TILASTOTIETEELLISIÄ TUTKIMUKSIA
STATISTISKA UNDERSÖKNINGAR 17
STATISTICAL RESEARCH REPORTS

KIMMO VEHKALAHTI

RELIABILITY OF MEASUREMENT SCALES

Tarkkonen's general method supersedes Cronbach's alpha

SUOMEN TILASTOSEURA
STATISTISKA SAMFUNDET I FINLAND
THE FINNISH STATISTICAL SOCIETY

HELSINKI 2000

Kimmo Vehkalahti

Licentiate of Social Sciences

RELIABILITY OF MEASUREMENT SCALES

Tarkkonen's general method supersedes Cronbach's alpha

Academic dissertation to be publicly discussed, with permission
of the Faculty of Social Sciences at the University of Helsinki,
in the small hall (pieni juhlasali), University main building,
on the 25th of November, 2000, at 10 o'clock.

HELSINKI 2000

The author's address:
Department of Statistics
P.O. Box 54 (Unioninkatu 37)
FIN-00014 University of Helsinki, Finland

http://www.helsinki.fi/people/Kimmo.Vehkalahti
Kimmo.Vehkalahti@helsinki.fi

Created from cover to cover using Survo,
an environment for creative processing
of text and numerical data.
http://www.helsinki.fi/survo/

Printed version
ISBN 952-91-2818-5
ISSN 0356-3499
TUMMAVUOREN KIRJAPAINO OY
VANTAA 2000

PDF version
ISBN 952-91-2819-3
http://ethesis.helsinki.fi/
HELSINGIN YLIOPISTON VERKKOJULKAISUT
HELSINKI 2000

STATISTICAL RESEARCH REPORTS 17
Published by the Finnish Statistical Society

RELIABILITY OF MEASUREMENT SCALES

Tarkkonen's general method supersedes Cronbach's alpha

by

Kimmo Vehkalahti

Department of Statistics, University of Helsinki, Finland

HELSINKI 2000

Abstract

Kimmo Vehkalahti: Reliability of measurement scales: Tarkkonen's general method supersedes Cronbach's alpha. Dissertation, University of Helsinki, 116pp. Finnish Statistical Society, Helsinki 2000. ISBN 952-91-2818-5

Methods for assessing the reliability of measurement scales were investigated. Reliability, which is defined as the ratio of the true variance to the total variance, is an important property of measurement. In order to estimate the reliability, the concepts of measurement model and measurement scale are required. The model specifies the structure of the measurement, and the scale, which is a combination of the measured items, represents a realization of the theoretical notions.

The focus was on two measures of reliability: Cronbach's alpha, which is widely applied, and Tarkkonen's more general measure. Both measures are founded on the same definition of reliability, but they imply different assumptions about the model and the scale. Cronbach's alpha is based on the classical true score model of psychometrics, while Tarkkonen's measure belongs to a general framework of modelling the measurement.

The measures were examined theoretically and by extensive Monte Carlo simulation experiments implemented in Survo environment. Cronbach's alpha is shown to be a restricted special case of Tarkkonen's measure. According to the simulation experiments, the statistical properties of Tarkkonen's measure proved to be acceptable. Conversely, the study revealed additional evidence concerning the unsuitability of Cronbach's alpha as a measure of reliability.

The results suggest that Tarkkonen's measure of reliability should supersede Cronbach's alpha in all applications.

Keywords: reliability, measurement model, measurement scale, simulation, factor analysis.

Acknowledgements

I would like to express my deepest gratitude to Professor Seppo Mustonen. He has guided me in my studies inspiringly and teached me the principles and practice of statistical computing. Participating in his Survo project has been important to me. I have learned that a huge variety of problems can be solved efficiently and in style within the unique editorial interface of Survo. Also this research is strongly grounded on using Survo in various ways.

I would like to extend my appreciation to Docent Lauri Tarkkonen who instructed me already before my student years. Taking his advice, I went to the entrance examination of statistics. His lectures have given me a direction for research with a deeper understanding of the questions of measurement and sailing.

I wish to thank the people at the Department of Statistics, especially Professor Hannu Niemi, for giving me an opportunity to concentrate mainly on the research work. It has been a privilege for me to combine working and studying in such a fruitful way.

Professors Göte Nyman and Hannu Niemi acted as pre-examiners of my thesis. I would like to thank them for presenting critical comments and suggestions.

My sincerest appreciation goes to Marjut Schreck. Working with her in the Survo project has been for me something that is not only statistically significant.

I wish to thank Jyrki Rantalainen for teaching me the art of programming and documentation. Those skills gave excellent grounds for my university studies.

Research work is often lonely. Luckily, I have wonderful friends, who make me close my notebook for a while. Thanks to Hessu, Jari, Krisu, Lade, and Perttu for our weekly floor ball games, where the goal is simply to score a goal.

Since the early school times, I have had two special friends, Kalle and Vsa. Years and decades have not faded our friendship, although we all have specialized in totally different fields. Special friends deserve special thanks.

Completing my studies would have been impossible without my families. I am grateful to my mother Miira, who has always been a stimulating example of a researcher. I am also grateful to my father Matti for supporting my ambitions.

Especially I would like to thank my brother Herkko. His incredible persistence in his own studies has given me additional strength. I dedicate this work to him.

Finally, I wish to give my warmest thanks and hugs to my treasure, Sirpa. Her personal encouragement, love, and amazing energy have been so vital to me.

Vuosaari, October 2000 Kimmo Vehkalahti

Contents

1. Introduction

The purpose of this study is to investigate modelling of measurement, and the methods for assessing the reliability of measurement scales. The scales are combinations, such as weighted sums, of the measured items. Reliability is defined as the ratio of the true variance to the total variance.

The study focuses on two measures of reliability: Cronbach's alpha, which is widely applied with traditional scales, and Tarkkonen's more general measure for composite scales. Both measures are based on the same definition of reliability, but they imply different assumptions. The properties of the measures are examined theoretically and by Monte Carlo simulation experiments. Examples of applications are also demonstrated. The objective is to show that Tarkkonen's measure supersedes Cronbach's alpha in all applications.

1.1 Background

Measurement is an essential concept in science. The conclusions of empirical studies are based on values measured on research objects. It is therefore crucial to assess the quality of the measurements. The most important property of measurement is validity. Broadly stated, validity is concerned with whether a measuring instrument measures what it is supposed to measure in the context in which it is to be applied. But, the measurements should also be reliable, in the sense that the researchers can rely on the accuracy of the measuring instrument.

In statistical research, two sources of uncertainty require attention: the sampling variation and any empirical errors, including the measurement errors. The traditional statistical models are mainly focused on the sampling errors, and the theory of sampling is well known. However, the measurement errors are often treated with neglect, and typically included in the sample variation. This is rather vague, since the measurements themselves contain errors regardless of the sampling procedure. Sometimes there is no sampling involved at all, but the measurement errors are still there.

This vagueness has been present for a long time. It began already from a controversy involving two famous persons of the history of statistics:

Charles Spearman, a psychologist who invented factor analysis and the method of rank correlation, and *Karl Pearson*, an experienced and respected statistician who is best known of the product moment correlation method. Spearman had an idea of discriminating between the sampling error and the measurement error, but Pearson totally overlooked his ideas. Their colorful argumentation in scientific journals continued for several years – nearly a hundred years ago. In many studies today, the sampling errors and the measurement errors should be distinguished. That requires a method of modelling the measurement.

1.2 Modelling the measurement

Much of the literature on reliability originates in classical test theory from psychology, including the term *test*, which has been traditionally used to refer to psychological measures. A review of the central steps of development in reliability studies is provided in chapter 2. The basic concepts are presented here. For a more extended treatment of the subject, see the textbooks on psychometric methods, for example Gulliksen (1950), Guilford (1954), Horst (1966), Lord and Novick (1968), Guilford and Fruchter (1978), Crocker and Algina (1986), Cronbach (1990), Coolican (1994), or Nunnally and Bernstein (1994).

A fundamental equation of the classical test theory is

$$x_i = \tau_i + \varepsilon_i \,, \tag{1.1}$$

where x_i is the *i*th observed variable (or test score), τ_i is the latent *true score* that underlies x_i, and ε_i is the random measurement error defined as $x_i - \tau_i$. It is assumed that $cov\,(\tau_i, \varepsilon_i) = 0$ and $E\,(\varepsilon_i) = 0$. The model (1.1) is called the classical true score model.

The three major variations of the model (1.1) can be defined with two tests,

$$\begin{aligned} x_i &= \beta_i\tau_i + \varepsilon_i \quad \text{and} \\ x_j &= \beta_j\tau_j + \varepsilon_j \,, \end{aligned} \tag{1.2}$$

where ε_i and ε_j are uncorrelated and τ_i equals τ_j. Although different indices are used for the true scores, a single true score is assumed. Let us denote the variances of the measurement errors ε_i and ε_j by $\sigma^2_{\varepsilon_i}$ and $\sigma^2_{\varepsilon_j}$, respectively.

If $\beta_i = \beta_j = 1$ in (1.2) and $\sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon_j}$, the tests are *parallel*. (1.3)

If $\beta_i = \beta_j = 1$ in (1.2) but $\sigma^2_{\varepsilon_i} \neq \sigma^2_{\varepsilon_j}$, the tests are *tau-equivalent*. (1.4)

If $\beta_i \neq \beta_j$ in (1.2) and $\sigma^2_{\varepsilon_i} \neq \sigma^2_{\varepsilon_j}$, the tests are *congeneric*. (1.5)

The parallel model (Spearman 1904a) is the basic model of measurement in psychometrics. The tau-equivalent model (Novick and Lewis 1967) and the congeneric model (Jöreskog 1971) are slightly more general representations of it. Let us consider the parallel model (1.3).

According to definition, the reliability of test x_i is

$$\rho_{x_i x_i} = \frac{\sigma^2_{\tau_i}}{\sigma^2_{x_i}} \quad , \tag{1.6}$$

the ratio of the true score's variance to the observed variable's variance. It equals the squared correlation of the observed variable and the true score:

$$\rho^2_{x_i \tau_i} = \frac{[cov\,(x_i\,,\,\tau_i)]^2}{\sigma^2_{x_i}\,\sigma^2_{\tau_i}} = \frac{\sigma^2_{\tau_i}}{\sigma^2_{x_i}} = \rho_{x_i x_i} \quad , \tag{1.7}$$

as well as the correlation of two parallel tests:

$$\rho_{x_i x_j} = \frac{cov\,(x_i\,,\,x_j)}{\sigma_{x_i}\,\sigma_{x_j}} = \frac{cov\,[(\tau_i + \varepsilon_i)\,(\tau_j + \varepsilon_j)]}{\sigma_{x_i}\,\sigma_{x_j}} = \frac{\sigma^2_{\tau_i}}{\sigma^2_{x_i}} = \rho_{x_i x_i} \quad . \tag{1.8}$$

The results (1.7) and (1.8) follow from the assumptions of the parallel model.

Usually, tests are combined to form a composite test. A considerable number of methods have been proposed to estimate the reliability of composite or single tests. The four most common ones are known as test-retest, alternative forms, split-half, and Cronbach's alpha. They all have several drawbacks caused by unrealistic assumptions. Especially the assumption of a uni-dimensional true score is a condition, which is not flexible enough to fit the circumstances encountered in practice (Bollen 1989, 206–218).

A general framework for modelling the measurement was introduced by *Lauri Tarkkonen* in 1987. His model is a generalization of the common factor analysis model, and it brings the concept of reliability to the context of multidimensional, composite measurement scales. As its special cases, Tarkkonen's framework contains most of the models, scales and reliability measures presented in psychometric test theory, including the classical true score model and Cronbach's alpha.

1.3 Plan of the study

The aim of this study is to investigate the properties of Tarkkonen's framework, compare it with other approaches presented in the literature and demonstrate the principles of modelling the measurement. Some attention is given to structural validity, which is closely related to the modelling, but the primary object of research is reliability.

The literature review in chapter 2 gives a brief history of the studies on reliability measures and factor analysis during the 20th century. For historical reasons, the research on these issues has been concentrated on the fields of psychology and the social sciences. However, the challenge of reliable measurement is common to all of science.

Tarkkonen's general framework of modelling the measurement, including the method to assess the reliability of measurement scales, is presented in chapter 3. A new matrix form of the reliability measure is derived.

The general framework is compared to other approaches, like the factor analysis model, in chapter 4. Cronbach's alpha is a special case of Tarkkonen's reliability measure. More accurate proof of this result is established. A new result shows that alpha can not exceed Tarkkonen's measure.

Examples of applications are demonstrated in chapter 5.

The principles of Monte Carlo simulation of the measurement model as well as the design of the experiments in this study are described in chapter 6.

The analyses and results of the simulation experiments are presented in chapter 7.

The conclusions of the study are discussed in chapter 8.

> It has been suggested to make the measurements so efficient, that the correction will not be needed. But how are we to tell whether our measurements really are efficient enough, expect by trying with the correction formula? The suggestion is like telling a man to brush his coat until it is clean but never look whether it is so.
>
> – Charles Spearman, 1910.

2. Milestones of the psychometric literature

2.1 Spearman-Brown formula

The starting point of estimating the reliability of measurement is due to *Charles Spearman* at the turn of the 20th century. The concept of correlation was known, based on the work of *Bravais*, *Galton* and *Pearson* during the 19th century, but Spearman was the first to consider various hidden underlying causes affecting the true correlation. His idea was that the "accidental" errors of measurement could be estimated by the size of the discrepancies between successive measurements of the same thing. He proposed a formula for correction for attenuation in finding the true relation between two variables (Spearman 1904a).

These ideas gave theoretical grounds for measuring the general intelligence or ability, and formed the basis for factor analysis (Spearman 1904b). Spearman's Theory of Two Factors corresponds to the one-factor case in the modern terminology. For a comprehensive discussion of Spearman's work concerning the origin and development of factor analysis, see Bartholomew (1995).

In his paper *Correlation calculated from faulty data*, Spearman gave the expression *reliability coefficient*, meaning "the coefficient between one half and the other half of several measurements of the same thing" (Spearman 1910, 281). He developed an enhanced form of the correction formula, and a number of specific cases of it. The most important case concerns the reliability of an average of p dichotomous items (variables), which are assumed to measure the same thing equally well. If the composite is denoted by u, the formula is

$$\rho_{uu} = \frac{p\,\rho_{xx}}{1 + (p-1)\,\rho_{xx}}, \tag{2.1}$$

where ρ_{xx} is the reliability of a single item (see chapter 1.2). As (2.1) shows, the reliability of the composite gets higher, if the number of the items is increased.

It is notable that the formulas were derived with explicit references to the true scores and the measurement errors (although the terms used at that time were different). Also the assumptions were clearly stated, and the model was properly formulated. In this respect, Spearman was probably influenced by the style of *George Udny Yule*'s proof of the correction formula (Spearman 1910, 272).

Spearman (1910) also suggested a practical application of the formula (2.1), known as the split-half method. The procedure was to divide the measurements into two groups, typically by selecting the odd and the even numbered items. The idea was to calculate the correlation between the groups and then apply the correction formula. Later, the method became widely applied and studied (Brownell 1933; Rulon 1939; Mosier 1941; Cronbach 1946, 1947; Lyerly 1958).

The formula (2.1) is called *Spearman-Brown formula*, as also *William Brown* derived it – in a footnote to a table (Brown 1910, 299). Brown questioned the assumptions of the formula. He got support from *Karl Pearson*, who had heavily criticized Spearman's methods from the beginning. Many others followed them, for example *Truman Kelley* (1921, 1924, 1925) and *Karl Holzinger* (1923). Despite the critique, the Spearman-Brown formula was used in the psychological, educational and sociological research for decades. No real alternative appeared to exist, as the following quotation indicates:

> I know of no better simple way of securing an estimate of reliability of a college entrance test than to split it into halves and use the Spearman-Brown formula and though there are hazards in doing this I certainly think that such an estimate is very much better than none at all (Kelley 1924, 200).

A clear problem in the early measures of reliability was caused by the concept of item reliabilities. In theory, they were assumed to be known. In practice, there was no unique way for determining them, which caused serious limits for any general solutions, even decades later. On the other hand, simplicity was constantly an important issue, because the calculations were done mostly by hand. For the same reason, the test items were usually dichotomous.

Spearman's concern of the measurement errors established the need for a measure of reliability especially in psychology and the social sciences. For a review of Spearman's contributions to the psychometric test theory, see Levy (1995).

2.2 Multiple-factor analysis

By the 1930s it had become quite apparent that "Spearman's Two-Factor Theory was not always adequate to describe a battery of psychological tests" (Harman 1967, 4). Researchers became more conscious of the multidimensionality of the tests. It was noticed that the internal consistency of a test, inherited from Spearman's theory and implied by the method known as item analysis, was not a sufficient criterion for creating reliable tests (Handy and Lentz 1934, Zubin 1934, Richardson 1936, Mosier 1936).

The possibility of extracting several factors directly from a matrix of correlations was explored, and thus arose the concept of multiple-factor analysis. The actual term is generally attributed to *Louis Leon Thurstone* (1931), although the idea was already present in the work of *Maxwell Garnett* (1919). Thurstone, in his book *Vectors of the Mind* (1935), emphasized the multiple factorial nature of human abilities, thus making a clear break with the Spearman tradition. Thurstone adopted the modern terminology by referring to Spearman's theory as a single-factor rather than a two-factor model. Thurstone's concentration on many factors gave rotation in the factor space a central place and led him to formulate the notion of *simple structure* as a rotation criterion (Harman 1967, 4; Bartholomew 1995, 214).

Harold Hotelling (1933) presented the method of principal components, which had its origins in the early work of Pearson (1901). Hotelling had a critical view of factor analysis. According to Thurstone, Hotelling dismissed the fundamental concepts of multiple-factor analysis because they had not been formulated in terms of current statistical theory (Bartholomew 1995, 215).

In a sense, the psychometric research was split into two paths. The modern path followed Thurstone, while the other one stuck on the Spearman-Brown tradition.

Godfrey Thomson, one of the leading figures in British psychology, was engaged in a debate with Spearman for nearly 20 years. The debate began when Thomson discredited Spearman's claims of uniqueness of his method. Later, Thomson did much to explore the relationships between the different methods and tried hard to bring factor analysis and principal components together (Sharp 1997, 165–166, 172; Bartholomew 1995, 215). Thomson's book *The Factorial Analysis of Human Ability* (1939) is considered as the most remarkable prewar writing on factor analysis. A chapter of maximum likelihood estimation of the factor loadings, written by Thomson's Edinburgh colleague, *D. N. Lawley*, was added in later editions of the book. This was to form important links with the postwar era (Bartholomew 1995, 215).

Factor analysis was born before its time, and it had to mark time until the technology caught up (Bartholomew 1995, 216–217). Meanwhile, a great variety of *ad hoc* methods were implemented.

2.3 Kuder-Richardson formula 20

A collection of new reliability measures was introduced by *G. F. Kuder* and *M. W. Richardson* (1937). The aim was to get rid of the difficulties caused by the usage of the Spearman-Brown formula and the split-half method. The authors offered several choices for different situations, involving various assumptions and approximations. The central assumption was that the matrix of the inter-item correlations is of rank 1. This corresponds to the situation that all items are measuring the same factor.

Of the several formulas by Kuder and Richardson, one was preferred. On the basis of its number in the original article, this particular formula is referred to as *Kuder-Richardson formula* 20. Due to its importance, the formula deserves special attention. Thus we derive it here, following the lines of Kuder and Richardson (1937, 152–158).

The measurement model is not explicitly specified, but all items are supposed to measure the same thing. Implicitly this leads to the classical true score model (1.1). The items x_i are measured on dichotomous scale. Let then $u = x_1 + ... + x_p$ be a scale made up of p unit-weighted items, and let $v = y_1 + ... + y_p$ be a corresponding hypothetical scale. Two scales are needed, since the estimation of reliability is based on the correlation of two equivalent scales. Now, the correlation of u and v is given by

$$\rho_{uv} = \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} \rho_{x_i y_j} \sigma_{x_i} \sigma_{y_j}}{\sqrt{\sum_{i=1}^{p} \sigma_{x_i}^2 + 2 \sum_{i<j} \sum \rho_{x_i x_j} \sigma_{x_i} \sigma_{x_j}} \sqrt{\sum_{i=1}^{p} \sigma_{y_i}^2 + 2 \sum_{i<j} \sum \rho_{y_i y_j} \sigma_{y_i} \sigma_{y_j}}} \,. \tag{2.2}$$

If equivalence is defined as interchangeability of items x_i and y_i , for $i = 1 , ..., p$, then the two expressions in the denominator of (2.2) are identical. Hence, according to the definition of reliability (*cf*. 1.8), the correlation between the two equivalent scales u and v gives the reliability of the scale u in the form

$$\rho_{uu} = \frac{\sum_{i=1}^{p} \rho_{x_i x_i} \sigma_{x_i}^2 + 2 \sum_{i<j} \sum \rho_{x_i x_j} \sigma_{x_i} \sigma_{x_j}}{\sum_{i=1}^{p} \sigma_{x_i}^2 + 2 \sum_{i<j} \sum \rho_{x_i x_j} \sigma_{x_i} \sigma_{x_j}} \,. \tag{2.3}$$

Assume that the inter-item correlation matrix is of rank 1, which implies that the correlations $\rho_{x_i x_j}$ are equal. The common correlation is denoted by the symbol ρ_{xx}, standing also for the unknown item reliabilities. The formula (2.3) now becomes

$$\rho_{uu} = \frac{\rho_{xx} \left(\sum_{i=1}^{p} \sigma_{x_i} \right)^2}{\sigma_u^2} \,, \tag{2.4}$$

where

$$\sigma_u^2 = \rho_{xx} \left(\sum_{i=1}^{p} \sigma_{x_i} \right)^2 - \rho_{xx} \sum_{i=1}^{p} \sigma_{x_i}^2 + \sum_{i=1}^{p} \sigma_{x_i}^2 \,. \tag{2.5}$$

We need to get rid of the item reliabilities ρ_{xx}, since they are not operationally determinable except by use of certain assumptions or crude approximations. As Kuder and Richardson (1937, 157) put it: *we are in addition willing to assume equal standard deviations of items.*

Thus, let us assume that $\sigma_{x_i} = \sigma_x$, for $i = 1 , ..., p$. Then, (2.5) becomes

$$\sigma_u^2 = \rho_{xx}\, p^2\sigma_x{}^2 - \rho_{xx}\, p\sigma_x{}^2 + p\sigma_x{}^2 \ , \tag{2.6}$$

and we can solve ρ_{xx} from (2.6) as

$$\rho_{xx} = \frac{\sigma_u^2 - p\sigma_x{}^2}{(p-1)\, p\sigma_x{}^2} \ . \tag{2.7}$$

Hence, from (2.4),

$$\rho_{uu} = \frac{\rho_{xx}\, p^2\sigma_x{}^2}{\sigma_u^2} \ , \tag{2.8}$$

and substituting (2.7) in (2.8) yields

$$\rho_{uu} = \frac{(\sigma_u^2 - p\sigma_x{}^2)\, p^2\sigma_x{}^2}{(p-1)\, p\sigma_x{}^2\, \sigma_u^2}$$

$$= \frac{p}{p-1}\left(1 - \frac{p\sigma_x{}^2}{\sigma_u^2}\right) . \tag{2.9}$$

The formula (2.9) is known as Kuder-Richardson formula 20, or KR-20 for short. It has a very simple form, which was clearly one of the aims in developing it. This is reflected in the following quotation:

> Any one of the formulas will give a unique estimate of the coefficient in all situations to which it is applicable. In certain cases, the commonly calculated parameters of the test score distribution will afford, in two minutes of time, a fairly good estimate of the reliability coefficient (Kuder and Richardson 1937, 153).

During the derivation, the correlations and the standard deviations were assumed to be equal. If the assumptions are not met, the figures obtained are claimed to be underestimates (Kuder and Richardson 1937, 159). The assumptions are identical to those of the Spearman-Brown formula. The crucial difference of the formulas is that the formula (2.9) hides the item reliabilities, which were present in (2.1).

If we write (2.6) in the form

$$\sigma_u^2 = p\sigma_x^{\,2}\,[\,1 + (p-1)\,\rho_{xx}\,]\ , \tag{2.10}$$

and substitute it into (2.8), we end up with

$$\rho_{uu} = \frac{\rho_{xx}\,p^2\sigma_x^{\,2}}{p\sigma_x^{\,2}\,[\,1 + (p-1)\,\rho_{xx}\,]} = \frac{p\,\rho_{xx}}{1 + (p-1)\,\rho_{xx}}\ ,$$

which equals the Spearman-Brown formula (2.1).

Kelley, who had criticized the Spearman-Brown formula in the 1920s, did not praise the work of Kuder and Richardson, either. He reminded that their assumptions are generally too restrictive. In particular he questioned the assumption that the correlation matrix is of rank 1 (Kelley 1942, 80).

Despite its restrictions, the Kuder-Richardson twenty, as it was also called, soon became a classic. It was actively used by the researchers but also criticized from the beginning *e.g.* by Dressel (1940) and later by Tucker (1949). Alternative forms of the formula were suggested *e.g.* by Horst (1953), but the most famous variation, known as *alpha*, was presented by Cronbach (1951). It will be reviewed in detail in chapter 2.5.

2.4 Lower bounds and maximum reliability

Forty years after Spearman had written about errors of measurement, *Louis Guttman* (1945) tried to unify the concept of reliability, stressing the reliability of a sum of a number of variables, which had been the original idea of Spearman. Guttman derived six different lower bounds to the reliability, for different situations. Common to all of them was that the estimate could be drawn from single trial. Since the need of two independent trials was avoided, the computations were easy.

Let u be an unweighted sum of the variables x_i, $i = 1, ..., p$. Of the lower bounds that Guttman suggested, the third one was termed *an intermediate lower bound*. It is given by

$$\lambda_3 = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^{p}\sigma_{x_i}^2}{\sigma_u^2}\right), \tag{2.11}$$

with $\lambda_3 \le \rho_{uu} \le 1$, where the equality on the left holds if and only if the variances and covariances are all equal (Guttman 1945, 274). The formula resembles the Kuder-Richardson formula 20 (2.9), which is a pure coincidence according to Guttman. Indeed, the derivations differ significantly, but the basic assumptions are the same. As the inequality above says, the true value of the reliability is underestimated, if the assumptions are not met. This is the same conclusion that Kuder and Richardson made, nearly ten years earlier.

Charles Mosier (1943) developed a reliability measure for a composite variable $u = \mathbf{a}'\mathbf{x} = a_1 x_1 + \ldots + a_p x_p$, where a_i may be the weights obtained from a least-squares regression equation, or integers assigned on some arbitrary basis "to increase the validity with which the composite will predict some vaguely defined, unmeasured criterion" (Mosier 1943, 162). He derived the error variance and total variance of u , and expressed the reliability of u as a function of their ratio subtracted from one. Reducting to standard scores of x_i , the result was further simplified and finally presented in matrix form as

$$\rho_{uu} = \frac{\mathbf{a}'\mathbf{R}^*\mathbf{a}}{\mathbf{a}'\mathbf{R}\,\mathbf{a}} , \tag{2.12}$$

where $\mathbf{R}^*$ is the correlation matrix, whose diagonal terms are the item reliabilities, and $\mathbf{R}$ is the ordinary correlation matrix (Mosier 1943, 163–165). This formula occurs frequently in the literature, with small variations. It is specially used in the task of maximizing the reliability, which had already been considered by Thomson (1940). If no external criterion for weighting variables was available, a good choice was the one which had maximum reliability. Peel (1947) showed that the weights for such a composite are given by

$$(\mathbf{R}^* - \lambda\mathbf{R})\mathbf{a} = 0 , \tag{2.13}$$

where λ is the desired maximum reliability, corresponding to the largest root of the determinantal equation

$$|\mathbf{R}^* - \lambda\mathbf{R}| = 0 . \tag{2.14}$$

Green (1950) gave a transformation, which allowed the use of Hotelling’s (1933) or as well Lawley’s (1940) iterative methods in solving the equation (2.14).

Mosier’s formula (2.12) was useful in the task of maximizing the reliability, but it had two additional merits, as well. Firstly, unlike the traditional measures, it was defined for a weighted sum. Secondly, it was one of the first attempts to bring matrix notation into the discussion of reliability measures. The only drawback were the item reliabilities, which had been a stumbling block for a long time.

Robert Wherry and *Richard Gaylord* (1943) reviewed a large number of reliability coefficients. They noted that the internal consistency hypothesis is the basis of the two most common methods of measuring reliability: the split-half Spearman-Brown approach and the Kuder-Richardson formulas, and that in both of them, a single factor is assumed among the items. They stated that the Kuder-Richardson formula tends to underestimate the true reliability by a certain ratio, when the number of factors is greater than one. Their conclusion was that "the unsatisfactory conditions result from the blind assumption of a single factor" (Wherry and Gaylord 1943, 250–260).

At that time, multiple-factor analysis was already known, but not widely applied because of the large number of calculations required. Nevertheless, some researchers thought that the determination of the reliability coefficients resulting from a factor analysis "justified the expenditure of considerable labor" (Davis 1945, 58–60). The computation techniques had already improved since the beginning of the century, for example sums of squares and cross products could be computed using devices operating with punch cards (see *e.g.* Benjamin 1945).

2.5 Cronbach’s alpha

Lee Joseph Cronbach criticized the Kuder-Richardson methods as well as the split-half approach. He had repeatedly encountered the difficulty that the magnitude of the underestimate was unknown, and on more than one occasion, the Kuder-Richardson estimate had been a sizeable negative value:

> The Kuder-Richardson formula is not desirable as an all-purpose substitute for the usual techniques. This reopens the question whether a suitable estimate can be developed (Cronbach 1943, 488).

