Confidence intervals for maximal reliability of probability judgments

Kar Yin Lam* Alex J. Koning † Philip Hans Franses‡

Econometric Institute report EI 2007-09

Abstract

Subjective probabilities play an important role in marketing research, for example where individuals rate the likelihood that they will purchase a new to develop product. The tau-equivalent model can describe the joint behaviour of multiple test items measuring the same subjective probability. It improves the reliability of the subjective probability estimate by using a weighted sum as the outcome of the test rather than an unweighted sum. One can choose the weights to obtain maximal reliability.

In this paper we stress the use of confidence intervals to assess maximal reliability, as this allows for a more critical assessment of the items as measurement instruments. Furthermore, two new confidence intervals for the maximal reliability are derived and compared to intervals derived earlier in Yuan and Bentler (2002); Raykov and Penev (2006). The comparison involves coverage curves, a methodology that is new in the field of reliability. The existing Yuan-Bentler and Raykov-Penev intervals are shown to overestimate the maximal reliability, whereas one of our proposed intervals, the stable interval, performs very well. This stable interval hardly shows any bias, and has a coverage for the true value which is approximately equal to the confidence level.

1 Introduction and motivation

Marketing researchers often conduct tests to learn about individuals' preferences, opinion and attitudes. Opinions may often be expressed by subjective

*ERIM & Econometric Institute, Erasmus University Rotterdam, The Netherlands
†Econometric Institute, Erasmus University Rotterdam, The Netherlands
‡Econometric Institute, Erasmus University Rotterdam, The Netherlands

probabilities. A person's subjective probability is the degree of belief a person holds regarding a statement or an event, that is, it is a person's personal judgment about how likely a particular statement is or how likely a particular event is to occur. Subjective probabilities play a role in marketing, for example when potential consumers are asked how likely it is that they will purchase a new to develop product.

Most subjective probability scales are similar to psychological measurement scales, see Wallsten and Budescu (1983). In line with classical test theory, in subjective probability theory it is assumed that observed uncertainty is given by the sum of the true score and a random error. Typically, a test is composed of multiple items, where each item is an independent attempt to measure the same construct of interest T. The outcome of the test U is usually the unweighted sum of the individual item outcomes Y_j. We shall refer to T, U and Y_j as the true score, the test score, and the j^{th} item score, respectively.

The amount of agreement between the test score U and the true score T, as captured by the squared correlation between U and T, is defined to be the reliability of the test. However, as T cannot be observed, we are unable to estimate the reliability of a test directly. Fortunately, we may use the dependence between the individual item scores in order to evaluate the reliability of a test indirectly.

The most widely used indirect technique for assessing reliability, Cronbach's coefficient alpha, see (Cronbach, 1951), compares the variance of the test score with the sum of the variances of the individual item scores. Cronbach's alpha should be regarded as a lower bound of the reliability of a test, and in certain special situations it coincides with reliability.

Despite its popularity, the interpretation of Cronbach's alpha in practice is quite arbitrary. Nunnally's thresholds, see Nunnally and Bernstein (1994), are often taken as recommendations regarding minimally acceptable reliability, although one may argue that it is rather subjective to compare alpha to an arbitrary threshold.

Moreover, such an approach does not take into account the accuracy of the estimated alpha. In this paper we advocate the use of confidence intervals to assess reliability, as the additional information included in the interval allows for a more critical assessment of the statistic.

In the literature we find two types of intervals for Cronbach's alpha. The first type is derived under the so-called parallel model. In the parallel model each of the item scores Y_j is assumed to be the sum of the true score T and a measurement error ϵ_j with population mean zero and common population variance ψ. Moreover, the random variables $T, \epsilon_1, \ldots, \epsilon_k$ are assumed

to be independent. As a consequence, the parallel model imposes strong restrictions on the population covariance matrix of the item scores. In the parallel model, Cronbach's alpha coincides with reliablity. Thus, confidence intervals for Cronbach's alpha may be viewed as confidence intervals for reliability itself rather than confidence intervals for a lower bound to reliability. The statistical theory needed to derive confidence intervals for Cronbach's alpha in the parallel model appeared in Kristof (1963); Feldt (1965), but it remained relatively unnoticed. The coverage of this theory in van Zyl et al. (2000) revived the attention for confidence intervals for Cronbach's alpha, see (Iacobucci and Duhachek, 2003; Duhachek et al., 2005).

The second type of intervals for Cronbach's alpha relies on the so-called saturated model. This model does not impose any restriction on the population covariance matrix of the item scores.

It should be stressed that between the extremely restrictive parallel model on the one hand, and the extremely permissive saturated model on the other hand, other models exist. As both extreme cases have their problems, in this paper we restrict ourselves to a plausible intermediate model, the so-called tau-equivalent model. The tau-equivalent model can be obtained from the parallel model by relaxing the assumption that the measurement errors ϵ_j have common population variance ψ. Thus, the respective population variances $\psi_1, \ldots, \psi_k$ of the measurement errors $\epsilon_1, \ldots, \epsilon_k$ may differ. In other words, the tau-equivalent model allows nonhomogeneous error variances. By measuring the probability that a same event will occur with different methods, one can say something about the quality of the measured probability. Hence, one can understand that in case of subjective probabilities, a tau-equivalent model is more plausible than the extremely restrictive parallel model and also more informative than the extremely permissive saturated model. In this paper we propose new methods to establish confidence bounds for this tau-equivalent model.

Several methods allow for improving reliability by using a weighted sum W rather than an unweighted sum U as the outcome of the test. One may show that choosing the weight for Y_j equal to $1/\psi_j$ yields a test score $W = \sum_{j=1}^{k} Y_j/\psi_j$ which has maximal reliability, which is

$$\varphi = 1 - \frac{1}{1 + \sum_{\ell=1}^{k} \lambda^2/\psi}.$$

The quantity φ is referred to as the maximal reliability of $Y_1, \ldots, Y_k$.

We remark that under the parallel model the optimal weights have a common value $1/\psi$, which implies that the unweighted sum U yields maximal reliability. Thus, in the parallel model maximal reliability coincides with reliability and Cronbach's alpha.

In Yuan and Bentler (2002); Raykov and Penev (2006) different expressions for the standard deviation of the maximum likelihood estimator $\hat{\varphi}$ of the maximal reliability φ are given, from which confidence intervals for φ are readily derived. In this paper we compare the Yuan-Bentler and Raykov-Penev intervals to each other, and to our two newly proposed intervals.

The structure of this paper is as follows. In Section 2 we present the four confidence intervals, and show that their coverage is asymptotically equal to the requested confidence level. Proofs are relegated to the Appendix. In Section 3 we apply the four confidence intervals to real data involving measures of subjective probability. Section 4 discusses the results of an extensive simulation experiment. In Section 5 conclusions are drawn.

