Measurement and Research Department Reports 98-2

Estimating the Reliability of a Test from a Single Test Administration

N.D. Verhelst

Cito

Measurement and Research Department Reports 98-2

Estimating the Reliability of a Test from a Single Test Administration

N.D. Verhelst

Cito
Arnhem, 1998

Abstract

The article discusses methods of estimating the reliability of a test from a single test administration. In the first part a review of existing indices is given, supplemented with two heuristics to approximate Guttman's λ_4 and a new similar coefficient. Special attention is given to the greatest lower bound, to its meaning as well as to the problems in computing it. In the second part the relation between Cronbach's α and the reliability is studied by means of a factorial model for the item scores. This part gives some useful formulae to appreciate the amount with which the reliability is underestimated when α is used as its estimand. In the last part, the sampling distribution of the indices is investigated by means of two simulation studies, showing that the indices exhibit severe bias, the direction of which depends partly on the factorial structure of the test. For three indices the bias is modeled. The model describes the bias accurately for all cases studied in the simulation studies. It is shown how this bias correction may be applied in the case of a single data set.

key words: Reliability, Cronbach's α, split half, greatest lower bound, tau-equivalence, bias, attenuation paradox.

1. Introduction

In textbooks on test theory several methods for determining the reliability of a test are usually presented. The main methods are test-retest reliability, parallel test reliability, the split half method and the method of internal consistency. Although differences between these methods are discussed at length, sometimes essential characteristics of the methods are not or poorly discussed. We mention three of them.

- The reliability of a test is not a characteristic of the test alone, but should always be considered in relation to a certain population of testees. This follows directly from the definition of reliability, which is the ratio of the variance of the true scores to the variance of the observed scores. In a population where the true score is the same for everybody, the true score variance is zero and therefore the reliability of the test is zero.

- Although many formulae for 'computing' the reliability are published, it should be kept in mind that what is computed is not the reliability, but an estimate of the reliability which is always based on the responses to the items of a finite sample of persons. Computing the reliability twice on the responses of two independent samples, even when drawn from the same population, one will get almost surely two different results. What is needed therefore is, together with the estimate, a measure of the accuracy of that estimate, such as a standard error or a confidence interval.

- Perhaps the most widespread misunderstanding is the assumption that the four afore-mentioned methods of estimating the reliability are more or less equivalent, in the sense that any of them can be used interchangeably, and that the choice merely depends on practical considerations. This is not true, however. The test-retest method and the parallel method need two independent test administrations of the same or a parallel test to the same sample of persons. The correlation between the two test scores gives a consistent estimate of the reliability, without making any assumption about the relationship between the items in the test, such as parallelism or tau-equivalence. It is not even necessary that the measurement errors of the items are uncorrelated. The only assumption is that the measurement errors on the two occasions are independent. (Of course, when using the parallel method, it is assumed that two parallel tests are available, an assumption which is difficult

to prove.) Using the split-half method or an index of internal consistency, such as the KR-20 or Cronbach's α, only one test administration has to be carried out, such that there is no direct evidence of the instability of the observed test scores. The consequence is that neither of these methods do give an estimate of the reliability of tests in general. The claim that a given test has reliability equal to one cannot be disproved using only the data from one test administration. The claim, however, that a test has reliability equal to zero, can be disproved using only one test administration. If the sum of all the covariances between the items is positive, then the reliability is larger than zero. More generally, the structure of the matrix of variances and covariances of the items contains information on the minimal value (or lower bound) of the reliability. Cronbach's α, for example, uses (some of) this information, and the only generally valid conclusion from this index is the inequality

$$\alpha \leq Rel, \tag{1.1}$$

where *Rel* is used here as the general symbol for reliability. So, interpreting α as an estimate of the reliability may lead to serious errors.

The purpose of the present paper is threefold. Firstly, a review will be given of other lower bounds, a number of which give always values at least as large as Cronbach's α. In connection with this overview, some simple heuristics will be discussed to compute or approximate some of these indices, since it is not always a simple problem to compute them. Secondly , it will be investigated for some special cases to what extent α underestimates the reliability. Thirdly, some of the statistical properties of α and the other indices will be investigated.

2. Lower bounds of the reliability

Guttman (1945) was the first to investigate systematically a number of lower bounds of the reliability of a test using a single test administration. He derived six indices, called λ_1 through λ_6, one of which, λ_1 is trivial and will not be discussed here. The overview starts with the best known and most widely used lower bound, viz., λ_3 , which is given by

$$\lambda_3 = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right], \tag{2.1}$$

where k is the number of items in the test, σ_i^2 is the (observed) variance of item i, and σ_X^2 is the observed test variance. This index is better known as Cronbach's coefficient α (Cronbach, 1951), and if all items in the test are binary, equals the well known coefficient KR-20 (Kuder and Richardson, 1937). Observing that

$$\sigma_X^2 = \sum_{i=1}^{k} \sigma_i^2 + \sum_{i \neq j} \sigma_{ij},$$

where σ_{ij} is the covariance between items i and j, (2.1) can be written equivalently as

$$\lambda_3 = \frac{k}{k-1} \times \frac{\sum_{i \neq j} \sigma_{ij}}{\sigma_X^2}. \tag{2.2}$$

Arranging the variances and covariances in a square table or matrix, the numerator in the second fraction of the right-hand side of (2.2) is the sum of all off-diagonal elements of the matrix, while the denominator is the sum of all the elements of the matrix.

The coefficient λ_2 is given by

$$\lambda_2 = \frac{\sum_{i \neq j} \sigma_{ij} + \left(\frac{k}{k-1} \sum_{i \neq j} \sigma_{ij}^2 \right)^{\frac{1}{2}}}{\sigma_X^2}. \tag{2.3}$$

It can be shown that $\lambda_2 \geq \lambda_3$ whence we have

$$\alpha = \lambda_3 \leq \lambda_2 \leq Rel, \tag{2.4}$$

meaning that λ_2 is at least as good as a lower bound for the reliability, so that one might wonder why it is not preferred generally. If the covariances between the items are heterogeneous, especially if some covariances are negative, λ_2 may be substantially larger than λ_3.

Before discussing the other indices proposed by Guttman, we digress shortly to the work of Ten Berge and Zegers (1978), who developed an infinite sequence of lower bounds $\mu_j, j = 0, 1, 2 \cdots$, and such that

$$\mu_0 \leq \mu_1 \leq \mu_2 \leq \cdots \leq Rel. \tag{2.5}$$

Although the inequalities (2.5) look very promising, and suggest that μ_∞ equals the reliability, this is not true. The increase of passing from j to $j+1$ is usually

very small if j is larger than 2. The general expression for the μ-coefficients will not be given; only the first three will be mentioned explicitly:

$$\mu_0 = \alpha = \lambda_3,$$

$$\mu_1 = \lambda_2,$$

$$\mu_2 = \frac{1}{\sigma_X^2}\left\{\sum_{i\neq j}\sigma_{ij} + \left[\sum_{i\neq j}\sigma_{ij}^2 + \left(\frac{k}{k-1}\sum_{i\neq j}\sigma_{ij}^4\right)^{1/2}\right]^{1/2}\right\}. \tag{2.6}$$

The indices discussed thus far all take the number of items k in the test as given, and the formulae are all based on the entries of the $k \times k$ covariance matrix. What one tries to find, however, is a lower bound of the reliability of the test score, which is the sum of the item scores. But nowhere in test theory a precise definition is given of an item. If a test consists of 20 questions where one can earn one or zero points, one can say that the tests consists of 20 items. But one can also claim that the test consists of two items where one can earn zero to ten points by, for example, counting the number of correctly answered questions on the even numbered and odd numbered questions respectively. In both cases the test score will be identical, but in the former case the covariance matrix will be a 20×20 matrix and in the latter case it will be a 2×2 matrix. One can use both matrices to compute one of the indices discussed thus far, and every index is a lower bound of the reliability, such that one can choose the largest one. As an example consider a test consisting originally of three items with covariance matrix

$$\begin{bmatrix} 0.25 & 0.04 & 0.09 \\ 0.04 & 0.24 & 0.08 \\ 0.09 & 0.08 & 0.21 \end{bmatrix}.$$

Coefficient α computed on this matrix equals

$$\alpha = \frac{3}{2} \times \frac{0.42}{1.12} = 0.56.$$

Taking the first and second items together the covariance matrix of the test is now

$$\begin{bmatrix} 0.57 & 0.17 \\ 0.17 & 0.21 \end{bmatrix},$$

and coefficient α equals

$$\alpha = \frac{2}{1} \times \frac{0.34}{1.12} = 0.61,$$

which is a greater lower bound than α computed for the original three items. Denote by P a partition of the original items into two non-empty subsets, and by $\alpha(P)$ the coefficient computed on the 2×2 covariance matrix of the two subtests, then Guttman's coefficient λ_4 is defined as

$$\lambda_4 = \max_P [\alpha(P)] .$$

Notice that it is by no means necessary that the two subsets in the partition have an equal number of items, as is sometimes erroneously concluded (see also Jackson, 1979). As will be shown in the examples section, coefficient λ_4 often gives very good results, exceeding by far the value of α computed on the original items. But the computation of λ_4 is not simple. With k items there are $2^{k-1} - 1$ partitions in two subsets. With 40 items, this means that 5.5×10^{11} partitions are possible, and checking them all would not be realistic. Even if on a very fast computer 100,000 coefficients could be computed in one second, the total computing time would be more than 1,500 hours (more than two months), and for every additional item, the computing time doubles. In the section on heuristics, a procedure will be discussed which gives a good approximation to λ_4.

A drawback of coefficients as λ_2, λ_3 and the μ-coefficients of Ten Berge and Zegers is that they are symmetric in the items. Denoting the mean of the $k(k-1)$ squared covariances as $\overline{\sigma^2}$, λ_2 can be written as

$$\lambda_2 = \frac{\sum_{i \neq j} \sigma_{ij} + \left(k^2 \overline{\sigma^2}\right)^{1/2}}{\sigma_X^2}. \tag{2.7}$$

If there is substantial variation in the covariances, the effect of the large covariances on the average squared covariance will be attenuated by the small ones. Guttman was able to derive a lower bound, λ_5, of the reliability which, for the squared covariances, uses only one column of the covariance matrix:

$$\lambda_5 = \frac{\sum_{i \neq j} \sigma_{ij} + 2\left(\max_j \sum_{i \neq j} \sigma_{ij}^2\right)^{1/2}}{\sigma_X^2}. \tag{2.8}$$

From a comparison of (2.7) and (2.8), it follows that $\lambda_5 > \lambda_2$ if there is a column in the covariance matrix where the sum of the squared covariances is larger than $\frac{k^2}{4}\overline{\sigma^2}$.

The last index discussed by Guttman uses linear regression theory. Every item can be used as a criterium in a linear regression with the remaining $k-1$ items

as regressors or predictors. Let ε_i denote the residual variance when predicting item i from the remaining items. Then

$$\lambda_6 = 1 - \frac{\sum_i \varepsilon_i}{\sigma_X^2} \tag{2.9}$$

is a lower bound of the reliability.

In an effort to develop other and better lower bounds than the ones presented by Guttman, Jackson and Agunwamba (1977, p. 574) were able to make the lower bound λ_2 sharper. Let g denote the value of $j(\neq i)$ for which σ_{ig}^2/σ_g^2 is largest, and similarly let σ_{hj}^2/σ_h^2 be the largest value of σ_{ij}^2/σ_i^2, $(h \neq j)$, and define

$$d_{ij}^2 = \sigma_i^2 \sigma_j^2 \max(r_{ij}^2,\ r_{ig}^2 \times r_{hj}^2), \tag{2.10}$$

where r^2 denotes the squared correlation, then

$$\lambda_7 = \frac{\sum_{i \neq j} \sigma_{ij} + \left(\frac{k}{k-1} \sum_{i \neq j} d_{ij}^2\right)^{1/2}}{\sigma_X^2}. \tag{2.11}$$

Since it follows directly from (2.10) that $d_{ij}^2 \geq \sigma_{ij}^2$, it is clear that $\lambda_7 \geq \lambda_2$, so that from (2.4) we have the ordering

$$\alpha = \lambda_3 \leq \lambda_2 \leq \lambda_7 \leq Rel. \tag{2.12}$$

Although the definition of d_{ij}^2 is a bit complicated, its computation is easy and fast, so that there is no reason, except maybe historical interest, to keep computing Cronbach's α or Guttman's λ_2. Unfortunately, however, λ_7 is not the sharpest possible lower bound for the reliability. In fact, examples can be constructed where any of the indices $\lambda_4, \lambda_5, \lambda_6, \lambda_7$ or a μ -coefficient of sufficient high order is larger than the others (see Woodhouse and Jackson, 1977 and Ten Berge and Zegers, 1978).

In their effort to find the greatest lower bound for the reliability from a covariance matrix, Woodhouse and Jackson (1978) found an interesting result that applies in some cases. If in a 2×2 covariance matrix, the following inequality is fulfilled:

$$\sigma_i^2 < |\sigma_{ij}| < \sigma_j^2, \tag{2.13}$$

then the index

$$WJ2.6 = \frac{(\sigma_i^2 + \sigma_{ij})^2}{\sigma_i^2 \sigma_X^2} \tag{2.14}$$

is a lower bound for the reliability. Moreover if $k = 2$, and (2.13) is fulfilled, then (2.14) is the greatest lower bound. (The name of the index derives from the initials of the authors and the formula number in the article.) The generalization of this result to k items is difficult. Some aspects of the problem are discussed in the next section.

3. The general theory of lower bounds

In deriving his results, Guttman used one basic axiom of classical test theory, viz., the experimental independence of the measurement errors, meaning that the measurement errors of the same person on different items are independent and the measurement errors of two different persons on the same or different items are independent as well. If this axiom is fulfilled, the observed variance-covariance matrix of the item scores X can be decomposed as

$$\Sigma_X = \Sigma_T + U, \tag{3.1}$$

where Σ_T is the covariance matrix of the true scores and U is a diagonal matrix containing the variances of the measurement errors on the main diagonal. The reliability of the test score X is given by

$$Rel = 1 - \frac{\mathrm{tr}(U)}{\mathbf{1}'\Sigma\mathbf{1}}, \tag{3.2}$$

where tr(.) is the trace function and $\mathbf{1}$ is a vector with all elements equal to 1. But since U is not observed, (3.2) cannot be computed. Therefore it is impossible to know the reliability of a test from a single test administration. On the other hand, it is also not correct to state that from a single test administration no information about the reliability becomes available. The reason is that the matrices U and Σ_T are not completely arbitrary, because from test theory, they must have certain characteristics. These characteristics or restrictions are:

1. The elements on the main diagonal of U must be nonnegative since they are variances;

2. The matrix T must be positive semidefinite (or Gramian), since it is a covariance matrix.

Since X is observed, it follows from (3.1) that choosing U determines Σ_T. So computing the right-hand side of (3.2) for a matrix U fulfilling the two aforementioned restrictions gives a possible value of Rel. The smallest possible value of

Rel is obtained by choosing a matrix U with a trace as large as possible and which fulfills the two restrictions. This value of *Rel* is the greatest lower bound (glb.) of the reliability and all the indices discussed in the preceding section are not larger than this bound. Finding the greatest lower bound is not easy. For a detailed discussion of this problem, see Ten Berge, Snijders and Zegers (1981), who developed an algorithm, in the form of an iterative procedure, to find the matrix U with the largest trace. The algorithm, however, does not guarantee that the corresponding Σ_T matrix is Gramian. If it is not, the algorithm has to start from another configuration of initial values. The main difficulty with this algorithm is that the decision whether Σ_T is Gramian or not has to be taken on the basis of numerical results, and small negative eigenvalues may result from rounding errors or from stopping too early with the iterative procedure. The latter possibility may be checked by continuing the iterative procedure and checking whether the negative eigenvalues decrease in magnitude. The consequence is that the whole algorithm is rather time consuming. For a test of 80 items, finding the glb may take several minutes on a Pentium computer (130 MHz).

4. Some heuristics to find lower bounds

In discussing the index λ_4, it was argued that partitioning the items in two non-empty subsets does not alter the test score, but does alter the covariance matrix of the resulting pair of combined items. And α, computed on this 2×2 matrix, is a lower bound of the reliability. Of course, this argument is also valid if the original items are partitioned into more than two subsets. One could compute α on all possible partitions of the original items in $2, 3, \ldots, k-1$ subsets. The number of different partitions is, however, very large. For $k = 15$, this number is $1,382,958,544$, as opposed to $2^{14} - 1 = 16,383$ partitions in exactly two subsets. Technically, the number of partitions of k objects into p non-empty subsets is a so-called Stirling number of the second kind, denoted $S_k^{(p)}$. The number of partitions of 15 objects into 2 or more non-empty subsets is given by

$$\sum_{p=2}^{15} S_{15}^{(p)}.$$

The Stirling numbers themselves are defined by the recurrence relation

$$S_k^{(p)} = S_{k-1}^{(p-1)} + pS_{k-1}^{(p)}, \tag{4.1}$$

which can be computed, using the simple results

$$S_k^{(1)} = S_k^{(k)} = 1, \tag{4.2}$$

and

$$S_k^{(k+1)} = 0. \tag{4.3}$$

For $p \approx k/2$, the Stirling numbers grow very fast as k increases. So, a complete enumeration is not realistic, even in cases with small k. Therefore we have to look for ways to find a partition which can find the highest possible α or one quite close to it in a very short time. Two such methods or heuristics will be discussed. The first searches a partition in the set of all partitions; the second searches a partition into two subsets with the aim to find λ_4. These heuristics are discussed in turn.

Heuristic 1

Suppose that in a set of k items a combined item is formed by summing the scores of items g and h, and this item is labeled Y. Then it is easily seen that

$$\sigma^2(Y) = \sigma_g^2 + \sigma_h^2 + 2\sigma_{gh}, \tag{4.4}$$

$$\sigma_{Yi} = \sigma_{gi} + \sigma_{hi}, \quad (i \neq g, h). \tag{4.5}$$

Denoting the sum of the covariances in the original case as R and in the case with the combined item as R^*, it follows from (4.4) and (4.5) that

$$R^* = R + 2\sigma_{gh}. \tag{4.6}$$

Denoting Cronbach's α in the original case as α, and in the case with the combined item as α^*, and remembering that in the latter case the number of items is $k-1$, it follows readily that

$$\alpha^* > \alpha \Leftrightarrow \sigma_{gh} < \frac{R}{2\,(k-1)^2}. \tag{4.7}$$

Of course there may be more than one pair (g, h) for which the inequality in the right hand side of (4.7) is fulfilled. Therefore two versions of the heuristic are proposed. In the deterministic version the pair is chosen for which the difference $\alpha^* - \alpha$ is largest; in the probabilistic version a pair is chosen at random from the pairs for which inequality (4.7) is fulfilled. Once the merging of the selected items is carried out, a test with $k-1$ items is defined on which a new merging may be applied. The process stops if no more pairs can be found which fulfill

(4.7). If the covariance matrix is given, checking (4.7) and computing the merging process (4.4 and 4.5) are very fast operations, but to do the next merging a lot of administrative operations must be performed to rebuild the covariance matrix, making the heuristic relatively expensive. Applying the random version of the heuristic where the whole process of merging is repeated a large number of times is relatively time consuming, and may become prohibitive in routine applications.

Two further remarks are in order with this heuristic. In the first place, the maximum over the set of all partitions may be smaller than the greatest lower bound, and in the second place, the heuristic does not guarantee that all possible partitions can be reached. In fact, the heuristic always starts from the finest partition and proceeds to a merging of two items only if coefficient α increases. There may be a point in the merging process where α cannot raise by merging one single pair of items but may raise by merging two pairs at the same time. Consider the following covariance matrix for $k = 4$:

$$\begin{bmatrix} 1 & 0.24 & 0.36 & 0.30 \\ 0.24 & 1 & 0.30 & 0.40 \\ 0.36 & 0.30 & 1 & 0.40 \\ 0.30 & 0.40 & 0.40 & 1 \end{bmatrix}.$$

For this matrix coefficient α is $2/3 = 0.667$. Checking inequality (4.7) reveals that α can increase only if there is a covariance smaller than $4/18 = 0.222$. So the heuristic will stop here, but combining items 1 and 4 and items 2 and 3 gives a covariance between the two combined items of 1.40 and a coefficient α of 0.70, which for this case also equals λ_4.