The formulas had been in use for some years, especially number 20 (2.9) and its simplified counterpart, number 21. The principal advantages claimed for those formulas were ease of calculation, uniqueness of estimate (compared to split-half methods), and conservatism. Cronbach thought that "while conservatism has advantages in research, in this case it leads to difficulties" (Cronbach 1943, 487).

In 1951, Cronbach came up with the symbol α for the first time. He even used the term "Kuder-Richardson formula α", and noted that according to another, forecoming article on the subject, "α is the mean of all possible split-half coefficients" (Cronbach and Warrington 1951, 179).

The particular article was published in the subsequent number of *Psychometrika*. The names Kuder and Richardson were dropped away. Since then, the formula has been referred to as *Cronbach’s alpha*. The article, *Coefficient alpha and the internal structure of tests*, has become probably the most referred paper in the psychometric literature.

In that paper, Cronbach took the formula of alpha as given, but forgot the assumptions. Referring to earlier derivations by Kuder and Richardson (1937), Hoyt (1941) and Guttman (1945), he claims that *making the same assumptions but imposing no limit on the scoring pattern, will permit one to derive the formula in the form*

$$\alpha = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^{p} \sigma_{x_i}^2}{\sigma_u^2}\right), \tag{2.15}$$

where p is the number of items x_i , and $u = x_1 + x_2 + ... + x_p$ (Cronbach 1951, 299).

It is true that the original assumption of dichotomous variables can be extended to more general scales. But, the rigid assumptions of equal variances and equal correlations of the items are hidden. Instead of p equal variances, Cronbach uses a sum of the observed variances. In that sense, alpha is algebraically identical to the Guttman's formula given by (2.11). But, Guttman derived it as a lower bound only, reminding that the equality holds only if the variances and covariances are all equal. Cronbach does not make explicit assumptions. Instead, he writes:

> Since each writer offering a derivation used his own set of assumptions [...], the precise meaning of the formula became obscured. The original derivation unquestionably made much more stringent assumptions than necessary [...]. In this paper, we take formula [of alpha] as given, and make no assumptions regarding it. Instead, we proceed in the opposite direction, examining the properties of alpha and thereby arriving at an interpretation (Cronbach 1951, 299).

Proceeding in the opposite direction has continued in the psychometric literature for decades. Every now and then there is a new article on the properties of Cronbach's alpha. The formula is applied in various application fields, and lots of interpretations has been made. A large number of conclusions are based on the values of alpha. However, it is nearly impossible to find a situation that conforms to the assumptions of alpha. The obtained values are thus underestimates, a fact that was stated already by the original authors of the formula in the 1930s.

The value of alpha gets even negative, if the sum of all item covariances is negative. It was recommended that formulas of the Kuder-Richardson type should not be used in those circumstances (Cronbach and Hartmann 1954, 345).

Although the drawbacks of Cronbach's alpha should have been known for a long time, it is famous and widely used. One of the reasons why it became so popular, was probably the fact that it is easy to calculate. In the 1950s, that was still an important issue.

2.6 Exploring the limits of alpha

It was soon noticed that the unweighted sum is not general enough for practical needs. Hence, alpha was developed for a weighted scale as well, even though the original assumptions were violated. Let $\mathbf{a} = (a_1, a_2, ..., a_p)$ be a vector of weights, and let $\mathbf{\Sigma}$ be the covariance matrix of the items x_i . *Frederic Lord* (1958) showed that maximizing alpha of the form

$$\alpha = \frac{p}{p-1}\left(1 - \frac{\mathbf{a}'\,diag\,(\mathbf{\Sigma})\,\mathbf{a}}{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}\right) \tag{2.16}$$

is equivalent to maximizing the quadratic form $\mathbf{a}'\,diag\,(\mathbf{\Sigma})\,\mathbf{a}$, when another quadratic form, $\mathbf{a}'\mathbf{\Sigma}\mathbf{a}$ is held constant. The maximum is obtained by weighting the standardized items by the loadings on their first principal component. The result is seen from the characteristic equation of the corresponding correlation matrix.

In 1962, Cronbach and *Hiroshi Azuma* noted that the interpretation of the internal-consistency reliability coefficients had become increasingly confused, due to conflicting assumptions and alternative derivations. They compared several formulas and evaluated how well they served various purposes. They found alpha "satisfactory in most situations", and "even highly satisfactory if the test was a random sample from a pool of dichotomous items whose content represented a single factor, and whose mean intercorrelations were within a normal range". This range was not specified accurately, but typical correlations seemed to be quite low, 0.3 and below (Cronbach and Azuma 1962).

In the following year Cronbach, together with *Nageswari Rajaratnam* and *Goldine Gleser*, was willing to change the course, by giving a completely new definition for alpha. One of the aims was to discard the restrictive assumptions of the classical models and measures. Cronbach thus admitted that alpha is too limited in its original form. He trusted that "obscurities and inconsistencies in the choice of formulas would be eliminated by the new development" (Cronbach, Rajaratnam, and Gleser 1963, 154–155).

The new formulation, called the theory of generalizability, was based on additive analysis of variance models and the intraclass correlation. The concepts were somewhat more general than before, although not completely new, since the reliability issues based on the analysis of variance had been considered already earlier (Hoyt 1941, Burt 1955). However, Cronbach *et al.* (1963) did not generalize the measurement model, and thus the reliability coefficients derived were essentially the same as before. Because of its complexity and the lack of procedures for estimating many of its parameters, "generalizability theory has not been brought to practical status" (Weiss and Davison 1981, 634), and "a cautious approach to its use still appears warranted" (Jones and Applebaum 1989, 31).

A fatal step in the development was taken by *Henry Kaiser* and *John Caffrey* (1965), who tried to combine factor analysis and Cronbach's alpha, by developing a new method of factor analysis. Their suggestion, *alpha factor analysis*, is analogous to Rao's (1955) canonical factor analysis. However, the weighting of the items differs dramatically: instead of using as weights the inverses of the unique variances, Kaiser and Caffrey (1965) used inverses of the communalities, thus giving more weight for items with lower communality. This surely was not the goal − the goal was to develop a psychometric factor analysis − but the contradictory results could not be avoided, since the idea was based on maximizing the generalizability of the factors, that is, Cronbach's alpha.

Following Lord's (1958) treatment, Kaiser and Caffrey (1965) ended up to

$$\alpha = \frac{p}{p-1}\left(1 - \frac{1}{\lambda_i}\right), \tag{2.17}$$

where λ_i are the eigenvalues of the weighted correlation matrix of the items. The maximization procedure of alpha leads to the principal components of the standardized items, as Lord (1958) had earlier shown.

It is easy to see from (2.17) that α tends negative as soon as λ_i drops below one. Kaiser and Caffrey formulated this as a clear rule: only those alpha factors which have positive generalizability, *i.e.* the associated eigenvalues greater than one, should be accepted (Kaiser and Caffrey 1965, 11). Kaiser (1960) had earlier suggested the same procedure in the context of principal components. In the special case of one alpha factor, Kaiser and Caffrey (1965, 8) reasoned that it is "always perfectly generalizable".

P. M. Bentler (1968) criticized alpha factor analysis, and suggested instead alpha-maximized factor analysis *(alphamax)*, which is identical to Rao's (1955) canonical factor analysis. Both methods carrying the name of alpha are based on maximizing Cronbach's alpha. The difference is that in alphamax the variables are weighted by the inverses of the unique variances, not by the inverses of the communalities. Bentler (1968) showed that the maximization of alpha in the traditional form leads to principal components analysis, while maximizing the weighted alpha leads to factor analysis.

Melvin Novick and *Charles Lewis* (1967) specified the necessary and sufficient condition under which Cronbach's alpha is equal to the reliability of a composite measurement. The condition was named essential tau-equivalence, and it gave the exact form for the assumptions of parallel tests. Novick and Lewis derived alpha, and showed that it is a lower bound to the reliability. They also noted that "the lower bound will be a very bad one, except for cases in which tests are relatively homogeneous or long" (Novick and Lewis 1967).

In the 1970s, the interest of conceptualization and measurement of reliability increased especially in the social sciences. Factor analysis was extensively applied, and proved to be a proper statistical method, as indicated by Lawley and Maxwell (1971). The maximum likelihood estimation of factor loadings, developed 30 years earlier by Lawley (1940), came finally fully available, when *Karl Jöreskog* (1967) provided a better algorithm for the iteration procedure. The computers had an important role in the development. Several methods had been inapplicable because of the enormous amount of calculations required. Excerpts from the history of using computers in psychometrics can be found from Benjamin (1945) or Lefkowitz and Greene (1962), for instance.

David Heise and *George Bohrnstedt* (1970) suggested a reliability measure for composite variables in the context of factor analysis. They worked on the basis of the sample correlation matrix **R**. The basic equation of factor analysis is then

$$\mathbf{R} = \mathbf{F}\mathbf{F}' + \mathbf{U}^2 , \qquad (2.18)$$

where **F** is the factor matrix and $\mathbf{U}^2$ is a diagonal matrix of the unique variances.

The factors are assumed to be orthogonal. The reliability measure for a scale weighted by a vector $\mathbf{a} = (a_1, a_2, ..., a_p)$ is given by

$$\Omega = \frac{\mathbf{a}'(\mathbf{R} - \mathbf{U}^2)\,\mathbf{a}}{\mathbf{a}'\mathbf{R}\,\mathbf{a}} \quad , \tag{2.19}$$

where $\mathbf{R} - \mathbf{U}^2$ is a correlation matrix with communalities in the diagonal. In the case when the communalities are known, omega was claimed to be exactly equal to the reliability of a composite (Heise and Bohrnstedt 1970, 117). *Harry Harman* (1967), among others, has discussed the concepts of reliability, communality and specific variance in factor analysis. If the communalities can be estimated by the item reliabilities, then omega (2.19) is equal to Mosier's formula (2.12), which was presented nearly 30 years earlier.

David Armor (1974) criticized the assumptions of Cronbach's alpha and the methods of item analysis:

> The mathematical assumptions for alpha reliability are often not met; the usual steps of item analysis – throwing out "bad" items to enhance alpha reliability – may not in fact produce optimal alpha reliability (Armor 1974, 18).

Armor worked with principal components and came up with

$$\theta = \frac{p}{p-1}\left(1 - \frac{1}{\lambda_1}\right) , \tag{2.20}$$

where λ_1 is the largest eigenvalue of the sample correlation matrix (Armor 1974, 28). It may be noted that the formula (2.20) is identical to Kaiser and Caffrey's (1965) alpha in (2.17), but this fact seemed to be ignored by Armor. Although Armor referred to Bentler (1968), it seems that he did not catch Bentler's critique concerning the use of principal components.

The relations of alpha and omega were studied by *Kent Smith* (1974). His conclusion was that "if the research design limits one to internal-consistency estimates of reliability, then omega is clearly the choice". He thought that Heise and Bohrnstedt chose their symbol with *a bit of frivolity* (Smith 1974, 507).

Vernon Greene and *Edward Carmines* (1979) summarized the variations of Cronbach's alpha, concluding that Armor's theta (2.20) is equal to a maximized alpha, and the alpha related to Bentler's (1968) alpha-maximized factor analysis is equal to a maximized omega. Greene and Carmines stressed that the condition of essential tau-equivalence stated by Novick and Lewis (1967) is neither necessary nor sufficient condition for alpha to be equal to the true reliability, except in the case of equal weights (Greene and Carmines 1979).

2.7 The greatest lower bound

The statistical properties of Cronbach's alpha have been studied by several authors, beginning from Lord (1955). The effect of the test length, *i.e.* the number of the items in a scale, to the reliability had been studied earlier by Angoff (1953). Kristof (1963) went further by dividing the scale into several homogeneous parts to better conform to the strict assumptions. In a way, the idea resembled the ancient split-half methods. The effect of dividing the scale was studied among others by Cronbach, Schönemann, and McKie (1965), Kristof (1972, 1974), Feldt (1975), and Raju (1977, 1979).

The sampling distribution, statistical significance, interval estimation and group differences of alpha were studied by Feldt (1965), Payne and Anderson (1968), Pandey and Hubert (1975), Sedere and Feldt (1977), and Callender and Osburn (1979). Raju (1982) studied the test homogeneity and maximum alpha, although he referred to the formula as KR-20. The effect of sampling model on inference with alpha was recently studied by Barchard and Hakstian (1997).

Since the assumptions were generally never met, the values of the reliability measures were claimed to be lower bounds only. Guttman's (1945) work with various different lower bounds to reliability gave a basis for a new research line in the beginning of the 1970s, namely seeking the greatest lower bound to reliability. Series of papers on this subject were written by Bentler (1972), Jackson and Agunwamba (1977), Woodhouse and Jackson (1977), Woodward and Bentler (1978, 1979), Bentler and Woodward (1980, 1983, 1985), ten Berge, Snijders, and Zegers (1981), Gilmer and Feldt (1983), and Shapiro (1985). But, the improper definition of the concepts lead to inconsistent results, which finally forced Cronbach to give a general comment on internal consistency of tests:

> This paper originated in a paradox. Bentler and Woodward (1980, 1983) reasoned that a certain internal-consistency analysis promises "the greatest lower bound to reliability". Their illustrative coefficients, however, were strangely low. [...] In these instances, incorrect choice of unit of analysis undercut an otherwise brilliant technical development (Cronbach 1988, 63).

Common to these paths of research have been that usually factor analysis is not even mentioned in the papers. The research is concentrated around the reliability coefficients themselves, without a clear comprehension of how to use them and for what purpose. For example, *Michelle Liou* (1989) claims that "the maximum likelihood approach has the unfortunate disadvantage that it is not feasible without using a computer" (Liou 1989, 153–154). She proposes new formulations for techniques presented by Gilmer and Feldt (1983), with "a significant advantage that the corresponding reliability coefficients can be attained with a hand calculator" (Liou 1989, 154). This resembles the goals and developments of the 1920s and 1930s. The following quotation is telling:

> Somewhere during the three-quarter century history of classical test theory the real purpose of reliability estimation seems to have been lost (Weiss and Davison 1981, 633).

2.8 Analysis of covariance structures

Karl Jöreskog had contributed significantly to the estimation methods of the maximum likelihood factor analysis (Jöreskog 1967, Jöreskog and Lawley 1968), and extended his ideas to the confirmatory factor analysis (Jöreskog 1969). He also proposed a general method for the analysis of covariance structures, with applications such as factor analysis, variance component models and linear structural relationships (Jöreskog 1970).

One of the special cases covered was an analysis of sets of measurements that are assumed to measure the same thing. A set of test scores $x_1, ..., x_p$ with true scores $\tau_1, ..., \tau_p$ was said to be congeneric if every pair of true scores τ_i and τ_j had unit correlation (Jöreskog 1970, 242). The concepts of parallel tests and tau-equivalent tests are special cases of congeneric tests.

Jöreskog (1971) elaborated the analysis of congeneric tests, also discussing the question of reliability of the measurements. However, as the congeneric tests are restricted to the one-factor case, practical applications are hard to find. *Charles Werts*, *D. R. Rock* and *Robert Linn*, with Jöreskog (1978) tried to enhance the procedure and find a way to estimate the reliability of a factorially more complex composite. Their model gives a set of *p* observed variables by

$$\mathbf{x} = \mathbf{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon} \ , \tag{2.21}$$

where $\boldsymbol{\tau}$ is a vector of order p of the true scores, $\mathbf{B}$ is a $p \times p$ identity matrix and $\boldsymbol{\varepsilon}$ is a vector of order p of the errors of measurement on the p variables. According to classical test theory assumptions $E\,(\boldsymbol{\tau}) = E\,(\mathbf{x})$, $E\,(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $E\,(\boldsymbol{\tau}'\boldsymbol{\varepsilon}) = 0$. It is also assumed that $cov\,(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$ (diagonal) and $cov\,(\boldsymbol{\tau}) = \boldsymbol{\Gamma}$.

The true score vector $\boldsymbol{\tau}$ is assumed to have an underlying factor model with k common factors. Let

$$\boldsymbol{\tau} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\eta} \ , \tag{2.22}$$

where $\boldsymbol{\xi}$ is a vector of order k of common factors, $\boldsymbol{\eta}$ is a vector of order p of the unique factors and $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of factor loadings. It is assumed that $E\,(\boldsymbol{\eta}) = \mathbf{0}$, $E\,(\boldsymbol{\xi}\boldsymbol{\eta}') = \mathbf{0}$, $cov\,(\boldsymbol{\eta}) = \boldsymbol{\Psi}$ (diagonal), and $cov\,(\boldsymbol{\xi}) = \boldsymbol{\Phi}$ (Werts, Rock, Linn, and Jöreskog 1978).

From the above assumptions the covariance matrix of the *p* observed variables is given by

$$\boldsymbol{\Sigma} = \mathbf{B}\boldsymbol{\Gamma}\mathbf{B}' + \boldsymbol{\Theta} = \mathbf{B}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})\,\mathbf{B}' + \boldsymbol{\Theta} \,. \tag{2.23}$$

The identity matrix **B** was introduced "for ease of comparison to procedures discussed" in their paper (Werts *et al.* 1978, 933), but apparently it does not have any role in those discussions. The reason for keeping it in the model was probably an issue of compatibility: the general model (Jöreskog 1970) is exactly of the form (2.23).

The reliability of a composite variable $u = \mathbf{a}'\mathbf{x}$ is given by Werts *et al.* (1978) as

$$\rho_{uu} = \frac{\mathbf{a}'\boldsymbol{\Gamma}\,\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\,\mathbf{a}} \quad , \tag{2.24}$$

which may be written, using (2.23), in an alternative form

$$\rho_{uu} = \frac{\mathbf{a}'\boldsymbol{\Lambda\Phi\Lambda}'\mathbf{a} + \mathbf{a}'\boldsymbol{\Psi}\,\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\,\mathbf{a}} \quad . \tag{2.25}$$

The original formula (2.24) looks decent, but the discrepancy is revealed by the formula (2.25), which is not given in their paper. The effect of the unique variance is included in the true variation, suggesting the unique factors to be as important as the common factors. This is a contradictory issue, which has been discussed among others by Harman (1967). The unique factors cause problems of identification, although they may be used in confirmatory models. Usually it is better to include the unique variance in the error variance.

2.9 Present state and conclusions

Despite the general model provided by Jöreskog (1970), and the reliability measure of Werts *et al.* (1978) based on that model, the research has continued to focus on minor modifications of old coefficients, most often Cronbach's alpha:

> What is needed is a measure of reliability that is virtually the same as Cronbach's alpha when distributions are normal, but which is not overly affected when in fact distributions are slightly non-normal instead (Wilcox 1992, 241).

Rand Wilcox (1992) has considered robust generalizations of alpha, without criticizing the assumptions. *Brian Reinhardt* (1996) has conducted a collection of small-scale simulations with a conclusion that alpha can be negative. Since this is by no means a new result, the point of Reinhardt's study is questionable. Also *Donald Zimmerman, Bruno Zumbo* and *Coralie Lalonde* (1993) have studied Cronbach's alpha through use of computer simulation. Their conclusion is that under violation of certain assumptions, alpha gives biased values, independently of the number of observations and the number of items. That does not sound like a new result, either.

S. F. Blinkhorn (1997) gives some telling remarks on the development:

> For example, Wilcox (1992) points out how coefficient alpha is vulnerable to modest numbers of outlying observations, and may be substantially inflated as compared with more robust estimators of reliability in circumstances which are plausible for the practical use of tests. It is doubtful whether this result has touched the consciousness of more than a handful of people, or ever will, and its impact on practice has so far been negligible (Blinkhorn 1997, 178).

John Fleishman and *Jeri Benson* (1987) have used LISREL models (Jöreskog and Sörbom, 1983) to evaluate measurement models and scale reliability. They apply various unidimensional models, such as the congeneric model by Jöreskog (1971), and conclude that the assumption of uncorrelated measurement errors is perhaps most frequently violated in psychometric practice (Fleishman and Benson 1987, 937). *Sven-Eric Reuterberg* and *Jan-Eric Gustafsson* (1992) have combined confirmatory factor analysis estimated by LISREL to reliability estimated by Cronbach's alpha in a case of congeneric tests. *Donald Bacon, Paul Sayer* and *Murray Young* (1995) have studied Cronbach's alpha as well as Heise and Bohrnstedt's omega and even Armor's theta in the context of structural equation modelling. They claim that alpha is neither accurate nor a useful decision aid, and that a more useful tool is provided by a weighted omega (Bacon, Sayer, and Young, 1995). *Arthur Bedeian*, *David Day* and *Kevin Kelloway* (1997) even go back to the Spearman-Brown formula in the same context, to apply the correction for measurement error attenuation. There seems to be a constant need for proper methods of reliability estimation, but the tools are too restricted.

Since the 1930s, most studies on reliability have brought essentially nothing new to the discussion. The development has centered around one particular coefficient, having several names. Deriving it has been described as "one of the favorite indoor sports of psychometricians" (Kaiser and Michael 1977, 34). Blinkhorn (1997) reminds of the weakness of this favourite coefficient:

> No respectable essay on test theory can fail to note that coefficient alpha, or – to give it its pre-war identity for binary-scored tests Kuder-Richardson formula 20, has been derived dozens of times from different theoretical starting points. It is the apprentice psychometrician's favourite party trick. Alpha has become the universal reliability coefficient even if it is explicitly a lower bound, and possibly a very weak lower bound (Blinkhorn 1997, 182).

Exploring the limits of Cronbach's alpha continues. For example, *Jos ten Berge* and *Willem Hofstee* (1999) have examined alpha and its variations as reliabilities of unrotated and rotated principal components. It seems that they have not observed that the method of principal components is unsuitable for any reliability studies, since it lacks the concepts of statistical model and measurement errors.

Common to the development has been that the measurement model has not been criticized or generalized, or even considered in most of the cases. The classical true score model, inherited from Spearman's times, was mathematically formulated by Lord and Novick (1968). It has been blindly applied ever since. The weakest point of that model is the concept of true scores: the seemingly multidimensional true scores correlate perfectly with each other, thus reducing the true dimension to one.

The one-dimensional model has been accepted as such, but it is not sufficient for assessing the reliability of measurement scales. More general methods are needed.

3. General framework of modelling the measurement

We now present the general framework of modelling the measurement, which was introduced by *Lauri Tarkkonen* in his Ph.D thesis in 1987. The measurement model of the framework is a generalization of the common factor analysis model. In addition, it includes a general specification for the concept of measurement scale. An important part of the framework is the reliability measure. We expand the representation of this measure to a matrix form.

3.1 Measurement model

Let $\mathbf{x}$ be a vector of order p of the observed variables. The measurement model is

$$\mathbf{x} = \mathbf{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon} \ , \tag{3.1}$$

where $\boldsymbol{\tau}$ is a vector of order k of the true scores, $\mathbf{B}$ is a $p \times k$ pattern matrix which defines the relationship between $\mathbf{x}$ and $\boldsymbol{\tau}$, and $\boldsymbol{\varepsilon}$ is a vector of the measurement errors. The true scores and the measurement errors are not directly observable and must be estimated from the data. The assumptions are analogous to the classical test theory. It is assumed that $E\,(\boldsymbol{\tau}) = \boldsymbol{\mu}$, $E\,(\boldsymbol{\varepsilon}) = \mathbf{0}$, $cov\,(\boldsymbol{\tau}, \boldsymbol{\varepsilon}) = \mathbf{0}$, $cov\,(\boldsymbol{\tau}) = \boldsymbol{\Phi}$ and $cov\,(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$. Thus, the covariance structure of the observed variables is

$$cov\,(\mathbf{x}) = E\,(\mathbf{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon})\,(\mathbf{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon})' = \mathbf{B}\boldsymbol{\Phi}\mathbf{B}' + \boldsymbol{\Psi} = \boldsymbol{\Sigma} \tag{3.2}$$

(Tarkkonen 1987, 14–15). The covariance matrix $\boldsymbol{\Sigma}$ is assumed to be non-singular and positively definite:

$$rank\,(\boldsymbol{\Sigma}) = p \ \text{ and } \ \boldsymbol{\Sigma} > 0 \ . \tag{3.3}$$

3.2 Estimation of the parameters

Assuming multinormality, the parameters of the measurement model can be estimated from a sample covariance matrix by the maximum likelihood method. Some restrictive assumptions must be made, however, since in the general form, the measurement model includes

$$pk + \frac{k\,(k+1)}{2} + \frac{p\,(p+1)}{2}$$

parameters, while it is possible to identify only $\dfrac{p\,(p+1)}{2}$ of them.

The assumptions needed depend on the situation. In an exploratory approach, it is usually assumed that the measurement errors do not correlate, which decreases the number of parameters by

$$\frac{p\,(p-1)}{2} \quad \text{to} \quad pk + \frac{k\,(k+1)}{2} + p\,.$$

Assuming the true scores to be orthogonal and standardized leaves at most $pk + p$ parameters to be estimated. In this form, the measurement model (3.1) conforms to the traditional, orthogonal factor analysis model. Fine-tuning the assumptions by estimating some of the covariances of the measurement errors, or fixing any of the elements of the matrix **B**, moves the approach to the confirmatory direction (Tarkkonen 1987, 27–31).

3.3 Measurement scale

The measurement scale is a linear combination of the items. In general, we have m scales as a vector $\mathbf{u} = \mathbf{A}'\mathbf{x}$, where $\mathbf{A}$ is a $p \times m$ matrix of the weights. The case of one scale is denoted by $u = \mathbf{a}'\mathbf{x}$, where $\mathbf{a}$ is a vector of the weights.

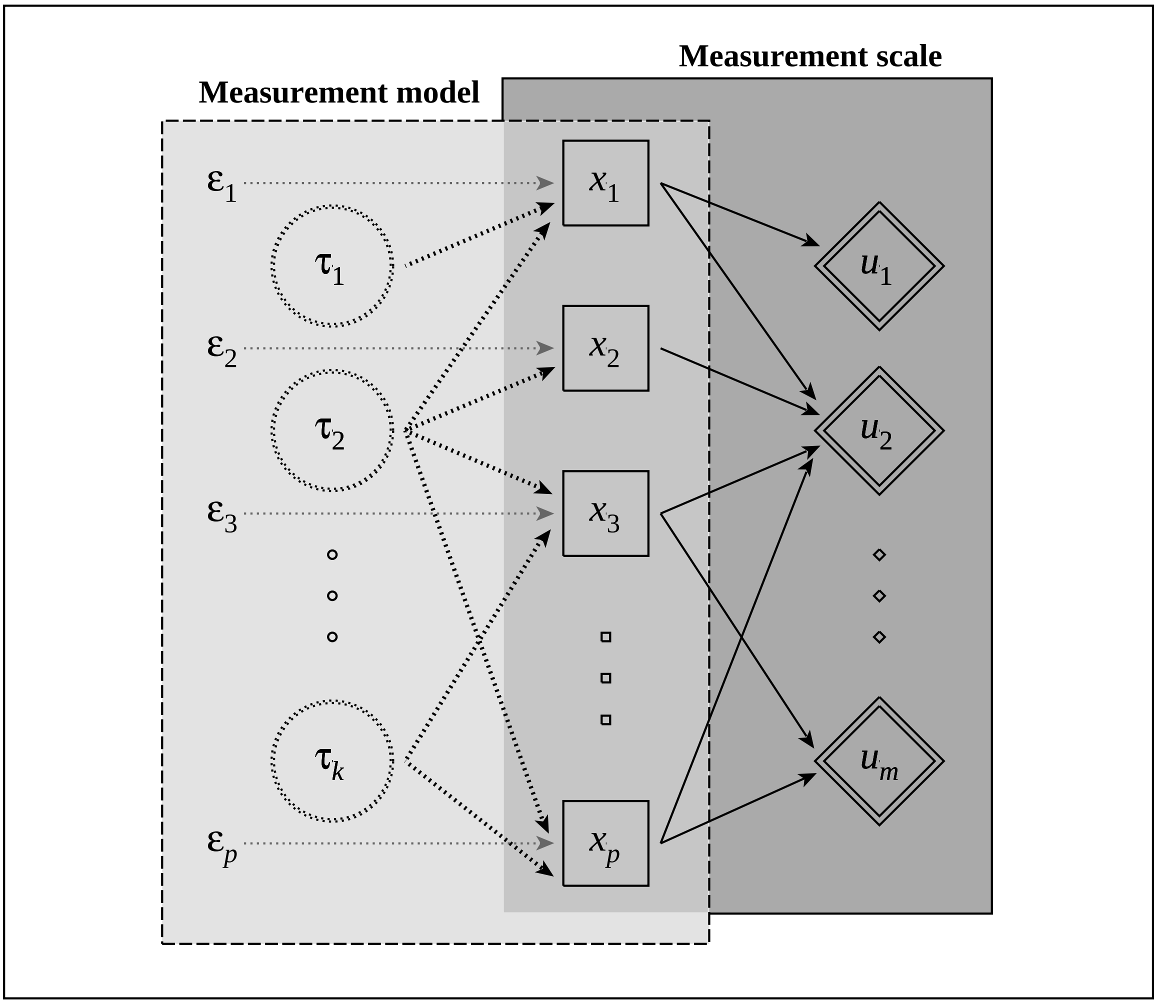

Figure 3.1. Measurement model and measurement scale (Tarkkonen 1987, 12).

It is important to distinguish between the concepts of the measurement scale and the measurement model. The model discriminates the underlying structure of the measurement from the use of the items (Tarkkonen 1987, 13). See Figure 3.1.