2 Confidence intervals for the tau-equivalent model

Let $\hat{\psi}_j$ and $\hat{\lambda^2}$ be the maximum likelihood estimators of ψ_j and λ^2 in the tau-equivalent model. Then, the maximum likelihood estimator of φ is given by

$$\hat{\varphi} = 1 - \frac{1}{1+\hat{\zeta}},$$

with

$$\hat{\zeta} = \sum_{j=1}^{k} \hat{\lambda^2}/\hat{\psi}_j.$$

The confidence intervals for φ involve

$$s^2 = 2 + 2\frac{\hat{\zeta}+1}{\hat{\zeta}-1}\frac{\hat{Q}}{1+\hat{Q}},$$

with

$$\hat{Q} = \frac{\hat{\zeta}-1}{\hat{\zeta}^2}\sum_{j=1}^{k}\frac{\left\{\hat{\lambda^2}/\hat{\psi}_j\right\}^2}{1+\hat{\zeta}-2\hat{\lambda^2}/\hat{\psi}_j}.$$

Choose $0 < \gamma < 1$, and determine $z_{(1-\gamma)/2}$ so as to satisfy

$$P\left(-z_{(1-\gamma)/2} < Z < z_{(1-\gamma)/2}\right) = \gamma.$$

We next present several asymptotic $100\gamma\%$ confidence intervals for the maximal reliability coefficient φ.

2.1 Available results

The first interval is proposed in general terms in Yuan and Bentler (2002). [1] By taking their computations for the tau-equivalent model, this yields that the endpoints of the Yuan-Bentler interval are given by

$$\hat{\varphi} \pm z_{(1-\gamma)/2} \left(1 - \hat{\varphi}\right) \frac{s}{\sqrt{n}} \tag{1}$$

In Raykov and Penev (2006) a "second-order" alternative to the Yuan-Bentler interval is proposed. In the tau-equivalent model, the endpoints of the Raykov-Penev interval are given by

$$\hat{\varphi} \pm z_{(1-\gamma)/2} \left(1 - \hat{\varphi}\right) \frac{s}{\sqrt{n}} \sqrt{1 + 2\frac{s^2}{n}}. \tag{2}$$

Note that the Yuan-Bentler interval is always contained in the Raykov-Penev interval.

2.2 The unstable interval

The unstable confidence interval has endpoints

$$1 - \frac{1 - \hat{\varphi}}{1 \pm z_{(1-\gamma)/2} \frac{s}{\sqrt{n}}} \tag{3}$$

In the Appendix it is shown that

$$\sqrt{n} \frac{\hat{\zeta} - \zeta}{\left(1 + \zeta\right) \sqrt{2 + 2\frac{\zeta+1}{\zeta-1} \frac{Q}{1+Q}}} \rightarrow_d N\left(0, 1\right), \tag{4}$$

as n tends to infinity. Here

$$\zeta = \sum_{j=1}^{k} \lambda^2 / \psi_j, \tag{5}$$

[1] It is critized in Raykov and Penev (2006) because it is "(a) rather laborious and tedious in routine behavioral research in need of interval estimation of maximal reliability, (b) involves taking by the researcher of multiple partial derivatives of this reliability coefficient with respect to model parameters, (c) has the inconvenient property that the number of these derivatives increases with increasing length k of the initial composite of interest (as could be repeatedly the case when one is involved in scale development and revision), and (d) can be viewed as based on a first-order approximation of maximal reliability as a function of model parameters."

$$Q = \frac{\zeta - 1}{\zeta^2} \sum_{j=1}^{k} \frac{\{\lambda^2/\psi_j\}^2}{1 + \zeta - 2\lambda^2/\psi_j}. \tag{6}$$

Note that s^2 is a consistent estimator of $\sqrt{2 + 2\frac{\zeta+1}{\zeta-1}\frac{Q}{1+Q}}$. As

$$\varphi = 1 - \frac{1}{1+\zeta},$$

it follows from (4) that

$$\lim_{n\to\infty} P\left(-z_{(1-\gamma)/2}\frac{s}{(1-\hat{\varphi})\sqrt{n}} < \hat{\zeta} - \zeta < z_{(1-\gamma)/2}\frac{s}{(1-\hat{\varphi})\sqrt{n}}\right) = \gamma,$$

which implies that

$$\hat{\zeta} \pm z_{(1-\gamma)/2}\frac{s}{(1-\hat{\varphi})\sqrt{n}}$$

are the endpoints of an asymptotic $100\gamma\%$ confidence interval for ζ. As φ is a monotone function of ζ, we obtain that

$$\begin{aligned} 1 - \frac{1}{1 + \hat{\zeta} \pm z_{(1-\gamma)/2}\frac{s}{(1-\hat{\varphi})\sqrt{n}}} &= 1 - \frac{1}{\frac{1}{(1-\hat{\varphi})} \pm z_{(1-\gamma)/2}\frac{s}{(1-\hat{\varphi})\sqrt{n}}} \\ &= 1 - \frac{1-\hat{\varphi}}{1 \pm z_{(1-\gamma)/2}\frac{s}{\sqrt{n}}} \end{aligned}$$

are the endpoints of an asymptotic $100\gamma\%$ confidence interval for φ. This interval coincides with our unstable interval in (3).

The unstable interval is related to the Yuan-Bentler interval and Raykov-Penev interval. As the derivative of $1 - (1+\zeta)^{-1}$ with respect to ζ equals $(1+\zeta)^{-2}$, combining (4) with the delta method yields that

$$\sqrt{n}\frac{\hat{\varphi} - \varphi}{(1-\varphi)\sqrt{2 + 2\frac{\zeta+1}{\zeta-1}\frac{Q}{1+Q}}} \to_d N(0,1). \tag{7}$$

It follows that

$$\lim_{n\to\infty} P\left(-z_{(1-\gamma)/2}(1-\hat{\varphi})\frac{s}{\sqrt{n}} < \hat{\varphi} - \varphi < z_{(1-\gamma)/2}(1-\hat{\varphi})\frac{s}{\sqrt{n}}\right) = \gamma,$$

which implies that

$$\hat{\varphi} \pm z_{(1-\gamma)/2}(1-\hat{\varphi})\frac{s}{\sqrt{n}}$$

are the endpoints of an asymptotic $100\gamma\%$ confidence interval for φ. This interval coincides with the Yuan-Bentler interval in (1).

In Raykov and Penev (2006) properties of the normal distribution and the second order approximation

$$\hat{\varphi} - \varphi = \frac{\hat{\zeta} - \zeta}{(1+\zeta)^2} - \frac{\left(\hat{\zeta} - \zeta\right)^2}{(1+\zeta)^3},$$

are used to obtain an asymptotic confidence interval for $\hat{\varphi} - \varphi$. This yields

$$\lim_{n\to\infty} P\left(-z_{(1-\gamma)/2}\,(1-\hat{\varphi})\,\frac{s}{\sqrt{n}}\sqrt{1+2\frac{s^2}{n}} < \hat{\varphi} - \varphi < z_{(1-\gamma)/2}\,(1-\hat{\varphi})\,\frac{s}{\sqrt{n}}\sqrt{1+2\frac{s^2}{n}}\right) = \gamma,$$

which implies that

$$\hat{\varphi} \pm z_{(1-\gamma)/2}\,(1-\hat{\varphi})\,\frac{s}{\sqrt{n}}\sqrt{1+2\frac{s^2}{n}}$$

are the endpoints of an asymptotic $100\gamma\%$ confidence interval for φ. This interval coincides with the Raykov-Penev interval in (2).

