariance of .028 (see Equation 19). It follows from Equation 20 that items with means of .036 and .932 cannot fit even under the assumption of essential tau-equivalence. All problems in the second experiment are as shown in Table 1.

Results

Results of both studies are displayed in Table 3. The error rates for Study 1 demonstrate that CAST was the correct test for the 15-item application; these rates were considerably smaller than the nominal values. The test retained its minimum strength (α,β) under the violated assumptions. The ASN was generally approximately 53% of the FSN (see Table 1). In the (.01,.01) problems this result was approximately 46%, and the sample number did not exceed the FSN in approximately 98% of the trials. The ASN in the (.05,.05) problems was approximately 59% of the FSN, and the sample number did not exceed the FSN in approximately 89% of the trials. The same result was observed in Table 2. Therefore, the efficiency of the CAST was not greatly affected by violation of the assumptions.

The results for Study 2 in Table 3 demonstrate that the CAST was also a reasonably correct test for the 30-item instrument with a skewed distribution. The Type II error rates tended to exceed the nominal values, sometimes by approximately 1%. The efficiency figures were fairly similar to those listed above. Lower relative efficiency was found for problems in which the fixed sample size was already small. In general, the ASN was approximately 54% of the FSN.

Discussion

Both the FTST and CAST proved to be efficient reliability tests with minimum strength (α,β)—the CAST was also efficient when its assumptions were violated. The robustness of the FTST was not evaluated; it may be inferred from analytical considerations (Hawkins, 1989) that its robustness to

Table 3
Estimated Error Rate (EER), Average Sample Number (ASN), Probability Sample Number Less than Fixed (PLF) for Combinations of ζ_0 and ζ_1, and Varying Nominal Error Rates (α, β) for CAST in Studies 1 and 2

	Study 1				Study 2			
Statistic	ζ_0 = .848 ζ_1 = .956	ζ_0 = .474 ζ_1 = .848	ζ_0 = .848 ζ_1 = .900	ζ_0 = .769 ζ_1 = .848	ζ_0 = .848 ζ_1 = .956	ζ_0 = .474 ζ_1 = .848	ζ_0 = .848 ζ_1 = .900	ζ_0 = .769 ζ_1 = .848
α = .05, β = .05								
EER	.0060	.0115	.0010	.0005	.0190	.0565	.0350	.0490
ASN	13	8	85	67	13	9	77	63
PLF	.8050	.9380	.8775	.9320	.7665	.8740	.8740	.9100
α = .05, β = .01								
EER	.0030	.0015	.0015	.0000	.0220	.0225	.0320	.0095
ASN	20	8	129	69	19	10	121	71
PLF	.8180	.9930	.9060	.9865	.7745	.9385	.8645	.9405
α = .01, β = .05								
EER	.0005	.0090	.0000	.0030	.0045	.0570	.0055	.0420
ASN	14	11	87	101	14	12	82	97
PLF	.9380	.9460	.9690	.9500	.8935	.8815	.9450	.9020
α = .01, β = .01								
EER	.0010	.0015	.0000	.0000	.0025	.0160	.0070	.0080
ASN	20	11	132	102	19	12	123	101
PLF	.9445	.9960	.9800	.9890	.9115	.9590	.9480	.9640

violated distributional assumptions was questionable. The present results demonstrated, however, that the CAST was almost as efficient as the FTST in the bivariate normal case. Furthermore, the CAST does not require the construction of split-halves (for $k > 2$), whereas the stepped-up reliability from split-halves (for $k = 2$) can also be estimated with alpha. The simulation experiments suggest that deviations from normality (i.e., dichotomous items) and from equal item error variances (i.e., tau-equivalent items) do not reduce the strength of the CAST. In addition, slight violations of essential tau-equivalence in the empirical example did not notably affect its strength. A similar result for its fixed analogue was obtained by Feldt (1965).

It may be asked, however, what will happen if essential tau-equivalence is severely violated, resulting in an instrument with congeneric items only or even a multifactor structure. It is well known (e.g., Lord & Novick, 1968; Raju, 1977) that alpha underestimates the actual reliability to a degree that is generally unknown in such a case, except when Raju's beta applies. Significant limits for robustness research are in fact provided by the model for alpha, because statistical properties of an almost useless reliability estimator are not very interesting. However, the appropriateness of alpha is often not known in advance. It may be advisable to compute both alpha and a factor-analytic estimate (Fleishman & Benson, 1987) with an economical method (Hagglund, 1982) as soon as the sequential test terminates. If the two estimates are far apart, the result of the sequential test is in doubt.

In general, 47% of the sample size can be saved by the application of CAST; this percentage may be even higher if the actual reliability is lower than the hypothesized value or higher than the alternative value (see Figure 1). Conversely, the sample number will increase if the actual reliability is between the hypothesized and alternative values. As this research has demonstrated, there is still a 5% to 15% risk that the sample number will exceed the FSN if the hypothesis or alternative is true. Ghosh (1967; see also Wetherill & Glazebrook, 1986, p. 24) gives some directions for a stopping rule that affects the error rates very little. This rule amounts to stopping the sequential test if $m > m'$, and performing a fixed-sample test in the sample at the point of termination, where m' is the 1% point of the sample size distribution. Although m' will exceed the FSN, it is seldom twice that number. As a rule of thumb, m' may be approximately 1.5 times the FSN. A suitable choice of hypothesis and alternative that minimizes the risk that the actual value falls between them is more important.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale NJ: Erlbaum.

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Proceedings of the Cambridge Philosophical Society, 48,* 290–299.

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

de Gruijter, D. N. M. (1988). Evaluating an item and an option statistic using the bootstrap method. *Tijdschrift voor Onderwijs Research, 13,* 345–352.

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

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

Fleishman, I., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement, 47,* 925–939.

Ghosh, B. K. (1967). Sequential analysis of variance under random and mixed models. *Journal of the American Statistical Association, 62,* 1401–1417.

Hagglund, G. (1982). Factor analysis by instrumental variables methods. *Psychometrika, 47,* 209–222.

Hall, W. J., Wijsman, R. A., & Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Annals of Mathematical Statistics, 36,* 575–614.

Hawkins, D. L. (1989). Using *u* statistics to derive the asymptotic distribution of Fisher's *z* statistic. *The American Statistician, 43,* 235–237.

Hoel, P. G. (1966). *Introduction to mathematical statistics.* New York: Wiley.

Kinderman, A. J., & Ramage, J. G. (1976). Computer generation of normal random variables. *Journal of the American Statistical Association, 71,* 893–896.

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.

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

Pradhan, M., & Sathe, Y. S. (1975). An unbiased estimator and a sequential test for the correlation coefficient. *Journal of the American Statistical Association, 70,* 160–161.

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42,* 549–565.

Seeger, P., & Gabrielsson, A. (1968). Applicability of the Cochran *Q* test and the *F* test for statistical analysis of dichotomous data for dependent samples. *Psychological Bulletin, 69,* 269–277.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Wetherill, G. B., & Glazebrook, K. D. (1986). *Sequential methods in statistics* (3rd ed.). London: Chapman and Hall.

Acknowledgments

The author thanks Dato de Gruijter and two anonymous reviewers for their comments on early versions of this article, Jan Neuvel for making available the data on which an empirical example was based, and Fred Marschall for designing the figures.

Author's Address

Send requests for reprints or further information to Mindert H. Eiting, Centre for Educational Research of the University of Amsterdam, Grote Bickersstraat 72, 1013 KS Amsterdam, The Netherlands.