Heuristic 2

Suppose that the set of k items is partitioned into two non-empty subsets. To indicate these subsets, define a k-vector $\mathbf{u}'=(u_1, \ldots, u_k)$ with $u_i \in \{-1, 1\}$ and where items i and j belong to the same subset if and only if $u_i = u_j$. Jackson and Agunwamba (1977) have shown that coefficient α associated with this partition is given by

$$\alpha(\mathbf{u}) = 1 - \frac{\mathbf{u}'\Sigma_X\mathbf{u}}{\sigma_X^2}. \tag{4.8}$$

It is clear that

$$\lambda_4 = \max_{\mathbf{u}} \alpha(\mathbf{u}) = 1 - \frac{\min_{\mathbf{u}} (\mathbf{u}'\Sigma_X\mathbf{u})}{\sigma_X^2}. \tag{4.9}$$

Now suppose that a certain partition with associated $\mathbf{u}$-vector is given, then we might try to reach a higher α by moving some items to the other subset, i.e.

changing their algebraic sign in the $\mathbf{u}$-vector. Suppose that for a number of items the sign will be changed, and that the items are reordered such that the unchanged items appear first and then the changed items. Correspondingly we partition the vector $\mathbf{u}$ as $(\mathbf{u}_1', \mathbf{u}_2')'$ and the covariance matrix as

$$\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

From (4.9) it follows that

$$\alpha(\mathbf{u}_1, -\mathbf{u}_2) > \alpha(\mathbf{u}_1, \mathbf{u}_2) \Leftrightarrow \mathbf{u}_1' \Sigma_{12} \mathbf{u}_2 > 0. \tag{4.10}$$

If the vector $\mathbf{u}_2$ contains only one element, inequality (4.10) is very easy to check. If a vector $\mathbf{u}$ is given, the heuristic consists of the following steps:

1. For each item: check (4.10). If α increases by changing the sign of u_i, do so.
2. If at least one change of sign has occurred in step 1, go back to step 1; else stop.

As was remarked in connection with Heuristic 1, it is not certain that λ_4 is found with this procedure, since only one element at a time is replaced, and the procedure stops if α does not increase by moving a single element. So one might try improvement by replacing two elements at a time also, and then three, and so on, but this is the same as a complete enumeration. A better method seems to be to apply the heuristic as described here from a number of different starting vectors and picking the largest result as the best approximation to λ_4.

In order to have an impression of the effect of the heuristics, the reliability of the Dutch mathematics central examination (denoted as MAVO-C) of 1997 is estimated from a sample of 2236 examinees ($k = 25$). The results are displayed in Table 4.1. The two heuristics are applied 500 times each. Notice that there are no special heuristics developed for $\mu_1, \mu_2, \lambda_5, \lambda_6$ and λ_7. In every application of the random version of Heuristic 1 these coefficients are computed for every partition encountered and the maximum value is retained. The maximum, minimum, average and standard deviation are taken over the 500 replications.

The highest coefficient found is the approximation to λ_4 (printed in bold face). The value of the glb is 0.803. The row 'Original' denotes the coefficients computed on the original 25×25 covariance matrix; the row 'Fixed' denotes the deterministic version of Heuristic 1. Although the difference between the highest index and the

Table 4.1: Examples of the different indices on a mathematics examination

			Heuristic 1				H.2
	μ_0	μ_1	μ_2	λ_5	λ_6	λ_7	λ_4
Original	.719	.730	.732	.717	.743	.737	
Fixed	.744	.750	.751	.737	.743	.750	
Max.	.774	.774	.774	.781	.744	.774	**.797**
Min.	.731	.735	.736	.718	.743	.737	.762
Average	.745	.749	.751	.744	.743	.750	.786
S.D.	.006	.006	.006	.012	.000	.006	.006

usually applied α on the original number of items is striking in this example, it is certainly not justified to state that λ_4 is the best index in general. In cases with other covariance structures, one of the other indices might do better. Therefore it is important to gain insight in the reasons why the coefficients are in general smaller than the reliability. This problem will be addressed in the next section.

5. The relation between α and the reliability

A standard result in classical test theory is the following (Lord and Novick, 1968, p. 90):

> Coefficient α (computed on k test scores) equals the reliability if and only if the k tests are mutually essentially τ-equivalent, i.e., $T_i = T_j + a_{ij}$, where T_i and T_j are the true scores on tests i and j. In all other cases, coefficient α is smaller than the reliability.

The purpose of the present section is to gain some insight in the relation between coefficient α and the reliability. Therefore a general model will be considered:

$$X_i = \sum_{g=1}^{p} a_{ig}\theta_g - \beta_i + \varepsilon_i, \tag{5.1}$$

where ε_i is a random variable representing the measurement error. It will be assumed that ε_i has finite variance σ_i^2 and expectation zero. The symbol θ_g represents a factor score on the g-th factor, and it is assumed that factor scores are associated with persons, and that they are constant during test administration.

The parameter β_i represents the difficulty of item i. The coefficient a_{ig} will be called the factor loading of item i on factor g. The variance of the measurement error may be dependent on the factor scores or independent of it, and the measurement error itself may be continuous or discrete. The only assumption that is made is experimental independence between the item responses. It follows directly from these assumptions and from (5.1) that for an arbitrary person

$$\mathcal{E}(X_i) = \sum_{g=1}^{p} a_{ig}\theta_g - \beta_i, \tag{5.2}$$

so that in a population where the factor scores θ_g can take arbitrary values, two measures (test scores or item scores) X_i and X_j are essentially τ-equivalent if and only if $a_{ig} = a_{jg}, (g = 1, \ldots, p)$.

To gain insight in the relation between α and the reliability, three cases will be considered. In the first case it is assumed that the covariance between the factor scores vanishes for all pairs of factors, i.e. the case of uncorrelated factor scores. In the second case, which is a bit artificial, it will be assumed that the correlation within each pair of factors is constant. In the third case, which presumably approaches real applications best, a general factor and a number of uncorrelated group factors, will be studied.

Case 1 uncorrelated factors

Suppose there are p uncorrelated factors, and assume that each item in the test has a non-zero loading on one factor only. Suppose, moreover, that there are exactly m items loading on each factor, half of which have a loading a_g, the other half having loading b_g. So the whole test consists of $k = pm$ items. To fix the unity of the factors, a population is considered where all factors have a variance equal to one. As an example consider the case where $m = 4$. The covariance matrix of the true sores $T_i, (i = 1, \ldots, 4)$ for the m items loading on the same factor is given by

$$\begin{bmatrix} a^2 & a^2 & ab & ab \\ a^2 & a^2 & ab & ab \\ ab & ab & b^2 & b^2 \\ ab & ab & b^2 & b^2 \end{bmatrix}.$$

The variance of the true scores $T = \sum_i T_i$ in the population is given by

$$\mathrm{Var}(T) = \left(\frac{m}{2}\right)^2 \left(\sum_g a_g^2 + \sum_g b_g^2 + 2\sum_g a_g b_g\right). \tag{5.3}$$

Denoting the sum of the covariances as R, it is easily verified that in the present case

$$R = \text{Var}(T) - \frac{m}{2}\sum_g a_g^2 - \frac{m}{2}\sum_g b_g^2. \tag{5.4}$$

The ratio $\frac{\alpha}{Rel}$ is given in general by

$$\frac{\alpha}{Rel} = \frac{k}{k-1} \times \frac{R}{\text{Var}(T)}. \tag{5.5}$$

Notice that the relationship between α and the reliability is not affected by the variances of the measurement errors. In the present case, this gives

$$\frac{\alpha}{Rel} = \frac{k}{k-1} \times \left[1 - \frac{\sum_g a_g^2 + \sum_g b_g^2}{\frac{m}{2}\left(\sum_g a_g^2 + \sum_g b_g^2 + 2\sum_g a_g b_g\right)}\right]. \tag{5.6}$$

To generate more specific cases, it is assumed that $a_g = a$ and $b_g = b$, and that both a and b differ from zero. Setting

$$b = ca,$$

(5.6) reduces to

$$\frac{\alpha}{Rel} = \frac{k}{k-1} \times \left[1 - \frac{1+c^2}{\frac{m}{2}(1+c)^2}\right], \tag{5.7}$$

if $c \neq -1$. If $c = -1$, the ratio is not defined because $\alpha = Rel = 0$. As an interesting side result, (5.7) can be used to show that α is not bounded from below:

$$\lim_{c \to -1} \frac{\alpha}{Rel} = -\infty.$$

Specializing further, and setting $c = 1$ yields a test consisting of p independent subtests, and within each subtest the m items are essentially τ-equivalent subtests. The ratio is given by

$$\frac{\alpha}{Rel} = \frac{p(m-1)}{mp-1} = \frac{k-p}{k-1}. \tag{5.8}$$

If $m = 1$, $\alpha = 0$ and the ratio is zero. As m increases, while p is kept constant, the ratio approaches one.

Case 2 correlated factors

To get relatively simple formulae, only the special case will be considered where the correlation between factors is constant, ρ say. The variance of the sum of the true scores is then given by

$$\mathrm{Var}(T) = \left(\frac{m}{2}\right)^2 \left[\sum_g a_g^2 + \sum_g b_g^2 + 2\sum_g a_g b_g + 2\rho \sum_{g<h} (a_g + b_g)(a_h + b_h)\right]. \quad (5.9)$$

Since (5.4) is also valid in case of correlated factors, the ratio $\frac{\alpha}{Rel}$ will be lower in this case than in the case of uncorrelated factors only if the last term in the right-hand side of (5.9) is negative. Assuming that in a realistic case, all factor loadings are nonnegative, the ratio $\frac{\alpha}{Rel}$ will grow with increasing correlations. Assuming again that for all factors $b_g = b$ and $a_g = a$, and setting as before $b = ca$, the ratio $\frac{\alpha}{Rel}$ is given by

$$\frac{\alpha}{Rel} = \frac{k}{k-1} \times \left[1 - \frac{1 + c^2}{\frac{m}{2}(1+c)^2 [1 + \rho(p-1)]}\right], \quad (5.10)$$

and in case $c = 1$, this reduces further to

$$\frac{\alpha}{Rel} = \frac{k}{k-1} \times \left[1 - \frac{1}{m[1 + \rho(p-1)]}\right]. \quad (5.11)$$

In Figures 1 and 2, the right-hand side of (5.10) is graphed as a function of c and ρ respectively, for $p = 4$. Notice that the value of (5.10) does not change if c is replaced by $1/c$, so in Figure 1, c ranges from zero to one.

Summarizing, two characteristics of the general model (5.1) influence the ratio α/Rel: the multidimensional structure of the underlying factors θ_g, i.e. the number of factors and their intercorrelations, and the variability of the factor loadings of the items. The ratio, however, is not affected by the measurement error and by the item difficulty.

Case 3 One general factor and group factors

To keep the resulting formulae simple, it will be assumed that all item loadings on the general factor are equal to a and that there are p group factors where m items have a nonzero loading; so $k = pm$. The factor loadings on the group factors will be assumed equal to b for the items loading on them. All factors are assumed to be uncorrelated. The variance of the true test score T is given by

$$\mathrm{Var}(T) = k^2a^2 + pm^2b^2, \quad (5.12)$$

and the sum of the covariances R by

$$R = k(k-1)a^2 + k(m-1)b^2, \tag{5.13}$$

whence

$$\frac{\alpha}{Rel} = \frac{k}{k-1} \times \frac{(k-1)a^2 + (m-1)b^2}{ka^2 + mb^2}. \tag{5.14}$$

Using, as before, the relation $b = ac$, gives

$$\frac{\alpha}{Rel} = \frac{k}{k-1} \times \frac{(k-1) + (m-1)c^2}{k + mc^2}. \tag{5.15}$$

An interesting case results when $m = 1$, i.e. when the group factors become specific factors (to be distinguished from measurement error, because they generate true score variance). In this case (5.15) reduces to

$$\frac{\alpha}{Rel} = \frac{k}{k + c^2}, \tag{5.16}$$

which has two interesting limits: if c is fixed, then $\frac{\alpha}{Rel}$ approaches unity as $k \to \infty$. On the other hand, if k is fixed and $c^2 \to \infty$ (which is the case if b is fixed and nonzero and the loading a on the general factor approaches zero), then the ratio $\frac{\alpha}{Rel}$ and therefore α itself goes to zero. The case where a equals zero is the case with uncorrelated factors which has been discussed as case 1.

The case with $m = 1$ is also interesting to demonstrate the impossibility of estimating the reliability of an arbitrary test on the basis of a single test administration and without making assumptions on parallelism or τ-equivalence or the like. Assume that the k items in the test load equally on the general factor with factor loading a ($\neq 0$), that each item loads on a specific factor with loading b and has an error variance equal to τ^2. Let $A = b^2 + \tau^2$. The observed covariance matrix is given by

$$\begin{pmatrix} a^2 + A & a^2 & \cdots & a^2 \\ a^2 & a^2 + A & \cdots & a^2 \\ \vdots & \vdots & \ddots & \vdots \\ a^2 & a^2 & \cdots & a^2 + A \end{pmatrix}.$$

It can be shown that the glb associated with this matrix corresponds to the case where the true variance of each item equals a^2, i.e.,

$$\text{glb} = \frac{k^2a^2}{k^2a^2 + kA} = \frac{ka^2}{ka^2 + A}.$$

and this is valid for all tests with $A = b^2 + \tau^2$. For a test with $b^2 = 0$, the glb equals the reliability, and for a test where $b^2 = A$, the reliability is one and the glb underestimates the reliability. In other words, within a single test administration specific factors and measurement error are confounded and cannot be disentangled.

To conclude this section, one further remark is in order. One might wonder whether the general model (5.1) is general enough to be applied in practical test situations, since in most tests the item scores are binary or have at most a very limited number of values. Nevertheless, there is no objection in principle to use this model in the case of binary items, as long as no further assumptions are added which are incompatible with the basic model and the data. Suppose X_i is binary, then it follows immediately from (5.2) that

$$p_i(\theta_1, \ldots, \theta_g) \equiv P(X_i = 1 \mid \theta_1, \ldots, \theta_g) = \mathcal{E}(X_i \mid \theta_1, \ldots, \theta_g) = \sum_{g}^{p} a_{ig}\theta_g - \beta_i. \quad (5.17)$$

If the item factor loadings and the item difficulty are given, it follows from (5.17) that the space of the factor scores must be bounded because $0 \leq p_i(\theta_1, \ldots, \theta_g) \leq 1$. So the general model applied to binary items is not compatible with the assumption of a multivariate normal distribution of the factor scores. But the model can be used and, in fact, is used when the covariance matrix calculated on binary items is submitted to a factor analysis. Although linear models for binary items are in general not very parsimonious with regard to the number of factors required to describe the covariances adequately, the results of the present section are valid also in the case of binary items.

6. Statistical properties

The results of the preceding sections all concern the relation between different indices and the reliability which can be derived from the observed covariance matrix, i.e. from a matrix which contains the variances of and covariances between observed scores in some population. But this matrix itself is in general not observed: only an estimate, based on the responses of n (randomly drawn) persons to k items is given, and the indices computed on this matrix are of course only estimates of their population analogs, and these estimates may be larger or smaller than the population value they estimate. Although the relation $\alpha \leqslant \text{Rel}$ is generally valid, the relation $\widehat{\alpha} \leqslant \text{Rel}$ is not true in general.

Theoretical results on the statistical properties of the discussed indices are sparse (Feldt, 1965; Kristof, 1963), and, moreover, are only valid under quite strict assumptions. Feldt, for example, derived the sampling distribution of coefficient α under the assumption that the items are essentially τ-equivalent, and that true scores and measurement errors are normally distributed. His main finding was that the ratio

$$\frac{1-\alpha}{1-\widehat{\alpha}} \sim F_{(n-1),(n-1)(k-1)}, \tag{6.1}$$

i.e., the ratio $(1-\alpha)/(1-\widehat{\alpha})$ follows the central F-distribution with $(n-1)$ and $(n-1)(k-1)$ degrees of freedom, where α is the population coefficient and $\widehat{\alpha}$ the sample estimate. This result may be used to define the $100(1-a)$ per cent confidence intervals for α. Denoting the degrees of freedom as $m_1(=n-1)$ and $m_2(=(n-1)(k-1))$, the confidence interval is given by

$$P\left[1-(1-\widehat{\alpha})F_{(1-\frac{1}{2}a)(m_1,m_2)} < \alpha \leqslant 1-\frac{(1-\widehat{\alpha})}{F_{(1-\frac{1}{2}a)(m_2,m_1)}}\right] = 1-a, \tag{6.2}$$

where $F_{(1-\frac{1}{2}a)(m_1,m_2)}$ is the $(1-\frac{1}{2}a)$ upper critical point in the F-distribution with m_1 and m_2 degrees of freedom. Notice the reversal in the order of the degrees of freedom in both expressions in (6.2); Feldt warns against several erroneous formulae published in the older literature.

From the properties of the F-distribution and from (6.1), it follows that

$$\frac{1-\widehat{\alpha}}{1-\alpha} \sim F_{(n-1)(k-1),(n-1)}, \tag{6.3}$$

and that, using standard results on the F-distribution,

$$\mathrm{Var}(\widehat{\alpha}) = \mathrm{Var}(1-\widehat{\alpha}) = (1-\alpha)^2 \frac{2\,(n-1)\,[k\,(n-1)-2]}{(k-1)\,(n-3)^2\,(n-5)} \tag{6.4}$$

if $n > 5$. For large n, (6.4) can be approximated by

$$\mathrm{Var}(\widehat{\alpha}) \cong (1-\alpha)^2 \frac{2k}{n\,(k-1)} \tag{6.5}$$

Notice that the right hand sides of (6.4) and (6.5) depend on α. A practical result may be obtained by replacing α by its sample estimate $\widehat{\alpha}$.

It follows also from (6.3) that

$$\mathcal{E}\left(\frac{1-\widehat{\alpha}}{1-\alpha}\right) = \frac{n-1}{n-3}, \quad (n > 3), \tag{6.6}$$

which means that $\widehat{\alpha}$ underestimates true α, but this bias disappears rapidly if n increases.

The problem which will be addressed in the present section is more general. If the population covariance matrix is known, we have a lot of coefficients which do better in estimating the lower bound of the reliability than Cronbach's α. But we do not know what the behavior of these coefficients is in case we have only an estimate of the covariance matrix. A number of these coefficients, like the glb or λ_4 are maximal function values defined on the covariance matrix as given. This means that these coefficients might tend to profit maximally from the sampling errors in the covariance matrix, such that the resulting estimate is positively biased, i.e. will tend to give larger estimates than would be obtained if the population covariance matrix were used to compute them. It is interesting of course to have a rather precise idea of this bias.

The method that will be followed to investigate the statistical properties of the coefficients consists of a simulation study, which will be described now. One simulation study consists of the following steps

- **step 1**. Choose k, the number of items and p, the number of factors. Fix the $k \times p$ matrix of factor loadings and the k error variances in the model defined by (5.1). It will be assumed throughout that the factors are uncorrelated and are normally distributed with zero mean and unit variance. Since only continuous responses will be considered in the present section, the difficulty parameters are irrelevant and can be fixed to any value.

- **step 2**. Given the model as defined in step 1, the population covariance matrix can be computed, as well as coefficient α and the reliability. All the coefficients discussed in the preceding sections can be computed on this covariance matrix. Notice that none of these can exceed the reliability.

- **step 3.** To obtain a sample covariance matrix, n response patterns are drawn following model (5.1) and the specifications of step 1, resulting in a $n \times k$ response matrix. The sample covariance matrix of the item responses is computed from this response matrix. This procedure is replicated r times and yields r sample covariance matrices.

- **step 4.** On each of the r sample covariance matrices, the coefficients discussed in the preceding sections are computed, yielding for each coefficient a random sample of r estimates. The distribution of these estimates is the main outcome of the simulation study.

Of course, this whole procedure can be applied to any factorial structure, so that it is not possible to give a 'complete' description of the statistical properties of the coefficients discussed. It is even impossible to give a good definition of what complete means. Therefore, only two examples will be given, which are deemed to be typical in research settings. The computer program to do the simulations is available on request from the author.

The two examples have a number of features in common. Both follow a simulation design with two facets. The first facet is the number of items which takes the values $10, 20$ and 40. The second facet is the number of respondents to the tests which take the values $200, 400, 600, 800, 1000, 1200, 1600, 2000$ and 2400, the latter number being approximately the sample size used at Cito to estimate the reliability of the central examinations taken in Dutch secondary education. In each cell of the design 500 replications were run, and within each run 100 restarts were carried out for the approximation of λ_4 (Heuristic 2) as well as for the random version of Heuristic 1. In each run all the coefficients discussed in the first section as well as the glb were computed, yielding for each coefficient an empirical distribution of 500 sample estimates. Of each distribution the mean and the standard deviation, the median and the 10th, 25th, 75th and 90th percentile are computed.

The results will be displayed graphically. It appeared from the output that in all cases the difference between the mean and the median estimates was negligible. To do right to the skewness of the distributions the median will be displayed in the interval $(P25, P75)$, i.e., the interval contains the middle 50% of the estimates.

6.1. Example 1: essential τ-equivalence

In this example the statistical properties of the coefficients were investigated in a situation were coefficient α equals the reliability. To this end a unifactorial test was constructed where all items had a loading of one on the factor, and all items had the same error variance. By manipulating the error variance the reliability can be altered. For the ease of the presentation of the results the simulation design was extended with a third facet, the reliability, which takes two values,

Table 6.1: Reliability of the six tests using Spearman-Brown

k	*Rel*	*Rel* with $k = 40$
10	0.7	0.903
20	0.7	0.824
40	0.7	0.700
10	0.9	0.973
20	0.9	0.947
40	0.9	0.900

0.7 and 0.9. Admittedly, the combination of this facet with the facet test-length may give rise to somewhat unrealistic situations: a test of 40 items having a reliability of 0.7 may be rejected for being too unreliable, while a test consisting of only 10 items and having a reliability of 0.9 may be deemed too optimistic in many situations. To appreciate the reality value of the constructed tests, their reliability is given under homogeneous lengthening to 40 items in table (6.1), using the Spearman-Brown formula.