2.3 The stable interval

The stable confidence interval, which we wish to advocate in the present paper, has endpoints

$$1 - (1-\hat{\varphi})\exp\left\{\pm z_{(1-\gamma)/2}\frac{s}{\sqrt{n}}\right\} \tag{8}$$

We believe that there is a serious drawback to the direct use of (4) in constructing confidence intervals for φ. Although we know that the left hand side of (4) converges in distribution to a standard normal random variable as the number of observations n becomes large, it may well have a very different distribution for a given value of n. Figure 1 illustrates that this is indeed the case.

In order to motivate our remedy for this problem, we assume in a first instance that the items of the test are drawn at random from a large test battery. That is, the values $\psi_1, \ldots, \psi_k$ are drawn independently from the distribution of some non-negative random variable, say E. In addition, we shall assume that the variance of E^{-1} is finite. It follows by (5) and (6) that the quantities ζ and Q depend on the number k of items drawn. One may show that

$$\lim_{k\to\infty} \frac{\zeta+1}{\zeta-1}\frac{Q}{1+Q} = 0. \tag{9}$$

It follows by (4) that the asymptotic variance of $\sqrt{n}\left(\hat{\zeta}-\zeta\right)$ is approximated by $2\left(1+\zeta\right)^2$ for large k. Thus, the variance of $\hat{\zeta}$ depends on ζ, and this dependence does not disappear when k tends to infinity. That is, the variance of $\hat{\zeta}$ is not stable. As the variance is proportional to $\left(1+\zeta\right)^2$ for k sufficiently large, the theory of variance stabilizing transformations suggests the use of $\ln\left(1+\hat{\zeta}\right)$ for constructing confidence intervals.

It can be shown that

$$\sqrt{n}\frac{\ln\left(1+\hat{\zeta}\right)-\ln\left(1+\zeta\right)}{\sqrt{2+2\frac{\zeta+1}{\zeta-1}\frac{Q}{1+Q}}}\rightarrow_d N\left(0,1\right). \tag{10}$$

Figure 2 illustrates that the standard normal distribution provides a far better fit to the distribution of the left hand side of (10) than to the distribution of the left hand side of (4).

It follows from (10) that

$$\lim_{n\rightarrow\infty}P\left(-z_{(1-\gamma)/2}\frac{s}{\sqrt{n}}<\ln\left(\frac{1+\hat{\zeta}}{1+\zeta}\right)<z_{(1-\gamma)/2}\frac{s}{\sqrt{n}}\right)=\gamma,$$

which implies that

$$\left(1+\hat{\zeta}\right)\exp\left\{\pm z_{(1-\gamma)/2}\frac{s}{\sqrt{n}}\right\}-1$$

are the endpoints of an asymptotic $100\gamma\%$ confidence interval for ζ. As φ is a monotone function of ζ, we obtain that

$$1-\frac{\exp\left\{\pm z_{(1-\gamma)/2}\frac{s}{\sqrt{n}}\right\}}{1+\hat{\zeta}}=1-\left(1-\hat{\varphi}\right)\exp\left\{\pm z_{(1-\gamma)/2}\frac{s}{\sqrt{n}}\right\}$$

are the endpoints of an asymptotic $100\gamma\%$ confidence interval for φ. This interval coincides with our stable interval in (8).

3 Illustration

In Wallsten and Budescu (1983) it is assumed that measures of subjective uncertainty can be written as the sum of two independent random variables, a fixed true measure and a variable error. Thus, if the subjective probability of a given event is measured by different methods, then a tau-equivalent model seems plausible.

In Ofir and Reddy (1996) the psychometric properties of three measures of subjective uncertainty are investigated. The 117 respondents were asked

	StckP	*StckL*	*StckC*
StckP	638.790		
StckL	562.214	620.501	
StckC	509.735	501.765	619.956

Table 1: Covariance matrix of a zero-to-hundred subjective probability scale and two seven point rating scales measuring the subjective probability of the event "The stock market will rise during 1991 by at least 10%", compiled from Appendix A in Ofir and Reddy (1996). There are 117 respondents.

to express the subjective probability of the event "The stock market will rise during 1991 by at least 10%" on each of the following measurement scales:

StckP a seven-point "probable" rating scale with categories Highly Improbable, Improbable, Somewhat Improbable, Equally Probable, Somewhat Probable, Probable and Highly Probable;

StckL a seven-point "likelihood" rating scale with categories Very Unlikely, Unlikely, Somewhat Unlikely, Equal Likelihood, Somewhat Likely, Likely, Very Likely.

StckC a subjective probability scale ranging from zero to one hundred.

The rating scales *StckP* and *StckL* were transformed to 0-100 scales by using the transformation $100(x-1)/6$, where x is the value on the seven-point scale.

Table 1 reports the covariance matrix of these three measures of the subjective probability of the event "The stock market will rise during 1991 by at least 10%." Indeed, the tau-equivalent model fits the data [$\chi^2 = 3.161$, df $= 2$, $P = 0.206$]. However, the parallel model is clearly rejected [$\chi^2 = 14.679$, df $= 4$, $P = 0.005$]. In the tau-equivalent model, the estimated true variance is 541.563, and the estimated error variances are 72.454, 67.065 and 162.962. Thus,

$$\frac{StckP}{72.454} + \frac{StckL}{67.065} + \frac{StckC}{162.962}$$

is an estimate of the weighted composite of *StckP*, *StckL* and *StckC* yielding maximal reliability.

The corresponding 95% confidence intervals for maximal reliability φ are found in Table 2. Note that the Yuan-Bentler and the Raykov-Penev intervals are symmetric around $\hat{\varphi}$, whereas the stable and the unstable interval

Interval	*Lower*	*Upper*
Unstable	0.9248181	0.9621855
YuanBentler	0.9330400	0.9663209
RaykovPenev	0.9325728	0.9667881
Stable	0.9299584	0.9638492

Table 2: Asymptotic 95% confidence intervals for maximal reliability φ, derived from the data in Tabel 2. The estimator $\hat{\varphi}$ takes the value 0.9496805.

are not. Moreover, the Raykov-Penev interval contains the Yuan-Bentler interval. The lower endpoint of the unstable interval is smaller than the lower endpoint of the stable interval, which in turn is smaller than the lower endpoint of the Raykov-Penev interval. Similarly, the upper endpoint of the unstable interval is smaller than the upper endpoint of the stable interval, which in turn is smaller than the upper endpoint of the Raykov-Penev interval.

4 Simulation results

The above illustration for actual data showed that there are differences between confidence intervals, but these do not necessarily have to be very large. To further our understanding of the differences between the four confidence bounds, we rely on simulation, to be discussed in this section.

In the simulations, we let k take the values $2, 3, 4, 5, 10, 15$, and n take the values $25, 50, 100, 200, 400$. We expect that the simulation results largely depend on the quantities μ_1 and μ_2, where

$$\mu_1 = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\psi_i}, \quad \mu_2 = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{\psi_i} - \mu_1\right)^2. \tag{11}$$

In order to be able to confirm this expectation, we generate the ψ_i's using various patterns. In particular, we choose

$$\frac{1}{\psi_i} = a + bg\left(\frac{i}{k+1}\right),$$

where g is one of the three following functions:

$$g_1(s) = s - \tfrac{1}{2},$$

$$g_2(s) = s^2 - \tfrac{1}{3},$$

$$g_3(s) = \begin{cases} -1 & \text{for } s < \tfrac{1}{2}, \\ +1 & \text{for } s \geq \tfrac{1}{2}, \end{cases} .$$

Here a and b are chosen so as to satisfy (11) for given μ_1 and μ_2. Throughout the simulations we set λ^2 equal to 1. As $\mu_1 = (1/k)\sum_{\ell=1}^{k} 1/\psi_\ell$, we may show that $\mu_1 = (k\lambda^2)^{-1}\varphi/(1-\varphi)$. In our simulations we set μ_1 equal to values which correspond to maximal reliability 0.60, 0.75, 0.90 and 0.95. Finally, we set μ_2 equal to $\frac{1}{2}\mu_1^2$, μ_1^2 and $2\mu_1^2$.