The scales weighted by the corresponding pattern elements of the model, denoted by $\mathbf{u} = \mathbf{B}'\mathbf{x}$, are termed true score images. They are used to assess the structural validity of the measurement model (Tarkkonen 1987, 14–22).

3.4 Reliability of measurement scales

According to definition, reliability is the ratio of the true score's variance to the observed variable's variance. For composite measurement scales, the concept of reliability is established on the same principle. Using (3.2), the variance of a measurement scale is divided into two parts, the variance generated by the true scores, and the variance generated by the measurement errors. Now, proceeding for multiple scales $\mathbf{u} = \mathbf{A}'\mathbf{x}$ we have

$$cov\,(\mathbf{u}) = cov\,(\mathbf{A}'\mathbf{x}) = \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{A} + \mathbf{A}'\boldsymbol{\Psi}\mathbf{A}\,. \tag{3.4}$$

The decomposition (3.4) is necessary for the assessment of reliability (Tarkkonen 1987, 21–23).

It is sufficient to consider the variances of the scales, given by the diagonal elements of the covariance matrices in (3.4). Multiplying the diagonal elements of $\mathbf{A}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{A}$ by the inverses of the diagonal elements of $\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}$ we obtain

$$\boldsymbol{\rho}_{\mathbf{u}} = diag\,(\mathbf{A}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{A}) \times [diag\,(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})]^{-1}\,, \tag{3.5}$$

where $\boldsymbol{\rho}_{\mathbf{u}}$ is an $m \times m$ diagonal matrix of the reliabilities of the scales $\mathbf{u}$.

For alternative assumptions concerning the measurement errors, (3.5) can be presented without the matrix $\boldsymbol{\Sigma}$ as

$$\begin{aligned}\boldsymbol{\rho}_{\mathbf{u}} &= diag\,(\mathbf{A}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{A}) \times [diag\,(\mathbf{A}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{A}) + diag\,(\mathbf{A}'\boldsymbol{\Psi}\mathbf{A})]^{-1} \\ &= \{\,[diag\,(\mathbf{A}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{A})]^{-1} \times [diag\,(\mathbf{A}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{A}) + diag\,(\mathbf{A}'\boldsymbol{\Psi}\mathbf{A})]^{-1}\,\}^{-1} \\ &= \{\,\mathbf{I} + diag\,(\mathbf{A}'\boldsymbol{\Psi}\mathbf{A}) \times [diag\,(\mathbf{A}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{A})]^{-1}\,\}^{-1}\,,\end{aligned} \tag{3.6}$$

where $\mathbf{I}$ is an $m \times m$ identity matrix. In the case of one scale, $u = \mathbf{a}'\mathbf{x}$, the formula (3.5) reduces to the original form given by Tarkkonen (1987, 24) as

$$\rho_{uu} = \frac{\mathbf{a}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}\,, \tag{3.7}$$

and correspondingly, (3.6) reduces to

$$\rho_{uu} = \frac{1}{1 + \dfrac{\mathbf{a}'\boldsymbol{\Psi}\mathbf{a}}{\mathbf{a}'\mathbf{B}\boldsymbol{\Phi}\mathbf{B}'\mathbf{a}}}\,. \tag{3.8}$$

The quadratic forms in (3.7) and (3.8) are variances of linear combinations, and thus non-negative. Namely,

$$\mathbf{a}'\mathbf{B\Phi B}'\mathbf{a} = var\,(\mathbf{a}'\mathbf{B\tau}) \geq 0\,,$$
$$\mathbf{a}'\mathbf{\Psi}\,\mathbf{a} = var\,(\mathbf{a}'\boldsymbol{\varepsilon}) \geq 0\,, \text{ and}$$
$$\mathbf{a}'\mathbf{\Sigma}\,\mathbf{a} = var\,(\mathbf{a}'\mathbf{x}) \geq 0\,. \tag{3.9}$$

Using the assumption (3.3) we can infer that

$$\mathbf{a}'\mathbf{\Sigma}\,\mathbf{a} > 0\,. \tag{3.10}$$

It then follows from (3.9) and (3.10) that

$$0 \leq \rho_{uu} \leq 1\,, \tag{3.11}$$

which means that ρ_{uu} fulfills the basic requirement for a reliability measure.

4. Comparisons with other approaches

The general framework of chapter 3 is now compared with other approaches presented in the literature, namely the covariance structure model, the factor analysis model and the classical true score model. In addition, the conditions required for the equivalence of the general reliability measure and Cronbach's alpha are sought without an explicit assumption of a measurement model. A new result stating that alpha can not exceed the general measure is proved.

4.1 Covariance structure model

When comparing the general measurement model (3.1) with alternative models appearing in the psychometric literature, the closest equivalent is found from a covariance structure based approach (Werts, Rock, Linn, and Jöreskog 1978), which was reviewed in chapter 2.8. The corresponding model is given by (2.21).

The essential difference between the models is in the definition of the true scores. Werts *et al.* (1978) defined each observed variable to have its own true score, which conforms to the classical true score model (1.1). A more general solution was then sought by defining a factor analysis model for the true score. Thus the model was brought under the general covariance structure modelling approach developed by Jöreskog (1970). To emphasize this, Werts *et al.* (1978) even included an unnecessary identity matrix **B** in their measurement model (*cf.* 2.21).

4.2 The factor analysis model

In practice, the most important application of the measurement model (3.1) is the common factor analysis model

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\eta}\,, \tag{4.1}$$

where $\mathbf{x}$ is a vector of order p of the observed variables, $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of factor loadings, $\boldsymbol{\xi}$ is a vector of order k of common factors and $\boldsymbol{\eta}$ is a vector of unique factors. Let us now consider the central properties of this model, seen from the general framework of modelling the measurement.

4.2.1 The factor model as a measurement model

By associating the common factors with the true scores, and interpreting the unique factors as measurement errors, we have a measurement model

$$\mathbf{x} = \mathbf{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon}\,, \tag{4.2}$$

where the matrix **B** is the factor matrix. The assumptions of the general model are applied, but according to the usual assumptions of factor analysis, the factors are assumed to be orthogonal and only transformed by a suitable oblique rotation, if necessary. Thus,

$$cov\,(\boldsymbol{\tau}) = \mathbf{I}\,. \tag{4.3}$$

The covariance structure of the observed variables then has the form

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Psi}\,, \tag{4.4}$$

where $\boldsymbol{\Psi}$ is assumed to be diagonal. The equation (4.4) is the fundamental equation of factor analysis.

4.2.2 Estimation of the parameters

The elements of **B** and $\boldsymbol{\Psi}$ are unknown parameters that have to be estimated from data. By using a suitable rescaling, consistent estimators can be found. Rescaling each variable so that its residual variance is equal to unity, transforms the covariance matrix $\boldsymbol{\Sigma}$ to

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Psi}^{-1/2}\,\boldsymbol{\Sigma}\,\boldsymbol{\Psi}^{-1/2}\,. \tag{4.5}$$

Similarly $\boldsymbol{\Sigma} - \boldsymbol{\Psi}$ becomes

$$\boldsymbol{\Sigma}^* - \mathbf{I} = \boldsymbol{\Psi}^{-1/2}\,(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\,\boldsymbol{\Psi}^{-1/2}\,. \tag{4.6}$$

The matrix $\boldsymbol{\Sigma}^* - \mathbf{I}$ may be expressed in the form

$$\boldsymbol{\Sigma}^* - \mathbf{I} = \boldsymbol{\Omega}\,\boldsymbol{\Delta}\,\boldsymbol{\Omega}'\,, \tag{4.7}$$

where $\boldsymbol{\Delta}$ is a diagonal matrix of order k, and $\boldsymbol{\Omega}$ is a $p \times k$ matrix satisfying

$$\boldsymbol{\Omega}'\boldsymbol{\Omega} = \mathbf{I}\,. \tag{4.8}$$

From (3.3) it follows that $\boldsymbol{\Delta}$ is uniquely determined. We may then define **B** uniquely by

$$\mathbf{B} = \boldsymbol{\Psi}^{1/2}\,\boldsymbol{\Omega}\,\boldsymbol{\Delta}^{1/2}\,, \tag{4.9}$$

and we have, as required by (4.4),

$$\begin{aligned}
\mathbf{B}\mathbf{B}' &= \boldsymbol{\Psi}^{1/2}\,\boldsymbol{\Omega}\,\boldsymbol{\Delta}^{1/2}\,\boldsymbol{\Delta}^{1/2}\,\boldsymbol{\Omega}'\boldsymbol{\Psi}^{1/2} \\
&= \boldsymbol{\Psi}^{1/2}\,(\boldsymbol{\Sigma}^* - \mathbf{I})\,\boldsymbol{\Psi}^{1/2} \\
&= \boldsymbol{\Psi}^{1/2}\,\boldsymbol{\Psi}^{-1/2}\,(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\,\boldsymbol{\Psi}^{-1/2}\,\boldsymbol{\Psi}^{1/2} \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Psi}\,.
\end{aligned} \tag{4.10}$$

Multiplying the equation (4.9) from left by $\mathbf{\Psi}^{-1/2}$ gives

$$\mathbf{\Psi}^{-1/2}\,\mathbf{B} \;=\; \mathbf{\Omega}\,\mathbf{\Delta}^{1/2}\,, \tag{4.11}$$

and we have

$$\mathbf{B}'\,\mathbf{\Psi}^{-1}\,\mathbf{B} \;=\; (\mathbf{\Psi}^{-1/2}\,\mathbf{B})'\,\mathbf{\Psi}^{-1/2}\,\mathbf{B} \;=\; \mathbf{\Delta}^{1/2}\,\mathbf{\Omega}'\mathbf{\Omega}\,\mathbf{\Delta}^{1/2} \;=\; \mathbf{\Delta}\,. \tag{4.12}$$

Thus the matrix $\mathbf{B}$ is chosen so that $\mathbf{B}'\,\mathbf{\Psi}^{-1}\,\mathbf{B}$ is a diagonal matrix. This choice is convenient for the maximum likelihood estimation (Lawley and Maxwell 1971).

4.2.3 Factor images

The reliability of a factor is not a sound concept, although it appears often in the literature. The factors are theoretical constructions. They should be distinguished from the measurement scales corresponding to the factors. These scales are weighted by the factor loadings and hence termed factor images (Tarkkonen 1987, 33). The factor images affect all scales and their reliabilities in the later phase of modelling, because the factor matrix $\mathbf{B}$ forms the basic structure of the model.

The reliabilities of the factor images provide additional, useful tools for assessing the structural validity of a factor model. The reliabilities are obtained by applying the formula (3.5) with $\mathbf{A} = \mathbf{B}$ and $\mathbf{\Phi} = \mathbf{I}$:

$$\begin{aligned}
\boldsymbol{\rho}_{\mathbf{u}} \;&=\; \mathit{diag}\,(\mathbf{A}'\,\mathbf{B}\mathbf{B}'\,\mathbf{A}) \times [\mathit{diag}\,(\mathbf{A}'\,\mathbf{\Sigma}\,\mathbf{A})]^{-1} \\
&=\; \mathit{diag}\,(\mathbf{B}'\,\mathbf{B}\mathbf{B}'\,\mathbf{B}) \times [\mathit{diag}\,(\mathbf{B}'\,\mathbf{\Sigma}\,\mathbf{B})]^{-1} \\
&=\; \mathit{diag}\,[(\mathbf{B}'\,\mathbf{B})^2] \times [\mathit{diag}\,(\mathbf{B}'\,\mathbf{\Sigma}\,\mathbf{B})]^{-1}\,.
\end{aligned} \tag{4.13}$$

A reliability measure for maximum likelihood factor analysis was developed by Tucker and Lewis (1973), but it is inapplicable for measurement scales. Instead, it is meant for assessing the reliability of the factor solution as a whole. A better approach is provided by the factor images.

4.2.4 Factor scores

In practice, estimates of the values corresponding to the factors are needed for each observation in the data. However, there is no unique way to calculate these factor scores, because the equations corresponding to the model (4.2) can not be solved with respect to the factors $\boldsymbol{\tau}$. The best way is to seek a least squares approximation by minimizing

$$f\,(\mathbf{A}) \;=\; E\,(\|\boldsymbol{\tau} - \mathbf{u}\|^2) \;=\; E\,[\;\mathit{tr}\,(\boldsymbol{\tau} - \mathbf{A}'\,\mathbf{x})'\,(\boldsymbol{\tau} - \mathbf{A}'\,\mathbf{x})\;] \tag{4.14}$$

with respect to the $p \times k$ matrix $\mathbf{A}$. The minimum is reached, when $\mathbf{A} = \mathbf{\Sigma}^{-1}\,\mathbf{B}$.

To show this, we follow Mustonen (1995), first simplifying (4.14) to get

$$\begin{aligned}
f\,(\mathbf{A}) \;&=\; E\,[\;\mathit{tr}\,(\boldsymbol{\tau} - \mathbf{A}'\,\mathbf{x})'\,(\boldsymbol{\tau} - \mathbf{A}'\,\mathbf{x})\;] \\
&=\; E\,[\;\mathit{tr}\,(\boldsymbol{\tau}\,\boldsymbol{\tau}' - \boldsymbol{\tau}\,\mathbf{x}'\,\mathbf{A} - \mathbf{A}'\,\mathbf{x}\,\boldsymbol{\tau}' + \mathbf{A}'\,\mathbf{x}\mathbf{x}'\,\mathbf{A})\;]
\end{aligned}$$

$$= tr\,(\mathbf{I} - \mathbf{B}'\mathbf{A} - \mathbf{A}'\mathbf{B} + \mathbf{A}'\mathbf{\Sigma}\mathbf{A})\,, \tag{4.15}$$

since the factors were assumed to be orthogonal by (4.3) and because $E\,(\mathbf{x}, \boldsymbol{\tau}') = \mathbf{B}$.

Using the Cholesky decomposition $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}'$ (since $\mathbf{\Sigma} > 0$) we then show that

$$\delta = f\,(\mathbf{A}) - f\,(\mathbf{\Sigma}^{-1}\,\mathbf{B}) \geq 0\,. \tag{4.16}$$

Substituting (4.15) to (4.16) with $\mathbf{A}$ and $\mathbf{\Sigma}^{-1}\,\mathbf{B}$ gives

$$\begin{aligned}
\delta &= tr\,(\mathbf{I} - 2\,\mathbf{A}'\mathbf{B} + \mathbf{A}'\mathbf{\Sigma}\mathbf{A}) - tr\,(\mathbf{I} - 2\,\mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B} + \mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{\Sigma}\,\mathbf{\Sigma}^{-1}\,\mathbf{B}) \\
&= tr\,(\mathbf{A}'\mathbf{\Sigma}\mathbf{A} - 2\,\mathbf{A}'\mathbf{B} + \mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B}) \\
&= tr\,[\,(\mathbf{A}'\mathbf{C})\,(\mathbf{C}'\mathbf{A}) - 2\,(\mathbf{A}'\mathbf{C})\,(\mathbf{C}^{-1}\,\mathbf{B}) + (\mathbf{B}'\mathbf{C}'^{-1})\,(\mathbf{C}^{-1}\,\mathbf{B})\,] \\
&= tr\,(\mathbf{L}\mathbf{L}' - 2\,\mathbf{L}\mathbf{F}' + \mathbf{F}\mathbf{F}')\,,
\end{aligned}$$

where $\mathbf{L} = \mathbf{A}'\mathbf{C}$ and $\mathbf{F} = \mathbf{B}'\mathbf{C}'^{-1}$. Then

$$\begin{aligned}
\delta &= tr\,(\mathbf{L}\mathbf{L}' - \mathbf{L}\mathbf{F}' - \mathbf{F}\mathbf{L}' + \mathbf{F}\mathbf{F}') \\
&= tr\,[\,(\mathbf{L} - \mathbf{F})\,(\mathbf{L} - \mathbf{F})'\,]\,.
\end{aligned}$$

We see that $\delta \geq 0$ and that $\delta = 0$ only if $\mathbf{L} = \mathbf{F}$. In other words, the minimum is reached when $\mathbf{A}'\mathbf{C} = \mathbf{B}'\mathbf{C}'^{-1}$, which gives the optimal matrix of weights for the factor scores as

$$\mathbf{A} = (\mathbf{B}'\mathbf{C}'^{-1}\,\mathbf{C}^{-1})' = (\mathbf{B}'\mathbf{\Sigma}^{-1})' = \mathbf{\Sigma}^{-1}\,\mathbf{B} \tag{4.17}$$

(Mustonen 1995, 90–91).

Making use of the identity

$$\mathbf{\Sigma}^{-1}\,\mathbf{B} = \mathbf{\Psi}^{-1}\,\mathbf{B}\,(\mathbf{I} + \mathbf{\Delta})^{-1}\,, \tag{4.18}$$

(Lawley and Maxwell 1971, 27), where $\mathbf{\Delta} = \mathbf{B}'\mathbf{\Psi}^{-1}\,\mathbf{B}$ by (4.12), we see that

$$\mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B} = \mathbf{B}'\mathbf{\Psi}^{-1}\,\mathbf{B}\,(\mathbf{I} + \mathbf{\Delta})^{-1} = \mathbf{\Delta}\,(\mathbf{I} + \mathbf{\Delta})^{-1} \tag{4.19}$$

is a diagonal matrix.

Finally, the reliabilities of the factor scores are obtained by applying the formula (3.5) with $\mathbf{A} = \mathbf{\Sigma}^{-1}\,\mathbf{B}$ and $\mathbf{\Phi} = \mathbf{I}$ as

$$\begin{aligned}
\boldsymbol{\rho}_{\mathbf{u}} &= diag\,(\mathbf{A}'\mathbf{B}\mathbf{B}'\mathbf{A}) \times [diag\,(\mathbf{A}'\mathbf{\Sigma}\mathbf{A})]^{-1} \\
&= diag\,(\mathbf{B}'(\mathbf{\Sigma}^{-1})'\mathbf{B}\mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B}) \times [diag\,(\mathbf{B}'(\mathbf{\Sigma}^{-1})'\mathbf{\Sigma}\,\mathbf{\Sigma}^{-1}\,\mathbf{B})]^{-1} \\
&= diag\,[(\mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B})^2] \times [diag\,(\mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B})]^{-1} \\
&= (\mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B})^2 \times (\mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B})^{-1} \quad (cf.\ 4.19) \\
&= \mathbf{B}'\mathbf{\Sigma}^{-1}\,\mathbf{B}\,.
\end{aligned} \tag{4.20}$$

4.3 Classical true score model

Using the notation of the general model (3.1), the classical true score model (1.1) is given by

$$\mathbf{x} = \mathbf{1}\tau + \boldsymbol{\varepsilon} \ , \tag{4.21}$$

where $\mathbf{x}$ is a vector of order p of the observed variables, τ is a scalar true score, $\mathbf{1}$ is a $p \times 1$ vector of ones and $\boldsymbol{\varepsilon}$ is a vector of the measurement errors.

The assumptions $E\,(\boldsymbol{\varepsilon}) = \mathbf{0}$, $var\,(\tau) = \phi$ and $cov\,(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$ give the covariance structure of the observed variables the form

$$\boldsymbol{\Sigma} = \mathbf{1}\phi\mathbf{1}' + \boldsymbol{\Psi} , \tag{4.22}$$

where $\boldsymbol{\Psi}$ is assumed to be a diagonal matrix.

The measurement scale is an unweighted sum

$$u = \mathbf{1}'\mathbf{x} , \tag{4.23}$$

and its variance is given by

$$var\,(u) = \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} = \mathbf{1}'\mathbf{1}\phi\mathbf{1}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Psi}\mathbf{1} = p^2\phi + tr\,(\boldsymbol{\Psi}) . \tag{4.24}$$

Thus, the reliability of the scale (4.23) is

$$\rho_{uu} = \frac{\mathbf{1}'\mathbf{1}\phi\mathbf{1}'\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} = \frac{p^2\phi}{p^2\phi + tr\,(\boldsymbol{\Psi})} = \frac{1}{1 + \dfrac{tr\,(\boldsymbol{\Psi})}{p^2\phi}} . \tag{4.25}$$

Assuming that $cov\,(\boldsymbol{\varepsilon}) = \psi\mathbf{I}$, we have the most simple model. Then,

$$var\,(u) = \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} = \mathbf{1}'\mathbf{1}\phi\mathbf{1}'\mathbf{1} + \mathbf{1}'\psi\mathbf{I}\mathbf{1} = p^2\phi + p\psi , \tag{4.26}$$

and the reliability becomes

$$\rho_{uu} = \frac{\mathbf{1}'\mathbf{1}\phi\mathbf{1}'\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} = \frac{p^2\phi}{p^2\phi + p\psi} = \frac{1}{1 + \dfrac{\psi}{p\phi}} . \tag{4.27}$$

Substitution of

$$\frac{\psi}{\phi} = \frac{1 - \rho_{xx}}{\rho_{xx}} \tag{4.28}$$

in (4.27) gives the Spearman-Brown formula (2.1).

4.4 Cronbach's alpha

Cronbach's alpha (2.15) is a special case of the general measure of reliability (3.5). These two measures are equivalent under a restricted, one-dimensional measurement model.

Since Cronbach (1951) took alpha as given, without specifying any measurement model, we seek the conditions for the equivalence in an analogous way, by setting the measures equal and making necessary assumptions as needed. This was done already by Tarkkonen (1987, 56–57), but as his proof was partially incomplete, we try to improve it here. As a new result we prove that alpha can not exceed the general measure. The equivalence then follows as the most restricted special case of this result.

To begin with, we write alpha (2.15), using matrix notation, as

$$\alpha = \frac{p}{p-1}\left(1 - \frac{tr\,(\mathbf{\Sigma})}{\mathbf{1}'\mathbf{\Sigma}\,\mathbf{1}}\right) , \tag{4.29}$$

since $\sigma_u^2 = var\,(\mathbf{1}'\mathbf{x}) = \mathbf{1}'\mathbf{\Sigma}\,\mathbf{1}$, where $\mathbf{\Sigma} = cov\,(\mathbf{x})$, and $\mathbf{1} = (1, 1, ..., 1)$.

As (4.29) indicates, alpha is easily computed as soon as the covariance matrix of the observed variables is known. Because the generalized version of alpha given by (2.16) even more clearly violates the original assumptions, we do not apply it here. Instead, we proceed with the reliability of the unweighted sum. For a reasonable comparison, we thus present the general measure in the form

$$\rho_{uu} = \frac{\mathbf{1}'\mathbf{B\Phi B}'\mathbf{1}}{\mathbf{1}'\mathbf{\Sigma}\,\mathbf{1}} , \tag{4.30}$$

and then set it equal to alpha, which results to

$$\frac{\mathbf{1}'\mathbf{B\Phi B}'\mathbf{1}}{\mathbf{1}'\mathbf{\Sigma}\,\mathbf{1}} = \frac{p}{p-1}\left(1 - \frac{tr\,(\mathbf{\Sigma})}{\mathbf{1}'\mathbf{\Sigma}\,\mathbf{1}}\right) . \tag{4.31}$$

Multiplying both sides by $\mathbf{1}'\mathbf{\Sigma}\,\mathbf{1} \geq 0$ gives

$$\mathbf{1}'\mathbf{B\Phi B}'\mathbf{1} = \frac{p}{p-1}\left(\mathbf{1}'\mathbf{\Sigma}\,\mathbf{1} - tr\,(\mathbf{\Sigma})\right) , \tag{4.32}$$

and by substituting $\mathbf{\Sigma} = \mathbf{B\Phi B}' + \mathbf{\Psi}$ (*cf.* 3.2) we obtain

$$\mathbf{1}'\mathbf{B\Phi B}'\mathbf{1} = \frac{p}{p-1}\left(\mathbf{1}'\mathbf{B\Phi B}'\mathbf{1} + \mathbf{1}'\mathbf{\Psi}\,\mathbf{1} - tr\,(\mathbf{B\Phi B}') - tr\,(\mathbf{\Psi})\right) . \tag{4.33}$$

If we assume that the measurement errors are uncorrelated, then the matrix $\mathbf{\Psi}$ is diagonal, and $\mathbf{1}'\mathbf{\Psi}\,\mathbf{1} = tr\,(\mathbf{\Psi})$, so we have

$$\mathbf{1}'\mathbf{B\Phi B}'\mathbf{1} = \frac{p}{p-1}\left(\mathbf{1}'\mathbf{B\Phi B}'\mathbf{1} - tr\,(\mathbf{B\Phi B}')\right) , \tag{4.34}$$

where all the terms depend on the covariance matrix $\mathbf{B\Phi B}'$.

In order to get further, we are obliged to make an essential restriction. It follows that the equation (4.34) holds if and only if the elements of the matrix $\mathbf{B\Phi B'}$ are equal, which means that the rank of the matrix $\mathbf{B\Phi B'}$ and thus the dimension of the model is inevitably one. To show this, we prove a more general result.

Lemma 4.1. For any non-negative definite $p \times p$ matrix $\mathbf{C}$,

$$\mathbf{1'C\,1} \geq \frac{p}{p-1}\left(\mathbf{1'C\,1} - tr\,(\mathbf{C})\right) , \tag{4.35}$$

and the equality holds if and only if $\mathbf{C} = \theta^2 \mathbf{11'}$, $\theta \in \Re$.

Proof. The proof is in two parts. Let us first consider the case when $rank\,(\mathbf{C}) = 1$. Then, $\mathbf{C} = \mathbf{cc'}$, where $\mathbf{c} = (c_1, c_2, ..., c_p)$. Thus,

$$\mathbf{C} = \begin{bmatrix} c_1^2 & c_1c_2 & . \;\; . \;\; . & c_1c_p \\ c_2c_1 & c_2^2 & . \;\; . \;\; . & c_2c_p \\ . & . & . & . \\ . & . & \;\;\; . & . \\ . & . & \;\;\;\;\;\; . & . \\ c_pc_1 & c_pc_2 & . \;\; . \;\; . & c_p^2 \end{bmatrix} ,$$

and (4.35) can be rewritten as

$$(c_1 + ... + c_p)^2 \geq \frac{p}{p-1}\left((c_1 + ... + c_p)^2 - (c_1^2 + ... + c_p^2)\right) \tag{4.36}$$

or

$$(c_1^2 + ... + c_p^2) + 2\,(c_1c_2 + ... + c_{p\text{-}1}c_p) \geq \frac{p}{p-1}\; 2\,(c_1c_2 + ... + c_{p\text{-}1}c_p) \tag{4.37}$$

which simplifies to

$$(c_1^2 + ... + c_p^2) - \frac{2}{p-1}\,(c_1c_2 + ... + c_{p\text{-}1}c_p) \geq 0 \;. \tag{4.38}$$

Multiplying by $(p-1)$ gives

$$(p-1)\,(c_1^2 + ... + c_p^2) - 2\,(c_1c_2 + ... + c_{p\text{-}1}c_p) \geq 0 \tag{4.39}$$

which is equivalent to

$$(c_1 - c_2)^2 + \ldots + (c_1 - c_p)^2 + (c_2 - c_3)^2 + \ldots + (c_{p\text{-}1} - c_p)^2 \;\geq\; 0 \qquad (4.40)$$

or simply

$$\sum_{i<}\sum_{j} (c_i - c_j)^2 \;\geq\; 0\,. \qquad (4.41)$$

Hence, it is seen that the sum is non-negative, and it is equal to zero only if $c_1 = c_2 = \ldots = c_p = \theta\,,\;\; \theta \in \Re$.

This proves the Lemma 4.1 in the case when $rank\,(\mathbf{C}) = 1$.

Next, we consider the case when $\mathbf{C}$ is a general $p \times p$ matrix, and $\mathbf{C} \geq 0$. Then there exists a spectral decomposition

$$\begin{aligned} \mathbf{C} \;&=\; \lambda_1 \mathbf{u}^{(1)}\mathbf{u}^{(1)\prime} \;+ \ldots +\; \lambda_p \mathbf{u}^{(p)}\mathbf{u}^{(p)\prime}\;,\;\; \lambda_i \geq 0 \\ &=\; \mathbf{C}_1 + \ldots + \mathbf{C}_p\,, \end{aligned} \qquad (4.42)$$

where

$$\mathbf{C}_i \;=\; \lambda_i \mathbf{u}^{(i)}\mathbf{u}^{(i)\prime}\;, \qquad (4.43)$$

with

$$tr\,(\mathbf{C}_i) \;=\; tr\,(\lambda_i \mathbf{u}^{(i)}\mathbf{u}^{(i)\prime}) \;=\; \lambda_i \mathbf{u}^{(i)\prime}\,\mathbf{u}^{(i)} \;=\; \lambda_i \qquad (4.44)$$

and

$$r\,(\mathbf{C}_i) \;=\; 1\,. \qquad (4.45)$$

So we have

$$\mathbf{1}'\,\mathbf{C}_i\,\mathbf{1} \;\geq\; \frac{p}{p-1}\left(\mathbf{1}'\,\mathbf{C}_i\,\mathbf{1} - tr\,(\mathbf{C}_i)\right)\;, \qquad (4.46)$$

and summing for $i = 1, 2, \ldots, p$, and using (4.44) we obtain

$$\mathbf{1}'\,\mathbf{C}\,\mathbf{1} \;\geq\; \frac{p}{p-1}\left(\mathbf{1}'\,\mathbf{C}\,\mathbf{1} - (\lambda_1 + \ldots + \lambda_p)\right)\;, \qquad (4.47)$$

which is equivalent to

$$\mathbf{1}'\,\mathbf{C}\,\mathbf{1} \;\geq\; \frac{p}{p-1}\left(\mathbf{1}'\,\mathbf{C}\,\mathbf{1} - tr\,(\mathbf{C})\right)\;. \qquad (4.35)$$

The equality in (4.46) can hold only for one i with $\lambda_i > 0$ since $\mathbf{u}^{(i)\prime}\mathbf{u}^{(j)} = 0$ for all $i \neq j$ and thus only one of the $\mathbf{u}^{(i)}$'s can be of the form $\theta\mathbf{1}$, $\theta \in \Re$. Assume that $\mathbf{u}^{(1)} = \theta\mathbf{1}$. Then the equality in (4.35) holds only if $\lambda_2 = \ldots = \lambda_p = 0$. But then $\mathbf{C} = \mathbf{C}_1 = \theta^2\,\mathbf{11}'$ and $rank\,(\mathbf{C}) = 1$.