The main results are displayed in three panels. The first panel summarizes the results of the glb-estimate, and the results of Heuristic 2, which tries to find λ_4. (See Figures 3 through 8) Because it was anticipated that these estimators would have a positive bias, two results were retained from the application of Heuristic 2. The first is the maximal value found over the 100 random starts of the heuristics procedure. This result is displayed as 'split-half (max.)'. To attenuate the effect of capitalizing on sampling error, also the average result over the 100 random starts was computed, and is displayed as 'split-half (ave.).' Although comparison of the Figures 3 through 8 is a bit difficult, due to the different scales used for the ordinate, the following conclusions are obvious

- All three indices systematically overestimate the reliability, and they do so in an invariant order, as necessarily follows from their definitions: glb $\geq$ split-half (max.) $\geq$ split-half (ave.).
- The bias increases with increasing test length, which may be explained by the fact that a larger covariance matrix offers more opportunities to capitalize on sampling error.
- The bias disappears rather slowly with increasing sample size, and is con-

siderable for the glb also in the case of the largest sample size used in the simulation study. Although the Figures 3, 4 and 6 suggest that the bias is absent or disappears rapidly for the average (over random starts) of Heuristic 2 results, this finding can certainly not be generalized to all cases, as appears clearly, for example, from Figure 7.

- It is a bit more difficult to appreciate the bias as a function of the reliability of the test. Because all indices are bounded, it is to be expected that bias will decrease as the reliability increases. If we change to another metric, however, this dependency could change considerably. A reasonable approach might be the following. Consider the true reliability as defined in the simulation study as the target in the construction phase of a test, and consider the indices as an estimate of the reliability resulting from a pilot study. Ignoring the bias, the difference between the estimate and the target may be used to save on the number of items in the final version of the test which should reach the target reliability. The final length as a proportion p of the pilot length is the effect of the bias. Using the Spearman-Brown formula, and glb as index, this proportion is the solution of the equation

$$Rel = \frac{p \text{ glb}}{1 + (p - 1) \text{ glb}}, \tag{6.7}$$

which is

$$p = \frac{Rel(1 - \text{ glb})}{\text{glb } (1 - Rel)}. \tag{6.8}$$

In Figure 9, p for the glb estimate is plotted against the sample size, for the three test lengths used and for the two values of the true reliability, showing that for all three test lengths the effect of the bias is more pronounced for the lower reliability. An advantage of using this metric is that it shows clearly the impact of the bias in terms of test design, and that the impact is dramatic if the sample size used in the pilot study is small, as it often is. Take for example the case where the target is 0.9 and $k = 40$, in a pilot study with a sample size of 200 students. The median glb estimate is 0.959, yielding a proportion of 0.385, or a test with 15 or 16 items. If such a test were built, it would have, using Spearman-Brown, a reliability of 0.776.

The second panel shows the effect of the design facets on the indices which are computed in a direct way from the sample covariance matrix, i.e. all the λ-

and μ-coefficients with the exception of λ_4. As may be expected from the theory treated in the first section, in the case of a test with essential τ-equivalent items, where the error variances are equal, coefficients which exploit unequal covariances (λ_5 and λ_7), or which in general can do better than coefficient α will give no effect here, because the reliability equals α. The sampling distributions of these indices were indeed very similar. The only index which does not have a clear-cut relation to α is λ_6, which is based on linear regression. Therefore only the results for coefficient α ($= \lambda_3 = \mu_0$) and for λ_6 are displayed. (See Figures 10 through 15). The results may be summarized as follows

- The sample estimate of α behaves neatly as predicted: no bias in any of the six conditions.
- The behavior of coefficient λ_6 is much stranger, and as far as known, it has never been noticed. It seems to be a non-increasing function of the sample size, decreasing first and then seemingly leveling off to a constant level, which seems to underestimate true reliability, at least in the cases with 10 and 20 items. In the case of 40 items it seems to converge to the correct value. If λ_6 is computed on the population covariance matrix, it also underestimates true reliability in the cases $k = 10$ and $k = 20$ with about the same amount as shown in the Figures, suggesting that λ_6 is an asymptotically (as $k \to \infty$ and $n \to \infty$) unbiased estimate, at least in the case with essentially τ-equivalent items, but shows a small negative bias for small k. The reason for this bias is not understood, but since it is very small (about -0.005 for $k = 10$), it will not be studied further.

In the third panel, the results of Heuristic 1 are displayed. Remember that this Heuristic combines items with the aim to maximize coefficient $\dot{\alpha}$. Only results for the fixed procedure will be given (i.e. the procedure where in each step that merging is applied for which the increase in α is largest), because the random procedure yielded almost the same results. As in the second panel, only the distribution of α and λ_6 will displayed. See Figures 16 through 21. The conclusions can be summarized as follows

- As was the case with the glb- and the λ_4-estimates (see first panel), the sampling error causes positive bias in the α-estimates. The bias decreases with sample size and increases with test length. The bias is less severe than for the glb- and λ_4-estimates. For this example, less bias is a good thing,

Table 6.2: Design for the second example

k	p	m	*Rel*	α	α_{APA}
10	5	2	0.700	0.562	0.549
20	5	4	0.824	0.746	0.696
40	5	8	0.903	0.862	0.805

but more in general, it means that Heuristic 1 performs worse than Heuristic 2, since it ends with an estimate which is based on some partition of the items, and which yields a lower value of the associated α than the one that is found in Heuristic 2.

- The λ_6-estimates in general follow the same pattern as the α-estimates, but are somewhat lower, meaning that for large samples they are negatively biased for small k. This negative bias is larger than in the case where λ_6 is based on the covariance matrix. But notice that Heuristic 1 was devised to maximize α, and not to maximize λ_6.

6.2. Example 2: One general factor and group factors

Although many tests are constructed with the intention to measure a single concept (factor), it is often found that a single factor does not explain all the common variance. A situation where a single factor does not suffice, and which has retained much attention in recent years (e.g. Wainer & Lewis, 1990; Thissen, Steinberg & Mooney, 1989) is the case of so-called testlets, where items in a test are organized in such a way that several items share some of the item material. For example, if in a reading test several questions are asked about the same text, these items together form a testlet, and usually the covariance between items of the same testlet is larger than can be explained by a single common factor. If it is assumed that testlets are statistically independent, the test scores can generally be explained by one common factor and a group factor for each testlet.

The design for the example is partly displayed in table 6.2. The starting point is a test of 10 items consisting of five testlets with two items each. The test has a reliability of 0.70, and for each item the loading on the general factor is 1, and the loading on the testlet factor is 0.5. Moreover, within each testlet the items have a different error variance, σ^2 and $2\sigma^2$ respectively. Using (5.12) σ^2 can be chosen such that the reliability is 0.7, and from (5.15) the value of α follows directly. The tests with 20 and 40 items are not homogeneous lengthenings of the short

test, but consist also of five testlets, each consisting of 4 or 8 items. Within each testlet half of the items have error variance σ^2 and the other half $2\sigma^2$; the factor loadings on the general factor and the testlet factor are 1 and 0.5 respectively. The column α_{APA} will be explained in the discussion section.

The other facet of the design is the sample size, which takes the same values as in the first example. Data were generated and indices were calculated in the same way as in the first example: for each cell of the design 500 runs and within each run 100 random restarts for the approximation of λ_4.(Heuristic 2). To save computing time, the random version of Heuristic 1 was not applied .

The results are displayed in much the same way as in example 1. In the first panel (Figures 22, 23 and 24) the glb estimate and the two estimates related to λ_4 are displayed. They show a similar pattern as in example 1 with two noticeable differences: the positive bias is slightly less than in the first example, and the averaged split-half estimate looks like an ideal compromise between the weakness of the heuristic (i.e., not finding the sample λ_4) and the bias associated with the λ_4-estimate. Even in the case of 40 items, the bias has disappeared almost completely with 1000 observations.

For the second panel (coefficients calculated on the covariance matrix as given), four coefficients were chosen: $\mu_0(=\alpha)$, $\mu_1(=\lambda_2)$, λ_6 and λ_7. Coefficients μ_2 and λ_5 did not differ much from μ_1, as could be expected. For μ_2, it was already stated by ten Berge and Zegers that higher order coefficients usually do not much influence the magnitude of the coefficient, and for λ_5 because the average covariance in each column of the covariance matrix is the same. On the other hand it is expected that λ_7 will do a good job, because it depends also on differences in inter-item correlations, and these differences are influential in this example because the error variance of the items is not constant. In the Figures 25, 26 and 27 the results are displayed, showing indeed the order specified in (2.12), but with clear differences. Moreover, the median of the distribution seems to be independent of the sample size, although there appears to be a slight downward trend for λ_7 in the case of 40 items. Clearly the three estimates are negatively biased when interpreted as estimates of the reliability. Calculating the increase in test length (substituting λ_7 for glb in (6.8)) at a sample size of 2400 gives 1.53, 1.32 and 1.27 times the original test length for 10, 20 and 40 items respectively to reach the target reliability (the true reliability of the test.)

The fourth index displayed is λ_6 which is uniformly better than the other three and gives for the case of 40 items certainly acceptable results: the bias is within a 0.01 band around the true value, and, as can easily be seen for the smaller values

of the sample size, has smaller confidence intervals. For the short test, however, it is only slightly better than λ_7.

The third panel displays results for these same indices, but now on a partitioned test following the procedures of Heuristic 1. See Figures 28, 29 and 30. The main results can be summarized as follows:

- The differences between the three indices α, λ_2 and λ_7 disappear. This means that the effect of Heuristic 1 consists in constructing a partition were essential τ-equivalence between the combined items is approximated. Notice that the design is constructed in such a way that full τ-equivalence can be reached by evenly dividing the items of each testlet in two parts. The Heuristic, however, is not able to find this partition systematically, because if it did, the three indices would coincide with λ_4, but they are systematically lower. Compare with the Figures of the first panel (22 to 24).

- Except for small sample size and for $k = 20$ or $k = 40$ the bias is negative. Its magnitude, however, decreases with increasing k, and for $k = 40$, it certainly is acceptable for all sample sizes studied. Notice, moreover, that with these estimates the confidence intervals are markedly shorter than in the case of an estimate based on the covariance matrix as given. Compare for example Figures 27 and 30.

- The behavior of λ_6 seems to be the opposite from its behavior in the first panel, giving the largest bias. On the other hand, comparing Figures 25, 26 and 27 with the Figures 28, 29 and 30 respectively, reveals that the behavior of λ_6 is fairly constant. This index seems unaffected by the partitions induced by Heuristic 1. It is not understood why this is the case.

7. Modeling the bias of the glb

The results of the two examples are disappointing in two respects: The coefficients derived by Guttman, and especially coefficient α, are inconsistent estimators of the reliability as soon as the requirement of essential τ-equivalence is not fulfilled. The only exception seems to be λ_4 - or its approximation by Heuristic 2 - but this index is positively biased, and, moreover, one should be careful to generalize too much from the two examples given: It is not true that λ_4 is a consistent estimator of the reliability in the general case. This can easily be seen by considering a test of two items, in which case $\lambda_4 = \alpha$. The second disappointment is the severe

positive bias of the theoretically superior glb. If the assumption of experimental independence is fulfilled and if the unique variances of all the items in the test are due to measurement error, the glb is a consistent estimator of the reliability, independent of the factorial structure of the items. But the examples have shown that it is positively biased and that the bias is persistent and serious also in the case of fairly large samples.

At present no theoretical results are known on the bias of the glb, but even without such results, it is possible to model the systematic decrease of the glb as a function of the sample size. Consider for example the Figures 3 through 8: The line connecting the median glb (as a function of the sample size) approximates a smooth curve, and it seems easy to extrapolate this curve towards its asymptote, which is the true reliability. Of course, in practice, it is not possible to sample a great number of times from the same population, but one could mimic the simulated data from the examples by a bootstrap technique. In the present section, first a model for the bias will be presented. Next, it will be investigated whether this model can be used in a realistic setting where one has only a single data set.

The following notation will be used. The estimate of the expected value of some index will be denoted by d_n, where n is the sample size. In the examples, d_n can be the median or average glb-estimate across replications, or in a real-life application the average glb resulting from a number of bootstrap samples all of size n. The corresponding expected value will be denoted by δ_n. The asymptote of δ_n will be denoted by ρ. Since it can be expected that the bias decreases with increasing sample size, we start with a fairly simple model:

$$\delta_n - \rho = \beta^n, \quad (0 < \beta < 1), \tag{7.1}$$

meaning that the bias reduces geometrically with rate β. Model (7.1) is equivalent to

$$\ln(\delta_n - \rho) = n \ln(\beta), \tag{7.2}$$

meaning that the logarithm of the bias is linear in the sample size. If ρ is known (as is the case in the simulation study), (7.2) can be easily checked by substituting d_n for δ_n in the formula. Scatter plot were constructed for the three cases from example 1, where the true reliability is 0.7 (corresponding to the Figures 3, 5 and 7). See Figure 31. The deviation from linear regression is clear: the decrease grows slower than linear in n. So we need to replace n in formula (7.1) by some suitable concave increasing function of n, such as the square root. To have a fairly general model, we choose as extension of (7.1):

$$\delta_n - \rho = \beta^{n^t}, \quad (0 < \beta, t < 1). \tag{7.3}$$

Taking logarithms twice in succession (and adapting the sign as necessary) yields

$$\ln(-\ln(\delta_n - \rho)) = t\ln(n) + \ln(-\ln(\beta)), \tag{7.4}$$

meaning that the log-log of the bias is linear in the logarithm of the sample size. If d_n is substituted for δ_n in the left hand side of (7.4), and the resulting values for the same data as in Figure 31, are plotted against $\ln(n)$, the scatter plot looks fairly linear (See Figure 32). The estimation of ρ, using a least squares loss function, however, turned out to be quite difficult, and the results were disappointing, yielding rather gross deviations from the theoretical value. So this approach was given up.

Another way to model the bias is to consider the ratio of the expected value of the statistic to the parameter. Since the glb is a consistent estimator of the reliability (under the conditions mentioned) we know that

$$\lim_{n\to\infty} \frac{\delta_n}{\rho} = 1, \tag{7.5}$$

which also means that the relative bias goes to zero in the limit:

$$\lim_{n\to\infty} \frac{\delta_n - \rho}{\rho} = 0. \tag{7.6}$$

In Figure 33 the relative bias (using d_n as an approximation to δ_n) is plotted against the sample size for the three aforementioned cases of example 1, clearly indicating a non-linear relationship. In Figure 34 the logarithm of the relative bias is plotted against the logarithm of the sample size, together with the linear regression lines, giving an excellent fit. The corresponding model is given by

$$\frac{\delta_n - \rho}{\rho} = \beta n^t. \tag{7.7}$$

Using the estimates d_n of δ_n makes it in principle possible to estimate the three parameters of the model: β, ρ and t. But as in the case of model (7.4) the estimation procedure turned out to be quite complicated, and the results were not very close to the theoretical values.

A closer look at Figure 34 revealed that the slope of the three regression lines was about -0.47, and suggested that a square root transformation of the sample size might do a good job as well. Therefore we chose as our final model:

$$\frac{\delta_n - \rho}{\rho} = \beta\frac{1}{\sqrt{n}}, \tag{7.8}$$

Table 7.1: Estimation of the reliability using linear regression

		ρ estimated from		
k	*Rel*	glb	max.	ave.
10	.700	.701	.696	.680
20	.700	.704	.701	.687
40	.700	.711	.710	.705
10	.900	.900	.900	.900
20	900	.901	.894	.891
40	900	.904	.898	.891
10	.700	.701	.702	.699
20	.824	.824	.819	.814
40	.903	.906	.894	.890

which suggests the quadratic loss function

$$F(\rho, \beta) = \sum_n [d_n - \rho\beta n^{-\frac{1}{2}} - \rho]^2 \tag{7.9}$$

from which it follows immediately that the estimator of ρ is the additive coefficient in a simple linear regression problem. The parameter β, which undoubtedly depends on the number of items and on the factorial structure of the model is not interesting for the present problem.

Using (7.9), ρ has been estimated in all cases treated in examples 1 and 2, using the average value of the glb, as well as the two indices derived from Heuristic 2 (and indicated in the figures as Split-half (max.) and Split-half (ave.) respectively). The results are given in Table 7.1, where the last three rows refer to example 2 and the others to example 1. The extrapolation is quite accurate for the glb and for the maximum found with Heuristic 2; the average value given by this heuristic tends to be a little bit lower in all cases but certainly can be considered reasonable.

Of course, the estimates of the expected value of the indices in the example are quite stable, being the averages over 500 replications. To check whether the extrapolation works well in realistic settings, a single data set with 2400 observations for the three cases of Example 2 was generated. For each data set three estimates of the reliability were computed, using model (7.8). In the first approach samples from the observed data set were drawn with replacement; in the second approach sampling without replacement was carried out. In each case

Table 7.2: Estimating the reliability from a single data set

k	*Rel*	repl.	no repl.	part.
	glb			
10	.700	.691	.708	.687
20	.824	.822	.814	.812
40	.903	.909	.903	.904
Spl-half (m)				
10	.700	.693	.711	.689
20	.824	.813	.810	.808
40	.903	.893	.890	.889
Spl-half (a)				
10	.700	.688	.716	.696
20	.824	.809	.812	.809
40	.903	.884	.884	.882

10 samples were drawn of each of the 11 sample sizes 200 (200) 2200. For each of the 110 samples thus constructed, the three indices, glb, Split-half (max) and Split-half (ave) were computed, and averaged over the 10 replications per sample size. The loss function (7.9) was then minimized using these averages, yielding a single estimate of the reliability. In the third approach the observed data set was partitioned randomly in 12, 8, 6, 4, 3 and 2 subsets of equal size giving sample sizes of 200, 300, 400, 600, 800 and 1200 respectively. The remaining part of the procedure was identical to the other two approaches, except that for application of (7.9) the three indices computed on the full sample of 2400 observations were used as well. The results are displayed in Table 7.2

Although one should be hesitant in drawing conclusions from a single example, Table 7.2 shows clearly that the reliability can be estimated with reasonable accuracy from a single data set, correcting for the bias present in the estimate based on the full data set. Much more research is needed to demonstrate the validity of (7.8) and to determine the sampling scheme optimal to producing an accurate sequence of indices for extrapolation. For, as good as the estimates in Table 7.2 may be, the true reliability is underestimated more often (21 times out of 27) than it is overestimated.

8. Discussion and conclusions

This section will be used to give a short summary of the findings in the preceding sections and to address a conceptual problem which has been ignored so far.

Two aspects in the problem of reliability estimation should be clearly distinguished. One is the choice of an index, and the other concerns the sampling distribution of the indices. As early as 1945 it was shown in the seminal paper by Guttman that coefficient α (his λ_3) is certainly not the best estimate: the inequality $\lambda_2 \geq \lambda_3$ is his. A number of important papers since then have shown that still better lower bounds could be found, culminating in the theory of the greatest lower bound, discussed thoroughly by ten Berge e.a. It is true that coefficient α is computed with an extremely simple formula, and it is completely understandable that it was used widely in the days that computations had to be done by hand. But these days are over. It is true that the computation of the glb is not trivially simple, but one good implementation in a computer program suffices to make it available to all practitioners in test theory. Moreover, in the present paper two simple Heuristics were presented which do better than the traditional coefficients and which are computationally much simpler than determining the glb. As was shown in the two examples, Heuristic 2 - an attempt to find λ_4- is quite successful. In all cases considered the differences between the medians of the glb estimates and the maximal split-half coefficients, found by Heuristic 2, were less than 0.01.

As to the assumptions needed to make the indices consistent estimators of the reliability, a marked contrast was shown between the case of coefficient α and the glb. The assumption of essential τ-equivalence needed for α was shown to be extremely hard: The factorial structure w.r.t. the common factors must be identical for all items. The glb on the other hand, is a consistent estimator of the reliability if the unique variance of each item is error variance, a much more lenient assumption than essential τ-equivalence.

For the other indices discussed in the preceding sections, very little is known about their relationship to the reliability, except the fact that they are all lower bounds. The second simulation study (Example 2) showed clearly that most of the indices (with the possible exception of λ_4) are inconsistent if the assumption of essential τ-equivalence is not fulfilled: the systematic deviation of the median estimates which are to a great extent independent of the sample size hardly allow another conclusion.

This could close the discussion, if the reliability of a test were estimated from a huge sample. But the theoretically superior glb appears to be seriously biased.

Although no theoretical results on this bias are available, this bias could be modeled quite accurately using the results of the simulation studies. The practical result is an extrapolation formula, which uses (average) glb-estimates from samples with varying sample sizes as input. It was shown, although only for three examples, that this extrapolation technique can also be applied with reasonable success if only one data set is available, by sampling repeatedly from this set. If this result is corroborated in cases with different factorial structures, fine tuning w.r.t sampling scheme and possibly theoretical results on the bias of the glb may lead to a superior estimate of the reliability from a single test administration.

Although similar results were found using the two indices issued by Heuristic 2, one should be careful in generalizing these results. In our search for a good model for the bias, it was assumed throughout that the index used was a consistent estimate of the reliability. This is the case for the glb if the unique variance is error variance; for λ_4 (or its approximation by Heuristic 2) this is not the case in general. For the two simulation studies, it is easily verified that there existed in all cases a partition of the items into two subtests which are essentially τ-equivalent, making λ_4 a consistent estimator of the reliability. In cases where λ_4 underestimates the reliability, a (correct) extrapolation formula will estimate true λ_4 (its population value), and as a consequence be inconsistent.