First, we investigate the extent in which the standard normal approximations (10) and (4) are valid for small to moderate sample sizes n. As Figure 3 illustrates, these approximations are truly asymptotic in nature when the number of test items k is equal to 2. When $k = 3$, the approximations are reasonably accurate, except for the sample size $n = 25$. Recall that the standard normal approximations (10) and (4) provide the probabilistic basis for all confidence intervals discussed in this paper. Hence, in the remainder of this section we shall restrict ourselves to situations in which $k \geq 3$ and $n \geq 50$.

Next, we examine the simulated coverage curves of the four intervals. The coverage of a confidence interval for a given hypothetical value of φ is defined as the probability that this hypothetical value is contained in the interval. The coverage curve is the curve that the coverage follows as the hypothetical value of φ ranges through an interval of possible values of φ. The ideal shape of the coverage curve is as follows:

- if the hypothetical value of φ equals the true value, then the coverage should be equal to the confidence level;
- if the hypothetical value of φ differs from the true value, then the coverage should be as low as possible. In particular, the coverage should be lower than the confidence level.

There are various ways in which a coverage curve may deviate from the ideal shape. In our examination of the performance of the confidence interval, we shall in particular consider the unbiasedness of a confidence interval. A confidence interval is called unbiased if for every hypothetical value of φ the coverage of the interval does not exceed the coverage for the true value of φ.

If an interval is unbiased, the next issue to consider is whether the coverage for the true value equals the confidence level. If this coverage is larger than the confidence interval, the unbiased confidence interval is called conservative; if this coverage is smaller than the confidence interval, the unbiased

confidence interval is called anti-conservative. Being conservative is considered less harmful than being anti-conservative.

Statistical testing theory yields an alternative way to interpret coverage curves. Note that for every confidence interval for φ, there exists a related statistical test of the null hypothesis $H_0 : \varphi = \varphi_0$; this test does not reject the null hypothesis if the hypothetical value φ_0 lies inside the confidence interval, and rejects the null hypothesis if the hypothetical value φ_0 is outside the confidence interval. Now, if the true value of φ differs from the hypothetical value φ_0, then the coverage of the hypothetical value φ_0 is equal to the probability of a type II error of the related statistical test of $H_0 : \varphi = \varphi_0$. Thus, by subtracting the coverage curve from 1, we in fact obtain the power curve of the related test. In fact, the description of the ideal shape of the coverage curve given above is a direct translation of generally accepted rules involving the ideal shape of the power curve.

Evaluating the coverage not only under the null hypothesis, but also under the alternative hypothesis provides a much more comprehensive view of the behaviour of the various confidence intervals.

The simulated coverage curves depend on the sample size n, the true value of the maximal reliability φ, the number of items k and on the quantities μ_1 and μ_2.

- For fixed values of n, φ, k, μ_1 and μ_2, there is little difference between the coverage curves.

- The sample size n has a positive effect on the performance of all confidence intervals, see Figure 4. That is, the confidence intervals perform better for large n.

- The true value of the maximal reliability has a positive effect on the coverage of all confidence intervals, see Figure 5. That is, the coverage of the confidence intervals becomes higher when the true value of φ approaches 1. If the true value of the maximal reliability is small, the confidence intervals are anti-conservative. When the true value of φ approaches 1, the coverage of the true value increases. As a result, the anti-conservatism diminishes in most situations. However, in some situations (especially $k = 3$) the anti-conservatism turns into conservatism.

- The number of items k has a positive effect on the performance of all confidence intervals, see Figure 6. That is, the confidence intervals perform better for large k.

The simulated coverage curves yield the following general findings with respect to the differences in performance between the four confidence intervals.

- There is little difference between the Yuan-Bentler and the Raykov-Penev interval. Both intervals show a positive bias, and thus overestimate the the true maximal reliability.
- The unstable interval shows a negative bias, and thus underestimates the true maximal reliability.
- Except for extreme conditions (that is, a combination of a small sample size, a small number of items k and a high true value of the maximal variability), the stable interval hardly shows any bias, and has a coverage for the true value which is approximately equal to the confidence level.

Finally, we remark that we could not have detected the positive bias of the Yuan-Bentler and the Raykov-Penev without the construction of coverage curves, that is, evaluating the coverage under the null hypothesis as well as under the alternative hypothesis. We highly recommend the use of coverage curves in other studies on confidence intervals.

5 Conclusion

We have shown that a tau-equivalent model is plausible when measuring subjective probabilities, which play an important role in marketing research. To improve the reliability of the test we use a weighted sum of individual item scores rather than an unweighted sum. In principle, the weights may be chosen so as to obtain maximal reliability, and these optimal weights may be estimated from the data.

Next, we discussed the issue of estimating the maximal reliability. We stressed the use of confidence intervals rather than point estimators to assess maximal reliability, as the additional information included in the interval allows for a more critical assessment of the quality of the items as measurement instruments of the underlying subjective probability.

We have derived two new confidence intervals for maximal reliability and compared the performance of these intervals with earlier proposed intervals in Yuan and Bentler (2002) and Raykov and Penev (2006). To compare these intervals, we have used coverage curves, a methodology that seems new in the field of reliability. That is, we have not only considered the coverage of the true maximal reliability, but the coverage of hypothetical values which differ from the true maximal reliability as well.

It turns out that the Yuan-Bentler and the Raykov-Penev intervals are closely related to each other. In fact, the Yuan-Bentler interval is always contained in the alternative interval proposed in Raykov and Penev (2006). Both intervals show a positive bias. Interestingly, the similarity in behaviour of the two intervals is in contrast with the fierce criticism in Raykov and Penev (2006) with regard to the Yuan-Bentler interval. Moreover, it seems that the additional complexity of the Raykov-Penev interval does not pay off, as a clear advantage of using this interval over the Yuan-Bentler is lacking.

We have also examined the performance of the two new intervals proposed in this article. Though the unstable interval shows a considerable negative bias, the stable interval performs considerably well. Except for extreme conditions the stable interval hardly shows any bias, and has a coverage for the true value which is approximately equal to the confidence level. This shows the advantage of the use of a stabilization technique in constructing confidence intervals. In further applications in marketing we therefore recommend the use of this new stable confidence interval.

The advantage of stabilization should not only hold in the tau-equivalent model. This raises the issue whether variance stabilization is of use in the parallel model as well. In the parallel model an unstable interval is given in van Zyl et al. (2000), see also Iacobucci and Duhachek (2003). It would be interesting to compare those intervals using coverage curves.