This completes the proof of the Lemma 4.1, and we can present the result in a form of a theorem.

Theorem 4.1. If the measurement model is one-dimensional, and the scale is an unweighted sum, then

$$\rho_{uu} \geq \alpha\,,$$

and the equality holds only when all elements of the corresponding covariance matrix are equal.

Proof. Follows from the Lemma 4.1, by choosing $\mathbf{C} = \mathbf{B\Phi B}'$.

The equation (4.34) is thus satisfied as a special case of Theorem 4.1. If we choose $\mathbf{B} = \theta\mathbf{1}$ and $\mathbf{\Phi} = \phi$, for some $\theta, \phi \in \Re$, the equation becomes

$$\mathbf{1}'\theta\mathbf{1}\phi\mathbf{1}'\theta\mathbf{1} = \frac{p}{p-1}\left(\mathbf{1}'\theta\mathbf{1}\phi\mathbf{1}'\theta\mathbf{1} - tr\,(\theta\mathbf{1}\phi\mathbf{1}'\theta)\right)\,, \tag{4.48}$$

but, since $\mathbf{1}'\mathbf{1} = tr\,(\mathbf{11}') = p$, we have

$$\theta^2 p^2 \phi = \frac{p}{p-1}\left(\theta^2 p\phi\,(p-1)\right)\,, \tag{4.49}$$

and the result is seen immediately.

Hence, Cronbach's alpha is a special case of Tarkkonen's measure.

5. Examples of applications

In this chapter, we present examples of applications related to the reliabilities of measurement scales. The examples also demonstrate the way of working in the integrated environment of Survo (Mustonen 1992), which is used as the platform for the simulation experiments of this study.

Knowing the details of Survo is not essential for understanding the examples. The primary aim is to deal with the questions of reliability in empirical applications. However, since the interface of Survo differs from the mainstream, a brief description is provided below.

"Basically, everything in Survo is carried out in an *edit field* which corresponds to a spreadsheet but has also capabilities of a word processor. The user types text and commands in this working area. When a command is activated, the editor program passes the task to a suitable program module. The results are automatically written partly in the same edit field (in legible form) and partly into files (numerical results in double precision). The user may edit his/her own text and results and type and activate more commands." (Text quoted from Mustonen and Vehkalahti 1997).

The views of edit fields are displayed as *work schemes*. The output of Survo operations is presented mainly with black text on a grey background, while the activated commands appear as **white text on a black background**.

The first example demonstrates the process of modelling the measurement and shows how the reliabilities of various measurement scales are computed.

The second example briefly re-evaluates certain misleading experiments with artificial factor structures presented in the literature. The structural validity of the factor model is considered through the use of the reliabilities of the factor images.

The last example consists of two empirical applications, where the scales are analyzed separately, thus hiding their true multidimensional nature. Both cases are re-analyzed with more appropriate methods.

5.1 Measuring the physical capacity by decathlon scores

The data set DECA, available in each Survo installation, includes the names and scores of the 48 best athletes in decathlon in 1973 (see Scheme 5.1).

```
  1  1 SURVO 98  Tue Jan 04 08:05:27 2000                D:\V\    1000  100 0
*SAVE DECA / Decathlon scores as measures of physical capacity
*LOAD INDEX
*
*FILE STATUS DECA
*
*    Best athletes in decathlon in 1973
*                                          {lower_limit,upper_limit}
*FIELDS: (active)
*   1 SA-   8 Name     Name of athlete
*   2 NA-   2 Points   Total score                 (####) {7000,9000}
*   3 NA-   2 100m     100 meters run              (####) {500,1200}
*   4 NA-   2 L_jump   Long jump                   (####) {500,1200}
*   5 NA-   2 Shot_put                             (####) {500,1200}
*   6 NA-   2 Hi_jump  High jump                   (####) {500,1200}
*   7 NA-   2 400m     400 meters run              (####) {500,1200}
*   8 NA-   2 Hurdles  110 meters hurdles          (####) {500,1200}
*   9 NA-   2 Discus                               (####) {500,1200}
*  10 NA-   2 Pole_vlt Pole vault                  (####) {500,1200}
*  11 NA-   2 Javelin                              (####) {500,1200}
*  12 NA-   2 1500m    1500 meters run             (####) {400,1200}
*  13 NA-   2 Height   in centimeters              (###)  {160,210}
*  14 NA-   2 Weight   in kilograms                (###)  {50,120}
*END
*SURVO 84C data file DECA: record=128 bytes, M1=30 L=64  M=14 N=48
*
```

Scheme 5.1. Structure of the data set DECA.

Here we may think the decathlon as a measurement instrument for measuring the physical capacity of a man. The capacity consists of at least three factors: speed, force and strength. These could be measured for example by 100-metre sprint, shot put and 1500-metre run, respectively.

However, to create reliable measurement scales, we need to have some kind of redundancy of measurement and thus we use the scores of all events. We then have the maximum information available for further processing. The aim is to reduce the problem from ten dimensions to three only. This is best achieved by factor analysis. By fixing the number of factors *a priori*, we move towards a confirmatory factor analysis, but this is as it should be in most cases. This kind of approach could be called *exploratory with a proper concept of measurement.*

We begin by computing the correlation matrix. The names of the variables, which appear automatically as row and column labels of the matrix, are self-explaining in Scheme 5.2.

```
   1  1 SURVO 98  Tue Jan 04 08:07:06 2000            D:\V\   1000  100 0
*..........................................................................
*MASK=--AAAAAAAAAA--
*CORR DECA
*/LOADCORR / this gives the setup below
*..........................................................................
*LIMITS=-0.46,-0.364,-0.283,0.283,0.364,0.46,1 SHADOWS=7,5,1,0,1,5,7
*Limits: P=0.001 0.46   P=0.01 0.364   P=0.05 0.283
*LOADM CORR.M,12.12,CUR+1
*R(DECA)
*           100m L_jum Shot_ Hi_ju  400m Hurdl Discu Pole_ Javel 1500m
*100m       1.00  0.17 -0.03 -0.41  0.46  0.32  0.01  0.05 -0.22 -0.29
*L_jump     0.17  1.00 -0.03  0.00  0.13  0.30  0.02  0.06  0.15 -0.21
*Shot_put  -0.03 -0.03  1.00  0.16 -0.30  0.09  0.73 -0.20  0.02 -0.45
*Hi_jump   -0.41  0.00  0.16  1.00 -0.34 -0.04  0.22 -0.12  0.15 -0.15
*400m       0.46  0.13 -0.30 -0.34  1.00  0.18 -0.34  0.01 -0.10  0.30
*Hurdles    0.32  0.30  0.09 -0.04  0.18  1.00  0.05 -0.07 -0.15 -0.22
*Discus     0.01  0.02  0.73  0.22 -0.34  0.05  1.00 -0.18  0.14 -0.57
*Pole_vlt   0.05  0.06 -0.20 -0.12  0.01 -0.07 -0.18  1.00 -0.13  0.01
*Javelin   -0.22  0.15  0.02  0.15 -0.10 -0.15  0.14 -0.13  1.00 -0.07
*1500m     -0.29 -0.21 -0.45 -0.15  0.30 -0.22 -0.57  0.01 -0.07  1.00
*
```

Scheme 5.2. Correlation matrix of DECA.

The scores of shot put and discus are highly correlated, as one might expect. To take a closer view of the total structure, we compute the maximum likelihood factor analysis of three factors, according to our concept of the physical capacity.

```
   1  1 SURVO 98  Tue Jan 04 08:07:32 2000            D:\V\   1000  100 0
*
*FACTA CORR.M,3,CUR+1
*Factor analysis: Maximum Likelihood (ML) solution
*Factor matrix
*               F1      F2      F3    h^2
*100m      -0.298   0.875   0.176  0.886
*L_jump    -0.206   0.163  -0.112  0.082
*Shot_put  -0.456  -0.313   0.654  0.733
*Hi_jump   -0.144  -0.501  -0.061  0.275
*400m       0.300   0.617   0.035  0.471
*Hurdles   -0.227   0.283   0.058  0.135
*Discus    -0.582  -0.301   0.562  0.745
*Pole_vlt   0.016   0.115  -0.245  0.073
*Javelin   -0.064  -0.254  -0.058  0.072
*1500m      0.997   0.004   0.014  0.995
*
```

Scheme 5.3. Factor matrix of DECA, a three-factor maximum likelihood solution.

The highlighted loadings in Scheme 5.3 clearly reveal the basic structure, but the communalities (h^2) of certain variables are really poor. It seems that those variables represent quite technical events, like javelin and pole vault. In an exploratory sense, we are allowed to modify our concept by taking into account some extra factors. On the other hand, we could throw away the worst items, but then the structural validity would suffer. Why measure something in the first place and then abandon it later?

The best solution seems to require two more factors. One could think them as the technical components of the physical capacity, consisting of hand and feet skills. It must be noted, though, that five factors from ten variables is not very efficient or recommendable. In practice, we should have more indicators for the factors. All the same, this example serves as a compact demonstration of the methods.

The varimax rotation of the five-factor solution is presented with the columns sorted hierarchically and the loadings highlighted (see Scheme 5.4).

```
   1  1 SURVO 98  Tue Jan 04 08:07:48 2000                D:\V\    1000  100 0
  61 *
  62 *FACTA CORR.M,5
  63 *ROTATE FACT.M,5
  64 */LOADFACT / this gives the setup below
  65 *LIMITS=-0.7,-0.3,0.3,0.7,1
  66 *SHADOWS=7,1,0,1,7
  67 *SUMS=2 WIDE=1 POSDIR=1 COLUMNS=SORT
  68 *LOADM AFACT.M,12.123,CUR+1 / SORT=-1,0.3
  69 *A
  70 *             F3      F1      F2      F5      F4  Sumsqr
  71 *1500m     0.885   0.052  -0.137  -0.255  -0.031  0.871
  72 *100m     -0.322   0.875  -0.038   0.240  -0.120  0.942
  73 *400m      0.349   0.592  -0.096   0.199   0.006  0.521
  74 *Hi_jump  -0.106  -0.502   0.106   0.032   0.086  0.283
  75 *Shot_put -0.393  -0.148   0.747  -0.036  -0.110  0.747
  76 *Discus   -0.549  -0.141   0.660  -0.038   0.044  0.760
  77 *Pole_vlt -0.059   0.032  -0.320  -0.057  -0.097  0.119
  78 *Hurdles  -0.046   0.127   0.122   0.710  -0.193  0.574
  79 *L_jump   -0.101   0.062  -0.040   0.461   0.200  0.268
  80 *Javelin  -0.017  -0.150   0.109   0.013   0.791  0.661
  81 *Sumsqr    1.491   1.455   1.165   0.886   0.750  5.747
  82 *
```

Scheme 5.4. Rotated factor matrix of DECA, a five-factor solution sorted and highlighted.

The reliabilities are computed by the RELIAB module, programmed by the author, and presented in the Appendix A. It takes as parameters the correlation matrix (CORR.M) and the rotated factor matrix (AFACT.M). The results are shown for two different models, to test the assumption of the measurement error correlations (see Scheme 5.5). The matrices of residuals are also helpful in that sense, telling which measurement errors correlate mostly with each other. In this case, the differences between the results are negligible. The matrix of means, standard deviations and number of observations (MSN.M) is needed for Cronbach's alphas only.

The first three factor images are acceptable, but the fourth and the fifth factor are quite weak. The corresponding alpha values are very low, because we are far away from the assumptions of alpha (see Scheme 5.5).

According to Cronbach and Hartmann (1954), alpha should not be used here, since the sum of the item covariances is negative. This is most seriously reflected in the value of the traditional alpha: it tends negative, giving a reliability of -0.48 for the sum of the decathlon scores. It does not sound reasonable, however, as the sum of the scores is the scale which is used in decathlon ranking. Although this example might not be among the most typical ones from a psychometrician's point of view, for instance, the behaviour of alpha is unfortunately quite typical.

```
   1  1 SURVO 98  Tue Jan 04 08:08:03 2000                D:\V\    1000  100 0
*
*RELIAB CORR.M,AFACT.M,CUR+1 / MSN=MSN.M
*Reliabilities according to models E2 and E3:
*E2: errors do not correlate; E3: errors may correlate.
*F1\E2=0.8571  F1\E3=0.8564   (Cronbach’s alpha: 0.5598)
*F2\E2=0.8546  F2\E3=0.8534   (Cronbach’s alpha: 0.5974)
*F3\E2=0.9088  F3\E3=0.9089   (Cronbach’s alpha: 0.5746)
*F4\E2=0.6989  F4\E3=0.7000   (Cronbach’s alpha: 0.1916)
*F5\E2=0.7209  F5\E3=0.7205   (Cronbach’s alpha: 0.4369)
*Sum\E2=0.4815  Sum\E3=0.4829   (Cronbach’s alpha:-0.4826)
*LOADM RCOV.M,##.###,END+2 / Residual covariance matrix
*LOADM RCORR.M,##.###,END+2 / Residual correlation matrix
*LIMITS=-0.9,-0.2,-0.1,0.1,0.2,0.9,1 SHADOWS=7,8,1,0,1,8,7
*
```

Scheme 5.5. Reliabilities of the factor images and the unweighted sum of DECA.

The general reliability for the sum (Sum\E2) is about the same size as alpha but without the negative sign. Because the physical capacity is multidimensional, a plain sum can not necessarily be the optimal scale. A better alternative is provided by the factor scores. The matrix of coefficients is computed in Scheme 5.6 by the /FCOEFF sucro, and the factor scores are named according to the interpretation, by employing the column labels of the coefficient matrix.

```
   1  1 SURVO 98  Tue Jan 04 08:08:18 2000                D:\V\    1000  100 0
*.......................................................................
*/FCOEFF AFACT.M,MSN.M,FCOEFF.M
*Use FCOEFF.M for factor scores by LINCO <data>,FCOEFF.M(F1,F2,...)
*MAT FCOEFF.M(0,1)="Speed"
*MAT FCOEFF.M(0,2)="Force"
*MAT FCOEFF.M(0,3)="Stren"
*MAT FCOEFF.M(0,4)="Tech1" / technical skills: hands
*MAT FCOEFF.M(0,5)="Tech2" / technical skills: feet
*
```

Scheme 5.6. Computing and naming the factor scores of DECA.

The reliabilities of the factor scores are computed in the same way as earlier, but now weighting the true variation and the observed variation with the factor score coefficients. The most reliable scores seem to be speed and strength, while the technical scores are not so reliable. The alpha values are completely useless (see Scheme 5.7).

```
   1  1 SURVO 98  Tue Jan 04 08:08:32 2000                D:\V\    1000  100 0
*
*RELIAB CORR.M,AFACT.M,CUR+1 / MSN=MSN.M WEIGHT=FCOEFF.M
*Reliabilities according to models E2 and E3: (weighted by FCOEFF.M)
*E2: errors do not correlate; E3: errors may correlate.
*Speed\E2=0.9147  Speed\E3=0.9147   (Cronbach’s alpha:-0.1511)
*Force\E2=0.7664  Force\E3=0.7664   (Cronbach’s alpha:-0.0403)
*Stren\E2=0.8505  Stren\E3=0.8505   (Cronbach’s alpha: 0.1549)
*Tech1\E2=0.6681  Tech1\E3=0.6680   (Cronbach’s alpha: 0.0484)
*Tech2\E2=0.6316  Tech2\E3=0.6315   (Cronbach’s alpha: 0.1537)
*LOADM RCOV.M,##.###,END+2 / Residual covariance matrix
*LOADM RCORR.M,##.###,END+2 / Residual correlation matrix
*LIMITS=-0.9,-0.2,-0.1,0.1,0.2,0.9,1 SHADOWS=7,8,1,0,1,8,7
*
```

Scheme 5.7. Reliabilities of the factor scores of DECA.

The general reliabilities can also be computed from the formulas presented in chapter 3 by employing the matrix interpreter of Survo as follows.

We denote the correlation matrix by R and the rotated factor matrix by B. In the case of the factor images the scale coefficient matrix A is equal to the matrix B. Applying the formula (3.6) gives the reliabilities of the factor images as a diagonal matrix REL1 (Scheme 5.8).

```
   1  1 SURVO 98  Tue Jan 04 08:08:49 2000            D:\V\   1000  100 0
*.......................................................................
*MAT R!=CORR.M      / *R~R(DECA) S10*10
*MAT B!=AFACT.M     / *B~A 10*5
*MAT A=B           / *A~B 10*5
*MAT DIM A /* rowA=10 colA=5
*MAT I!=IDN(colA,colA)
*
*MAT REL1=INV(I+DIAG(A'*DIAG(R-B*B')*A)*INV(DIAG(A'*B*B'*A)))
*MAT LOAD REL1 ##.#### CUR+2
*
*MATRIX REL1
*INV(I+DIAG(B'*DIAG(R-B*B')*B)*INV(DIAG(B'*B*B'*B)))
*///           F1       F2       F3       F4       F5
*F1        0.8571  0.0000  0.0000  0.0000  0.0000
*F2        0.0000  0.8546  0.0000  0.0000  0.0000
*F3        0.0000  0.0000  0.9088  0.0000  0.0000
*F4        0.0000  0.0000  0.0000  0.6989  0.0000
*F5        0.0000  0.0000  0.0000  0.0000  0.7209
*
```

Scheme 5.8. Computing the reliabilities of the factor images using the matrix interpreter.

The reliabilities of the factor scores are obtained with the same formula and the same expression, replacing the matrix A with the factor score coefficients computed earlier. The constant included in the matrix A is not needed, so it is first removed (Scheme 5.9).

```
   1  1 SURVO 98  Tue Jan 04 08:09:04 2000            D:\V\   1000  100 0
*
*MAT F!=FCOEFF.M    / *F~FCOEFF 11*5
*MAT DIM F /* rowF=11 colF=5
*MAT REM Remove the constant from the first row of F:
*MAT A!=F(2:rowF,*)
*
*MAT REL2=INV(I+DIAG(A'*DIAG(R-B*B')*A)*INV(DIAG(A'*B*B'*A)))
*MAT LOAD REL2 ##.#### CUR+2
*
*MATRIX REL2
*INV(I+DIAG(A'*DIAG(R-B*B')*A)*INV(DIAG(A'*B*B'*A)))
*///        Speed   Force   Stren   Tech1   Tech2
*Speed     0.9147  0.0000  0.0000  0.0000  0.0000
*Force     0.0000  0.7664  0.0000  0.0000  0.0000
*Stren     0.0000  0.0000  0.8505  0.0000  0.0000
*Tech1     0.0000  0.0000  0.0000  0.6681  0.0000
*Tech2     0.0000  0.0000  0.0000  0.0000  0.6316
*
```

Scheme 5.9. Computing the reliabilities of the factor scores using the matrix interpreter.

The reliabilities of the factor scores can be computed even without the coefficient matrix, by applying the simple formula (4.20), but since it requires inverting the correlation matrix, the results are not identical. There is some noise in the non-diagonal elements of the matrix REL3 (Scheme 5.10), and the reliabilities are somewhat lower, compared with the matrix REL2 in Scheme 5.9.

```
   1  1 SURVO 98  Tue Jan 04 08:09:17 2000                D:\V\   1000  100 0
 152 *
 153 *MAT REL3=B'*INV(R)*B
 154 *MAT LOAD REL3 ##.#### CUR+2
 155 *
 156 *MATRIX REL3
 157 *B'*INV(R)*B
 158 *///           F1      F2      F3      F4      F5
 159 *F1        0.8989 -0.0463 -0.0240 -0.0491  0.0770
 160 *F2       -0.0463  0.7342 -0.1053 -0.0091 -0.0269
 161 *F3       -0.0240 -0.1053  0.8455 -0.0045 -0.0864
 162 *F4       -0.0491 -0.0091 -0.0045  0.6667 -0.0018
 163 *F5        0.0770 -0.0269 -0.0864 -0.0018  0.6161
 164 *
```

Scheme 5.10. Computing the reliabilities of the factor scores applying the formula (4.20).

With the reliabilities, we can assess the accuracy of the measurements, by calculating the standard error of measurement for each factor score. First, we have to compute the linear combinations and save them as new variables in the data set DECA. This takes place by the LINCO operation (see Scheme 5.11).

```
   1  1 SURVO 98  Tue Jan 04 08:09:28 2000                D:\V\   1000  100 0
 165 *.......................................................................
 166 *LINCO DECA,FCOEFF.M
 167 *CORR DECA / VARS=Speed,Force,Stren,Tech1,Tech2
 168 */COV
 169 *
 170 *LOADM MSN.M,(C6),CUR+1
 171 *MSN(DECA)
 172 *           mean stddev      N
 173 *Speed     0.000 0.9481     48
 174 *Force     0.000 0.8569     48
 175 *Stren     0.000 0.9195     48
 176 *Tech1     0.000 0.8166     48
 177 *Tech2     0.000 0.7850     48
 178 *
```

Scheme 5.11. Computing the factor score variables and their basic statistics.

The factor scores are usually centered. They are also roughly uncorrelated as we made an orthogonal factor rotation. The corresponding covariance matrix COV.M, computed with the /COV sucro on line 168 of Scheme 5.11, seems to be equal to the matrix REL3 (see Scheme 5.12). This is an implication of the simple form of the formula (4.20).

```
   1  1 SURVO 98  Tue Jan 04 08:09:40 2000              D:\V\    1000  100 0
 178 *
 179 *MAT LOAD COV.M ##.#### CUR+2
 180 *
 181 *MATRIX COV.M
 182 *&D*R(DECA)*&D
 183 */// 	       Speed   Force   Stren   Tech1   Tech2
 184 *Speed     0.8989 -0.0463 -0.0240 -0.0491  0.0770
 185 *Force    -0.0463  0.7342 -0.1053 -0.0091 -0.0269
 186 *Stren    -0.0240 -0.1053  0.8455 -0.0045 -0.0864
 187 *Tech1    -0.0491 -0.0091 -0.0045  0.6668 -0.0018
 188 *Tech2     0.0770 -0.0269 -0.0864 -0.0018  0.6162
 189 *
```

Scheme 5.12. Covariances of the factor score variables.

From the definition of reliability, it follows that the variance of the measurement errors is

$$\sigma_\varepsilon^2 = \sigma_u{}^2\,(1-\rho_{uu})\,, \qquad (5.1)$$

where $\sigma_u{}^2$ is the variance of the scale and ρ_{uu} is the reliability of the scale. The square root of (5.1) is defined as the standard error of measurement. We have the reliabilities as a diagonal matrix REL2, so we can form a column vector RLB of those diagonal elements (Scheme 5.13).

```
   1  1 SURVO 98  Tue Jan 04 08:10:23 2000              D:\V\    1000  100 0
 190 *......................................................................
 191 *MAT RLB=VD(REL2)   / *RLB~VD(INV(I+DIAG(A'*DIAG(R-B*B')*A)*INV(DIAG(A'*
 192 *MAT RLB(0,1)="reliab"
 193 *MAT LOAD RLB
 194 *MATRIX RLB
 195 *VD(INV(I+DIAG(A'*DIAG(R-B*B')*A)*INV(DIAG(A'*B*B'*A))))
 196 *///        reliab
 197 *Speed    0.914733
 198 *Force    0.766366
 199 *Stren    0.850474
 200 *Tech1    0.668087
 201 *Tech2    0.631571
 202 *
```

Scheme 5.13. Taking a column vector out of the diagonal matrix of reliabilities.

Extracting the second column of the matrix MSN.M, computed above by the CORR module, we have the standard deviations of the scales as another column vector STD. Transforming its values by applying the formula (5.1) and taking the square root, we have the standard errors of measurement as a column vector SEM (Scheme 5.14).

```
   1  1 SURVO 98  Tue Jan 04 08:10:34 2000              D:\V\    1000   100 0
 *
 *MAT STD=MSN.M(*,stddev)
 *MAT TRANSFORM STD BY RLB AND X#*SQRT(1-Y#)
 *MAT STD(0,1)="stderr"
 *MAT SEM=STD        / *SEM~T(STD_by_RLB_and_X#*SQRT(1-Y#)) 5*1
 *MAT LOAD SEM
 *MATRIX SEM
 *T(STD_by_RLB_and_X#*SQRT(1-Y#))
 *///        stderr
 *Speed    0.276848
 *Force    0.414179
 *Stren    0.355558
 *Tech1    0.470430
 *Tech2    0.476459
 *
```

Scheme 5.14. Computing the standard errors of measurement for the factor scores.

The most accurate score is Speed while the worst one is Tech2. We compute their basic statistics (see Scheme 5.15).

```
   1  1 SURVO 98  Tue Jan 04 08:10:52 2000              D:\V\    1000   100 0
 *...........................................................................
 *STAT DECA CUR+1 / VARS=Speed,Tech2
 *Basic statistics: DECA N=48
 *Variable: Speed
 *min=-2.070568 in obs.#15 (Avilov)
 *max=1.862117 in obs.#21 (Stroot)
 *mean=0        stddev=0.948094 skewness=-0.084115 kurtosis=-0.556675
 *lower_Q=-0.68     median=0.06     upper_Q=0.68
 *up.limit      f     % class width=0.4
 *    -2        1   2.1 *
 *    -1.6      1   2.1 *
 *    -1.2      2   4.2 **
 *    -0.8      7  14.6 *******
 *    -0.4      5  10.4 *****
 *     0        7  14.6 *******
 *     0.4     10  20.8 **********
 *     0.8      5  10.4 *****
 *     1.2      5  10.4 *****
 *     1.6      3   6.3 ***
 *     2        2   4.2 **
 *
 *Variable: Tech2
 *min=-1.616556 in obs.#41 (Brigham)
 *max=1.44267  in obs.#8 (Katus)
 *mean=0        stddev=0.784962 skewness=0.020944 kurtosis=-0.994043
 *lower_Q=-0.62     median=-0.016667 upper_Q=0.75
 *up.limit      f     % class width=0.2
 *    -1.6      1   2.1 *
 *    -1.4      0   0.0
 *    -1.2      2   4.2 **
 *    -1        2   4.2 **
 *    -0.8      3   6.3 ***
 *    -0.6      5  10.4 *****
 *    -0.4      5  10.4 *****
 *    -0.2      1   2.1 *
 *     0        6  12.5 ******
 *     0.2      4   8.3 ****
 *     0.4      3   6.3 ***
 *     0.6      3   6.3 ***
 *     0.8      2   4.2 **
 *     1        5  10.4 *****
 *     1.2      3   6.3 ***
 *     1.4      2   4.2 **
 *     1.6      1   2.1 *
```

Scheme 5.15. Basic statistics of the factor scores Speed and Tech2.

The frequency distributions are unfair, compared with the measurement accuracy. Let us plot histograms (see Figures 5.1 and 5.2) instead, using as class widths the rounded standard errors of measurements, namely 0.3 and 0.5 (Scheme 5.16).

```
   1  1 SURVO 98  Tue Jan 04 08:11:05 2000                D:\V\    1000  100 0
 261 *
 262 *..........................................................................
 263 *
 264 *GHISTO DECA Speed END+2 / Speed=-2.25(0.3)2.25
 265 *  YSCALE=0(1)13   XSCALE=-3(1)3   FIT=NORMAL
 266 *GHISTO DECA Tech2 END+2 / Tech2=-1.75(0.5)1.75
 267 *
```

Scheme 5.16. Plotting histograms of the factor scores Speed and Tech2.

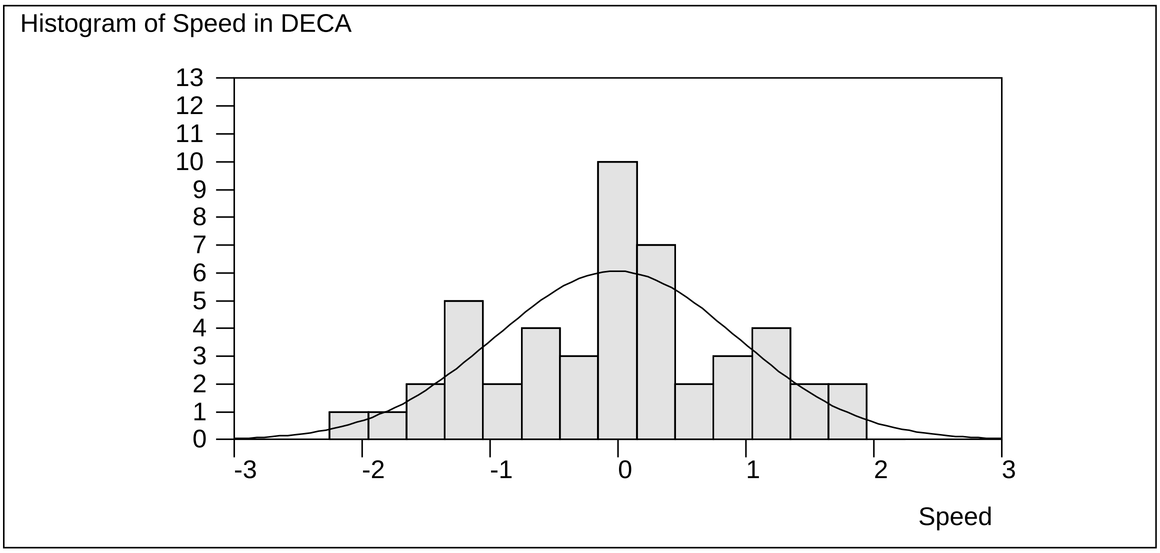

Figure 5.1. Histogram of the factor score Speed.