Notwithstanding the very promising results of the study reported in the previous sections, one might object against the approach used on conceptual grounds. To clarify such objections, we start with a quotation from the 1966 edition of the APA *Standards for Educational and Psychological Tests and Manuals* (Section D5.4) where it is stated that

> If several questions within a test are experimentally linked so that the reaction to one question influences the reaction to another, the entire group of questions should be treated preferably as an item when the data arising from application of split-half or appropriate analysis-of-variance methods are reported in the manual. (p. 30)

Sireci, Thissen and Wainer (1991) notice that the importance of the recommendation was diminished to 'very desirable ' in the 1974 edition, and disappeared altogether from the 1985 edition. They regret the disappearance of the recommendation and set out to demonstrate its wisdom. If we stick to the second example, this recommendation says that one should treat the testlet score as the basic observation, and base all estimates of the reliability on these scores. If we do so, we get for the example 2 the values listed in the rightmost column of Table

6.2, which are labeled as α_{APA}. The differences with the true reliabilities *Rel* are very large: to raise the reliability from 0.8 to 0.9 one has to construct a test which is 2.25 times as long as the original one. In terms of the heuristics, this result is to be understood as follows: The heuristics attempt to find a partition such that the equivalence classes are as similar (in factorial structure) as possible, with the consequence that the items within a class are as heterogeneous as possible. The APA-recommendation implies just the opposite: if the equivalence classes coincide with the testlets, the classes are as homogeneous as possible, while the testlet scores are as heterogeneous as possible, and as a consequence, coefficient α computed for testlet scores will be lower than α computed on the original item scores.

What about the wisdom of the APA recommendation? Consider two possible explanations for the existence of testlet associated factors in, for example, a test of reading comprehension. One is that for some students, reading errors in the introductory text which is common to all items in the testlet, will influence the score on all the items belonging to that testlet. As a consequence the item scores of that testlet will covary more than can be explained by the general factor. If these reading errors are caused by accidental events, which have nothing to do with the concept that is being measured (such as sudden noise in the class room), the testlet factor will not be stable. On retesting the factor associated with the testlet will be uncorrelated with the testlet factor of the first testing. This means that this factor has to be considered as a source of measurement error, which causes correlated errors in the item scores. If this explanation is correct and complete, one cannot but agree with the APA recommendation. The other explanation is that the testlet factors are stable. The subject matters in an authentic reading comprehension test will be heterogeneous rather than homogeneous. If one text is about history and another about sports, these texts may address different stable factors, which will show enduring influence in future testing. If such factors completely describe the extra covariation between the testlet items, the retest reliability will be equal (ignoring sampling errors) to the reliability as treated in example 2, and the use of α_{APA} will grossly underestimate it. Of course in real situations neither of the two explanations will suffice, and the truth will be somewhere in the middle. Unfortunately, the distance between the two extremes is quite large, and there is no possibility to decide between them on the basis of a single test administration.

9. References

Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika, 16, 279-334.*

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

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Jackson, P.H. (1979). A note on the relation between coefficient alpha and Guttman's "split-half" lower bounds. *Psychometrika, 44, 251-252.*

Jackson, P.W. & Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrika, 42, 567-578.*

Kristof, W. (1963). Statistical inferences about the error variance. *Psychometrika, 28, 129-143.*

Kuder, G.F. & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Sireci, S.G., Thissen, D. & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28, 237-247.*

Ten Berge, J.M.F., Snijders, T.A.B. & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika, 46,* 201-213.

Ten Berge, J.M.F. & Zegers, F.E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43, 575-579.*

Thissen. D, Steinberg, L. & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26, 247-260.*

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Woodhouse B. & Jackson, P.H. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: II. A search procedure to locate the greatest lower bound. *Psychometrika, 42, 579-591.*

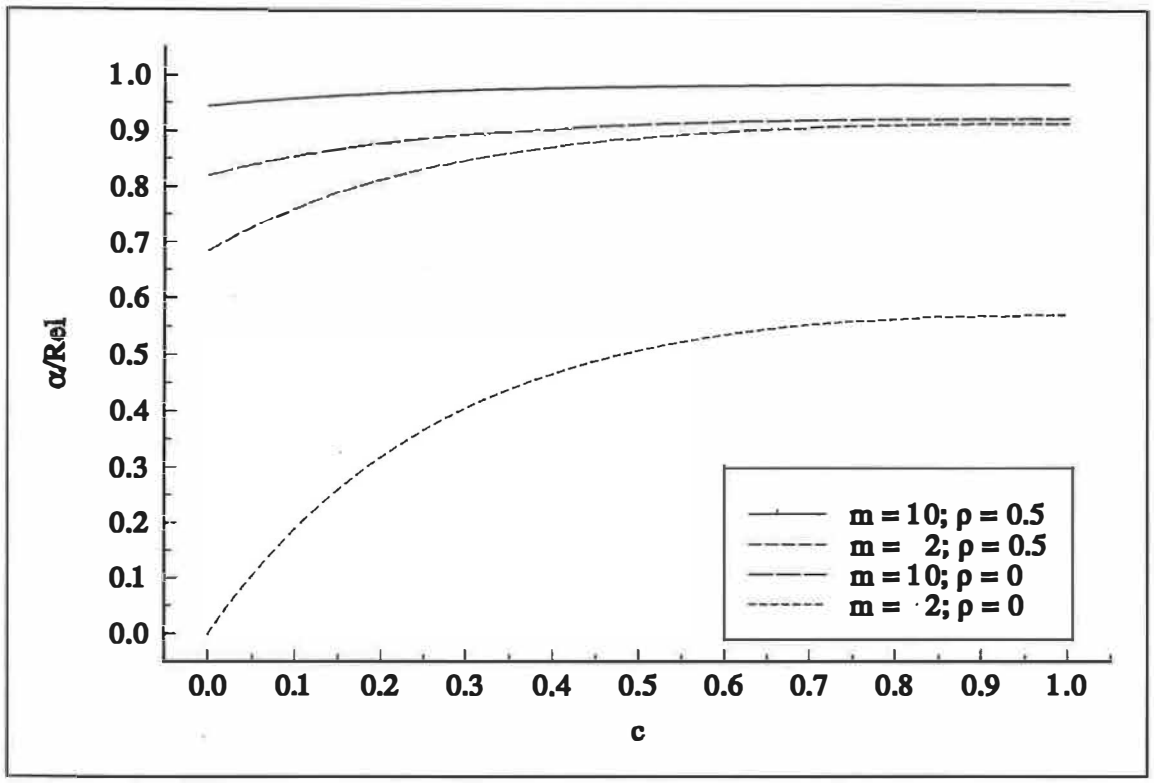

Figure 1: The ratio α/Rel as a function of c (p=4)

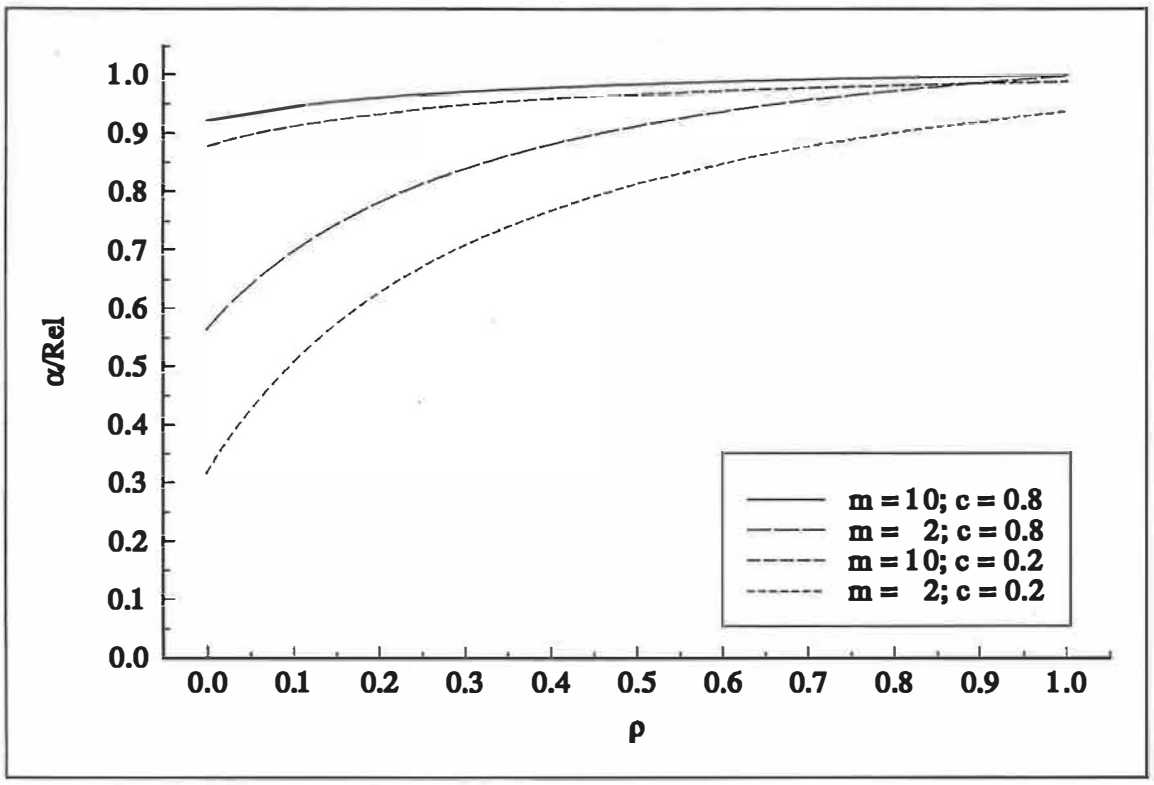

Figure 2: The ratio α/Rel as a function of ρ (p=4)