References

Browne, M. W. 1984. Asymptotically distribution-free methods for the analysis of covariance structures. *The British Journal of Mathematical and Statistical Psychology* **37**(1) 62–83.

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

Duhachek, Adam, Anne T Coughlan, Dawn Iacobucci. 2005. Results on the standard error of the coefficient alpha index of reliability. *Marketing Science* **24**(2) 294–301. doi:10.1287/mksc.1040.0097.

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

Hager, William W. 1989. Updating the inverse of a matrix. *SIAM Review* **31**(2) 221–239.

Iacobucci, Dawn, Adam Duhachek. 2003. Advancing alpha: measuring reliability with confidence. *Journal of Consumer Psychology* **13**(4) 478487.

Kano, Yutaka, Maia Berkane, Peter M. Bentler. 1990. Covariance structure analysis with heterogeneous kurtosis parameters. *Biometrika* **77**(3) 575–585.

Kristof, W. 1963. The statistical theory of stepped-up reliability when a test has been divided into several equivalent parts. *Psychometrika* **28** 221–238.

Nunnally, J. C., I. H. Bernstein. 1994. *Psychometric Theory, 3rd edition*. McGraw-Hill, New York.

Ofir, Chezy, Srinivas K. Reddy. 1996. Measurement errors in probability judgments. *Management Science* **42**(9) 1308–1325.

Raykov, Tenko, Spiridon Penev. 2006. Approximate standard error and confidence interval of maximal reliability for composites with congeneric measures. *Multivariate Behavioral Research* **41**(1) 15–28.

van Zyl, J. M., H. Neudecker, D.G. Nel. 2000. On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika* **65**(3) 271–280.

Wallsten, Thomas S., David V. Budescu. 1983. Encoding subjective probabilities: A psychological and psychometric review. *Management Science* **29**(2) 151–173.

Yuan, Ke-Hai, Peter M. Bentler. 2002. On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika* **67**(2) 251–259.

A Proofs

Proof of (4). The tau-equivalent model implies that the population covariance matrix of $Y_1, \ldots, Y_k$ is given by

$$\lambda^2 \iota \iota^T + \mathbf{\Psi},$$

where ι is the k-dimensional vector with each element having value 1, and $\mathbf{\Psi}$ the diagonal matrix with diagonal elements $\psi_1, \ldots, \psi_k$.

Let θ denote the $k+1$-dimensional parameter vector $(\psi_1, \ldots, \psi_k, \lambda^2)^T$, and let $\hat{\theta}$ denote the maximum likelihood estimator of θ. It follows from the theory of covariance structures, see (Browne, 1984; Kano et al., 1990), that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d N_k\left(\mathbf{0}, \mathbf{\Omega}\right), \tag{12}$$

with

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_{uu} & \mathbf{\Omega}_{uc} \\ \mathbf{\Omega}_{cu} & \mathbf{\Omega}_{cc} \end{pmatrix}.$$

Here, the $1 \times k$ matrix $\mathbf{\Omega}_{uc}$ and the $k \times 1$ matrix $\mathbf{\Omega}_{cu}$ are given by

$$\mathbf{\Omega}_{uc} = \mathbf{\Omega}_{cu}{}^T = -\mathbf{\Omega}_{uu}\left(\sum_{\ell=1}^{k}(1/\psi_\ell)\right)^{-2}\mathbf{\Psi}^{-2}\iota,$$

the 1×1 matrix $\mathbf{\Omega}_{cc}$ is given by

$$\mathbf{\Omega}_{cc} = 2\left(\frac{1+\zeta}{\sum_{\ell=1}^{k}(1/\psi_\ell)}\right)^2 + \left(\sum_{\ell=1}^{k}(1/\psi_\ell)\right)^{-4}\iota^T\mathbf{\Psi}^{-2}\mathbf{\Omega}_{uu}\mathbf{\Psi}^{-2}\iota,$$

and the $k \times k$ matrix $\mathbf{\Omega}_{uu}$ is given by

$$\mathbf{\Omega}_{uu} = 2\mathbf{\Psi}\left(\mathbf{D} + \mathbf{v}\mathbf{v}^T\right)^{-1}\mathbf{\Psi},$$

with

$$\mathbf{D} = \left(\mathbf{I} - \frac{2\lambda^2}{1+\zeta}\mathbf{\Psi}^{-1}\right), \qquad \mathbf{v} = \frac{\lambda^2}{\zeta}\sqrt{\frac{\zeta-1}{\zeta+1}}\mathbf{\Psi}^{-1}\iota. \tag{13}$$

The Sherman-Morrison-Woodbury formula, see Hager (1989), yields

$$\left(\mathbf{D} + \mathbf{v}\mathbf{v}^T\right)^{-1} = \mathbf{D}^{-1} - \left(1 + \mathbf{v}^T\mathbf{D}^{-1}\mathbf{v}\right)^{-1}\mathbf{D}^{-1}\mathbf{v}\mathbf{v}^T\mathbf{D}^{-1}.$$

It follows that

$$\mathbf{v}^T\mathbf{D}^{-1}\mathbf{v} = \frac{\zeta-1}{\zeta^2}\sum_{\ell=1}^{k}\frac{\{\lambda^2/\psi_\ell\}^2}{1+\zeta-2\lambda^2/\psi_\ell} = Q. \tag{14}$$

Now, note that we may view ζ as a function h of $\psi_1, \ldots, \psi_k, \lambda^2$. Define the vector $\dot{h}$ as $\left(\dot{h}_{\psi}^T, \dot{h}_{\lambda^2}^T\right)^T$, where

$$\dot{h}_{\psi} = \begin{pmatrix} \frac{\partial h}{\partial \psi_1} \\ \frac{\partial h}{\partial \psi_2} \\ \vdots \\ \frac{\partial h}{\partial \psi_k} \end{pmatrix} = \begin{pmatrix} -\lambda^2 \left(\psi_1\right)^{-2} \\ -\lambda^2 \left(\psi_2\right)^{-2} \\ \vdots \\ -\lambda^2 \left(\psi_k\right)^{-2} \end{pmatrix} = -\lambda^2 \mathbf{\Psi}^{-2} \iota,$$

$$\dot{h}_{\lambda^2} = \frac{\partial h}{\partial \lambda^2} = \sum_{\ell=1}^{k} \frac{1}{\psi_\ell}.$$

The delta method yields

$$\sqrt{n}\left(\hat{\zeta} - \zeta\right) \to_d N\left(0, \dot{h}^T \mathbf{\Omega} \dot{h}\right), \tag{15}$$

with

$$\dot{h}^T \mathbf{\Omega} \dot{h} = (1+\zeta)^2 \left\{ 2 + 2\frac{\zeta+1}{\zeta-1} \frac{\mathbf{v}^T \mathbf{D}^{-1} \mathbf{v}}{1 + \mathbf{v}^T \mathbf{D}^{-1} \mathbf{v}} \right\}, \tag{16}$$

where $\mathbf{D}$ and $\mathbf{v}$ are defined in (13). The combination of (14), (15) and (16) yields (4). □

Proof of (9). We have that the values $1/\psi_1, \ldots, 1/\psi_k$ are drawn independently from the distribution of the non-negative random variable E^{-1}, which has finite variance. Hence, the limits

$$m_1 = \lim_{k\to\infty} \frac{1}{k} \sum_{\ell=1}^{k} \frac{1}{\psi_\ell}, \tag{17}$$

$$m_2 = \lim_{k\to\infty} \frac{1}{k} \sum_{\ell=1}^{k} \left\{ \frac{1}{\psi_\ell} - \frac{1}{k} \sum_{i=1}^{k} \frac{1}{\psi_i} \right\}^2 \tag{18}$$

exist, and are finite. Moreover, one may show

$$\lim_{k\to\infty} \max_{i=1,\ldots,k} \frac{1/\psi_i}{\sum_{\ell=1}^{k} 1/\psi_\ell} = 0. \tag{19}$$

It immediately follows from (17) that

$$\lim_{k\to\infty} \frac{1}{k} \zeta = m_1.$$

Moreover, it follows from (17)–(19) that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{\ell=1}^{k} \frac{\left\{\lambda^2/\psi_\ell\right\}^2}{1 + \zeta - 2\lambda^2/\psi_\ell} = m_2 + (m_1)^2.$$

Now, (9) readily follows.