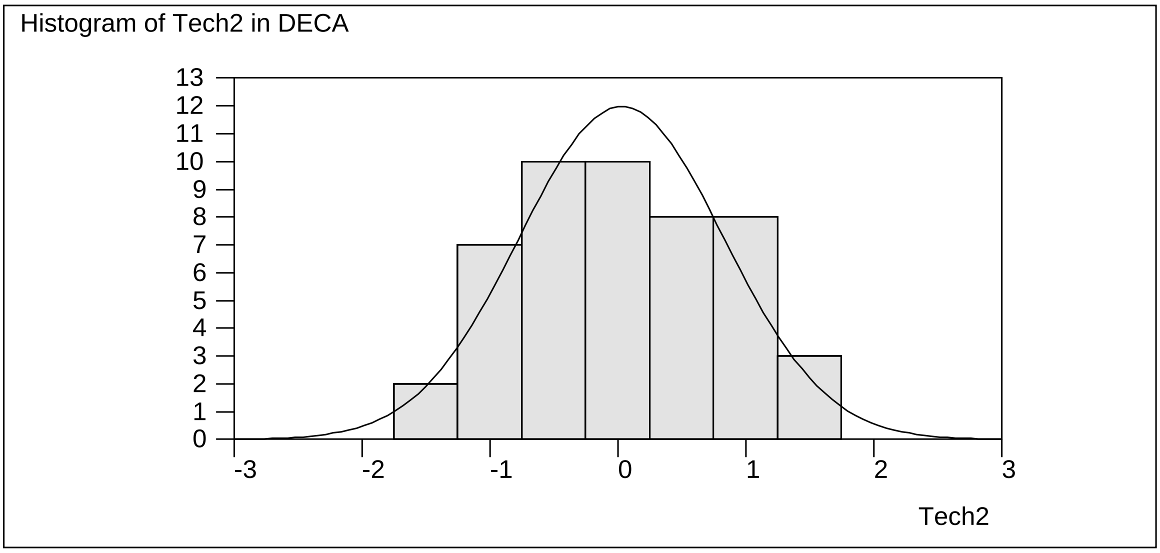

Figure 5.2. Histogram of the factor score Tech2.

The normal distribution fits rather well for both scores (Scheme 5.17).

```
   1   1 SURVO 98   Tue Jan 04 08:11:46 2000              D:\V\    1000   100 0
 *Frequency distribution of Speed in DECA: N=48
 *
 *Class midpoint  f     %   Sum     %     e       e    f     X^2
 *        -2.10    1   2.1     1   2.1   1.0
 *        -1.80    1   2.1     2   4.2   1.0
 *        -1.50    2   4.2     4   8.3   1.8
 *        -1.20    5  10.4     9  18.8   2.8    6.6    9     0.9
 *        -0.90    2   4.2    11  22.9   3.9
 *        -0.60    4   8.3    15  31.3   5.0    8.9    6     0.9
 *        -0.30    3   6.3    18  37.5   5.8    5.8    3     1.3
 *         0.00   10  20.8    28  58.3   6.0    6.0   10     2.6
 *         0.30    7  14.6    35  72.9   5.7    5.7    7     0.3
 *         0.60    2   4.2    37  77.1   4.9
 *         0.90    3   6.3    40  83.3   3.8    8.7    5     1.6
 *         1.20    4   8.3    44  91.7   2.7
 *         1.50    2   4.2    46  95.8   1.7
 *         1.80    2   4.2    48 100.0   1.0
 *         2.10    0   0.0    48 100.0   0.9    6.3    8     0.5
 *Mean=-0.012500 Std.dev.=0.948601
 *Fitted by NORMAL(-0.0125,0.8998) distribution
 *Chi-square=8.120 df=4 P=0.0873
 *
 *Frequency distribution of Tech2 in DECA: N=48
 *
 *Class midpoint  f     %   Sum     %     e       e    f     X^2
 *        -1.50    2   4.2     2   4.2   2.4
 *        -1.00    7  14.6     9  18.8   5.4    7.8    9     0.2
 *        -0.50   10  20.8    19  39.6   9.6    9.6   10     0.0
 *         0.00   10  20.8    29  60.4  11.8   11.8   10     0.3
 *         0.50    8  16.7    37  77.1  10.1   10.1    8     0.4
 *         1.00    8  16.7    45  93.8   5.9
 *         1.50    3   6.3    48 100.0   2.8    8.7   11     0.6
 *Mean=0.031250 Std.dev.=0.799780
 *Fitted by NORMAL(0.03125,0.6396) distribution
 *Chi-square=1.507 df=2 P=0.4708
```

Scheme 5.17. Frequency distributions of the factor scores Speed and Tech2.

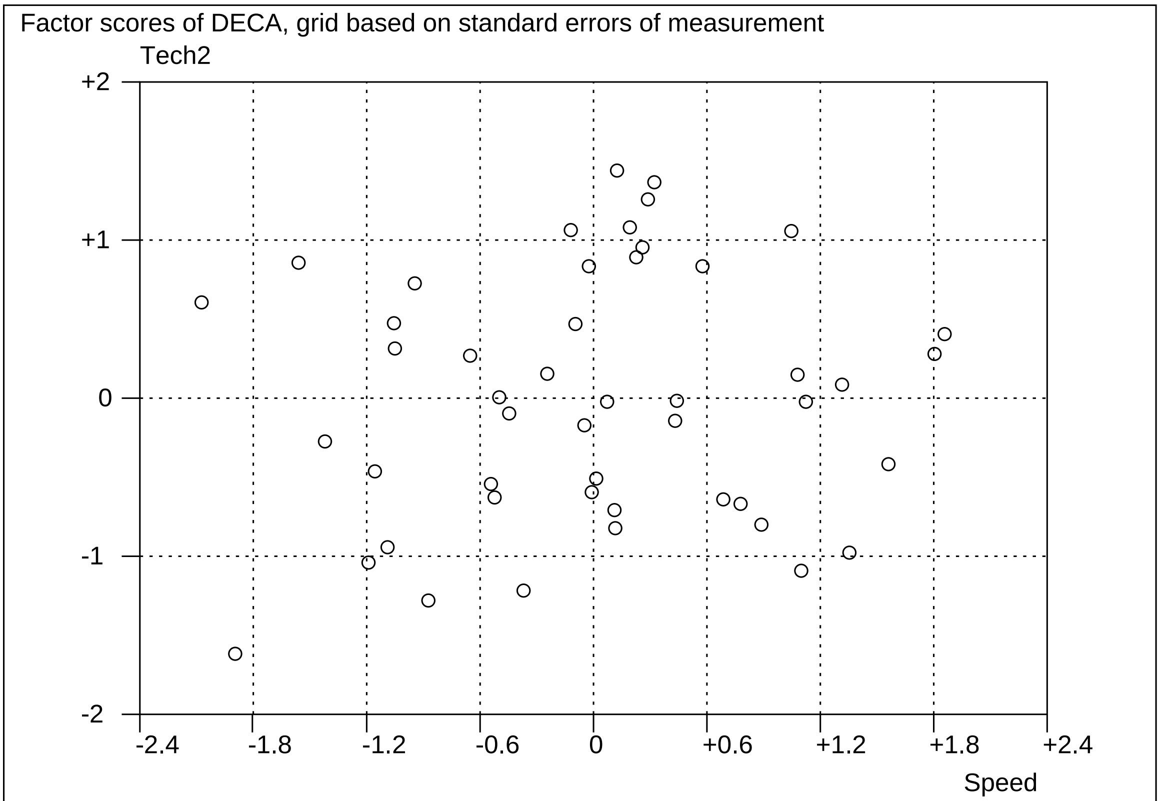

Figure 5.3. Scatter diagram of the factor scores Speed and Tech2.

The standard error of measurement gives the accuracy with which we can discriminate among the observations. A scatter diagram of Speed and Tech2 (Figure 5.3) gives an idea of this. The more accurate the measurement, the better we can make a clear distinction between the athletes.

Operating with the standard errors of measurement requires a proper measure of reliability. The previous example shows that Tarkkonen's measure is well suited for these purposes.

5.2 Assessing the structural validity of a factor model

In the following example, we consider a re-evaluation of misleading experiments with artificial factor structures that were originally carried out by *Ivor Francis* (1973) and quoted by *George Seber* (1984, 222–235) in a textbook of multivariate statistics. Some of the erratic conclusions of Seber have been rectified by Mustonen and Vehkalahti (1997). A more detailed analysis appears in Mustonen (1995, 106–112). There is not much to add in these studies, but it is still interesting to briefly consider the structural validity of the factor model, through the use of the reliabilities of the factor images.

The original factor pattern (see Scheme 5.18) is given as a matrix consisting of loadings of three common factors (F1, F2, F3) and the standard deviations of the unique factors (PSI).

```
   1  1 SURVO 98  Mon Feb 14 16:58:36 2000                D:\V\   1000  100 0
 *
 *MATRIX FRANCISV / MODEL V from Francis, I. (1973). Factor analysis:
 */// 	 F1    F2    F3   PSI  /  its purpose, practice, and packaged
 *X1    10     7     4    15     programs. Invited paper, American
 *X2    10     7     4    15     Statistical Association, New York,
 *X3    10     7     4    15     December 1973 (quoted by Seber [1984]).
 *X4    10     7     4    15
 *X5    10     7     0    15
 *X6    10     7     0    20
 *X7    10     7     0    20
 *X8    10     0     0    20
 *X9    10     0     0    20
 *X10   10     0     0    20
 *
 *MAT SAVE FRANCISV
 *MAT B!=FRANCISV(*,F1:F3)
 *MAT PSI!=DV(FRANCISV(*,PSI))
 *MAT S=B*B'+PSI^2                / covariance matrix
 *MAT D=DIAG(S)^(-0.5)
 *MAT R=D*S*D                     / correlation matrix
 *
```

Scheme 5.18. Saving and computing the basic matrices of model V.

Francis and Seber believe that the given pattern represents a simple structure which should be reproduced by factor analysis. This reasoning is incorrect, since the basic conditions of simple structure are not met. In addition, the correlation matrix (see Scheme 5.19) would seem to support only one factor.

```
   1  1 SURVO 98  Mon Feb 14 17:02:12 2000                D:\V\    1000   100 0
*
*MAT LOAD R #.### CUR+1
*MATRIX R
*DIAG(B*B'+PSI^2)^(-0.5)*(B*B'+PSI^2)*DIAG(B*B'+PSI^2)^(-0.5)
*///          X1    X2    X3    X4    X5    X6    X7    X8    X9   X10
*X1        1.000 0.423 0.423 0.423 0.390 0.322 0.322 0.226 0.226 0.226
*X2        0.423 1.000 0.423 0.423 0.390 0.322 0.322 0.226 0.226 0.226
*X3        0.423 0.423 1.000 0.423 0.390 0.322 0.322 0.226 0.226 0.226
*X4        0.423 0.423 0.423 1.000 0.390 0.322 0.322 0.226 0.226 0.226
*X5        0.390 0.390 0.390 0.390 1.000 0.329 0.329 0.231 0.231 0.231
*X6        0.322 0.322 0.322 0.322 0.329 1.000 0.271 0.191 0.191 0.191
*X7        0.322 0.322 0.322 0.322 0.329 0.271 1.000 0.191 0.191 0.191
*X8        0.226 0.226 0.226 0.226 0.231 0.191 0.191 1.000 0.200 0.200
*X9        0.226 0.226 0.226 0.226 0.231 0.191 0.191 0.200 1.000 0.200
*X10       0.226 0.226 0.226 0.226 0.231 0.191 0.191 0.200 0.200 1.000
*
```

Scheme 5.19. Correlation matrix of model V.

Following the experiments of Francis, we compute the maximum likelihood factor analysis of three factors (Scheme 5.20), but obviously the first factor is the only one we can argue about.

```
   1  1 SURVO 98  Mon Feb 14 17:23:40 2000                D:\V\    1000   100 0
*
*FACTA R,3,CUR+1
*Factor analysis: Maximum Likelihood (ML) solution
*Factor matrix
*              F1      F2      F3    h^2
*X1         0.643 -0.080 -0.051  0.423
*X2         0.643 -0.080 -0.051  0.423
*X3         0.643 -0.080 -0.051  0.423
*X4         0.643 -0.080 -0.051  0.423
*X5         0.618  0.010  0.129  0.398
*X6         0.510  0.008  0.106  0.271
*X7         0.510  0.008  0.106  0.271
*X8         0.378  0.236 -0.037  0.200
*X9         0.378  0.236 -0.037  0.200
*X10        0.378  0.236 -0.037  0.200
*
```

Scheme 5.20. Three-factor maximum likelihood solution of model V.

The reliabilities of the factor images in Scheme 5.21 reveal the rest: the second and the third factor are completely artificial, although with a huge number of observations, they could be possibly identified (Mustonen 1995, 111–112).

```
   1  1 SURVO 98  Mon Feb 14 17:27:50 2000                D:\V\    1000   100 0
*
*RELIAB R,FACT.M,CUR+1 / MSN=*
*Reliabilities according to models E2 and E3:
*E2: errors do not correlate; E3: errors may correlate.
*F1\E2=0.8240  F1\E3=0.8240   (Cronbach's alpha: 0.8048)
*F2\E2=0.2247  F2\E3=0.2247   (Cronbach's alpha:-0.0073)
*F3\E2=0.0807  F3\E3=0.0807   (Cronbach's alpha:-0.4236)
*Sum\E2=0.8095  Sum\E3=0.8095   (Cronbach's alpha: 0.7983)
*
```

Scheme 5.21. Reliabilities of the factor images and the unweighted sum of model V.

In practice, this kind of result would indicate a serious lack of the structural validity of the model. The correct number of factors in this case would be one. The reliability of the unweighted sum also reflects this, as it is quite close to the reliability of the first factor image (see Scheme 5.21).

Cronbach's alpha works in a satisfactory way in the case of the sum (Scheme 5.21), but the artificial dimensions cause serious trouble for alpha, when it is applied for the factor images.

Francis conducted several other experiments of the same kind. The major problem in all of them is that he misunderstood the simple structure principle in factor analysis. Ironically, the motivation behind those studies was that Francis had found factor analysis to be "the most misunderstood and misapplied statistical procedure" (Francis 1974, 9).

5.3 Separating the orientations and values

Jarmo Liukkonen and *Esko Leskinen* (1999) have studied the reliability and validity of scores from the children's version of the perception of success questionnaire. The data consisted of responses from 557 14-year-old Finnish male soccer players. The goal was to demonstrate alternative methods for analyzing the reliability and validity in the context of sport psychology. Also in that field, Cronbach's alpha has dominated recent studies (Liukkonen and Leskinen 1999).

The study concentrated on two particular subscales, the Task Orientation and the Ego Orientation. Because the authors found the scales to be orthogonal, they applied confirmatory factor analysis to the items of both scales separately. This was done to evaluate the item level reliabilities as well as the reliabilities of the factor scores (Liukkonen and Leskinen 1999, 651–660).

The referred article is an example of a rather typical application. The multidimensional nature of the trait under study is hidden by analyzing several unidimensional factors independently of each other. Although the confirmatory factor analysis would allow for more general methods, the assessment of the reliabilities is based on old-fashioned models which essentially assume one factor and scales that are assumed to be internally consistent. The authors put the main emphasis on the item analysis, reporting the reliabilities and validities of the single items, estimated from the confirmatory factor models. In addition, they report the reliabilities of the factor scores, defined through squared multiple correlations by Bollen (1989, 221). Cronbach's alphas are also reported, although they are admitted to be invalid (Liukkonen and Leskinen 1999, 655–660).

Let us now see, how the Task Orientation and the Ego Orientation scales are analyzed by using the exploratory factor analysis with proper tools of assessing the reliabilities of the measurement scales.

The perception of success is assumed to have two dimensions. Therefore we proceed with a two-factor maximum likelihood factor analysis from the correlation matrix, which is typed and saved in an edit field (Scheme 5.22).

```
 1  1 SURVO 98  Sun Feb 13 13:30:50 2000               D:\V\   1000  100 0
*
*MATRIX EGOTASK
*Liukkonen, J., & Leskinen, E. (1999). The reliability and validity of
* scores from the children’s version of the perception of success
* questionnaire. Educational and Psychological Measurement, 59, 651-664.
*Correlations among the items (n=553-557)
*/// E1   E2   E5   E6   E9   E12  T3   T4   T7   T8   T10  T11
aE1  1    -    -    -    -    -    -    -    -    -    -    -
*E2  .56  1    -    -    -    -    -    -    -    -    -    -
*E5  .49  .55  1    -    -    -    -    -    -    -    -    -
*E6  .36  .50  .44  1    -    -    -    -    -    -    -    -
*E9  .34  .42  .50  .43  1    -    -    -    -    -    -    -
*E12 .44  .58  .60  .54  .52  1    -    -    -    -    -    -
*T3  .05  .01  .01  -.04 .05  -.10 1    -    -    -    -    -
*T4  .04  .12  .16  -.02 .18  -.01 .49  1    -    -    -    -
*T7  .16  .13  .20  .08  .20  .06  .37  .47  1    -    -    -
*T8  .16  .16  .20  .01  .19  .07  .36  .53  .52  1    -    -
*T10 .01  .00  -.02 -.05 .04  -.16 .59  .43  .42  .39  1    -
bT11 .10  .15  .16  .02  .25  .13  .40  .53  .44  .49  .45  1
*
*AAA AAAA AAAA AAAA AAAA AAAA AAAA AAAA AAAA AAAA AAAA AAAA AAAA
*FORM a,b,b+2,L / fill the incomplete lines by ’-’
*MAT SAVE EGOTASK / save and complete the symmetric matrix
*MAT DIM EGOTASK /* rowEGOTASK=12 colEGOTASK=12
*MAT EGOTASK!=EGOTASK+EGOTASK’-IDN(rowEGOTASK,colEGOTASK)
*/MAKEMSN MSN.M EGOTASK,0,1,555 / create a matrix of moments
*
```

Scheme 5.22. Saving the basic matrices of the orientation scales.

We apply the varimax rotation, and present the factor matrix highlighted and sorted in Scheme 5.23. The structure appears to be quite simple.

```
 1  1 SURVO 98  Sun Feb 13 13:37:54 2000               D:\V\   1000  100 0
*
*FACTA EGOTASK,2 / ML factor analysis
*ROTATE FACT.M,2 / Varimax rotation
*/LOADFACT       / Presentation
*LIMITS=-0.7,-0.3,0.3,0.7,1
*SHADOWS=7,1,0,1,7
*SUMS=2 WIDE=1 POSDIR=1 COLUMNS=SORT
*LOADM AFACT.M,12.123,CUR+1 / SORT=-1,0.3
*Factor_matrix
*              F1     F2  Sumsqr
*E12        0.809 -0.069  0.659
*E5         0.747  0.117  0.571
*E2         0.741  0.080  0.555
*E6         0.643 -0.055  0.417
*E9         0.624  0.170  0.418
*E1         0.608  0.078  0.376
*T4         0.065  0.733  0.542
*T10       -0.110  0.682  0.477
*T11        0.153  0.674  0.478
*T8         0.148  0.670  0.470
*T3        -0.071  0.660  0.440
*T7         0.147  0.642  0.434
*Sumsqr     3.021  2.816  5.837
*
```

Scheme 5.23. Factor analysis and rotated factor structure of the orientation scales.

The reliabilities of the factor images are 0.8651 and 0.8463 (see Scheme 5.24). When the items are assumed to be standardized (as no more specific information was available in the paper), the reliabilities of the factor scores are about the same size, 0.8668 for the Ego scale and 0.8423 for the Task scale. The reliability of an unweighted sum would be 0.8650. Cronbach’s alphas are lower in each case.

```
  1  1 SURVO 98  Sun Feb 13 15:14:59 2000               D:\V\   1000  100 0
*
*RELIAB EGOTASK,AFACT.M,CUR+1 / MSN=MSN.M
*Reliabilities according to models E2 and E3:
*E2: errors do not correlate; E3: errors may correlate.
* F1\E2=0.8651   F1\E3=0.8653   (Cronbach’s alpha: 0.7847)
* F2\E2=0.8463   F2\E3=0.8463   (Cronbach’s alpha: 0.7706)
*Sum\E2=0.8650  Sum\E3=0.8638   (Cronbach’s alpha: 0.8045)
*
*/FCOEFF / create a matrix of factor score coefficients
*Use FCOEFF.M for factor scores by LINCO <data>,FCOEFF.M(F1,F2,...)
*MAT FCOEFF.M(0,1)="Ego"
*MAT FCOEFF.M(0,2)="Task"
*...................................................................
*RELIAB EGOTASK,AFACT.M,CUR+1 / MSN=MSN.M WEIGHT=FCOEFF.M
*Reliabilities according to models E2 and E3: (weighted by FCOEFF.M)
*E2: errors do not correlate; E3: errors may correlate.
* Ego\E2=0.8668               (Cronbach’s alpha: 0.7574)
*Task\E2=0.8423               (Cronbach’s alpha: 0.7568)
*
```

Scheme 5.24. Reliabilities of the orientation scales.

Liukkonen and Leskinen (1999) report somewhat higher alphas, but they are based on two independent, unidimensional scales. We can achieve the same results by analyzing the scales separately (see Scheme 5.25).

```
  1  1 SURVO 98  Sun Feb 13 15:58:13 2000               D:\V\   1000  100 0
*...................................................................
*MAT EGO=EGOTASK(E1:E12,E1:E12)   / separate the correlations
*MAT TASK=EGOTASK(T3:T11,T3:T11)  /   of the two scales
*FACTA EGO,1 / factor analysis of the Ego factor
*MAT FE=FACT.M      / *FE~F 6*1
*FACTA TASK,1 / factor analysis of the Task factor
*MAT FT=FACT.M      / *FT~F 6*1
*
*RELIAB EGO,FE,CUR+1           / MSN=*
* F1\E2=0.8573                 (Cronbach’s alpha: 0.8506)
*Sum\E2=0.8515                 (Cronbach’s alpha: 0.8495)
*RELIAB TASK,FT,CUR+1
* F1\E2=0.8372                 (Cronbach’s alpha: 0.8358)
*Sum\E2=0.8359                 (Cronbach’s alpha: 0.8356)
*
```

Scheme 5.25. Reliabilities of the orientation scales analyzed separately.

The alphas for the unweighted sums in Scheme 5.25 coincide with the figures given by Liukkonen and Leskinen (1999, 660). As usual, they are lower than the values of the general measure. When the scales are analyzed separately, the general reliabilities decrease a little, but not even remarkably. On the other hand, the alphas are clearly increased. This is probably the prime motivation of separating the scales. However, it is not a sound justification.

The idea of factoring the items separately or jointly was already demonstrated by Heise and Bohrnstedt (1970), who introduced the reliability coefficient omega (2.19) in the same paper. Omega was meant to be a better alternative for alpha, especially in the context of factor analysis. Let us briefly re-analyze their factor structure of the values of college students (Heise and Bohrnstedt 1970, 118–124), see Scheme 5.26.

```
   1  1 SURVO 98  Mon Feb 14 09:44:49 2000               D:\V\    1000  100 0
   1 *
   2 *Example of Religiosity and Fatalism by Heise & Bohrnstedt (1970).
   3 *MATRIX CRF ///
   4 a 0.928      -     -     -      -      -     -     -     -     -     -
   5 * 0.429  0.698     -     -      -      -     -     -     -     -     -
   6 * 0.438  0.370 0.795     -      -      -     -     -     -     -     -
   7 * 0.383  0.299 0.354 0.894      -      -     -     -     -     -     -
   8 * 0.428  0.340 0.358 0.396  0.671      -     -     -     -     -     -
   9 * 0.479  0.382 0.435 0.519  0.561  0.994     -     -     -     -     -
  10 * 0.114  0.112 0.139 0.010  0.009  0.109 0.636     -     -     -     -
  11 * 0.025 -0.016 0.003 0.005 -0.002  0.019 0.169 0.784     -     -     -
  12 * 0.046  0.018 0.003 0.000  0.024 -0.012 0.147 0.223 0.660     -     -
  13 *-0.025  0.005 0.003 0.000 -0.019  0.019 0.110 0.101 0.118 0.541     -
  14 b 0.046  0.015 0.004 0.004  0.004  0.049 0.170 0.120 0.134 0.245 0.635
  15 *
  16 *11.111 11.111 1.111 1.111 11.111 11.111 1.111 1.111 1.111 1.111 1.111
  17 *FORM a,b,b+2,L / fill the incomplete lines
  18 *MAT SAVE CRF   / save and complete the covariance matrix
  19 *MAT CLABELS Item TO CRF
  20 *MAT CRF!=CRF+CRF’-DIAG(CRF)
  21 *MAT RF=DIAG(CRF)^(-0.5)*CRF*DIAG(CRF)^(-0.5) / compute the correlations
  22 *
```

Scheme 5.26. Saving the basic matrices of the value scales.

Heise and Bohrnstedt used several different factoring methods. One of them was the maximum likelihood method, which we prefer here. The factor matrix, given in Scheme 5.27, is very simple, and no rotation is needed to reveal the structure.

```
   1  1 SURVO 98  Mon Feb 14 09:49:20 2000               D:\V\    1000  100 0
  22 *
  23 *FACTA RF 2 CUR+1
  24 *Factor analysis: Maximum Likelihood (ML) solution
  25 *Factor matrix
  26 *             F1      F2    h^2
  27 *Item1     0.694  0.030  0.482
  28 *Item2     0.647  0.020  0.419
  29 *Item3     0.659  0.017  0.435
  30 *Item4     0.639 -0.060  0.412
  31 *Item5     0.804 -0.084  0.654
  32 *Item6     0.792 -0.001  0.628
  33 *Item7     0.155  0.466  0.241
  34 *Item8     0.022  0.390  0.152
  35 *Item9     0.037  0.425  0.182
  36 *Item10    0.006  0.538  0.290
  37 *Item11    0.055  0.598  0.360
  38 *
```

Scheme 5.27. Maximum likelihood factor analysis of the value scales.

As the example included the covariances of the items, we can build the matrix of means, standard deviations and number of observations with the help of the variances in the diagonal of the matrix CRF (see Scheme 5.28). These computations are needed for the alphas, to compare them with the values given by Heise and Bohrnstedt.

```
  1  1 SURVO 98  Mon Feb 14 10:27:58 2000                D:\V\   1000  100 0
*
*MAT DIM CRF /* rowCRF=11 colCRF=11
*MAT MRF=ZER(rowCRF,3)
*MAT S=VD(CRF)       / *S~VD(CRF) 11*1
*MAT TRANSFORM S BY SQRT(X#)
*MAT N=CON(rowCRF,1,500) / (sample of 500 males)
*MAT MRF(1,2)=S
*MAT MRF(1,3)=N
*MAT MRF(0,1)="mean"
*MAT MRF(0,2)="stddev"
*MAT MRF(0,3)="N"
*
```

Scheme 5.28. Building the matrix of moments of the value scales.

The reliabilities of the factor images are 0.8643 and 0.6247 (Scheme 5.29). The corresponding omegas are 0.8626 and 0.6247 (Heise and Bohrnstedt 1970, 124), and thus they coincide with the general measure in this case. Cronbach’s alpha underestimates all reliabilities.

```
  1  1 SURVO 98  Mon Feb 14 10:30:56 2000                D:\V\   1000  100 0
*
*RELIAB RF,FACT.M,CUR+1 / MSN=MRF
*Reliabilities according to models E2 and E3:
*E2: errors do not correlate; E3: errors may correlate.
* F1\E2=0.8643   F1\E3=0.8627   (Cronbach’s alpha: 0.7868)
* F2\E2=0.6247   F2\E3=0.6241   (Cronbach’s alpha: 0.5349)
*Sum\E2=0.7928  Sum\E3=0.7908   (Cronbach’s alpha: 0.7387)
*
```

Scheme 5.29. Reliabilities of the value scales.

Heise and Bohrnstedt factored the items both separately and jointly, and with each factoring method – not counting the alpha factoring – the omegas were higher if factored jointly (Heise and Bohrnstedt 1970, 124).

```
  1  1 SURVO 98  Mon Feb 14 11:07:11 2000                D:\V\   1000  100 0
*..................................
*MAT REL=RF(Item1:Item6,Item1:Item6)
*MAT MR=MRF(Item1:Item6,*)
*MAT FAT=RF(Item7:Item11,Item7:Item11)
*MAT MF=MRF(Item7:Item11,*)
*
*FACTA REL 1
*RELIAB REL,FACT.M,CUR+1        / MSN=MR
* F1\E2=0.8632                  (Cronbach’s alpha: 0.8551)
*Sum\E2=0.8571                  (Cronbach’s alpha: 0.8550)
*
*FACTA FAT 1
*RELIAB FAT,FACT.M,CUR+1        / MSN=MF
* F1\E2=0.6246                  (Cronbach’s alpha: 0.6078)
*Sum\E2=0.6081                  (Cronbach’s alpha: 0.6070)
*
```

Scheme 5.30. Reliabilities of the value scales analyzed separately.