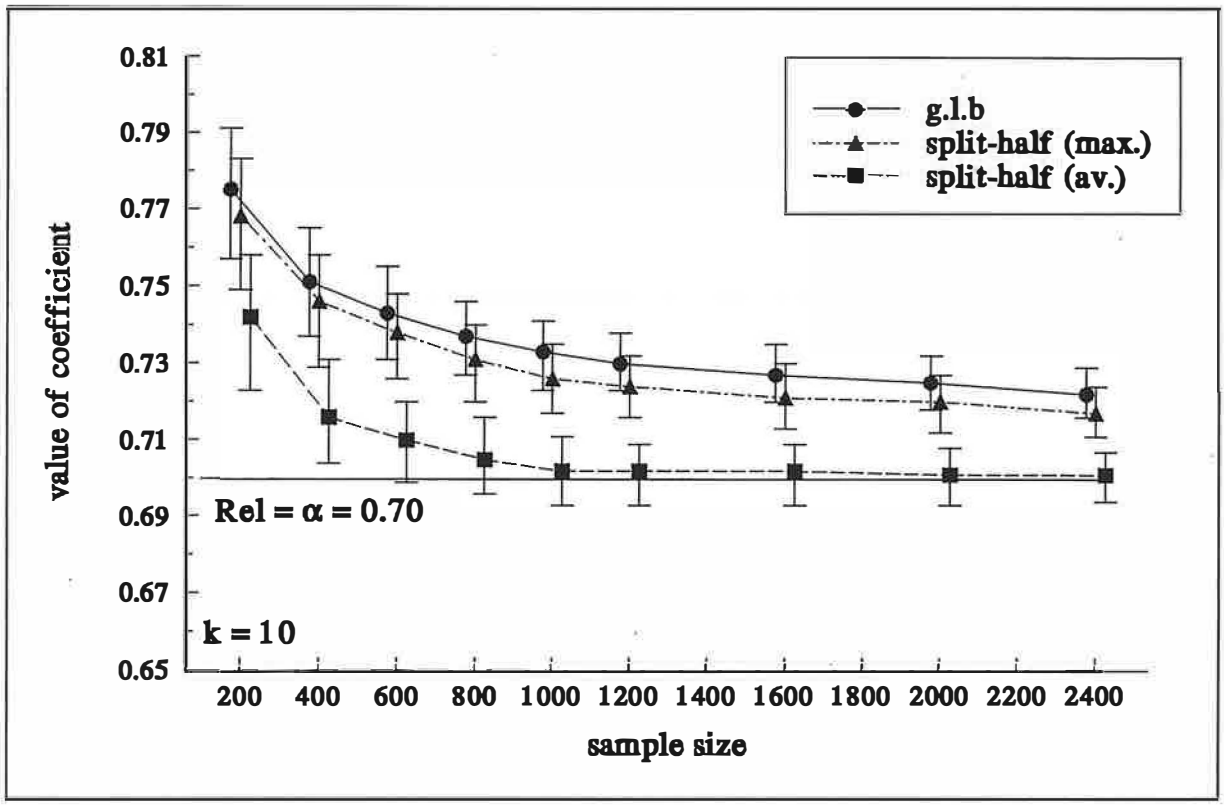

Figure 3: Example 1, first panel, $k=10$, $Rel=\alpha=0.7$

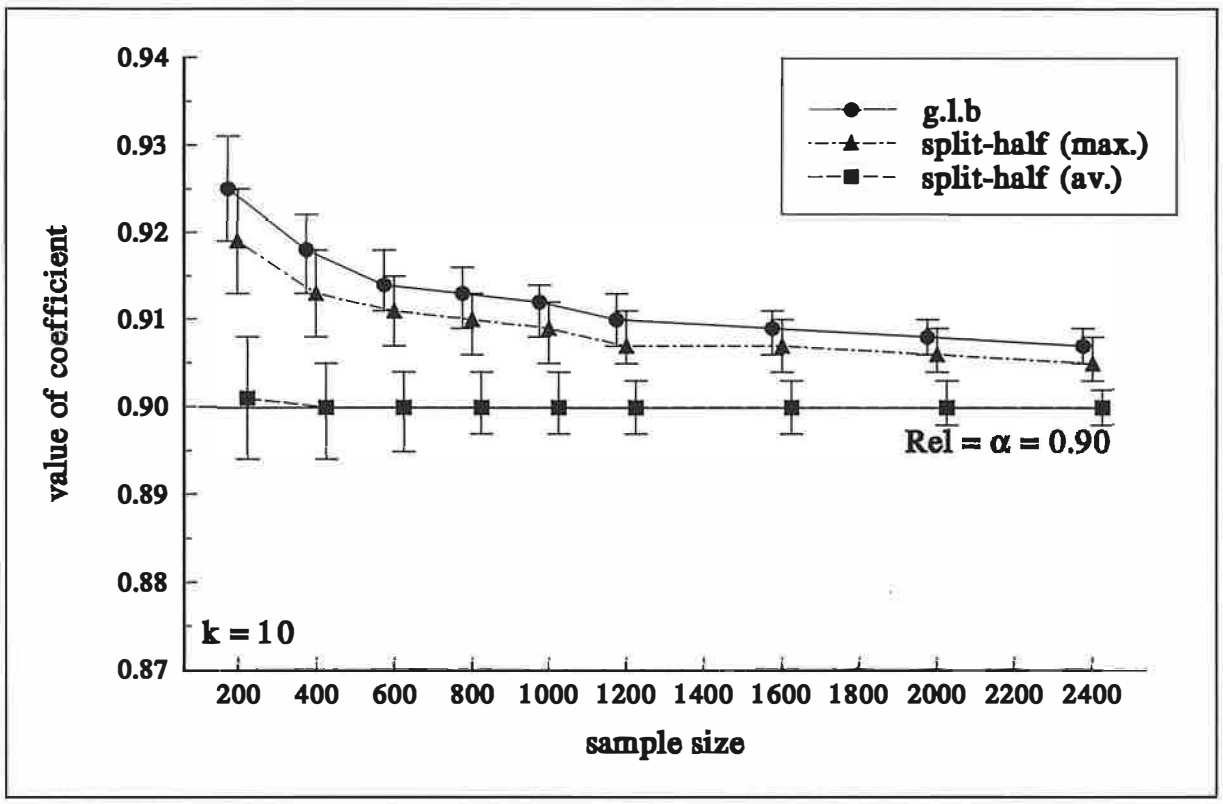

Figure 4: Example 1, first panel, $k=10$, $Rel=\alpha=0.9$

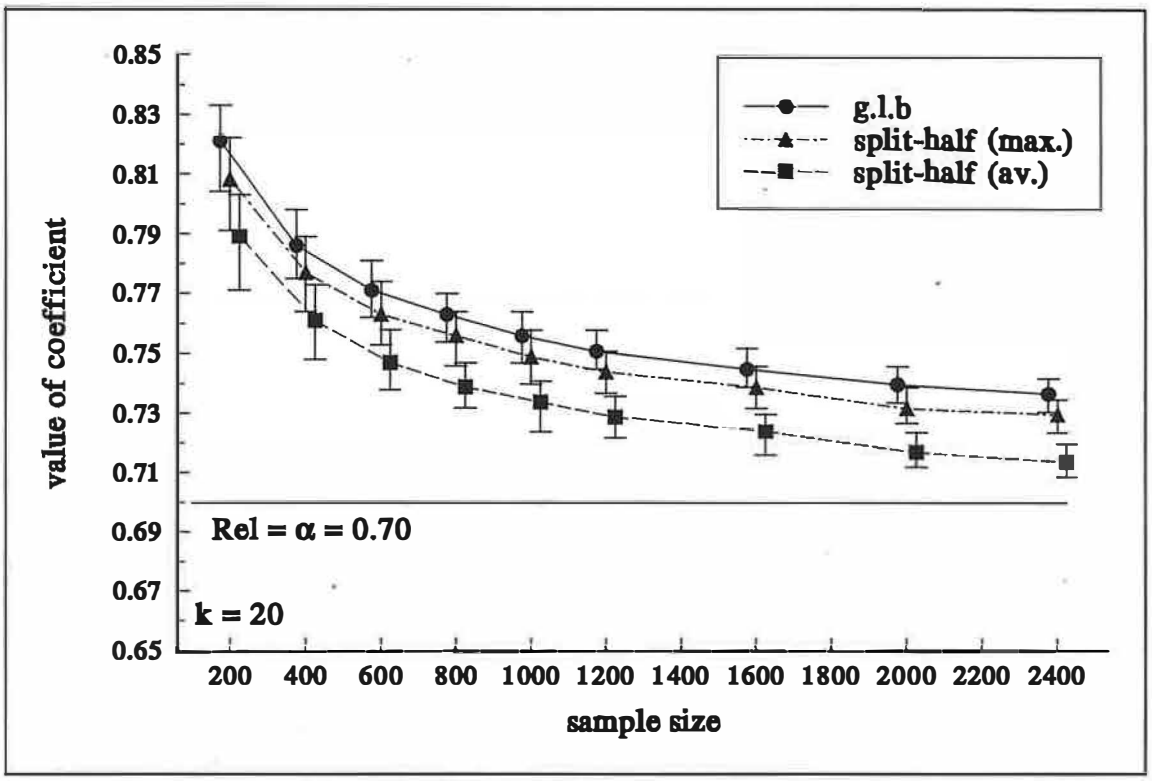

Figure 5: Example 1, first panel, k=20, Rel=α=0.7

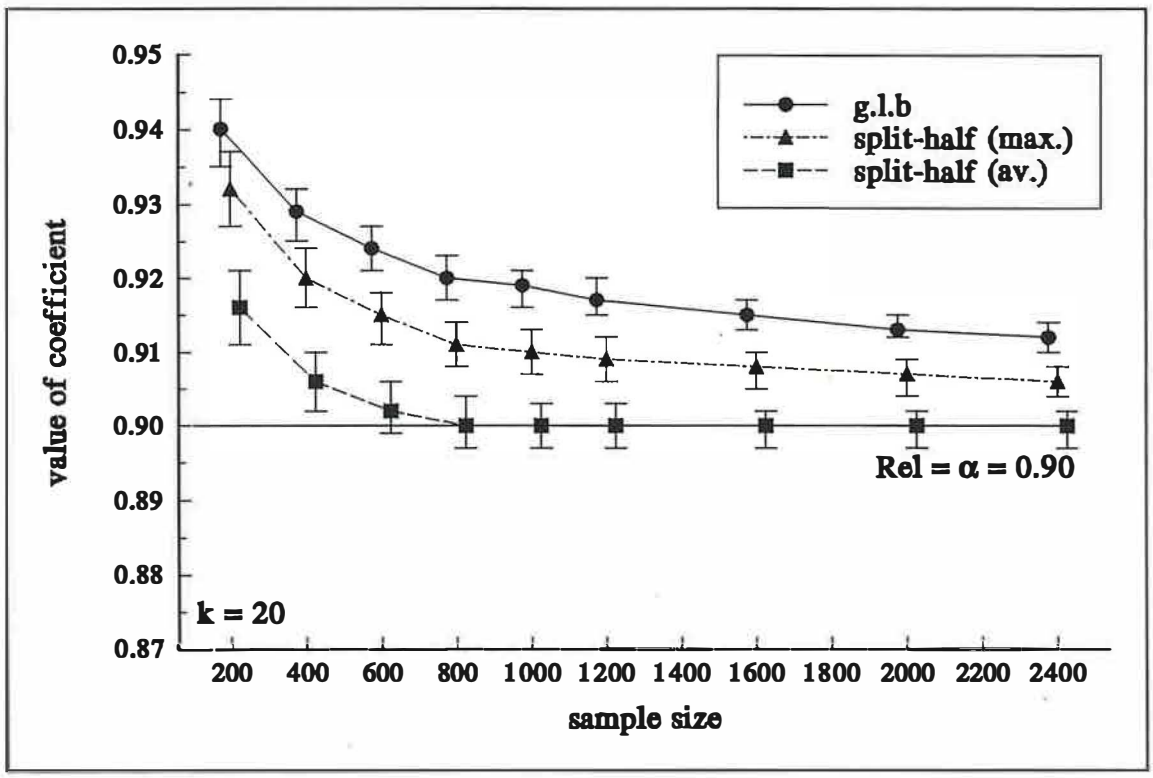

Figure 6: Example 1, first panel, k=20, Rel=α=0.9

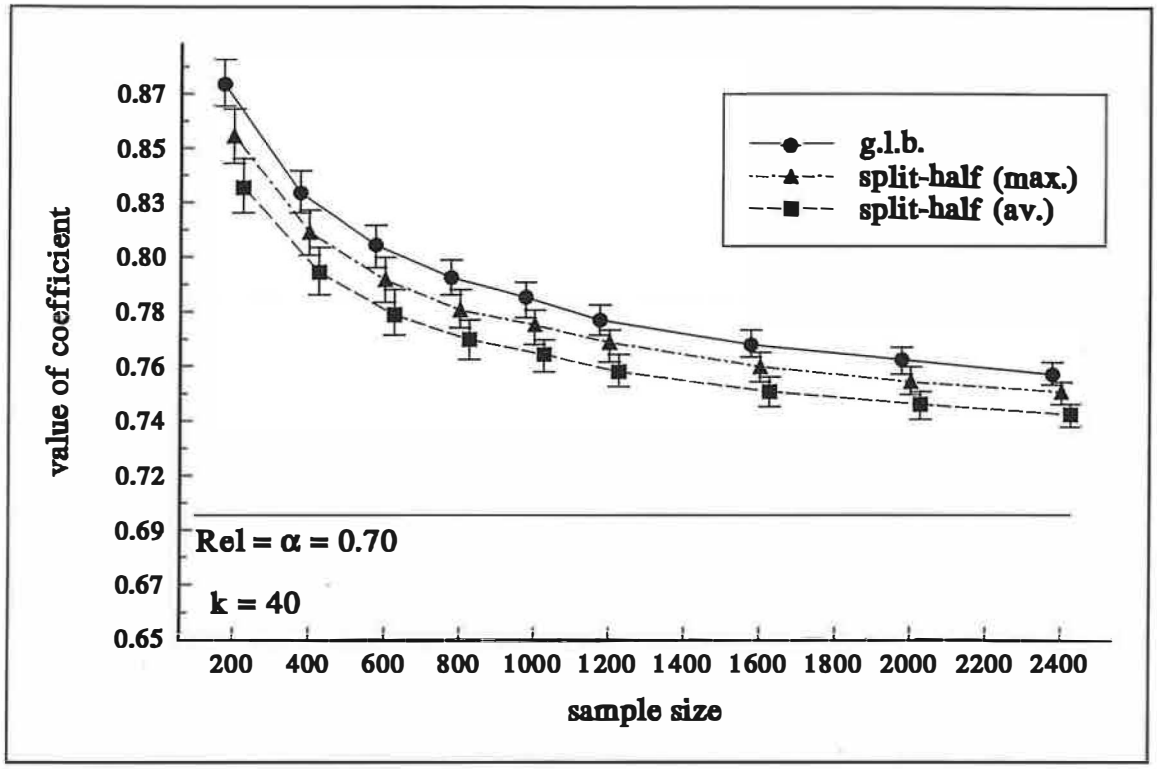

Figure 7: Example 1, first panel, k=40, Rel=α=0.7

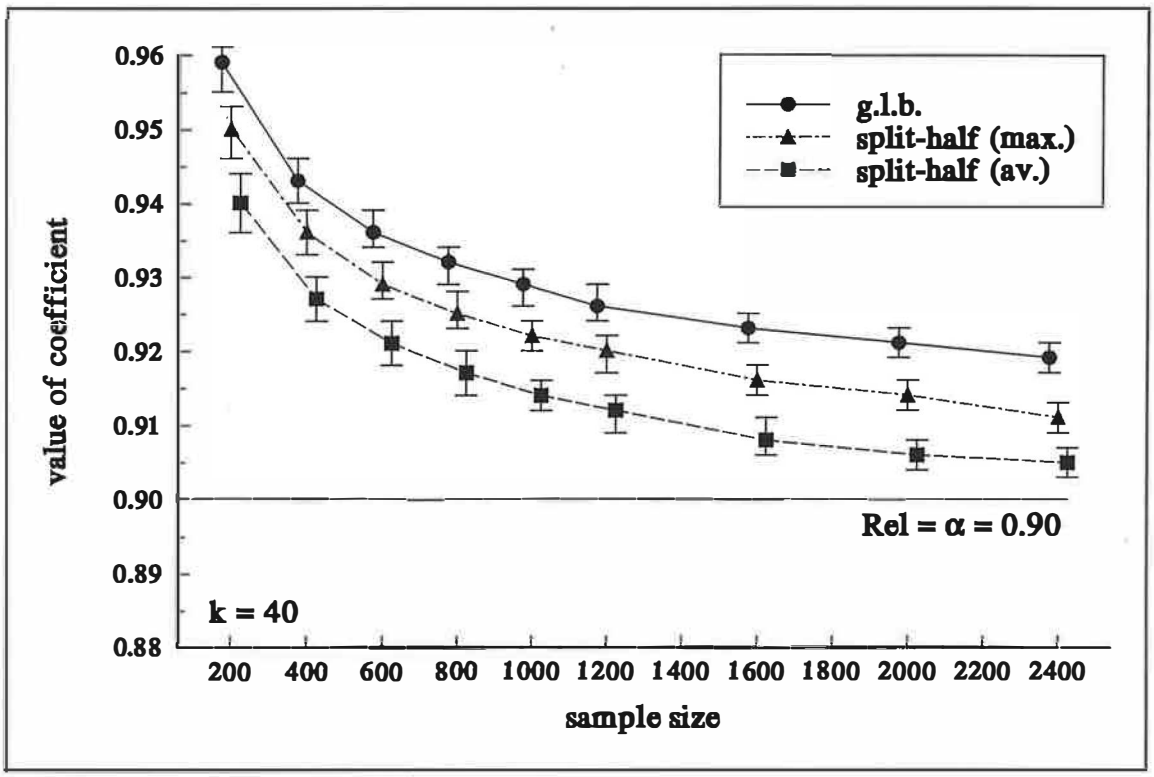

Figure 8: Example 1, first panel, k=40, Rel=α=0.9

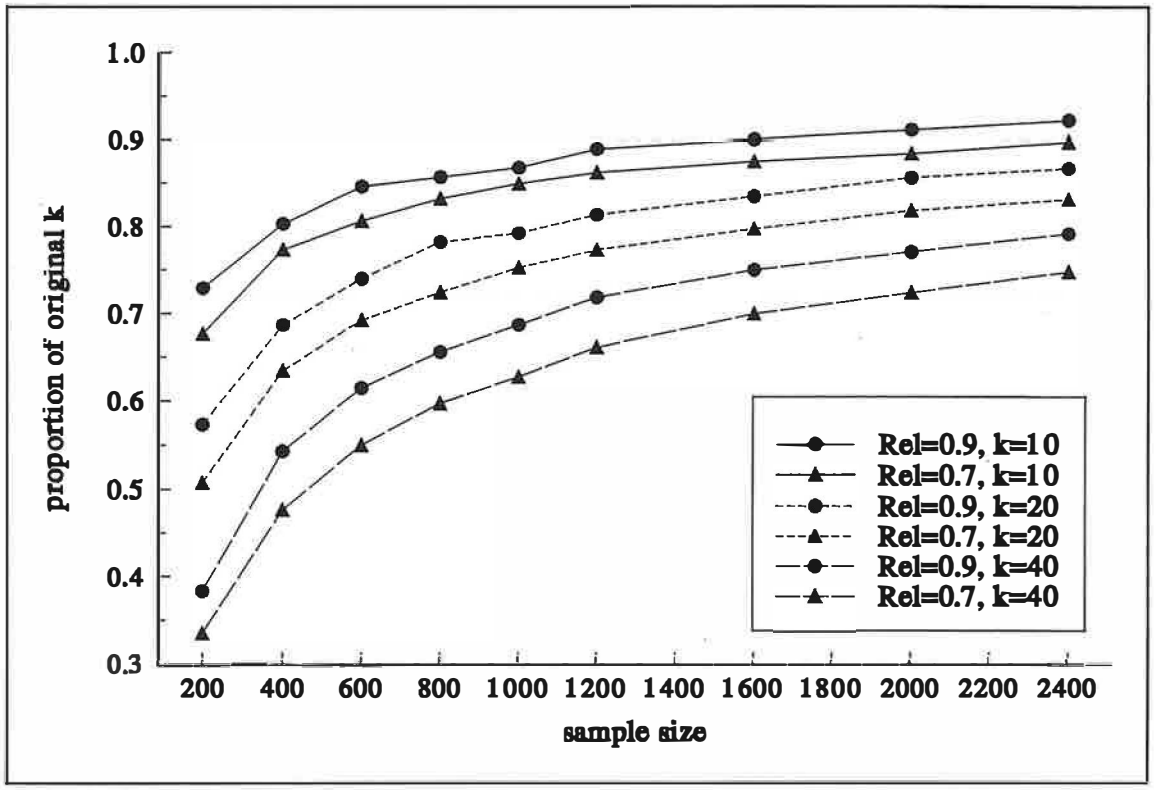

Figure 9: Example 1, first panel: effect of bias on target test length

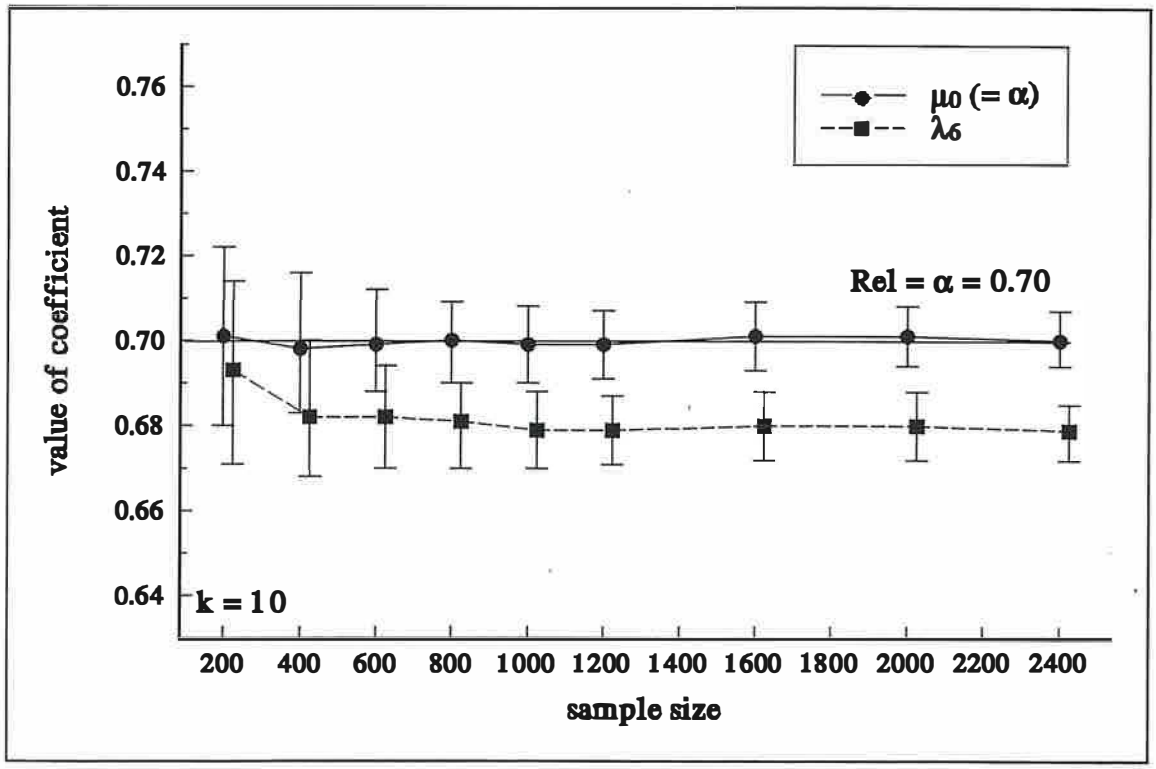

Figure 10: Example 1, second panel, $k=10$, $Rel=\alpha=0.7$

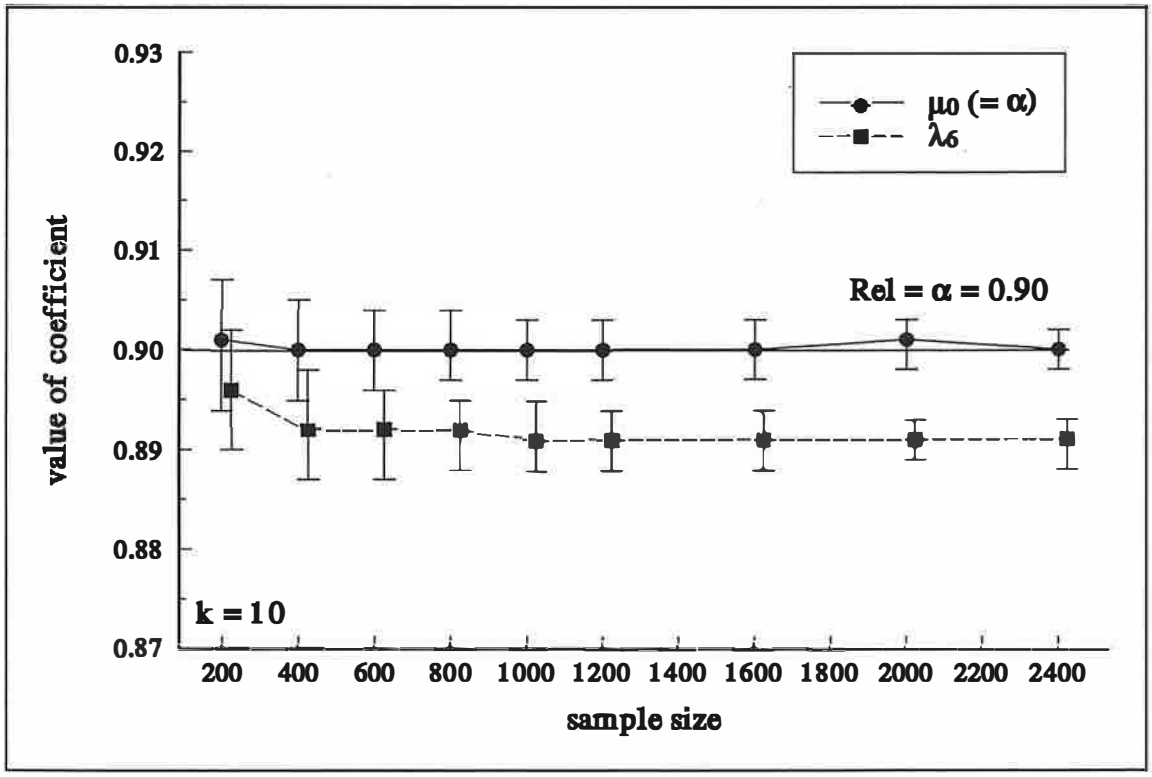