□

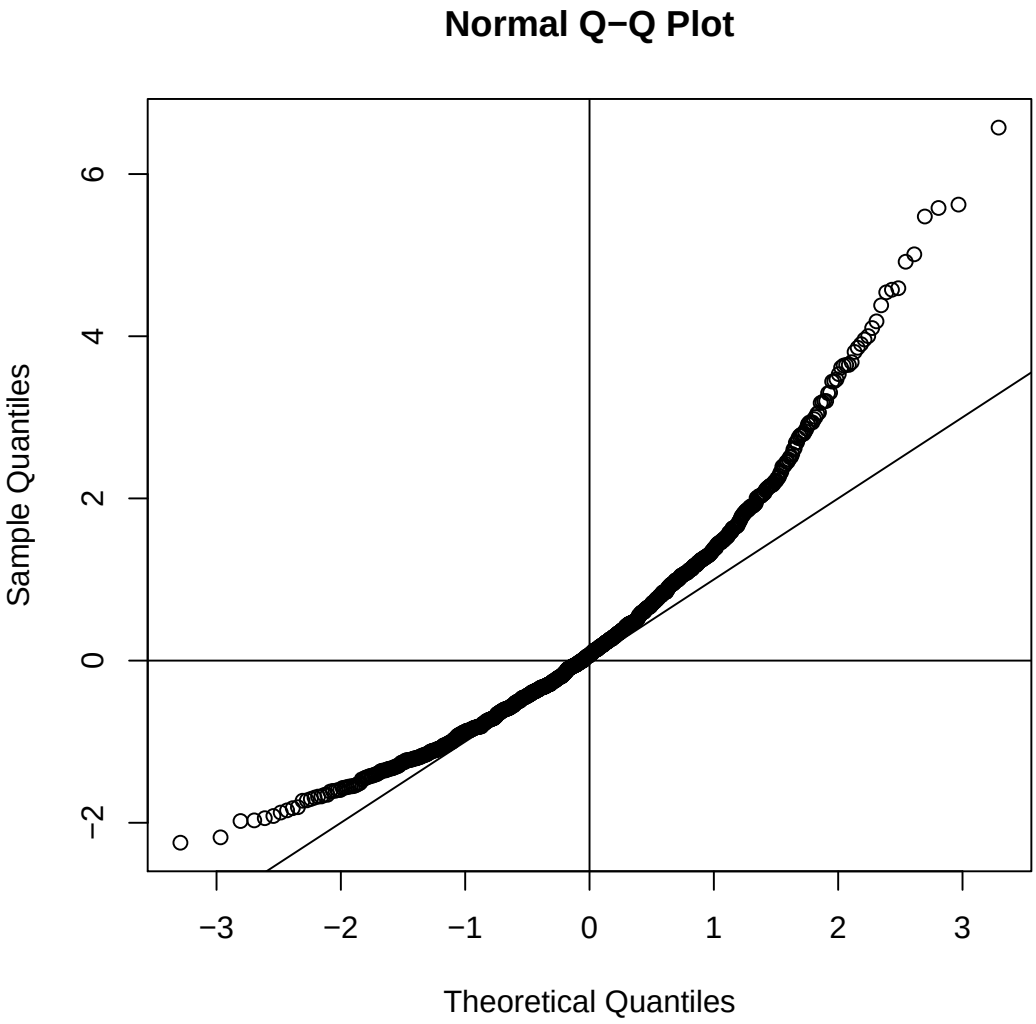

Figure 1: Normal probability plot of simulated values of the left hand side of (4) with $k = 6$, $\varphi = 0.6$, $n = 25$. The plot approaches the line with intercept 0 and slope 1 only in the center, there is a marked deviation in the tails.

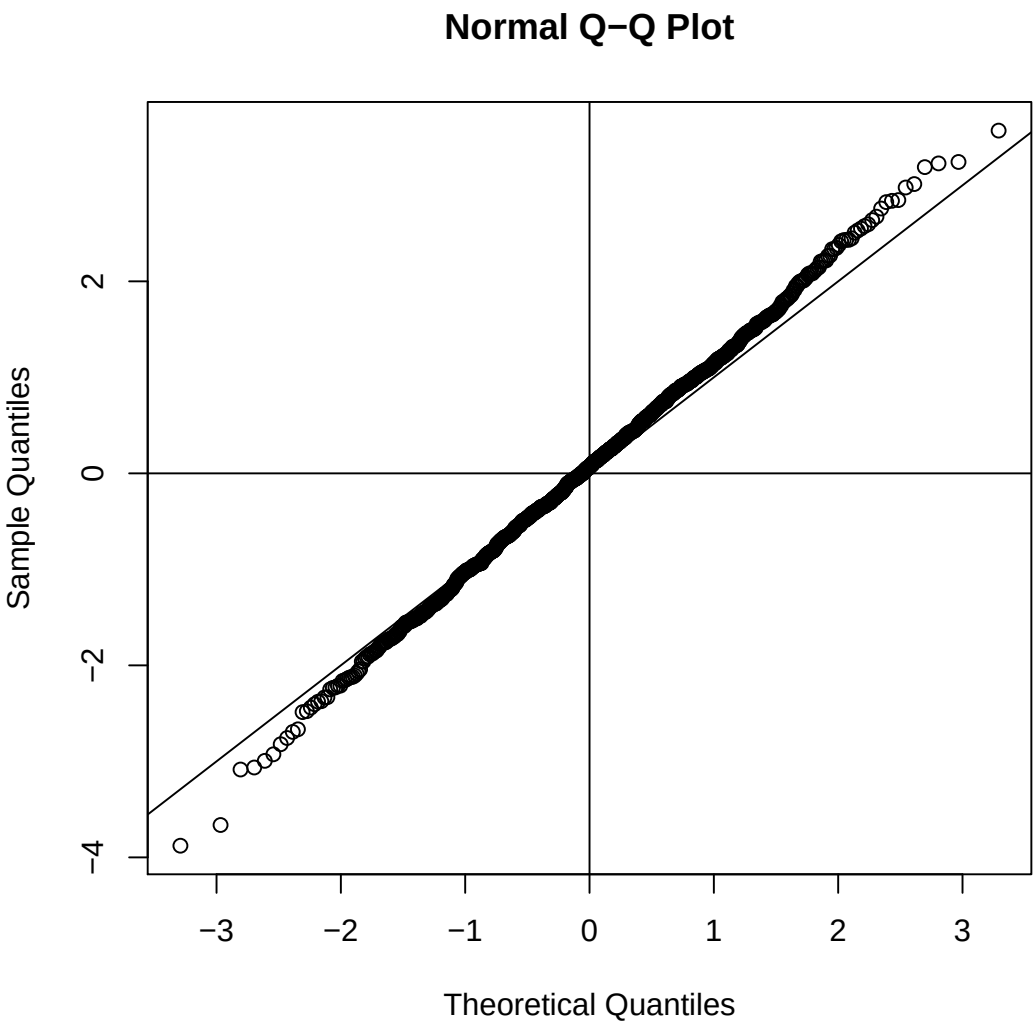

Figure 2: Normal probability plot of simulated values of the left hand side of (10), with $k = 6$, $\varphi = 0.6$, $n = 25$. The plot approaches the line with intercept 0 and slope 1, even in the tails.