Heise and Bohrnstedt (1970, 124) report Cronbach's alphas for both scales as 0.8550 and 0.6068, but they are based on computing unweighted sums separately, in the same way as Liukkonen and Leskinen (1999). Indeed, by separating the correlations and standard deviations, and factoring each scale with one factor, would give those figures (see Scheme 5.30).

Operating with one-factor solutions separately, using the assumptions of classical psychometric test theory, may give an illusion of Cronbach's alpha as a valid method to estimate the reliability of measurement scales. It must be noted, however, that the alpha values are still lower than the other ones, in any case.

The factor structures of the examples presented here are very simple, and only two-dimensional. Usually the number of items and the number of factors is much larger. Many items may have interpretable loadings on several factors, although the simple structure principle were used. Breaking the factor structure and analyzing the dimensions separately is by no means a sound procedure to follow.

6. Monte Carlo simulation of the measurement model

Statistical properties of the reliability measures were studied by Monte Carlo simulation. The true values of the measures were determined according to the measurement model (3.1), and compared with the values drawn from the simulated random samples.

It would be difficult to study the statistical properties analytically, as the number of the parameters affecting the outcome is rather large. On the other hand, simulation experiments can be designed so that some of the parameters are fixed, while a part of them are being varied.

6.1 Principles of the simulation experiments

In order to estimate the parameters of the measurement model, it is assumed that the true scores are orthogonal and that the measurement errors do not correlate. Hence, the model is

$$\mathbf{x} = \mathbf{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon}\,, \tag{6.1}$$

with assumptions

$$cov\,(\boldsymbol{\tau}) = \mathbf{I} \tag{6.2}$$

and

$$cov\,(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi} \quad \text{(diagonal).} \tag{6.3}$$

It is assumed that the observed variables follow a p-dimensional multinormal distribution, *i.e.*

$$\mathbf{x} \sim N\,(\mathbf{0}, \boldsymbol{\Sigma})\,, \tag{6.4}$$

where the structure of the covariance matrix $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}' + \boldsymbol{\Psi}\,. \tag{6.5}$$

Since the location of the scale does not affect reliability, the means are taken as zeros. Without losing generality, we can also standardize the observed variables.

From the equation (6.5) and the multinormality assumption (6.4) it follows that the estimation of the model (6.1) is achieved by the maximum likelihood factor analysis. Parameters to be varied across the experiments are the number of factors *(k)*, the number of variables *(p)* and the sample size *(N)*. The basic form of an experiment is constructed by fixing these parameters.

The covariance structure of an experiment is determined by choosing the $p \times k$ theoretical factor loadings, *i.e.* the elements of the matrix $\mathbf{B}$. The basic criteria is the simple structure principle presented by Thurstone (1935, 1947). The matrix $\mathbf{\Sigma}$ is obtained from the basic equation (6.5) by completing the diagonal elements of $\mathbf{BB}'$ (the communalities) to unities (variances of standardized variables). The remainders then correspond to the elements of the diagonal matrix $\mathbf{\Psi}$.

The true values of the reliabilities are computed for three scales, namely the factor images, the factor scores and the unweighted sum, by using the matrices referred above. The corresponding formulas are shown in Table 6.1.

Scale	Reliability formulas Tarkkonen's measure	 Cronbach's alpha
Factor images	(4.13)	(2.16)
Factor scores	(4.20)	(2.16)
Unweighted sum	(4.30)	(2.15)

Table 6.1. Measurement scales and the corresponding reliability formulas.

The formulas for computing Cronbach's alpha are valid only under the classical true score model (see chapter 4.3). When the assumptions are not met, the figures obtained are underestimates (Kuder and Richardson 1937, Novick and Lewis 1967, Greene and Carmines 1979). Nevertheless, those figures are taken as the true values of alpha for studying its statistical properties in certain situations, and for demonstrating the consequences of violating the assumptions.

To study the effect of the sample variation, multinormal samples are generated using the constructive definition of the multinormal distribution (Mustonen 1995, 15–18). The matrix $\mathbf{\Sigma}$ is decomposed into form

$$\mathbf{\Sigma} = \mathbf{CC}' \qquad (6.6)$$

by spectral decomposition, and the multinormal data values are generated by the formula

$$\mathbf{x} = \mathbf{Cv}\,, \qquad (6.7)$$

where $\mathbf{v} \sim N\,(\mathbf{0}, \mathbf{I})$ (Mustonen 1995, 19–20). The normal random deviates are obtained by the method of Box and Müller (1958). Each sample is based on different seeds of the pseudo random number generator. A combined Tausworthe generator by Tezuka and L'Ecuyer (1991) is applied.

For each sample, the sufficient statistics, *i.e.* the means, standard deviations and correlations are first computed. The maximum likelihood factor analysis is then carried out based on the sample correlation matrix. The number of factors is fixed within an experiment, which makes the factor analysis straightforward.

When the number of factors is greater than one, the estimated factors do not necessarily correspond to the theoretical factors, because of the rotational indeterminacy. To ensure correspondence, transformation analysis (Ahmavaara 1954) is conducted. Standard methods for factor rotation are not needed.

The goal of transformation analysis is to study the invariance of the factor structure, by seeking a $k \times k$ transformation matrix $\mathbf{L}_{12}$, so that

$$\mathbf{B}_1\mathbf{L}_{12} = \mathbf{B}_2 \,, \tag{6.8}$$

where $\mathbf{B}_1$ and $\mathbf{B}_2$ are $p \times k$ factor matrices. The matrix $\mathbf{L}_{12}$ may be obtained by the least squares solution

$$\mathbf{L}_{12} = (\mathbf{B}_1{}'\mathbf{B}_1)^{-1}\mathbf{B}_1{}'\mathbf{B}_2 \,, \tag{6.9}$$

but for orthogonal factors, the symmetric transformation analysis (Mustonen 1966) is preferable. Then, the transformation matrix is obtained directly from the singular value decomposition

$$\mathbf{B}_1{}'\mathbf{B}_2 = \mathbf{UDV}' \tag{6.10}$$

in the form

$$\mathbf{L}_{12} = \mathbf{UV}' \,. \tag{6.11}$$

Using the sample correlation matrix and the transformed factor matrix, the reliabilities of the factor images and the reliabilities of the unweighted sum of the observed variables are computed. The reliabilities of the factor scores are obtained by first computing the corresponding coefficient matrix. This takes place by the usual, shortened regression method of Ledermann (1938).

For each sample, the values of the reliabilities are saved. They form the basic data sets for studying the statistical properties of the reliability measures.

6.2 Designing simulation experiments in Survo

Let us now see, how the preceding principles are carried out in practice, using the matrix interpreter as well as statistical and other operations of Survo (Mustonen 1992).

We begin with a small-scale experiment, demonstrating step-by-step the stages described above. After that we are ready to show, how the required operations are glued together, with the help of the macro language of Survo, to build large-scale applications for Monte Carlo simulation experiments.

6.2.1 Defining the structure of the model

The heart of an experiment is the structure of the measurement model. It is defined here by the 7 × 2 factor matrix B on lines 2–9 of Scheme 6.1. The loadings have been highlighted to clearly display the two-dimensional structure. The editorial arithmetics has been employed to calculate the sums of the squares of the loadings. The communalities vary between 0.29 and 0.64, while the efficiencies of the factors (the columnwise sums of the squares) are 1.81 and 1.29.

```
   1  1 SURVO 98  Sat Jan 15 11:21:44 2000                D:\V\   1000  100 0
   1 *                                h2(x,y):=x^2+y^2
   2 *MATRIX B ///
   3 *  0.8  0.0  /  h2(0.8,0)=0.64
   4 *  0.8  0.0  /
   5 *  0.5  0.2  /  h2(0.5,0.2)=0.29
   6 *  0.5  0.2  /
   7 *  0.1  0.7  /  h2(0.1,0.7)=0.5
   8 *  0.1  0.6  /  h2(0.1,0.6)=0.37
   9 *  0.1  0.6  /   1.81  1.29
  10 *
```

Scheme 6.1. Defining the structure of the measurement model.

The matrices of means, standard deviations and correlations are formed by suitable commands of the matrix interpreter. The correlation matrix is computed by applying the formula (6.5). Nearly half of the commands on the lines 11–22 of Scheme 6.2 operate primarily with the names and the labels of the matrices, instead of the actual numerical elements. This "label arithmetics" is very useful, since it makes the results more readable, and essentially helps the documentation of the work. The technical structure of the matrix interpreter is defined in Mustonen (1989). We apply also recent updates and extensions which are described in Mustonen (1999).

```
   1  1 SURVO 98  Sat Jan 15 11:22:44 2000                D:\V\   1000  100 0
  10 *
  11 *MAT SAVE B
  12 *MAT CLABELS Image TO B
  13 *MAT RLABELS Item TO B
  14 *MAT DIM B /* rowB=7 colB=2
  15 *MAT REM p=rowB
  16 *MAT R=B*B'+(IDN(p,p)-DIAG(B*B'))
  17 *MAT M=ZER(p,2)
  18 *MAT M(1,2)=CON(p,1)
  19 *MAT NAME M AS M
  20 *MAT RLABELS FROM B TO M
  21 *MAT M(0,1)="Mean"
  22 *MAT M(0,2)="Stddev"
```

Scheme 6.2. Forming the matrices of means, standard deviations and correlations.

The results of the previous MAT commands are illustrated by loading the matrices into the edit field (see Scheme 6.3).

```
  1  1 SURVO 98  Sat Jan 15 11:23:44 2000              D:\V\   1000  100 0
 23 *MAT LOAD M END+2
 24 *MAT LOAD R ##.## END+2
 25 *
 26 *MATRIX M
 27 *///         Mean   Stddev
 28 *Item1          0        1
 29 *Item2          0        1
 30 *Item3          0        1
 31 *Item4          0        1
 32 *Item5          0        1
 33 *Item6          0        1
 34 *Item7          0        1
 35 *
 36 *MATRIX R
 37 *B*B'+IDN-DIAG(B*B')
 38 *///      Item1 Item2 Item3 Item4 Item5 Item6 Item7
 39 *Item1     1.00  0.64  0.40  0.40  0.08  0.08  0.08
 40 *Item2     0.64  1.00  0.40  0.40  0.08  0.08  0.08
 41 *Item3     0.40  0.40  1.00  0.29  0.19  0.17  0.17
 42 *Item4     0.40  0.40  0.29  1.00  0.19  0.17  0.17
 43 *Item5     0.08  0.08  0.19  0.19  1.00  0.43  0.43
 44 *Item6     0.08  0.08  0.17  0.17  0.43  1.00  0.37
 45 *Item7     0.08  0.08  0.17  0.17  0.43  0.37  1.00
 46 *
```

Scheme 6.3. Loading the matrices into the edit field.

6.2.2 Computing the true values of the reliabilities

Using the matrices B, R, and M, we compute the true values of the reliabilities of the factor images and the unweighted sum. According to our previous assumptions, we concentrate on the "E2" reliabilities (see lines 48–49 in Scheme 6.4). In the case of the true values, they are equal to the more general "E3" model, but there will be some differences when sample variation is added.

```
  1  1 SURVO 98  Sat Jan 15 11:24:44 2000              D:\V\   1000  100 0
 46 *
 47 *RELIAB R,B,CUR+1 / MSN=M
 48 *Reliabilities according to models E2 and E3:
 49 *E2: errors do not correlate; E3: errors may correlate.
 50 *Image1\E2=0.8044  Image1\E3=0.8044   (Cronbach's alpha: 0.6712)
 51 *Image2\E2=0.7063  Image2\E3=0.7063   (Cronbach's alpha: 0.5814)
 52 *Sum\E2=0.7784  Sum\E3=0.7784   (Cronbach's alpha: 0.7027)
 53 *LOADM RCOV.M,##.###,END+2 / Residual covariance matrix
 54 *LOADM RCORR.M,##.###,END+2 / Residual correlation matrix
 55 *LIMITS=-0.9,-0.2,-0.1,0.1,0.2,0.9,1 SHADOWS=7,8,1,0,1,8,7
 56 *
```

Scheme 6.4. Computing the true values of reliabilities of the factor images and the unweighted sum.

For computing the true reliabilities of the factor scores, we need the matrix of the factor score coefficients. It is here obtained by a sucro command /FCOEFF on line 57 of Scheme 6.5.

The column labels of the matrix are renamed on line 60, and then the contents of the matrix are checked by loading it into the edit field. This is not necessary, since the matrices are saved as separate files, and thus they are independent of the edit field. It is, however, important and useful when planning an experiment.

```
  1  1 SURVO 98  Sat Jan 15 11:25:44 2000                D:\V\   1000  100 0
 56 *
 57 */FCOEFF B,M,FCOEFF
 58 *Use FCOEFF for factor scores by LINCO <data>,FCOEFF(F1,F2,...)
 59 *
 60 *MAT CLABELS "Score" TO FCOEFF
 61 *MAT LOAD FCOEFF
 62 *MATRIX FCOEFF
 63 *///        Score1   Score2
 64 *Constant  0.00000  0.00000
 65 *Item1     0.42770 -0.08092
 66 *Item2     0.42770 -0.08092
 67 *Item3     0.12528  0.06336
 68 *Item4     0.12528  0.06336
 69 *Item5    -0.01249  0.43506
 70 *Item6    -0.00413  0.29513
 71 *Item7    -0.00413  0.29513
 72 *
```

Scheme 6.5. Computing and displaying the factor score coefficients.

The true reliabilities of the factor scores are computed by copying the RELIAB command from the line 47 and adding a specification WEIGHT, with the value of the coefficient matrix obtained above (see Scheme 6.6).

```
  1  1 SURVO 98  Sat Jan 15 11:26:44 2000                D:\V\   1000  100 0
 72 *
 73 *RELIAB R,B,CUR+1 / MSN=M WEIGHT=FCOEFF
 74 *Reliabilities according to models E2 and E3: (weighted by FCOEFF)
 75 *E2: errors do not correlate; E3: errors may correlate.
 76 *Score1\E2=0.8092  Score1\E3=0.8092   (Cronbach’s alpha: 0.5925)
 77 *Score2\E2=0.6860  Score2\E3=0.6860   (Cronbach’s alpha: 0.5107)
 78 *LOADM RCOV.M,##.###,END+2 / Residual covariance matrix
 79 *LOADM RCORR.M,##.###,END+2 / Residual correlation matrix
 80 *LIMITS=-0.9,-0.2,-0.1,0.1,0.2,0.9,1 SHADOWS=7,8,1,0,1,8,7
 81 *
```

Scheme 6.6. Computing the true values of reliabilities of the factor scores.

The true values of all measures are summarized later, in Table 6.2.

6.2.3 Working with a random sample

The structure of the measurement model was defined by the factor matrix B. The items were supposed to follow a multinormal distribution with zero means and unit variances, with a correlation matrix computed from the factor matrix according to the measurement model. To continue the experiment, we need to generate a random sample from this model, and see how the different reliability measures behave under the sample variation.

Using a ready-made MNSIMUL operation, we generate a multinormal sample of 100 observations as a Survo data file SAMPLE, and compute the sample statistics, given on lines 100–117 of Scheme 6.7. They are also saved as matrix files CORR.M and MSN.M. The RND specification on the line 98 selects a seed number for the pseudo random number generator.

```
   1  1 SURVO 98  Sat Jan 15 11:28:44 2000                D:\V\    1000   100 0
  97 *
  98 *MNSIMUL R,M,SAMPLE,100,0     /    RND=rand(123456789)
  99 *CORR SAMPLE CUR+1
 100 *Means, std.devs and correlations of SAMPLE  N=100
 101 *Variable  Mean        Std.dev.
 102 *Item1    -0.053779    1.071454
 103 *Item2     0.089077    0.938499
 104 *Item3     0.069334    0.967513
 105 *Item4     0.064618    0.891189
 106 *Item5     0.046029    0.971521
 107 *Item6     0.102824    1.055998
 108 *Item7    -0.058433    0.886997
 109 *Correlations:
 110 *             Item1   Item2   Item3   Item4   Item5   Item6   Item7
 111 * Item1        1.0000  0.6618  0.4204  0.3456  0.1929  0.0087  0.0936
 112 * Item2        0.6618  1.0000  0.3441  0.3847  0.1813  0.1135  0.1373
 113 * Item3        0.4204  0.3441  1.0000  0.2424  0.2463  0.0596  0.1746
 114 * Item4        0.3456  0.3847  0.2424  1.0000  0.1245 -0.0257  0.1315
 115 * Item5        0.1929  0.1813  0.2463  0.1245  1.0000  0.4647  0.2310
 116 * Item6        0.0087  0.1135  0.0596 -0.0257  0.4647  1.0000  0.2585
 117 * Item7        0.0936  0.1373  0.1746  0.1315  0.2310  0.2585  1.0000
 118 *
```

Scheme 6.7. Generating a multinormal random sample and computing the sample statistics.

Based on the sample correlation matrix CORR.M, we compute the maximum likelihood factor analysis of two factors (see Scheme 6.8). The solution is saved as a new matrix file FACT.M.

```
   1  1 SURVO 98  Sat Jan 15 11:29:44 2000                D:\V\    1000   100 0
 118 *
 119 *FACTA CORR.M,2,CUR+1
 120 *Factor analysis: Maximum Likelihood (ML) solution
 121 *Factor matrix
 122 *             F1     F2    h^2
 123 *Item1     0.831 -0.176  0.722
 124 *Item2     0.775 -0.066  0.604
 125 *Item3     0.499  0.030  0.250
 126 *Item4     0.440 -0.085  0.201
 127 *Item5     0.337  0.555  0.421
 128 *Item6     0.169  0.732  0.565
 129 *Item7     0.206  0.307  0.137
 130 *
```

Scheme 6.8. Computing the maximum likelihood factor analysis.

The solution resembles the theoretical matrix B, but it might look totally different as well. This is where the transformation analysis is needed. In an orthogonal case it is simply a question of applying the singular value decomposition. This is achieved on the line 131 of Scheme 6.9 by a ready-made sucro command /TRAN-SYMMETR, which gives as a result two matrices: the transformation matrix and the corresponding residual matrix. The latter one is extremely interesting in practice, although it is not used in these simulation experiments.

```
   1  1 SURVO 98  Sat Jan 15 11:30:44 2000             D:\V\   1000  100 0
*
*/TRAN-SYMMETR FACT.M,B
*MAT LOAD L.M,###.###,END+2  / Transformation matrix
*MAT LOAD E.M,###.###,END+2  / Residual matrix
*
*MATRIX L.M
*Transformation_matrix
*///       Image1  Image2
*F1         0.970   0.243
*F2        -0.243   0.970
*
*MATRIX E.M
*Residual_matrix
*///       Image1  Image2
*Item1      0.049   0.032
*Item2     -0.033   0.125
*Item3     -0.023  -0.049
*Item4     -0.052  -0.175
*Item5      0.091  -0.080
*Item6     -0.115   0.152
*Item7      0.025  -0.252
*
```

Scheme 6.9. Computing the symmetric transformation analysis.

Multiplying the factor matrix FACT.M from the right by the transformation matrix L.M gives the nearest rotation corresponding to the original factor structure. The matrix is saved as AFACT and loaded into the edit field (see Scheme 6.10).

```
   1  1 SURVO 98  Sat Jan 15 11:31:44 2000             D:\V\   1000  100 0
*.....................................................................
*MAT AFACT=FACT.M*L.M           / *AFACT~F*Transformation_matrix 7*2
*MAT LOAD AFACT,###.###,CUR+1
*MATRIX AFACT
*F*Transformation_matrix
*///       Image1  Image2
*Item1      0.849   0.032
*Item2      0.767   0.125
*Item3      0.477   0.151
*Item4      0.448   0.025
*Item5      0.191   0.620
*Item6     -0.015   0.752
*Item7      0.125   0.348
*
```

Scheme 6.10. Multiplying the factor matrix by the transformation matrix.

The reliabilities are computed by replacing the theoretical matrices by their sample estimates in Scheme 6.11.

```
   1  1 SURVO 98  Sat Jan 15 11:32:44 2000              D:\V\   1000  100 0
*
*RELIAB CORR.M,AFACT,CUR+1 / MSN=MSN.M
*Reliabilities according to models E2 and E3:
*E2: errors do not correlate; E3: errors may correlate.
*Image1\E2=0.8065  Image1\E3=0.8039   (Cronbach's alpha: 0.6547)
*Image2\E2=0.6950  Image2\E3=0.6940   (Cronbach's alpha: 0.4941)
*Sum\E2=0.7498  Sum\E3=0.7410   (Cronbach's alpha: 0.6737)
*LOADM RCOV.M,##.###,END+2 / Residual covariance matrix
*LOADM RCORR.M,##.###,END+2 / Residual correlation matrix
*LIMITS=-0.9,-0.2,-0.1,0.1,0.2,0.9,1 SHADOWS=7,8,1,0,1,8,7
*
```

Scheme 6.11. Computing the reliabilities of the factor images and the unweighted sum from the sample.

The factor score coefficients are computed accordingly (see Scheme 6.12). The "Constant" row of the coefficient matrix takes care of the centering, if the scores are computed as new variables into the data by the LINCO operation. It has no use in the reliability computations.

```
   1  1 SURVO 98  Sat Jan 15 11:33:44 2000              D:\V\   1000  100 0
*
*/FCOEFF AFACT
*Use FCOEFF.M for factor scores by LINCO <data>,FCOEFF.M(F1,F2,...)
*
*MAT CLABELS "Score" TO FCOEFF.M
*MAT LOAD FCOEFF.M,###.###,CUR+1
*MATRIX FCOEFF.M
*FCOEFF
*///       Score1  Score2
*Constant  -0.013  -0.071
*Item1      0.507  -0.075
*Item2      0.358   0.029
*Item3      0.110   0.042
*Item4      0.112  -0.013
*Item5      0.019   0.343
*Item6     -0.069   0.530
*Item7      0.012   0.141
*
```

Scheme 6.12. Computing and displaying the factor score coefficients from the sample.

The reliabilities of the factor scores are obtained in Scheme 6.13 by a copy of the previous RELIAB command, with a proper WEIGHT specification added.

```
   1  1 SURVO 98  Sat Jan 15 11:34:44 2000              D:\V\   1000  100 0
*
*RELIAB CORR.M,AFACT,CUR+1 / MSN=MSN.M WEIGHT=FCOEFF.M
*Reliabilities according to models E2 and E3: (weighted by FCOEFF.M)
*E2: errors do not correlate; E3: errors may correlate.
*Score1\E2=0.8221  Score1\E3=0.8220   (Cronbach's alpha: 0.5478)
*Score2\E2=0.6799  Score2\E3=0.6799   (Cronbach's alpha: 0.3940)
*LOADM RCOV.M,##.###,END+2 / Residual covariance matrix
*LOADM RCORR.M,##.###,END+2 / Residual correlation matrix
*LIMITS=-0.9,-0.2,-0.1,0.1,0.2,0.9,1 SHADOWS=7,8,1,0,1,8,7
*
```

Scheme 6.13. Computing the reliabilities of the factor scores from the sample.

We now have the observed values of the measures based on one random sample, together with the true values and differences (see Table 6.2).

Scale	**Tarkkonen's measure**			**Cronbach's alpha**		
	True	Sample	diff.%	True	Sample	diff.%
Factor image 1	0.8044	0.8065	+0.26	0.6712	0.6547	-2.46
Factor image 2	0.7063	0.6950	-1.60	0.5814	0.4941	-15.02
Factor score 1	0.8092	0.8221	+1.59	0.5925	0.5478	-7.54
Factor score 2	0.6860	0.6799	-0.89	0.5107	0.3940	-22.85
Unweighted sum	0.7784	0.7498	-3.67	0.7027	0.6737	-4.13

Table 6.2. True values of reliabilities and sample estimates.

The differences seem to be negligible, except in the case of certain Cronbach's alphas. Nothing can be infered from a single sample, however. The variation in the observed values of the reliabilities must be studied by repeatedly generating new samples and making the computations that were described above.

6.2.4 Repeating the experiment

Repeating the experiment a few times is simple: just go back and re-activate the commands one by one, changing the random number seed each time. Omitting the unnecessary output makes the command schemes even more simple. For example, generating the multinormal sample, computing the sufficient statistics and factoring the correlation matrix is achieved by just three commands. Nothing is printed in the edit field, as all input and output is handled through data files and matrix files. The rotation by transformation analysis is issued directly using the matrix interpreter by applying the formulas (6.10) and (6.11) on lines 304–306 of Scheme 6.14.

```
   1  1 SURVO 98  Sat Jan 15 11:36:44 2000            D:\V\   1000   100 0
 300 *
 301 *MNSIMUL R,M,SAMPLE,100,0 / RND=rand(2000001)
 302 *CORR SAMPLE
 303 *FACTA CORR.M,2
 304 *MAT B1B2=FACT.M'*B          / *B1B2~F'*B 2*2
 305 *MAT SVD OF B1B2 TO U,D,V
 306 *MAT L12=U*V'       / *L12~Usvd(F'*B)*Vsvd(F'*B)' 2*2
 307 *MAT AFACT=FACT.M*L12        / *AFACT~F*Usvd(F'*B)*Vsvd(F'*B)' 7*2
 308 *
```

Scheme 6.14. Making the central computations in a condensed form.

When a simulation experiment has been built up to this phase, it is time to consider proper methods for repeating the experiment automatically. In Survo, there are several possibilities. The whole experiment could be programmed in C language as a new Survo module (Mustonen 1989). However, re-programming or connecting existing modules would be a waste of time. Instead, the macro language of Survo (Mustonen 1988) allows the user to combine the operations of Survo as sucros. Certain sucros, such as /FCOEFF were already employed in the previous working schemes.

As an example, we make an experiment of 50 replications with a sample size of 100, applying the /RLBSIMUL sucro, which is described in detail in Appendix A. In Scheme 6.15, the sucro is activated with appropriate parameters. The results will be saved as a data file B10050.

```
   1  1 SURVO 98  Sat Jan 15 11:37:44 2000                D:\V\    1000  100 0
 328 *
 329 */RLBSIMUL B,-,100,B10050,50,20000100
 330 *
```

Scheme 6.15. Conducting a repeated simulation experiment.

A snap shot of the edit field when the /RLBSIMUL sucro is running, is presented in Scheme 6.16. Although the commands include lines for output (CUR+1 on the lines 6, 11 and 12), they are merely used for controlling and checking the program flow (see the sucro code for details). The results are handled directly through matrix files, data files and text files, in full precision. The RELIAB operation (see Appendix A) includes a special OUTFILE specification for the simulation purposes, giving the results in a more convenient form for further processing.

```
   1  1 SURVO 98  Sat Jan 15 11:38:44 2000                D:\V\    1000  100 0
   1 *Experiment B10050: computing 5/50...
   2 *OUTSEED=C:\TMP\RELSEEDS.DAT
   3 *MNSIMUL %R,%MSN,&TEST,100,1 / TYPES=8
   4 *INSEED=C:\TMP\RELSEEDS.DAT
   5 *CORR &TEST / PRIND=0 FAST=1
   6 *FACTA CORR.M,2,CUR+1
   7 *MATRUN C:\TMP\SYMTRANS.MTX
   8 *MATRUN C:\TMP\ROTATION.MTX
   9 *MATRUN C:\TMP\FSCORES.MTX
  10 *FILE DEL C:\TMP\RELOUT.
  11 *RELIAB CORR.M,&FACT,&RFACT,CUR+1 / MSN=MSN.M OUTFILE=C:\TMP\RELOUT
  12 *RELIAB CORR.M,&FACT,&RFACT,CUR+1 / WEIGHT=FCOEFF.M
  13 *FILE DEL C:\TMP\RELOUT
  14 *FILE SAVE C:\TMP\RELOUT TO C:\TMP\RELOUT
  15 *MAT SAVE DATA C:\TMP\RELOUT TO C:\TMP\RELOUT
  16 *MAT C:\TMP\RELOUT=C:\TMP\RELOUT'       / *C:\TMP\RELOUT~C:\TMP\RELOUT'
  17 *MAT C:\TMP\RELOUT(1,0)="20000105"
  18 *FILE DEL C:\TMP\RELOUT
  19 *FILE SAVE MAT C:\TMP\RELOUT TO C:\TMP\RELOUT
  20 *SEED1=939053036 SEED2=47664685
  21 *VAR SEED1,SEED2 TO C:\TMP\RELOUT
  22 *FILE COPY C:\TMP\RELOUT TO B10050
  23 *
```

Scheme 6.16. View of an ongoing simulation experiment.

The first line of the field in Scheme 6.16 tells the number of the current replicate. No other interaction is provided by the sucro. It is optimized for efficiency. This experiment consisting of 50 replications took 3 minutes and 18 seconds on a 133 MHz Pentium PC. That makes less than 4 seconds per replicate.

When the dimensions (number of factors and number of variables) grow, the time is spent mostly in the factor analysis phase. Generating the random sample is extremely fast even with large dimensions.

After the experiment has completed, the new data file B10050 includes 50 observations of 21 variables. Those ten that we have analyzed here, are selected by the MASK specification and sorted in the same order than earlier by the CORR operation (see Scheme 6.17). Let us take a look at the statistics of the observed reliabilities.

```
  1  1 SURVO 98  Sat Jan 15 11:39:44 2000             D:\V\   1000  100 0
*
*MASK=---AAE---BBF---CC--DD
*CORR B10050
*MAT LOAD MSN.M(*,mean:stddev),##.#### CUR+1
*MATRIX MSN.M
*MSN(B10050)
*///         mean  stddev
*F1\E2     0.8205  0.0319
*F2\E2     0.7206  0.0532
*a\F1      0.6741  0.0388
*a\F2      0.5629  0.0553
*%1\E2     0.8405  0.0494
*%2\E2     0.7145  0.0619
*a\%1      0.5250  0.1296
*a\%2      0.4372  0.0946
*Sum\E2    0.7746  0.0381
*a\Sum     0.6898  0.0517
*
```

Scheme 6.17. Computing and displaying the means and standard deviations of the observed reliabilities.