Figure 11: Example 1, second panel, k=10, Rel=α=0.9

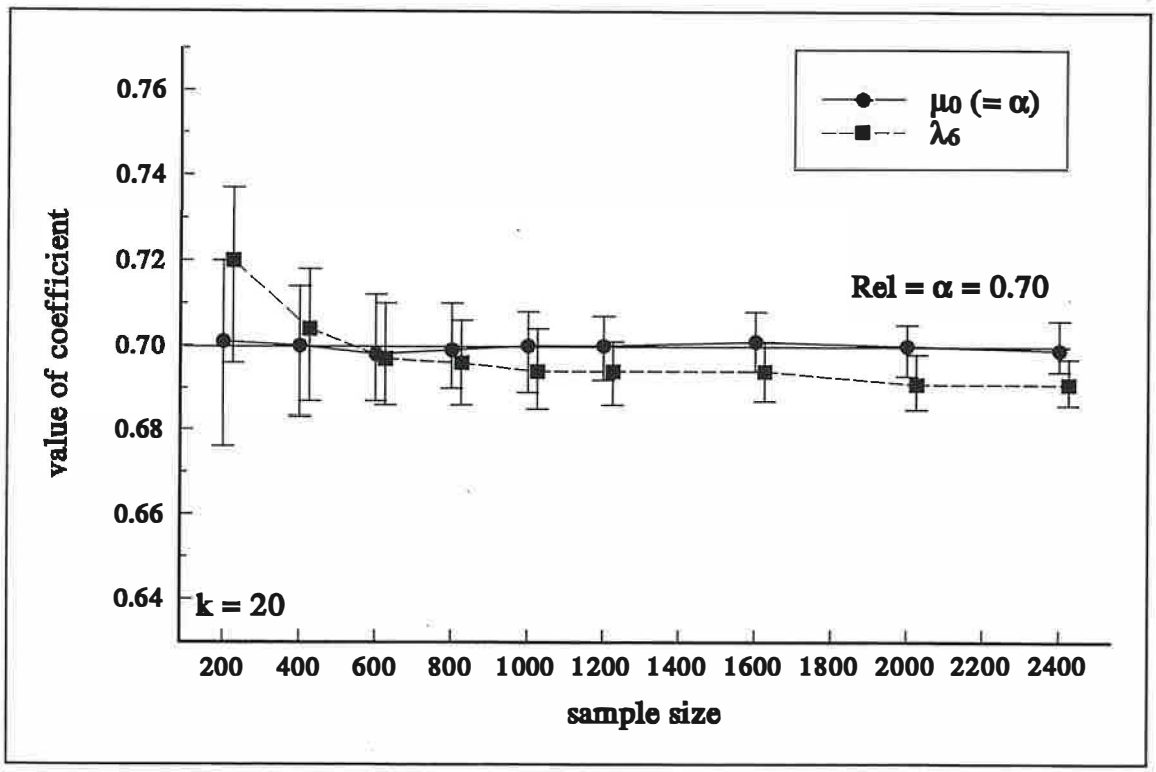

Figure 12: Example 1, second panel, k=20, Rel=α=0.7

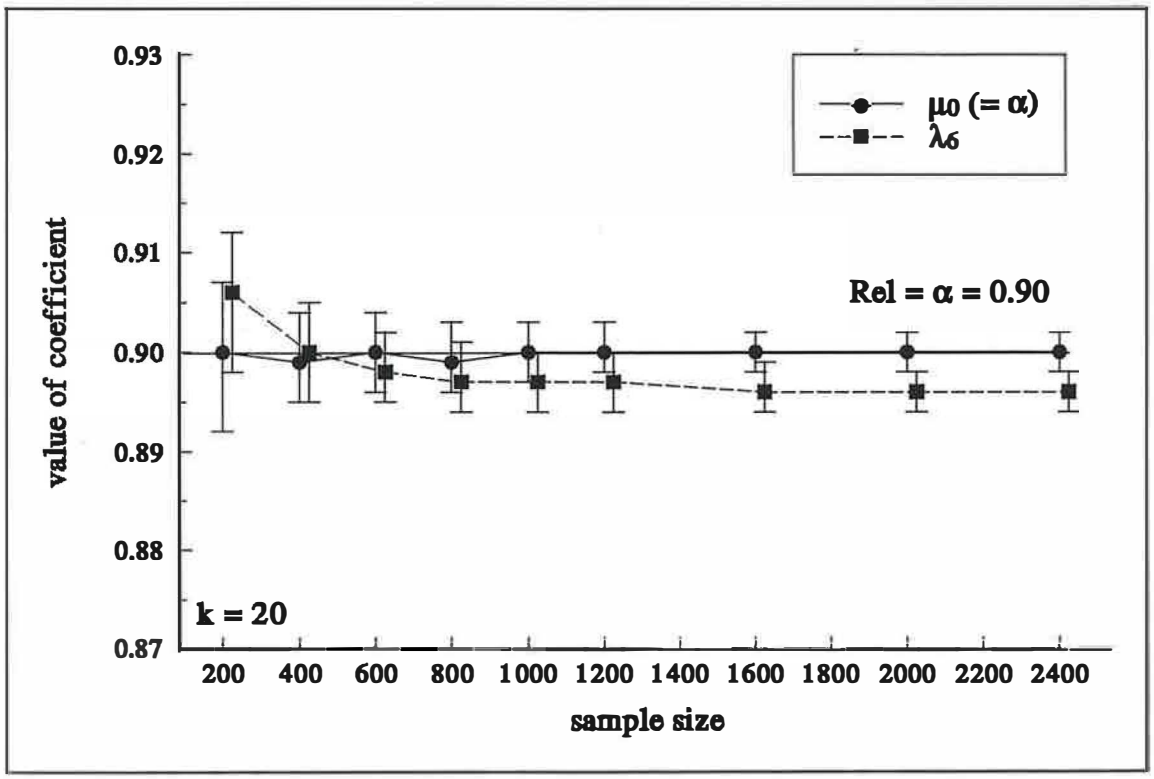

Figure 13: Example 1, second panel, k=20, Rel=α=0.9

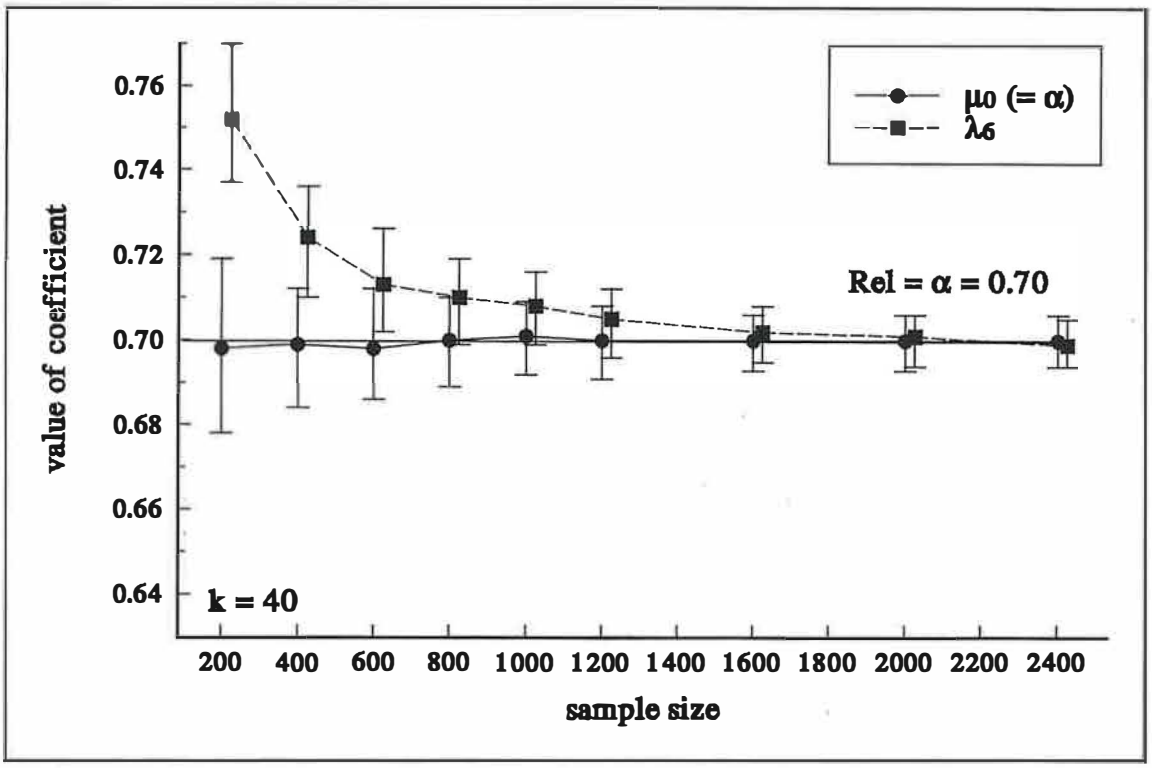

Figure 14: Example 1, second panel, k=40, Rel=α=0.7

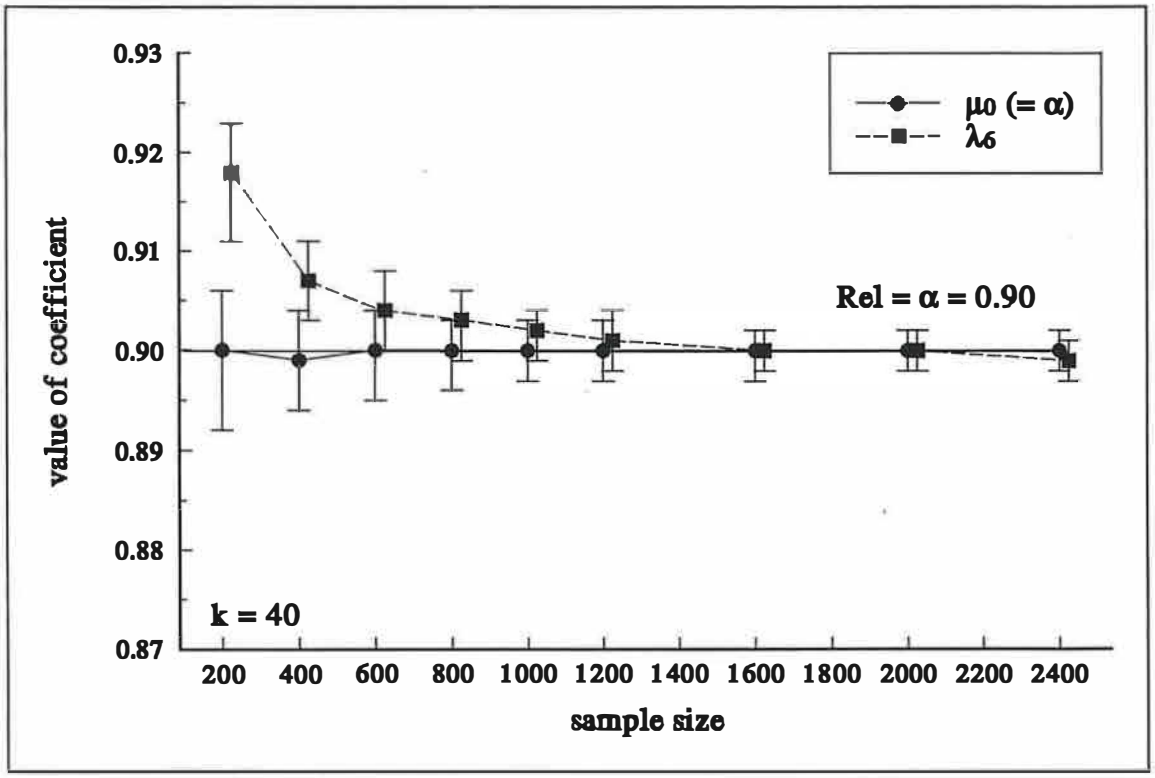

Figure 15: Example 1, second panel, k=40, Rel=α=0.9

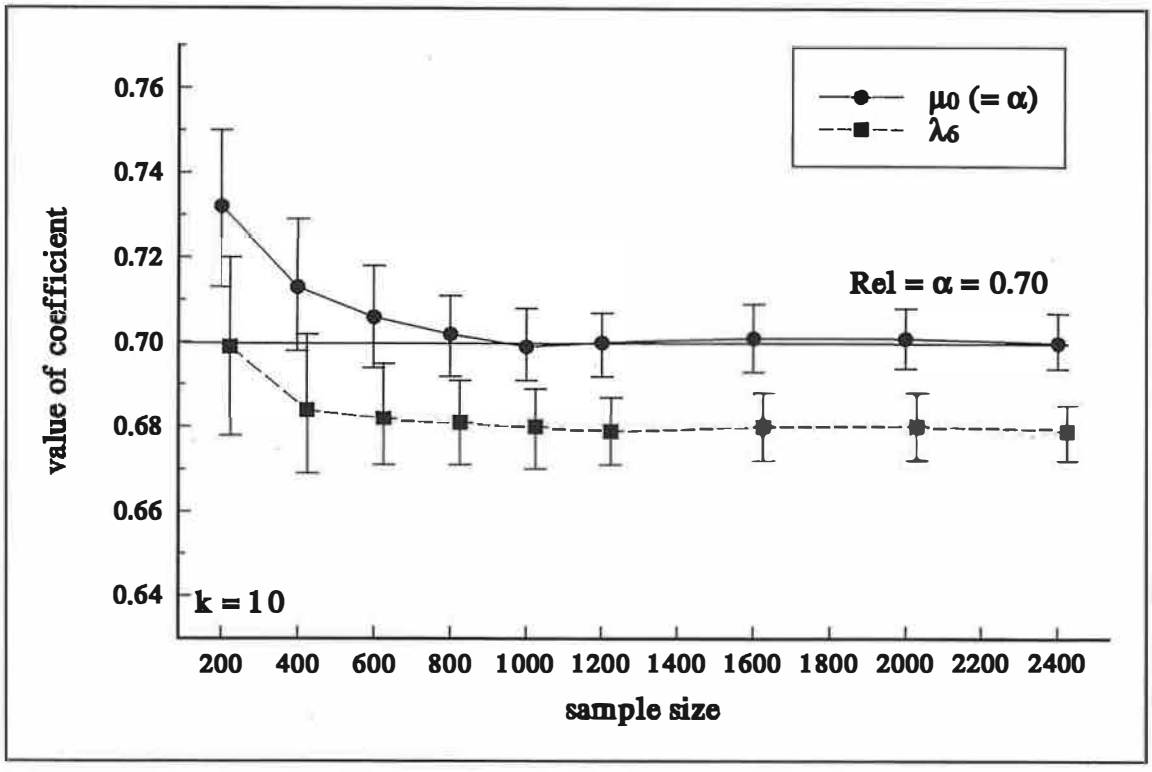

Figure 16: Example 1, third panel, k=10, Rel=α=0.7

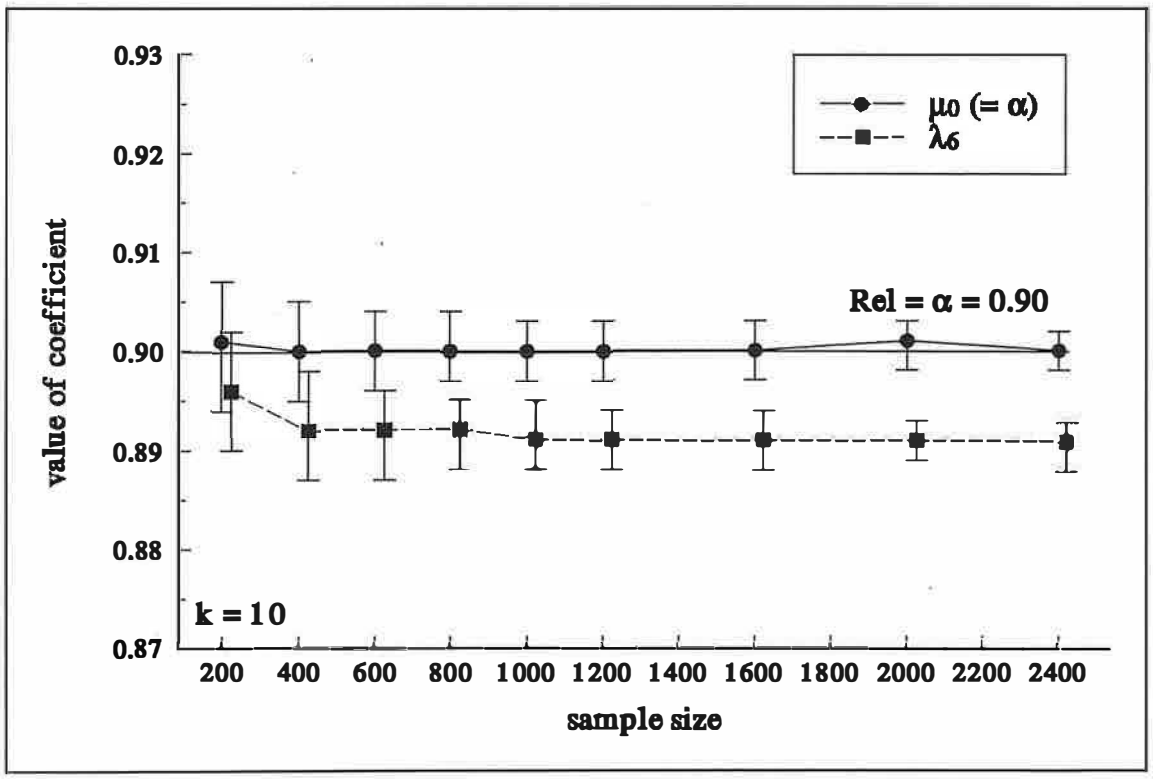

Figure 17: Example 1, third panel, k=10, Rel=α=0.9

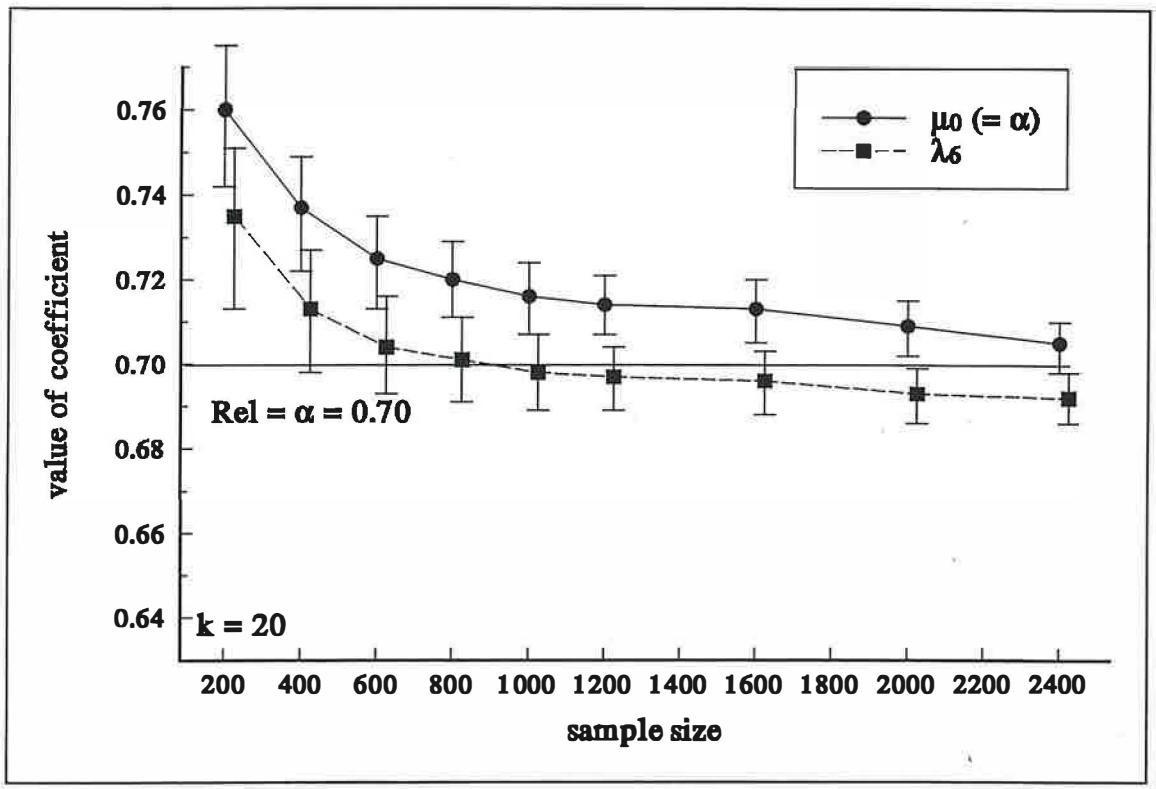

Figure 18: Example 1, third panel, k=20, Rel=α=0.7

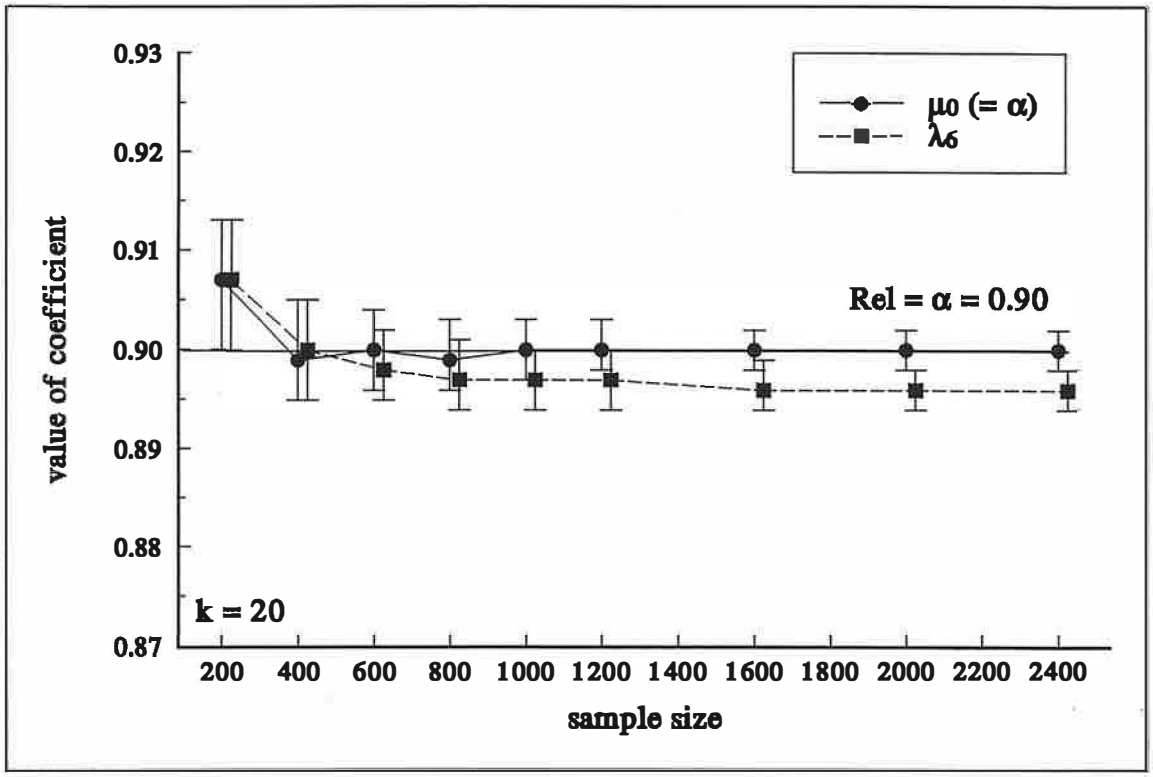

Figure 19: Example 1, third panel, $k=20$, $Rel=\alpha=0.9$

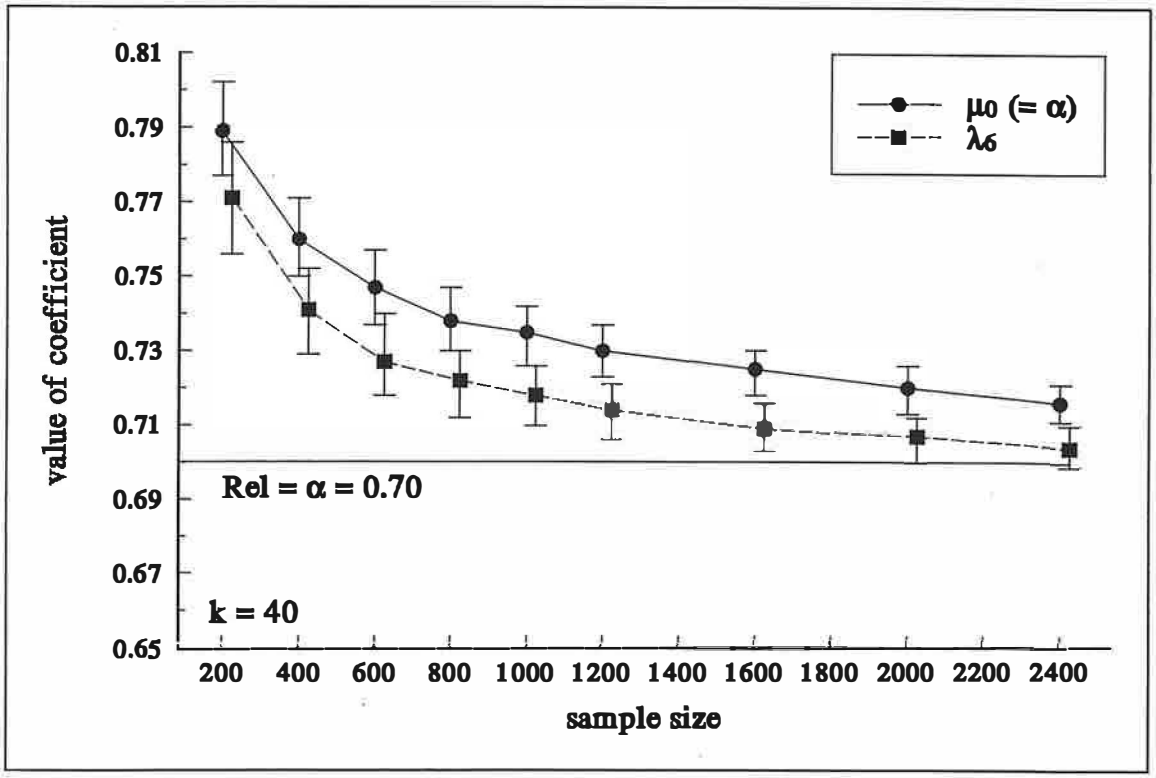