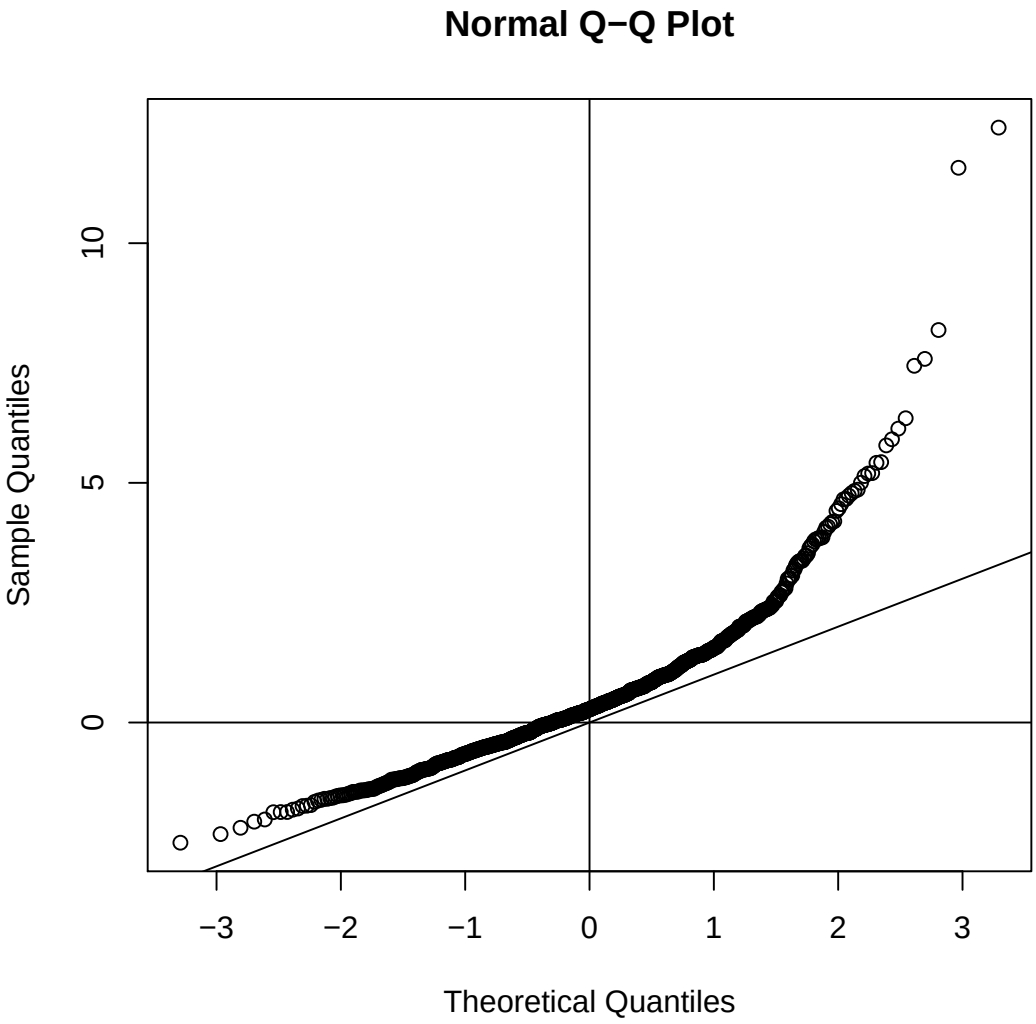

Figure 3: Normal probability plot of simulated values of the left hand side of (10), with $k = 2$, $\varphi = 0.95$, $n = 400$. Although the sample size n is large, the plot clearly deviates from the line with intercept 0 and slope 1.

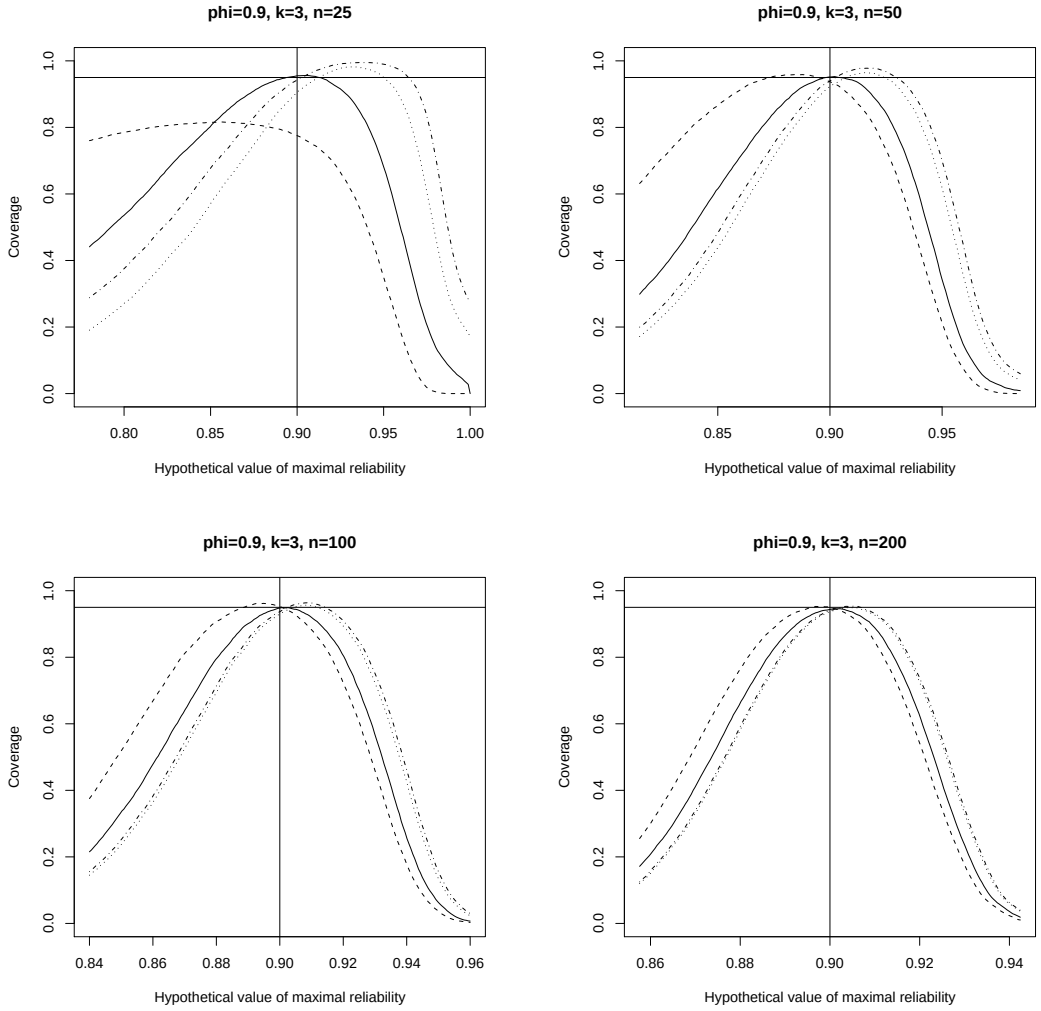

Figure 4: Coverage curves for $\varphi = 0.90$, $k = 3$ and $n = 25, 50, 100, 200$.