The average deviations from the true values are quite acceptable, except in the case of Cronbach’s alphas of the factor scores. This is seen from Table 6.3.

Scale	**Tarkkonen’s measure**			**Cronbach’s alpha**		
	True	Mean	diff.%	True	Mean	diff.%
Factor image 1	0.8044	0.8205	+2.00	0.6712	0.6741	+0.43
Factor image 2	0.7063	0.7206	+2.02	0.5814	0.5629	-3.18
Factor score 1	0.8092	0.8405	+3.87	0.5925	0.5250	-11.39
Factor score 2	0.6860	0.7145	+4.15	0.5107	0.4372	-14.39
Unweighted sum	0.7784	0.7746	-0.49	0.7027	0.6898	-1.84

Table 6.3. True values of reliabilities and means of 50 sample estimates.

Tarkkonen’s measure seems to slightly overestimate the reliability of factor images and factor scores. Cronbach’s alpha seems to underestimate especially the reliability of the factor scores. According to the standard deviations of Scheme 6.17, there is less variation in the observed values of Tarkkonen’s measure, in all scales.

Although alpha seems to manage rather well compared to its true values in the case of the factor images and the unweighted sum, it gives about 10–40% lower values than Tarkkonen’s measure. The underestimation occurs because the assumptions of alpha are violated.

6.3 Experimental settings

Above, the experiment was repeated 50 times, which is insufficient for drawing reliable conclusions. Also the effect of the sample size and the structure of the model have to be investigated. Hence, to find suitable frames for the experiments, several test runs were conducted. It seems that it is sufficient to repeat a single experiment 500 times. The total number of replicates is significantly higher, when the variations of the parameters are taken into account.

The sample size is an essential parameter. If large samples are used, the estimates are expected to come closer to their true values. On the other hand, it is interesting to see how the reliability measures behave with smaller samples. In practical applications, the size of the sample is often far from large.

The simulation experiments of this study are organized as entities, which are called experimental settings. Each setting consists of experiments conducted on eight different models. The models are constructed by defining the factor structures. The primary loadings (see Table 6.4) are used to determine the structure in all experiments. In certain experimental settings, some secondary loadings are added for nuisance purposes.

Model	1	2	3	4	5	6	7	8
Primary loading	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
Communality	0.36	0.42	0.49	0.56	0.64	0.72	0.81	0.90

Table 6.4. Primary loadings and the corresponding communalities of the models 1–8 in the experimental settings.

A primary loading appears only once per item, and the possible secondary loadings are negligible in any case. Thus, the communalities are determined mainly by the primary loadings. In a reasonable factor model the communalities should be over 0.5, which is not the case in the models 1–3. Thus, those models are included merely for seeing what kind of implications are to be expected when the assumptions behind the models are violated.

The method of maximum likelihood estimation requires a reasonable number of observations to be used. There are lots of suggestions for a minimum sample size. It depends also on the number of items, but generally 100 can be taken as a good minimum. In addition, three smaller sample sizes are used, for the same reasons as the worse models above. The maximum sample size is 300, which seemed to be large enough in this context. Hence, the experiments are repeated 500 times with 14 sample sizes from 40 to 300, making as such a total of 7000 replicates within each model (see Table 6.5).

Model	N													
	40	60	80	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**	**280**	**300**
1	500	500	500	500	500	500	500	500	500	500	500	500	500	500
2	500	500	500	500	500	500	500	500	500	500	500	500	500	500
3	500	500	500	500	500	500	500	500	500	500	500	500	500	500
4	500	500	500	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**
5	500	500	500	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**
6	500	500	500	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**
7	500	500	500	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**
8	500	500	500	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**	**500**

Table 6.5. Design of an experimental setting: 500 replicates in each cell.

The highlighted areas in Table 6.5 display the partition of the simulation experiments within an experimental setting. The analyses are conducted using models 4–8 and sample sizes 100–300. The three first models and sample sizes are only used to obtain extra information from the behaviour of the reliability measures, when various assumptions are violated.

Altogether five experimental settings are defined. In three of them, the one-factor case is explored. Multiple factors are considered in two settings. In the following, the settings are described in detail.

6.3.1 One-dimensional models

When all items are supposed to be equally good indicators of the underlying dimension, we have the most restricted one-factor case, where the true values of the reliability measures coincide. This is the special case of Theorem 4.1, where the elements of the covariance matrix are equal. It is analyzed by two experimental settings. In the first one, there are 10 items, having constant primary loadings (see Table 6.4), while in the second one the number of items is increased to 25. The true values of the reliability measures in these two settings are presented in Table 6.6.

Model	1	2	3	4	5	6	7	8
Setting I	0.8491	0.8798	0.9057	0.9278	0.9467	0.9630	0.9771	0.9893
Setting II	0.9336	0.9482	0.9600	0.9698	0.9780	0.9849	0.9907	0.9957

Table 6.6. True values of reliabilities in the experimental settings I and II.

As the Spearman-Brown formula (2.1) indicates, the reliability is increased when more items are added. This can be seen from Table 6.6, as the only difference between the settings I and II is the number of items. In both settings, the items are perfectly internally consistent. This property is disturbed in the third setting, by adding some secondary loadings as nuisance parameters to the models. This alone violates the assumptions of Cronbach's alpha, although the models are still unidimensional. The primary loadings are equal to those of the setting I, but in addition there are five loadings of 0.3 and -0.3 present in each model, giving a total of 20 items in the setting III.

According to Theorem 4.1, alpha can not exceed the general measure, and since the conditions for the equivalence no more exist, the values of alpha should be less than the corresponding values of the general measure. These results are reflected by the true values (see Table 6.7).

Model	Factor image		Factor score		Unweighted sum	
	ρ_{uu}	α	ρ_{uu}	α	ρ_{uu}	α
1	0.8664	0.8500	0.8687	0.8462	0.6990	0.6438
2	0.8896	0.8699	0.8925	0.8653	0.7396	0.6841
3	0.9102	0.8874	0.9138	0.8820	0.7753	0.7195
4	0.9285	0.9028	0.9326	0.8966	0.8067	0.7507
5	0.9446	0.9164	0.9494	0.9094	0.8344	0.7782
6	0.9590	0.9284	0.9643	0.9207	0.8588	0.8024
7	0.9717	0.9390	0.9776	0.9307	0.8804	0.8238
8	0.9830	0.9484	0.9894	0.9395	0.8996	0.8428
ρ_{uu} = Tarkkonen's measure, α = Cronbach's alpha						

Table 6.7. True values of reliabilities in the experimental setting III.

In all one-dimensional settings, it is interesting to analyze how the result of Theorem 4.1 works with sample estimates.

6.3.2 Multidimensional models

The properties of the measures are studied also with multidimensional models. The primary interest concerns Tarkkonen's measure, as the assumptions of Cronbach's alpha are seriously violated. Still, the values of alpha are computed in all cases, to give a chance to discuss the implications of its errorneous usage.

The experimental setting IV consists of a simple structure of two factors and 30 items. The primary loadings (see Table 6.4) appear on 20 items on the first, and on 10 items on the second factor. The rest of the items have secondary loadings of 0.2 on the first, and -0.1 on the second factor.

The true values of the reliabilities are displayed in Table 6.8. As the values for the factor images and the corresponding factor scores coincide (in four decimal places), the same figures represent both scales.

Model	Factor image 1 Factor score 1		Factor image 2 Factor score 2		Unweighted sum	
	ρ_{uu}	α	ρ_{uu}	α	ρ_{uu}	α
1	0.9236	0.9088	0.8633	0.7995	0.9193	0.8999
2	0.9399	0.9246	0.8914	0.8261	0.9362	0.9160
3	0.9534	0.9378	0.9154	0.8488	0.9503	0.9295
4	0.9647	0.9488	0.9360	0.8683	0.9622	0.9409
5	0.9742	0.9581	0.9536	0.8850	0.9724	0.9507
6	0.9824	0.9660	0.9689	0.8995	0.9812	0.9590
7	0.9893	0.9728	0.9822	0.9121	0.9888	0.9661
8	0.9953	0.9787	0.9937	0.9230	0.9953	0.9724

ρ_{uu} = Tarkkonen's measure, α = Cronbach's alpha

Table 6.8. True values of reliabilities in the experimental setting IV.

The experimental setting V consists of a simple structure of three factors and 35 items. The primary loadings (see Table 6.4) appear on 20 items on the first, 10 on the second, and on 5 items on the third factor. The rest of the items have zero loadings.

The true values of the reliabilities are displayed in Table 6.9. Again the values for the factor images and the corresponding factor scores coincide.

Model	Factor image 1 Factor score 1		Factor image 2 Factor score 2		Factor image 3 Factor score 3		Unweighted sum	
	ρ_{uu}	α	ρ_{uu}	α	ρ_{uu}	α	ρ_{uu}	α
1	0.9184	0.8981	0.8491	0.7866	0.7377	0.6075	0.8940	0.8590
2	0.9360	0.9154	0.8798	0.8151	0.7853	0.6467	0.9165	0.8805
3	0.9505	0.9296	0.9057	0.8391	0.8277	0.6816	0.9351	0.8984
4	0.9626	0.9413	0.9278	0.8596	0.8654	0.7127	0.9507	0.9134
5	0.9726	0.9512	0.9467	0.8771	0.8989	0.7403	0.9639	0.9261
6	0.9812	0.9595	0.9630	0.8922	0.9287	0.7648	0.9750	0.9368
7	0.9884	0.9666	0.9771	0.9052	0.9552	0.7866	0.9846	0.9460
8	0.9946	0.9727	0.9893	0.9166	0.9789	0.8061	0.9928	0.9539

ρ_{uu} = Tarkkonen's measure, α = Cronbach's alpha

Table 6.9. True values of reliabilities in the experimental setting V.

The technical implementation of the experimental settings I–V is described in Appendix A.

7. Analyses and results

7.1 General comments

The results of the Monte Carlo simulations are based on the experimental settings I–V. The structures of the settings are defined in chapter 6, and the technical implementation is described in Appendix A. The simulated data consist of reliabilities from $5\times8\times14\times500=280000$ factor analyses, which are based on total of 47.6 million multinormal observations. If the weakest models (1–3) and the smallest sample sizes (40, 60, and 80) are left out, there are still $5\times5\times11\times500=$ 137500 records to be analyzed, 27500 in each setting.

The analysis of the data is divided in two parts, depending on the dimensionality of the models. The one-factor case, with three experimental settings, forms the first part. The second part deals with the multiple factors.

Cronbach's alpha and Tarkkonen's measure may be compared only in the one-factor case. Strictly speaking, the comparisons are valid only within the settings I and II, where the assumptions of alpha do hold. Nevertheless, since the assumptions of alpha are nearly always violated in practice, it is reasonable to consider certain comparisons within other settings, as well.

The unweighted sum is a traditional scale, which deserves to be analyzed carefully, although in practice, it is recommendable to work with the factor scores instead. With the unweighted sum, it is natural to apply Cronbach's alpha in its original form (2.15). When making comparisons between alpha and Tarkkonen's measure, it is reasonable to use the same scale and apply the formula (4.30) for Tarkkonen's measure. Then, the setup corresponds to Theorem 4.1, which states that Cronbach's alpha can not exceed Tarkkonen's measure. The simulated data give an opportunity to see how the result works in various situations.

Besides by comparisons with Cronbach's alpha, Tarkkonen's measure is explored by studying its statistical properties in the multidimensional settings.

7.2 Sampling variation and bias

In order to analyze the statistical properties of the reliabilities, suitable measures of deviation are needed. The sampling variation is measured by the mean squared error (MSE) and its square root, which is termed root mean square (RMS). The precision of the reliabilities is assessed by the amount of bias. In the following formulas, the true value of a reliability measure is denoted by ρ. An observed value in an experiment that is repeated *n* times, is denoted by r_i, and the mean of the observed values is denoted by $\bar{r}$.

To make the comparisons fair across different models, the measures are presented relative in respect to the corresponding true values, and thus computed as

$$bias = \frac{(\bar{r} - \rho)}{\rho} \quad , \tag{7.1}$$

and

$$MSE = \frac{\sum_{i=1}^{n} (r_i - \rho)^2}{n\,\rho} \quad . \tag{7.2}$$

The bias (7.1) shows the average deviation from the true value, whereas the mean squared error (7.2) is more concerned with the total sampling variation.

It is also interesting to analyze the proportion of bias to the total variation. In those comparisons the bias is taken squared, thus ignoring the direction. The proportion is computed with the formula

$$\frac{bias^2}{MSE} = \frac{n\,(\bar{r} - \rho)^2}{\sum_{i=1}^{n} (r_i - \rho)^2} \quad , \tag{7.3}$$

but since squared values are not easily comparable with the original values of the measures, a square root of (7.3) is taken, giving the proportion of the absolute bias to the root mean square as

$$\frac{bias}{RMS} = \frac{\sqrt{n}\;|\,\bar{r} - \rho\,|}{\sqrt{\sum_{i=1}^{n} (r_i - \rho)^2}} \quad . \tag{7.4}$$

7.3 One-dimensional models

The one-dimensional models are represented by the experimental settings I, II, and III. The true values of the reliabilities in these settings are displayed in chapter 6, in Tables 6.6 and 6.7.

7.3.1 Internal consistency

In the experimental settings I and II, the internal consistency is perfect. In practice, it is hard to find items that measure the same thing equally well. In the setting III, this difficulty is simulated by adding certain secondary loadings as nuisance parameters to the models. When there are some weaker items, the internal consistency is broken. It has a remarkable effect on Cronbach's alpha, compared to Tarkkonen's measure of the same scale (see Figure 7.1). The profile of Tarkkonen's measure reveals the influence of the model structure and the sample size: better models and larger samples increase the precision. There is remarkably more variation in the values of alpha, even with the best models and largest samples.

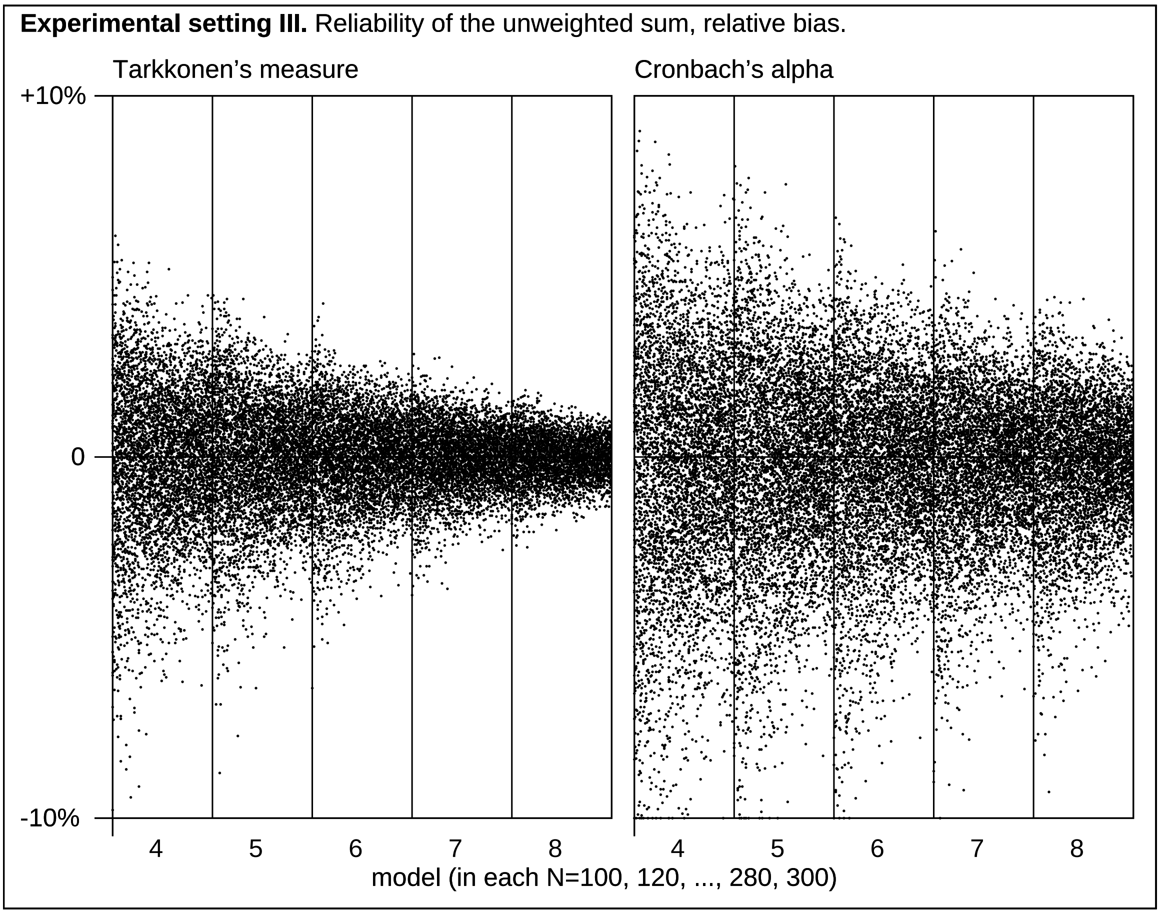

Figure 7.1. Relative bias of the reliabilities in the experimental setting III.

In Figure 7.1, the bias is taken on the level of the observed reliabilities, to display the total variation and the extreme cases. In 58 cases (which are gathered along the -10% line in Figure 7.1), alpha is even more inaccurate than it seems, the largest deviation being -17%.

The average bias of the measures is quite negligible, however (see Table 7.1). The values are obtained with the formula (7.1) using the means computed from 5500 observations in each model, corresponding to Figure 7.1. It is known that Cronbach's alpha underestimates the true reliability when its assumptions are violated. Tarkkonen's measure behaves better, although there is also a slight negative bias present, which seems to decrease when the model is improved. The bias of alpha, on the other hand, remains on the same level also with the best models.

Model	**Factor image**		**Factor score**		**Unweighted sum**	
	ρ_{uu}	α	ρ_{uu}	α	ρ_{uu}	α
4	0.0006	0.0002	0.0011	-0.0016	-0.0006	-0.0031
5	0.0004	0.0004	0.0008	-0.0014	-0.0002	-0.0025
6	0.0000	0.0001	0.0003	-0.0016	-0.0002	-0.0023
7	0.0000	0.0002	0.0002	-0.0016	-0.0001	-0.0022
8	0.0000	0.0002	0.0001	-0.0015	-0.0001	-0.0021
ρ_{uu} = Tarkkonen's measure, α = Cronbach's alpha						

Table 7.1. Average bias of the reliability measures in the experimental setting III.

The difference in variation, which is obvious from the Figure 7.1, is reflected by the standard deviations in Table 7.2.

Model	**Factor image**		**Factor score**		**Unweighted sum**	
	ρ_{uu}	α	ρ_{uu}	α	ρ_{uu}	α
4	0.0075	0.0087	0.0078	0.0082	0.0194	0.0318
5	0.0055	0.0068	0.0059	0.0062	0.0149	0.0269
6	0.0039	0.0052	0.0042	0.0046	0.0113	0.0230
7	0.0022	0.0036	0.0025	0.0029	0.0079	0.0188
8	0.0010	0.0024	0.0012	0.0016	0.0056	0.0165
ρ_{uu} = Tarkkonen's measure, α = Cronbach's alpha						

Table 7.2. Standard deviations of bias in the experimental setting III.

Tarkkonen's measure behaves well also with the factor image and the factor score scale. According to Table 7.1, the reliabilities of those scales are accurate, although they still have a slight bias, now on the positive side. However, when the model is improved, the bias disappears. In contrast, Cronbach's alpha seems not as good. Especially the factor score scale causes trouble, as the bias remains on the same level along the models.

The relative bias of Tarkkonen's measure when applied to the factor image and the factor score is shown in Figure 7.2.

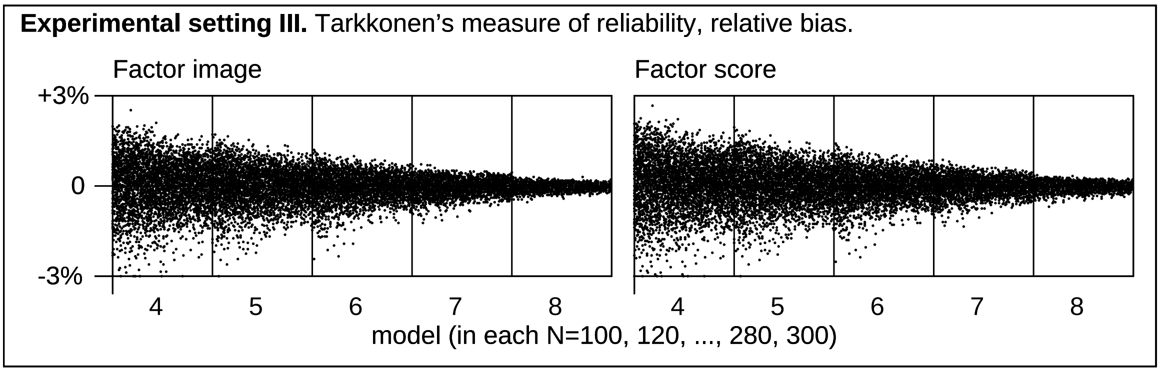

Figure 7.2. Relative bias of Tarkkonen's measure in the experimental setting III.

The profiles of the pictures are sharper, compared to Figure 7.1, and the amount of bias is significantly smaller.

7.3.2 The inequality result

In the experimental settings I and II, the true values of all measures coincide, and the special case of Theorem 4.1 holds, stating that also Cronbach's alpha and Tarkkonen's measure coincide. This fact does not change when random samples from the models 4–8 are considered (see Figure 7.3). In both cases, all 27500 observations are almost on a straight line.

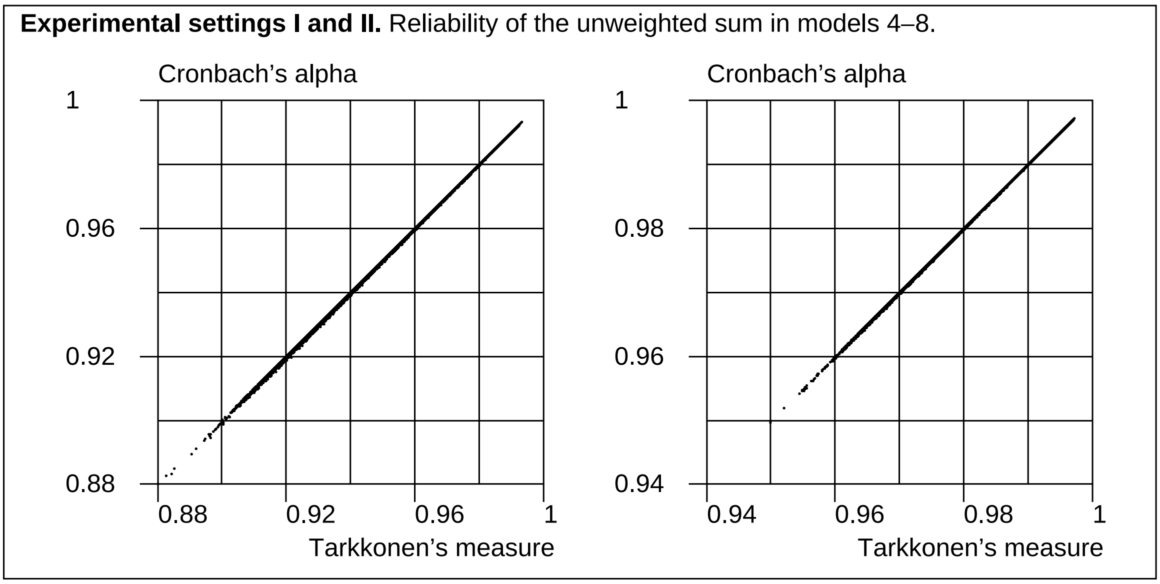

Figure 7.3. Cronbach's alpha and Tarkkonen's measure in the experimental settings I and II, models 4–8.

Generally, Theorem 4.1 states that Cronbach's alpha can not exceed Tarkkonen's measure. But, if the assumptions do not hold, the opposite might be true. In the experimental settings I and II, that seems to happen rarely. The value of alpha exceeds Tarkkonen's measure in less than 1% of those cases where the assumptions of the factor model are not completely satisfied. The frequencies of such cases in the setting I are listed in Table 7.3 (*cf.* Table 6.5).

Model	N													
	40	60	80	100	120	140	160	180	200	220	240	260	280	300
1	43	26	14	10	6	15	8	10	10	4	8	10	8	6
2	12	7	10	5	5	7	2	3	4	2	2	1	2	5
3	9	8	2	3	3	1	0	1	1	0	1	0	0	1
4	2	2	1	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 7.3. Frequencies of cases where Cronbach's alpha exceeds Tarkkonen's measure in the experimental setting I.

The result works perfectly with adequate models and sample sizes. It could be possibly said that if one gets an alpha higher than the general measure, there is probably something wrong either in the data or in the model.

The same results for the experimental setting II are summarized in Table 7.4. The proportion of cases is even smaller, since the number of items in the models is larger. The theorem works even with smaller sample sizes. Also model 3, where the true communality (see Table 6.4) is just below 0.5, behaves rather well.

Model	N													
	40	60	80	100	120	140	160	180	200	220	240	260	280	300
1	14	15	5	4	5	3	5	10	3	3	3	4	6	5
2	6	5	1	3	2	2	2	3	0	4	3	0	2	2
3	2	0	1	0	0	0	0	1	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 7.4. Frequencies of cases where Cronbach's alpha exceeds Tarkkonen's measure in the experimental setting II.

In the experimental setting III, the inequality result works already with the true values of the measures (see Table 6.7). The true values of alpha for the unweighted sum are 6–7% lower than the values of Tarkkonen's measure. The sample estimates follow this difference, as is seen from the left hand side of Figure 7.4.

The graph on the right hand side of Figure 7.4 displays the corresponding bias of the measures, relative to the true values. The bias of Cronbach's alpha is larger both in negative and positive directions (*cf.* Figure 7.1).

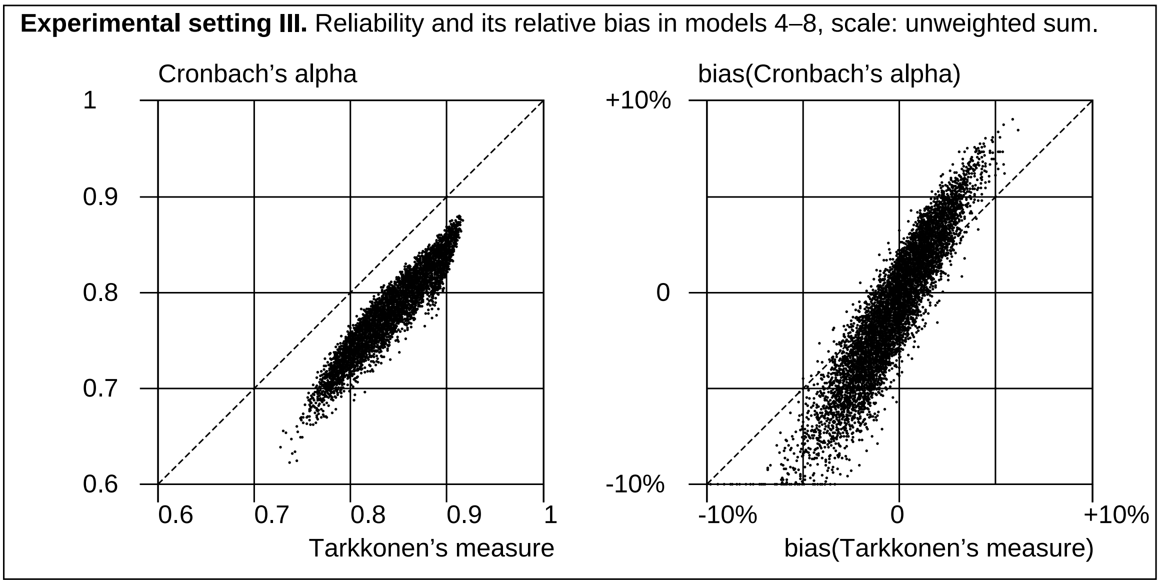

Figure 7.4. Reliabilities and their relative bias in the experimental setting III, models 4–8.

7.3.3 Proportion of bias

According to a working rule in the sample survey theory, the effect of bias on the accuracy of an estimate is negligible if the bias is less than one tenth of the standard deviation of the estimate (Cochran 1977, 14–15). In an analogous way, we consider the proportion of the absolute bias to the root mean square, computed with the formula (7.4).

An analysis of the reliability of the unweighted sum in this respect is displayed in Figures 7.5 and 7.6. The proportion of the bias is mostly less than 10% in Tarkkonen's measure, and more than 10% in Cronbach's alpha. The black dots denote the values exceeding the 10% level. The experimental setting II gives essentially the same results.