Figure 20: Example 1, third panel, $k=40$, $Rel=\alpha=0.7$

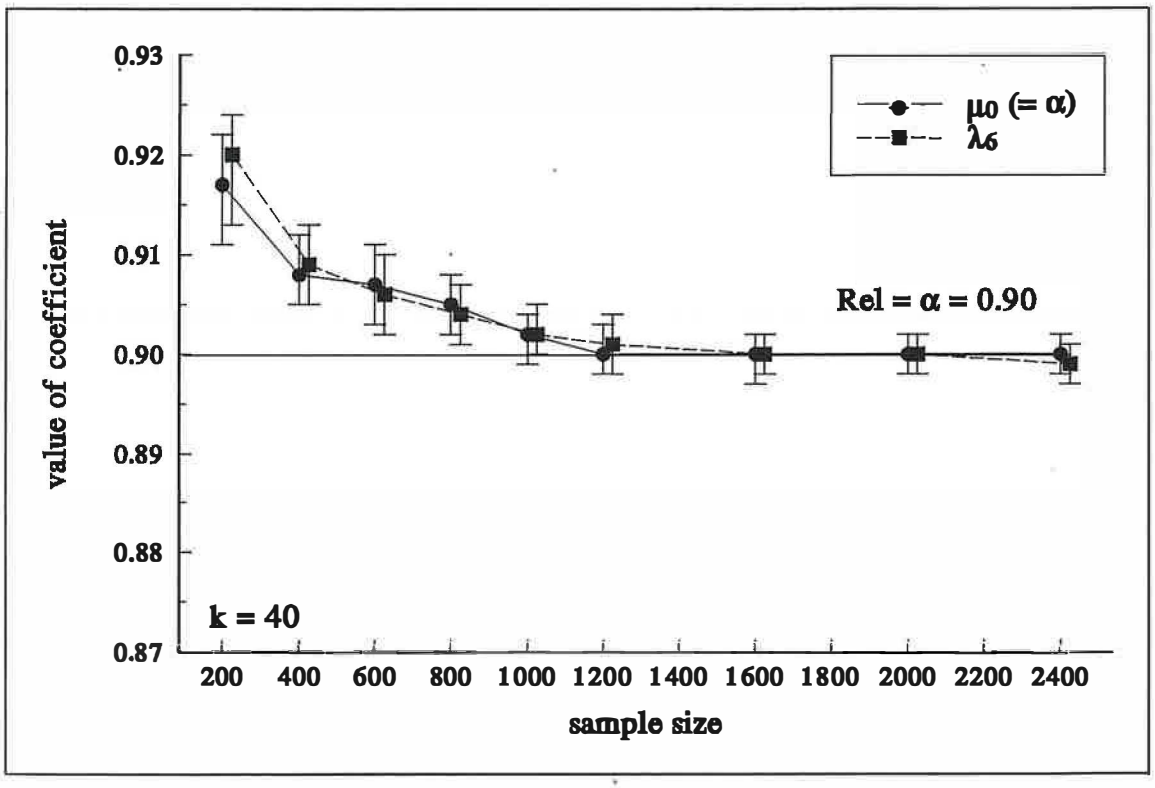

Figure 21: Example 1, third panel, $k=40$, $Rel=\alpha=0.9$

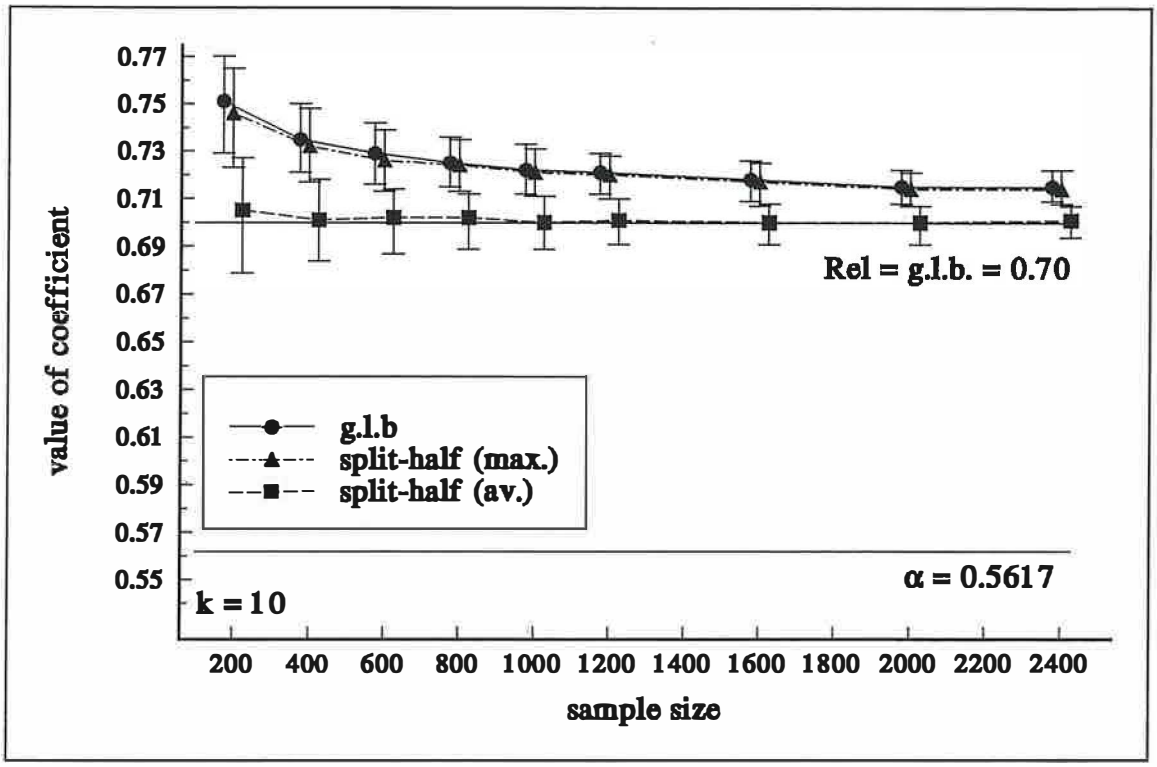

Figure 22: Example 2, first panel, $k=10$, $\alpha<Rel=0.7$

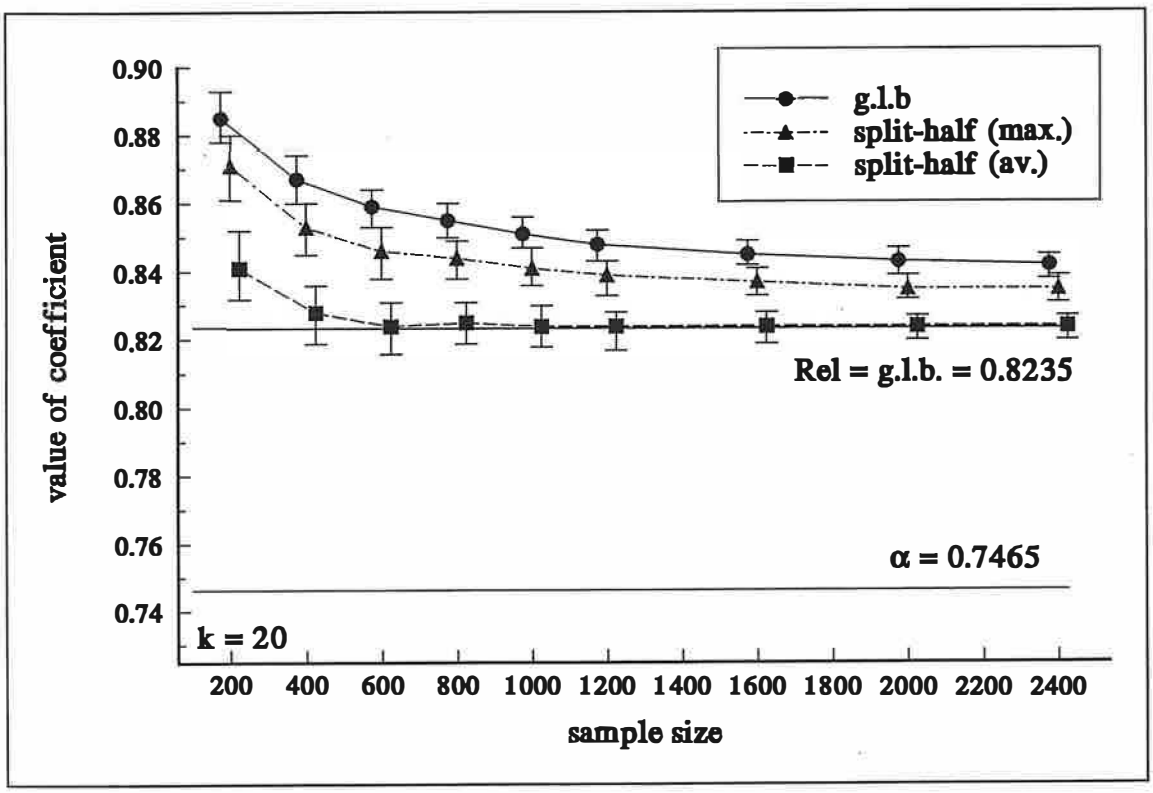

Figure 23: Example 2, first panel, $k=20$, $\alpha<\text{Rel}=0.82$

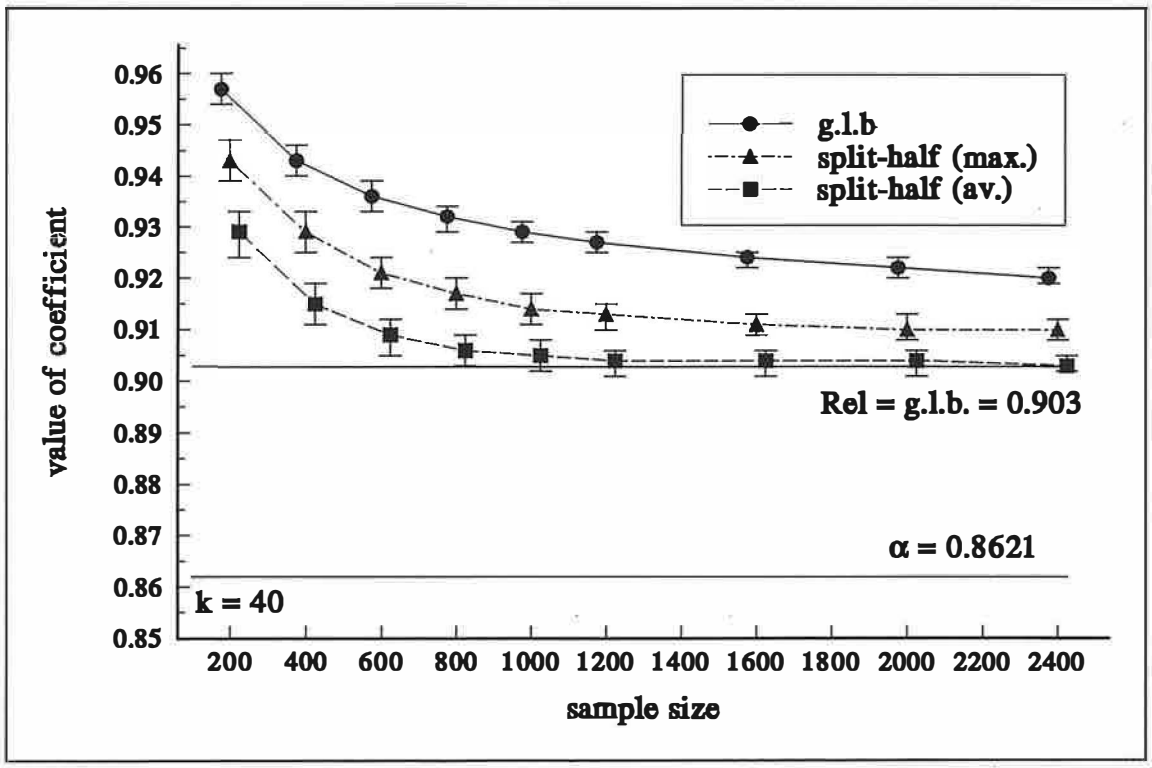

Figure 24: Example 2, first panel, $k=40$, $\alpha<Rel=0.9$

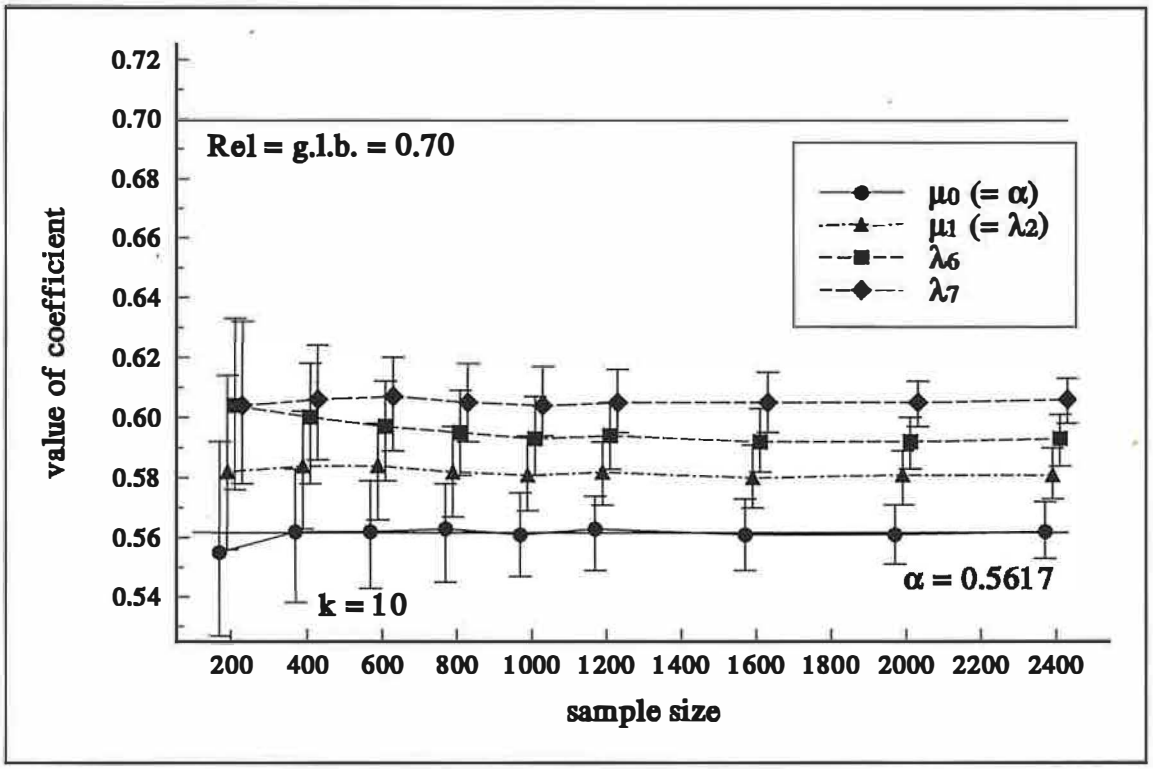

Figure 25: Example 2, second panel, $k=10$, $\alpha < Rel = 0.7$

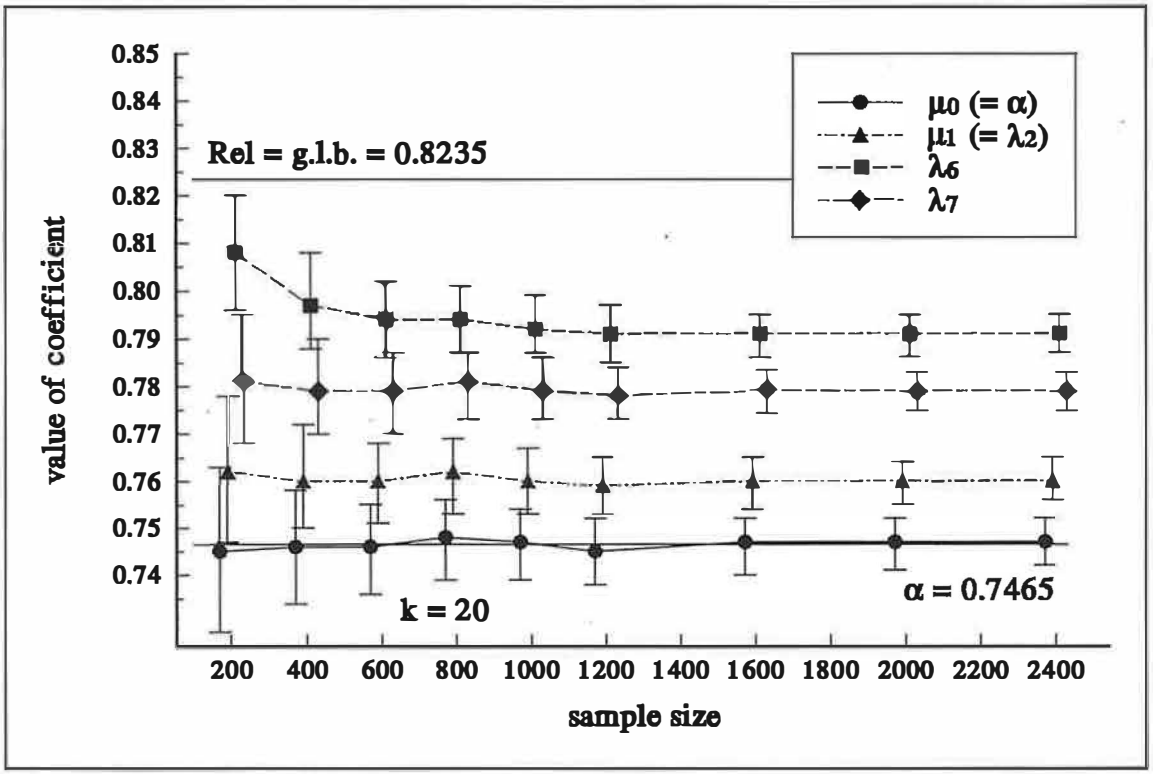

Figure 26: Example 2, second panel, $k=20$, $\alpha < Rel = 0.82$

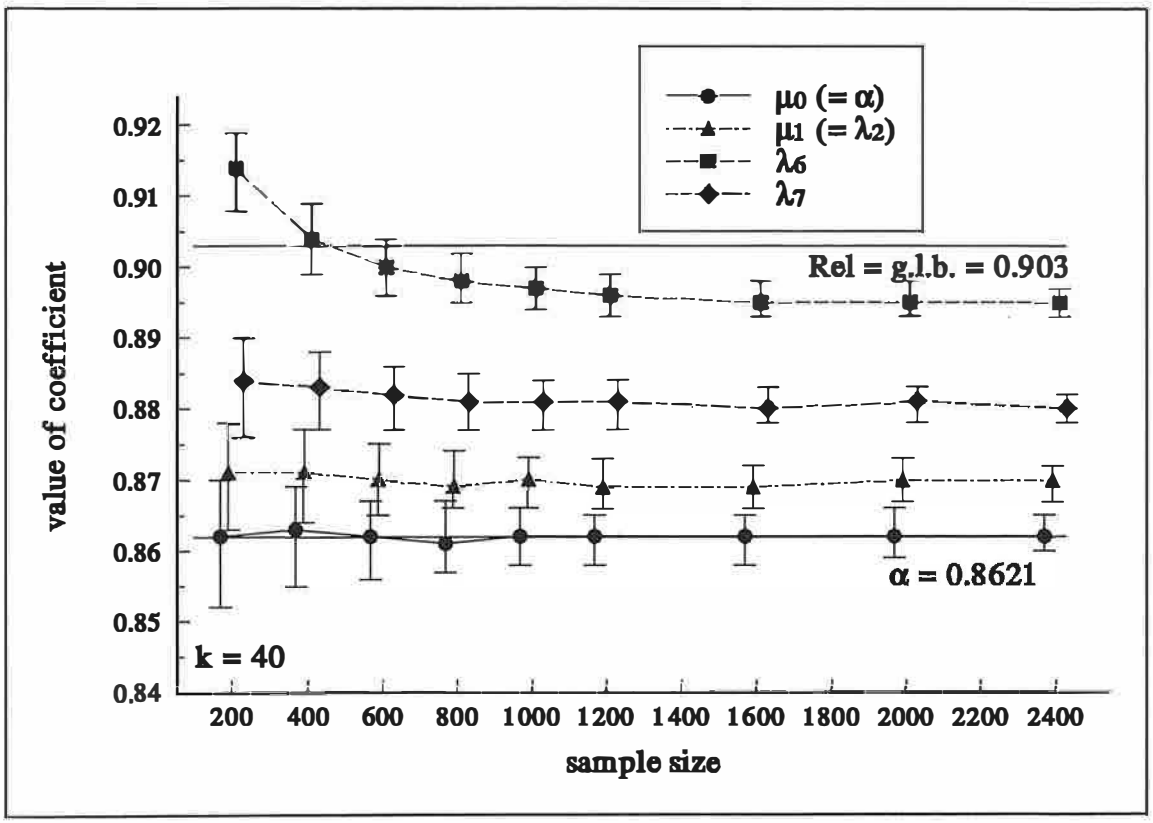

Figure 27: Example 2, second panel, k=40, $\alpha < Rel$=0.9

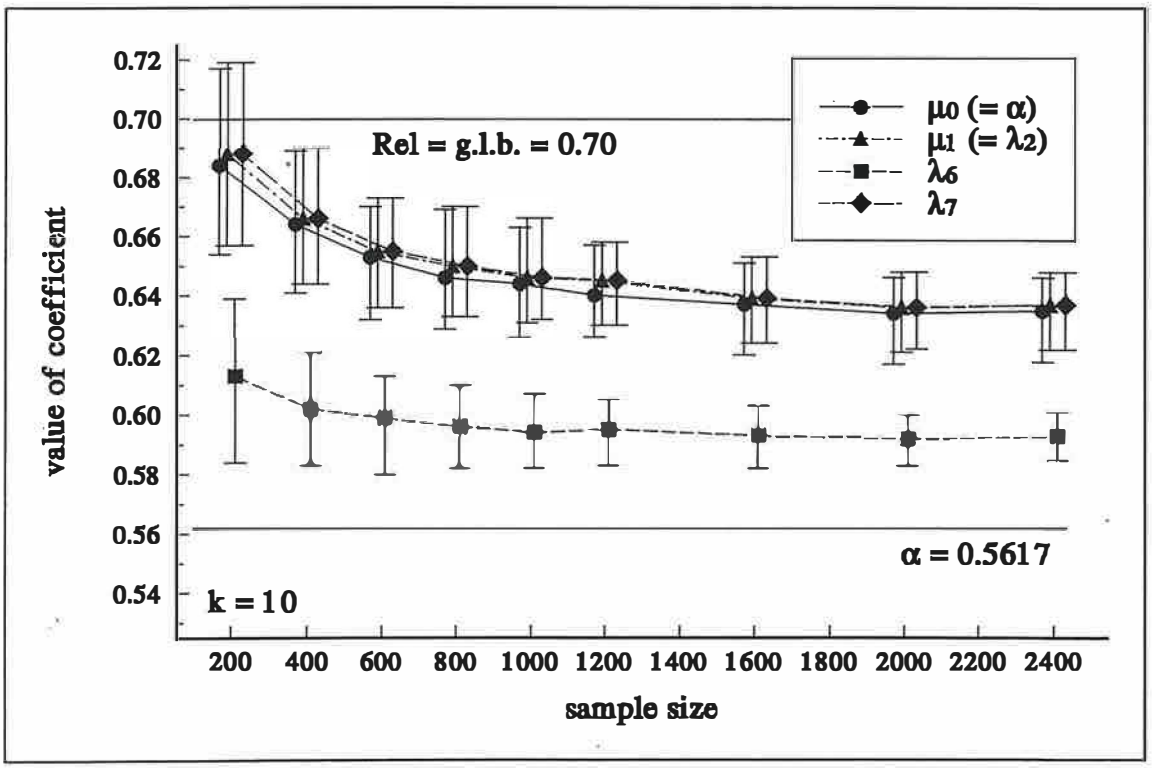