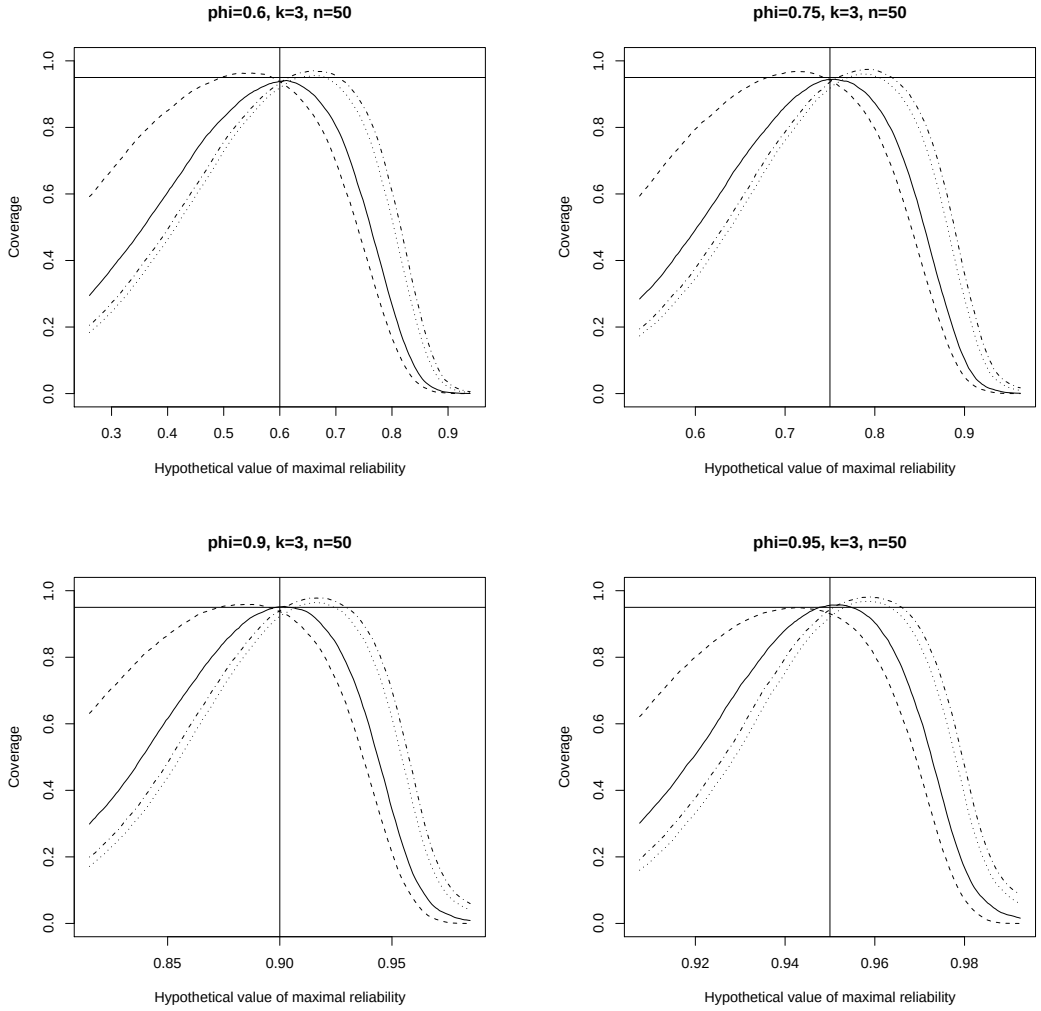

Figure 5: Coverage curves for $\varphi = 0.60, 0.75, 0.90, 0.95$, $k = 3$ and $n = 50$.

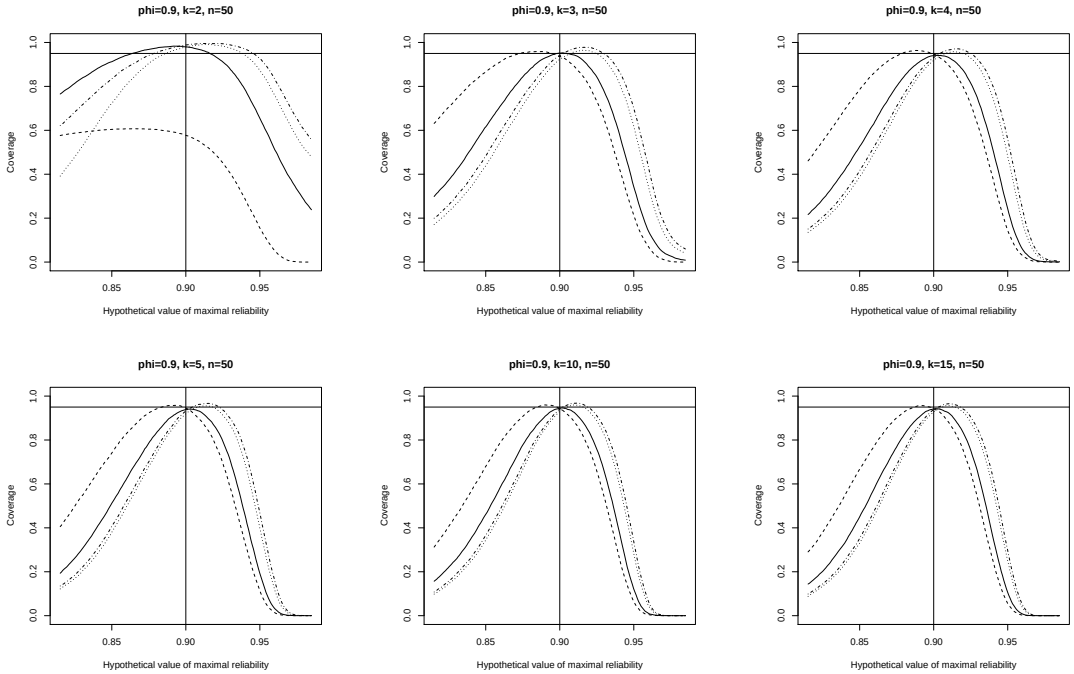

Figure 6: Coverage curves for $\varphi = 0.90$, $k = 2, 3, 4, 5, 10, 15$ and $n = 50$.