Figure 28: Example 2, third panel, k=10, $\alpha < Rel$=0.7

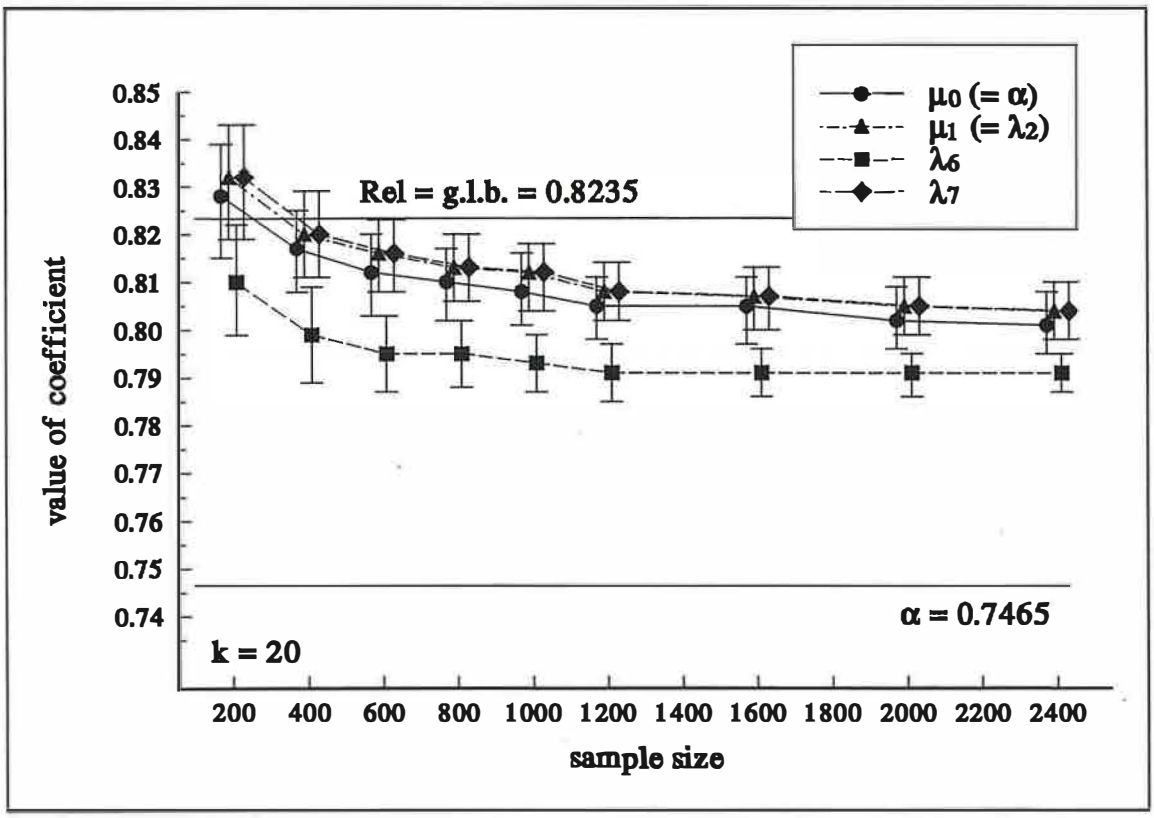

Figure 29: Example 2, third panel, $k=20$, $\alpha < Rel = 0.82$

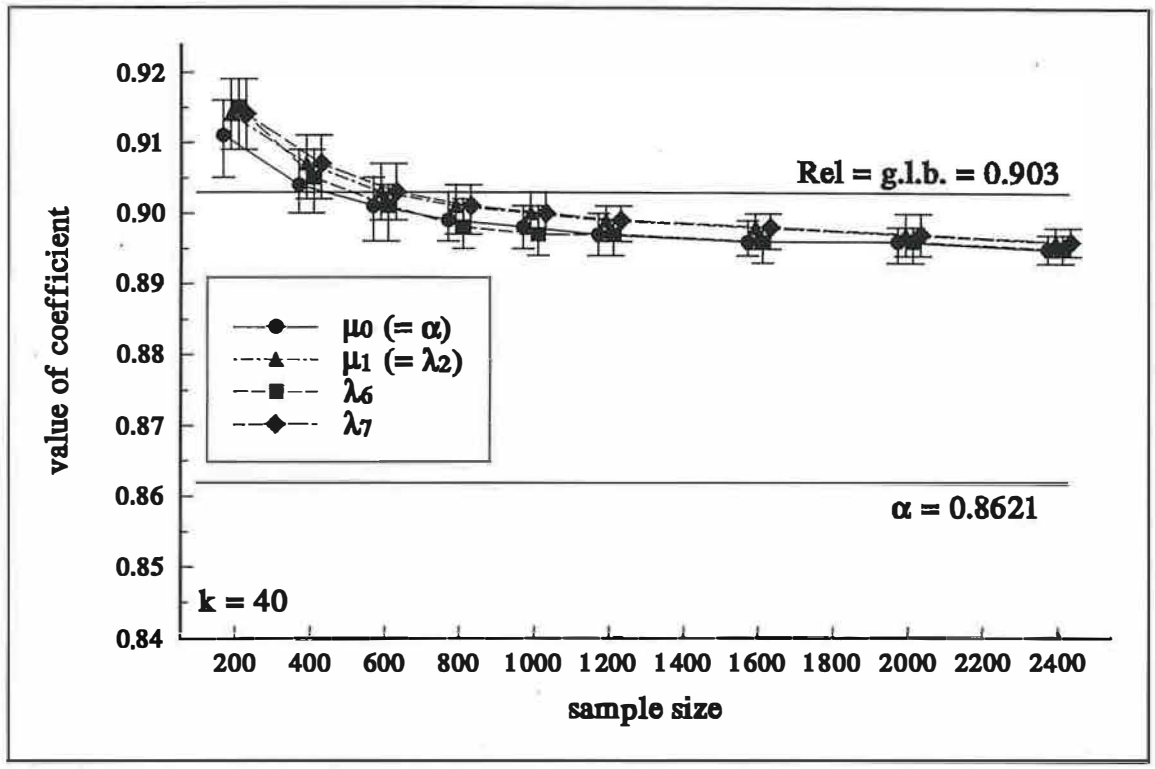

Figure 30: Example 2, third panel, $k=40$, $\alpha < Rel = 0.9$

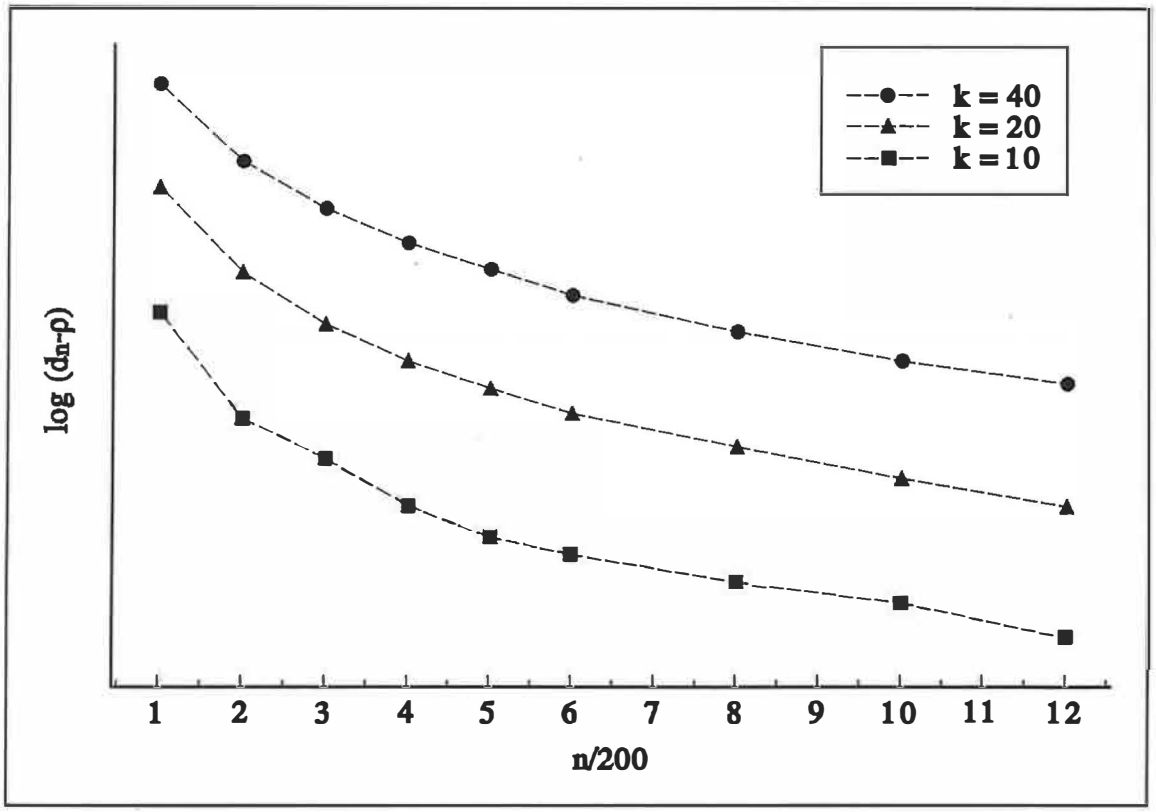

Figure 31: Logarithm of the bias as a function of the sample size

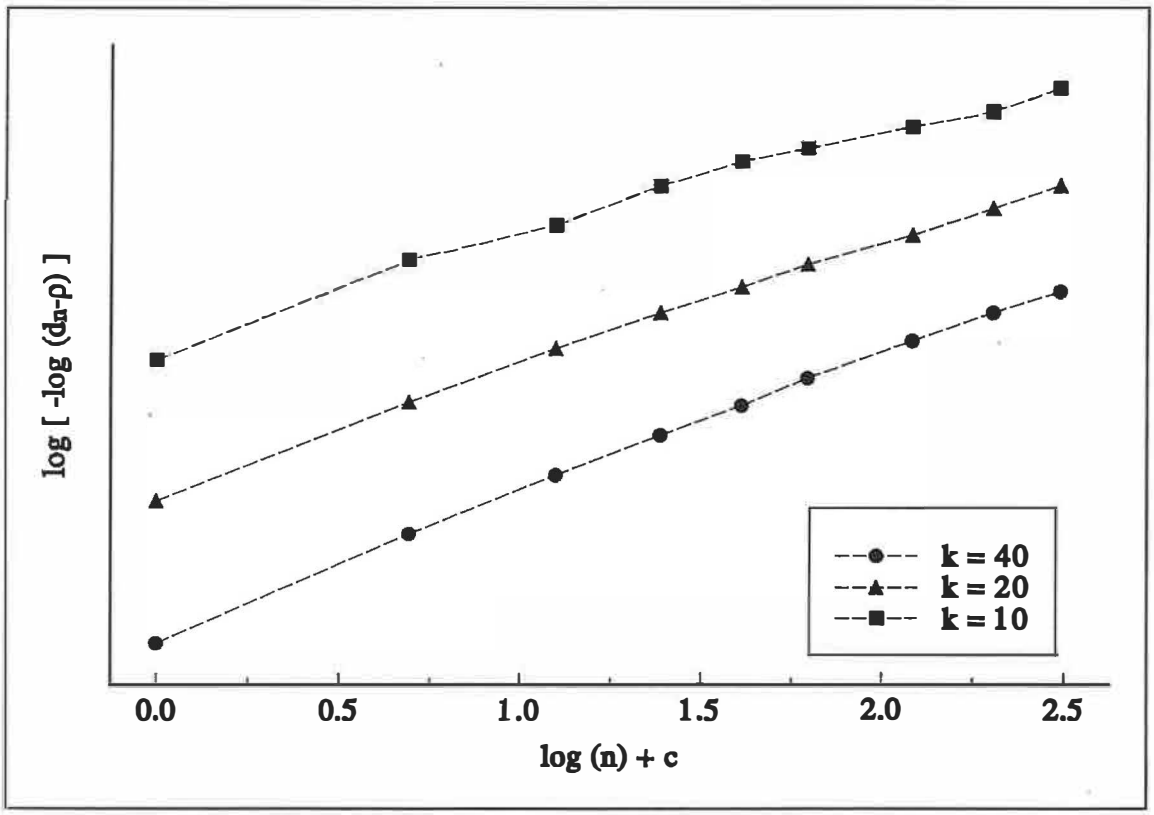

Figure 32: Log[-log(bias)] as a function of log(n)

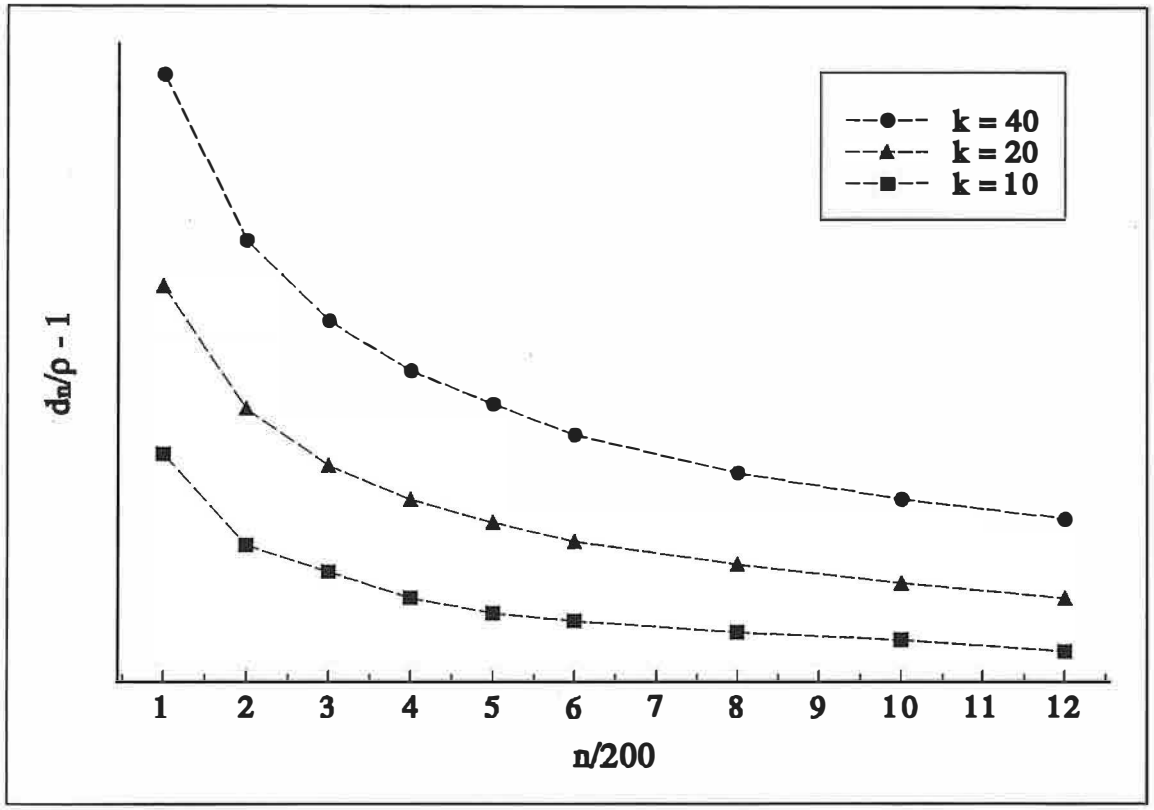

Figure 33: Relative bias as a function of the sample size

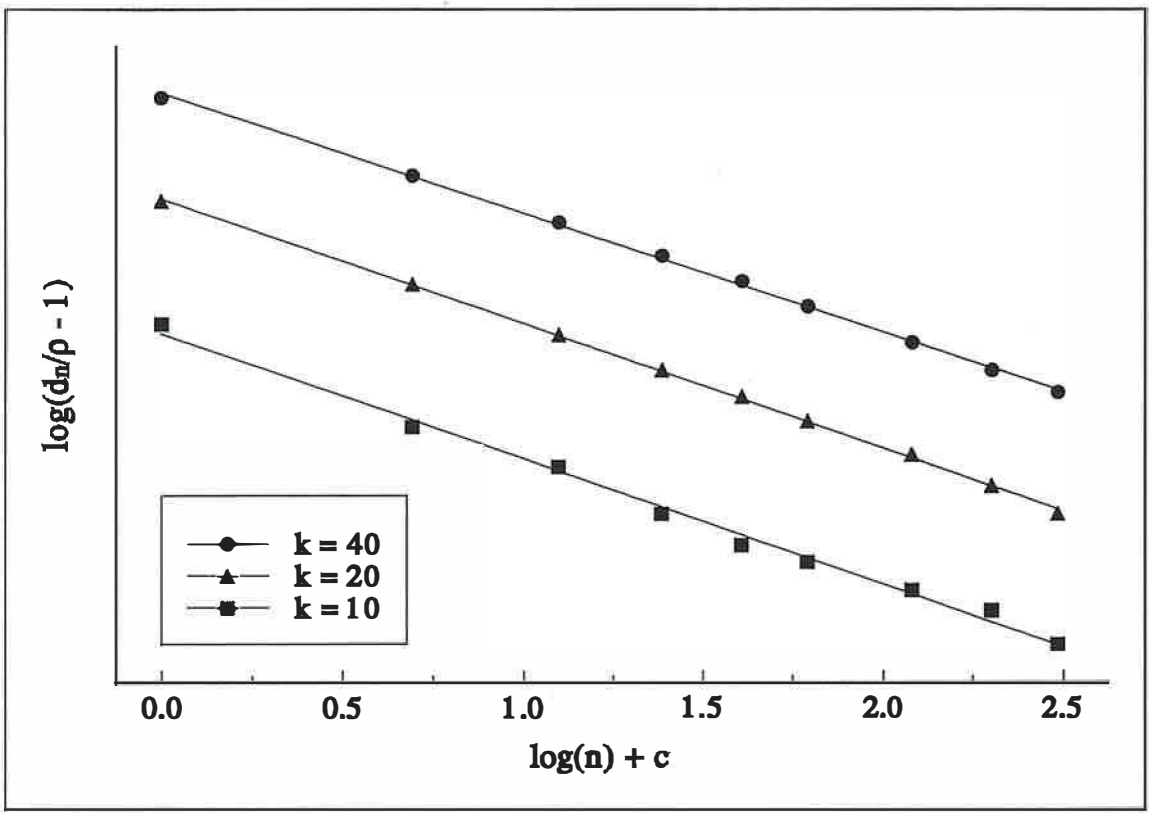

Figure 34: Logarithm of the relative bias as a function of log(n) with linear regression lines

Recent Measurement and Research Department Reports:

98-1 T.J.H.M. Eggen. Item Selction in Adaptive Testing with the Sequential Probability Ratio Test.