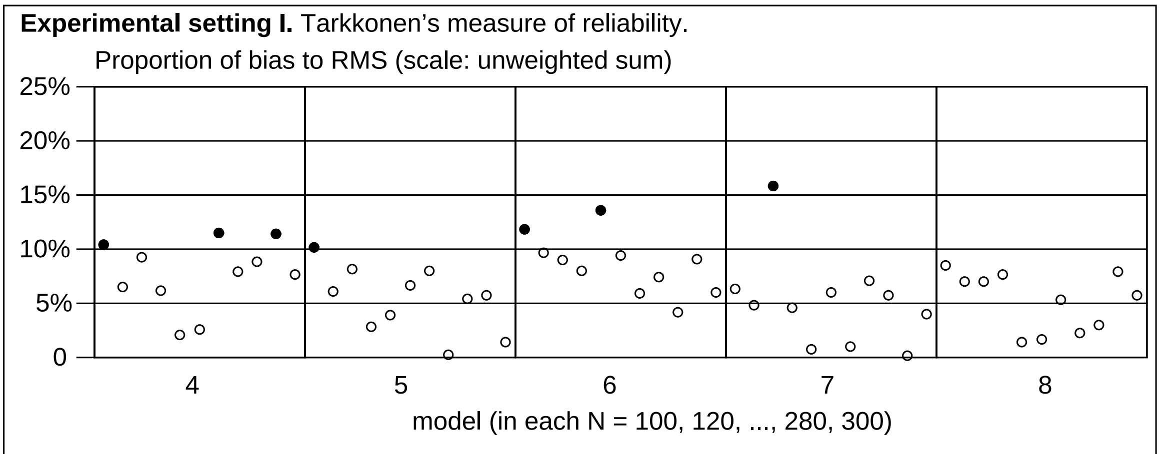

Figure 7.5. Proportion of bias to RMS in Tarkkonen's measure in the experimental setting I.

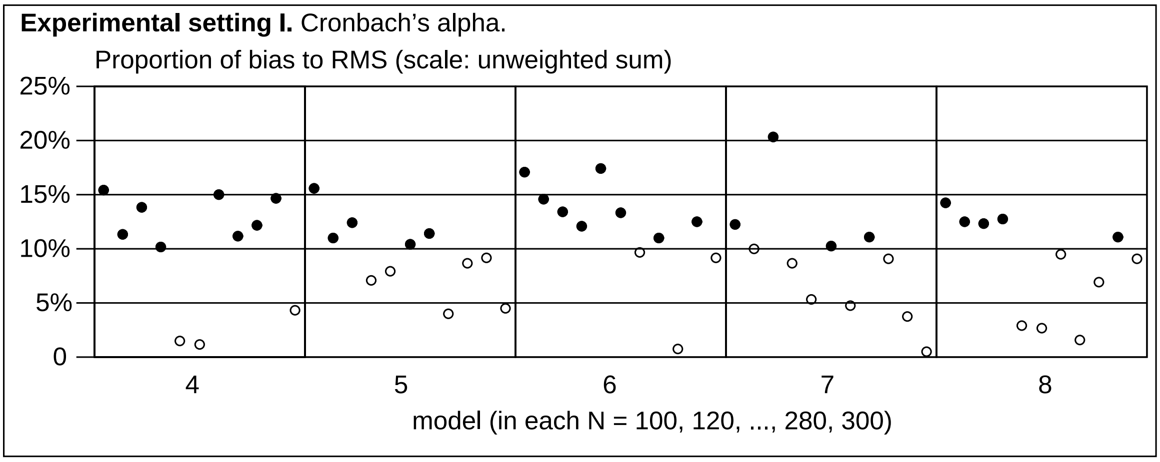

Figure 7.6. Proportion of bias to RMS in Cronbach's alpha in the experimental setting I.

The deviation from the internal consistency in the experimental setting III does not affect Tarkkonen's measure, since the proportion of the bias is still less than 10% in nearly all cases (Figure 7.7). Cronbach's alpha, on the other hand, gets even more biased. In some cases, the bias goes beyond 20% (Figure 7.8).

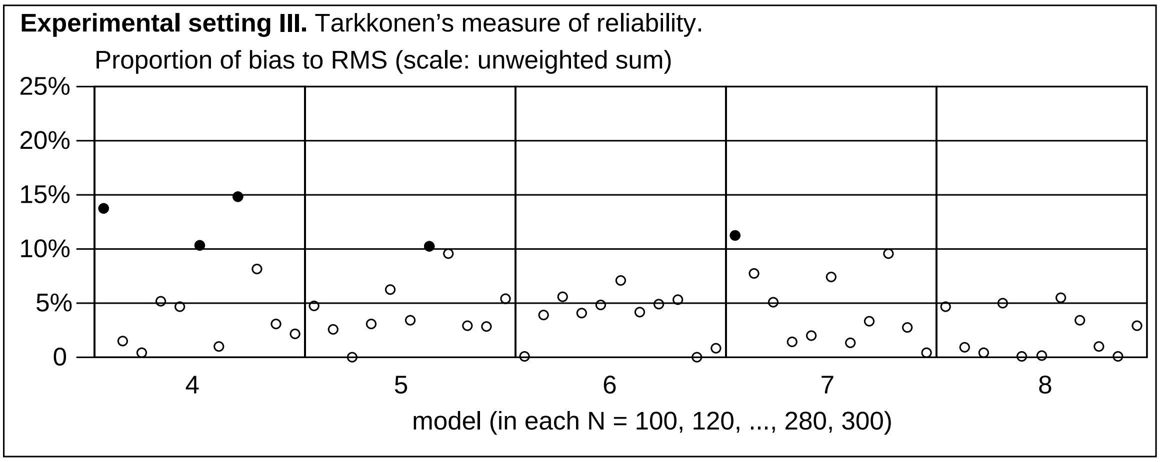

Figure 7.7. Proportion of bias to RMS in Tarkkonen's measure in the experimental setting III.

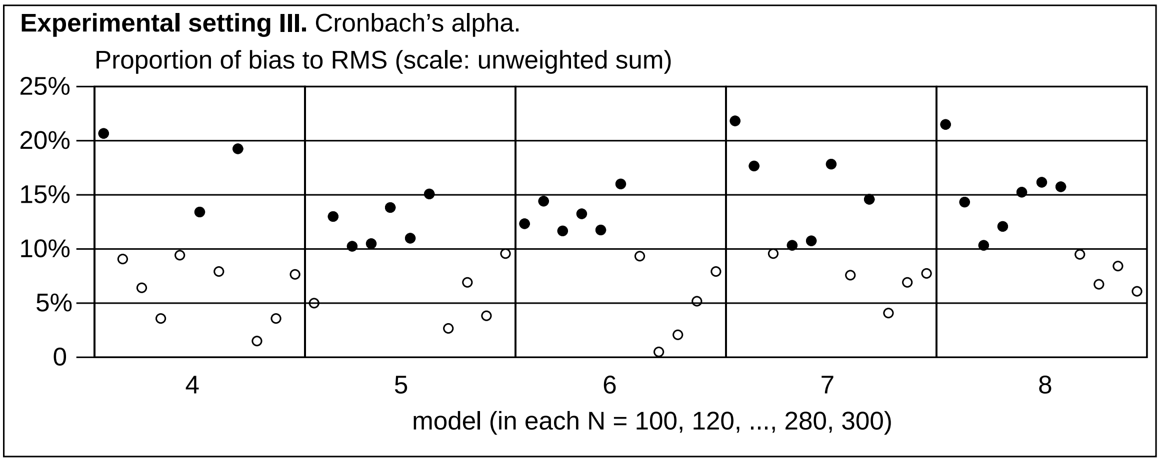

Figure 7.8. Proportion of bias to RMS in Cronbach's alpha in the experimental setting III.

7.4 Multidimensional models

The multidimensional models are represented by the experimental settings IV and V. The true values of the reliabilities in these settings are displayed in chapter 6, in Tables 6.8 and 6.9, respectively.

7.4.1 Stretching the assumptions

When the measurement model is truely multidimensional, the unweighted sum of the items can not be an optimal scale. A better alternative is provided by the factor scores, which can be computed for each dimension. For Tarkkonen's measure, the multidimensionality is natural, while the one-dimensional model is only a special case. Although the assumptions of Cronbach's alpha do not allow more than one dimension, we display certain results concerning it, since it is so common to use alpha beyond its limitations.

Figure 7.9 shows the relative bias of Tarkkonen's measure when applied to the first factor image and its factor score in the experimental setting IV. The bias is negligible, and it is decreased both by the quality of the model and by the sample size.

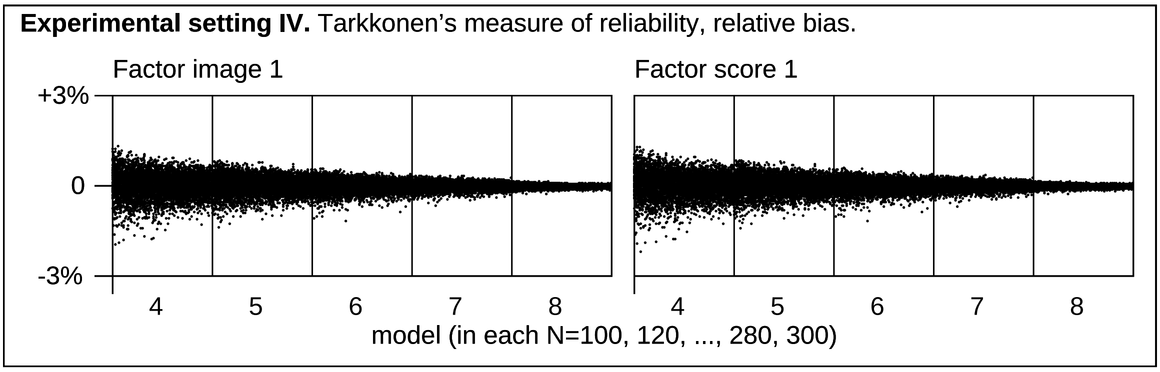

Figure 7.9. Relative bias of Tarkkonen's measure in the experimental setting IV.

The unweighted sum is the traditional scale, although not so clever in this case. However, from Figure 7.10 it is seen that there is no problem with Tarkkonen's measure, while Cronbach's alpha behaves quite suspiciously.

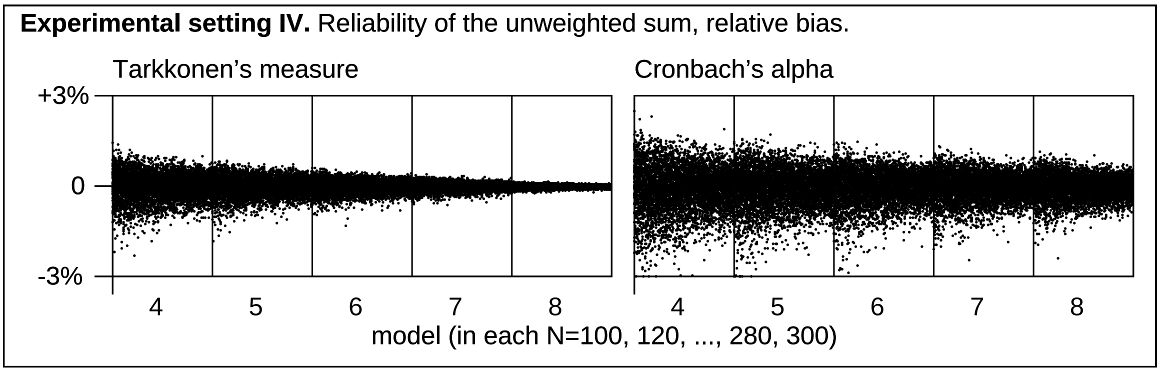

Figure 7.10. Relative bias of Tarkkonen's measure and Cronbach's alpha in the experimental setting IV.

A better scale in practice would be the factor scores. Looking at the second factor score in the experimental setting IV, and especially the proportion of the absolute bias to the root mean square, reveals more suspicious properties of Cronbach's alpha (see Figure 7.11). The proportion of bias increases when the model gets better, although it decreases within models with larger samples. This is probably just an indication of violating the assumptions. In practice, anything can happen when inappropriate methods are used.

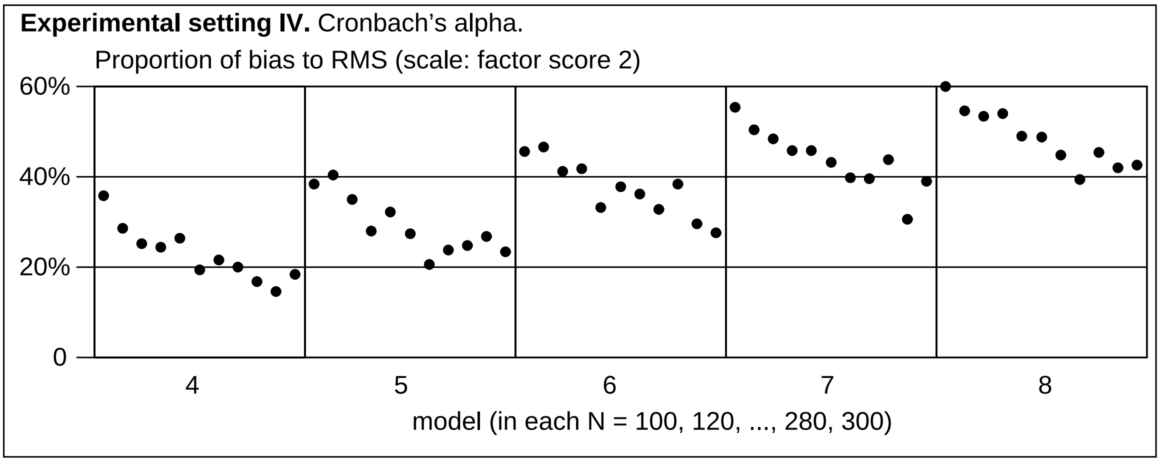

Figure 7.11. Proportion of bias to RMS in Cronbach's alpha in the experimental setting IV.

Tarkkonen's measure gives satisfactory results in the corresponding situation (see Figure 7.12). The proportion of bias remains mostly under 10% (denoted by the white dots).

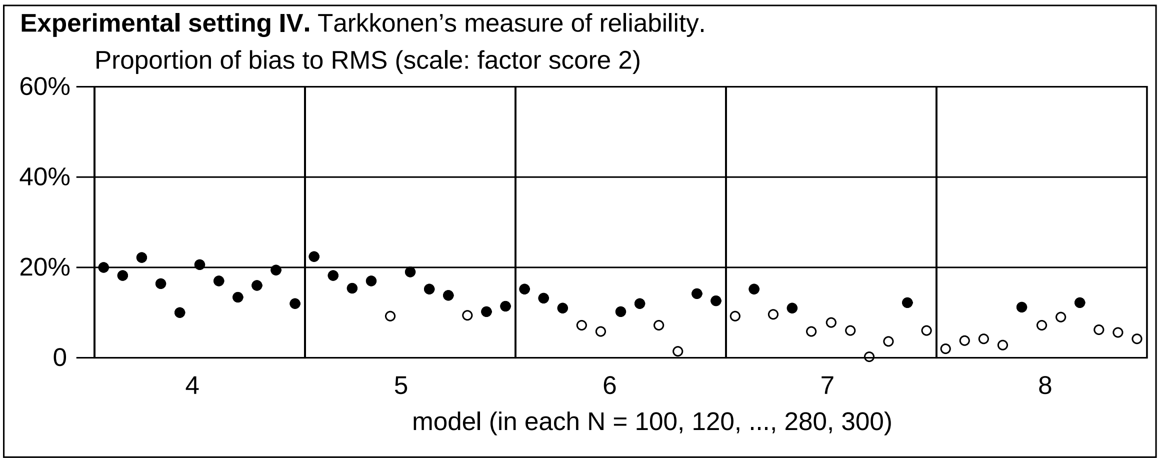

Figure 7.12. Proportion of bias to RMS in Tarkkonen's measure in the experimental setting IV.

7.4.2 Sampling distribution

Finally we consider the sampling distribution of Tarkkonen's measure. We restrict ourselves to the factor images in the experimental setting V, and its model 6, where the theoretical non-zero factor loadings are 0.85 each. It represents a rather accurate measurement.

The sample size has a clear effect on the distribution. Figure 7.13 shows how the shape of the distribution changes, when more observations are used during the simulation. The observed values in the histogram are quite far away from the frequency curve with samples of 100 observations. With a sample size of 200 the normal distribution fits perfectly (χ^2=4.977, *df*=7, *p*=0.66), but the case of 300 takes the distribution too skewed (χ^2=28.74, *df*=5, *p*=0.00).

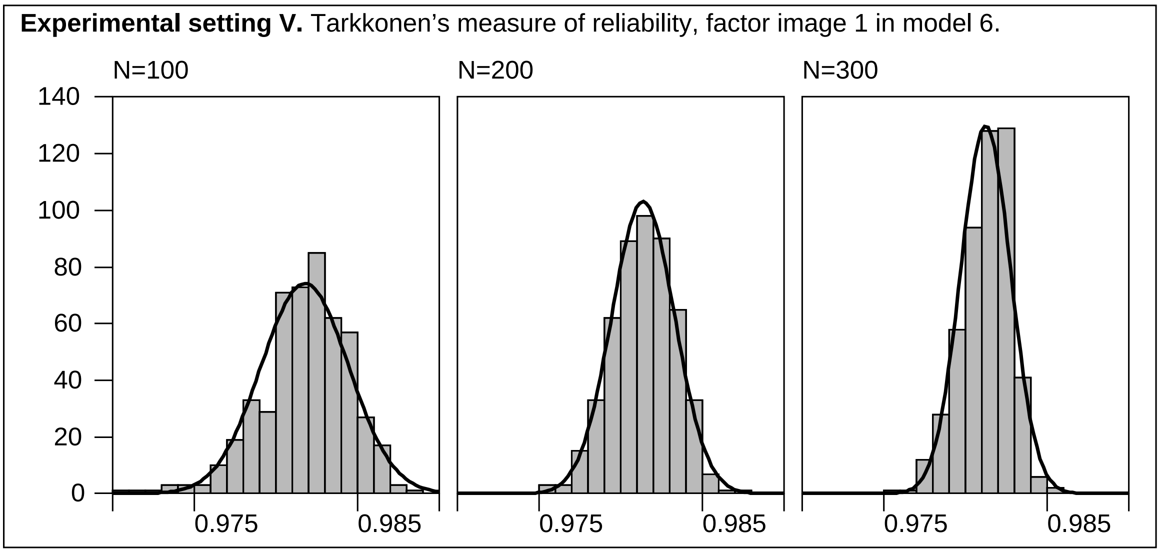

Figure 7.13. Tarkkonen's measure of reliability of the factor image 1 in the experimental setting V, model 6 with selected sample sizes.

Let us widen the view by including the other factor images, and aggregating the observations of all sample sizes (100, 120, ..., 280, 300) together. Hence, we have 5500 observations of each factor image in model 6. The corresponding histograms are presented in Figure 7.14. None of the reliabilities of the factor images seem to be normally distributed (χ^2=102.4, 123.2, and 83.77, with *df*=10, 11, and 10, respectively).

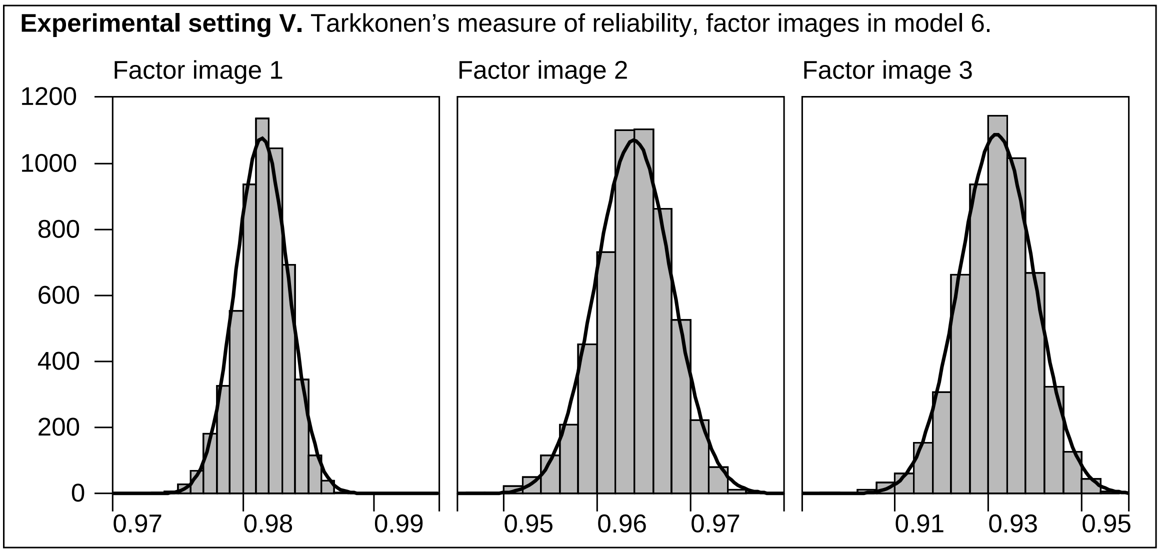

Figure 7.14. Tarkkonen's measure of reliability of the factor images 1, 2, and 3 in the experimental setting V, model 6.

The skewness is likely to increase when the observed values are approaching the theoretical upper limit of one. By using a suitable transformation, the skewness can be smoothed and the distribution relaxed to a wider area. After some experiments, a double logarithm (with suitable sign changes) was applied. The fit gets somewhat better (χ^2=34.33, 37.53, and 46.56, with *df*=12, 12, and 13, respectively), but it is still away from the normal distribution (see Figure 7.15).

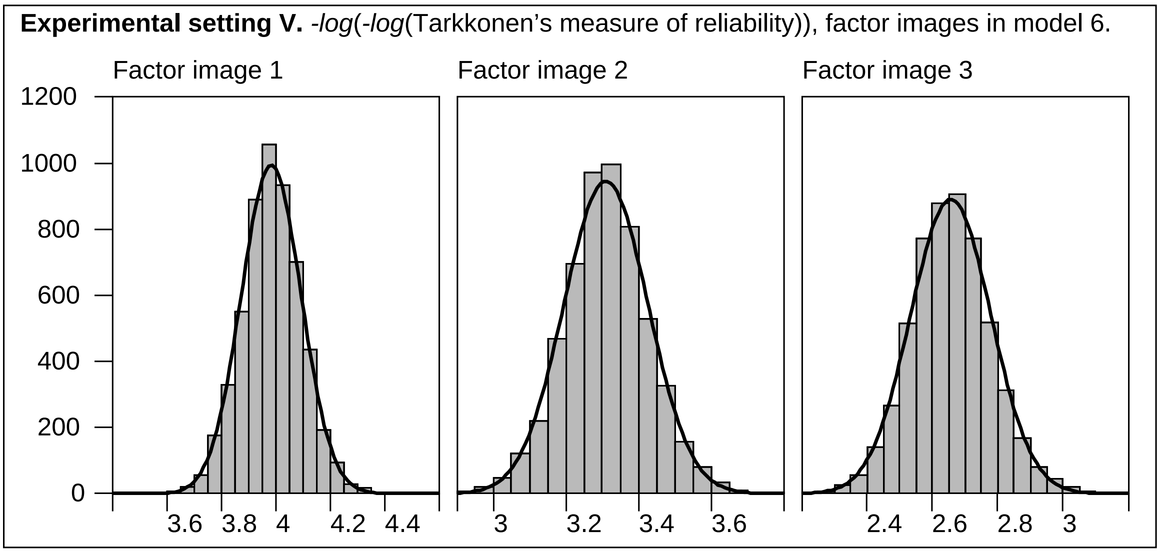

Figure 7.15. Tarkkonen's measure of reliability of the factor images 1, 2, and 3 transformed by *-log(-log())* in the experimental setting V, model 6.

The corresponding analyses with other models gave essentially the same results. Often the distribution is quite close to normal. In some cases the fit is very good. There is certain skewness which increases with larger samples. It might be caused by a slight bias, which is, however, unimportant in practice.

8. Discussion

In the following, we present a short summary of the reliability studies, followed by the conclusions of this study and the suggestions for further research.

8.1 Summary

Reliability is a well defined, and an important property of measurement, in addition to various aspects of validity. Two concepts are needed in order to assess the reliability: the measurement model, which specifies the structure of the measurement, and the measurement scale, which is a combination of the measured items, and represents a realization of the theoretical notions.

This study has concentrated on two measures of reliability, Cronbach's alpha and Tarkkonen's measure. Both are clearly based on the same definition of reliability, but they imply different assumptions about the model and the scale.

The classical true score model assumes each observed item to be a sum of a latent true score and a random measurement error. The model was inherited from the development of factor analysis by Spearman (1904a), and it was much later mathematically formulated by Lord and Novick (1968). The model and its variations are inapplicable in most practical situations. The restrictions are caused by unrealistic assumptions made about the true scores and the measurement errors. This model has been used, often implicitly, as the basis for the reliability studies. Although the multidimensional nature of psychological tests has been well-known since the 1930s, and a great variety of multivariate statistical methods has been developed, the classical true score model has remained in use essentially in its original form. Attempts to generalize the model have been unsatisfactory.

The method of multiple factor analysis, provided by Thurstone (1931), could have formed real grounds for modelling of measurement, but unfortunately the poor computational possibilities restricted its usage for decades. To enable the required calculations, the test items were commonly dichotomous. At that time, a variety of reliability measures were developed by Kuder and Richardson (1937), whose work has had an enormous effect on the later studies. Especially their formula 20, which was extended and renamed to alpha by Cronbach (1951), became a universal reliability measure, although it was soon proved to be a weak lower bound only (Novick and Lewis 1967).

The early measures were properly based on the definition of reliability, but the models and scales were too simple. Cronbach’s alpha was an enhancement compared to the other methods, but when computers began to be generally available, finally smoothing the way for factor analysis, there should have been a replacement for alpha. Perhaps the best suggestion was provided by Heise and Bohrnstedt (1970), envolving a factor-analytic point of view. The measure of Werts, Rock, Linn, and Jöreskog (1978) was notable, but it lacked a proper handling of the measurement scale. These procedures were not adopted as common methods.

The alternative suggestions have been inadequate for two main reasons. Firstly, the assumptions of the measurement model have not been questioned. Secondly, a clear distinction between the model and the measurement scale has not been made. Tarkkonen (1987) was the first, who brought the concept of reliability to the context of multidimensional, composite measurement scales. His general framework allows modelling of a large variety of applications, based on realistic assumptions. The reliability measure is derived according to the definition of reliability, leading to a ratio of variances, which are generated by the true scores and the observed variables. As its special cases, Tarkkonen’s framework contains most of the models, scales and reliability measures presented in psychometric test theory, including the classical true score model and Cronbach’s alpha.

8.2 Conclusions

For historical reasons, the research on reliability and factor analysis has been concentrated on the fields of psychology and the social sciences. A common view is that measurement error is more of a problem there than in the physical sciences. However, this is only partially true, since examples from unreliable measurements could be drawn from all of science. Also, regardless of the application field, the traits and phenomena are seldom one-dimensional. Their analysis requires methods that are capable to correctly handle multiple dimensions.

Cronbach’s alpha has become an obscure formula, which is calculated as a matter of routine. Even in the context of confirmatory factor analysis and structural equations modelling, alpha is generally applied, although its basic assumptions allow only one dimension. This restriction has been circumvented by analyzing the dimensions separately. But, because alpha is based on the internal consistency approach, where all items should be equally good, another bypass is needed. The traditional procedure is to compute a sum of selected items only, for example of those which have the highest loadings on a factor. In other words, some of the items are discarded, often on purpose so that alpha is increased. Some statistical programs even courage this procedure by providing *"alpha if item deleted"* statistics.

The problem here is, that while playing with the reliability, the validity may be lost. There is no point in discarding a part of the data to get a better alpha, as stated already by Armor (1974). Nevertheless, this is how these things are still applied.

Within Tarkkonen's framework, no circumventions are needed, since the concepts of measurement model and the measurement scale are defined on a perfectly general level. The framework allows assessing the structural validity of the model, as well as the reliabilities of multidimensional, composite scales.

The simulation experiments of this study prove that also the statistical properties of Tarkkonen's reliability measure are acceptable: the sample estimates are accurate, if reasonable models are used. The slight bias is not an appreciable disadvantage of the method. Even in those few rare instances, where the assumptions of Cronbach's alpha do hold, Tarkkonen's measure behaves better. It does not leave any objective reason for using alpha. The conclusion is that Tarkkonen's measure evidently supersedes Cronbach's alpha in all applications.

Some attention should be paid to the modelling process in general. Estimating the parameters of the model requires adequate number of observations. Working with too small data sets may cause bias to the estimated parameters, which is then reflected in the reliabilities. The level of measurement is also important in practice. The usual Likert scale items can well be used, but their discrete nature necessarily brings more error to the measurement. That error is best compensated with proper modelling having enough redundancy.

For future work, it is important that Tarkkonen's contribution is published, for example in *Psychometrika*. Further research could then be directed toward bringing the framework into the traditional, statistical models which do not include the measurement error component. This would mean generalizing such methods as linear regression analysis, or multivariate methods related to principal components, *e.g.* correspondence analysis.

This should lead to changes in the procedures of assessing the quality of measurements. It should also sharpen the statistical models, to better manage with the different sources of uncertainty, not only the sampling variation but also the measurement errors. This may require slight changes in the way of thinking about statistical problems. It might not always be easy, as the old argumentation of Spearman and Pearson reminds.

References

Ahmavaara, Y. (1954). Transformation analysis of factorial data. *Annales Academiae Scientiarum Fennicae,* Series B, Vol. 88, 2.

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika,* **18**, 1–14.

Armor, D. J. (1974). *Theta reliability and factor scaling.* In Costner, H. L. (ed.), Sociological Methodology. Jossey-Bass, San Francisco. 17–50.

Bacon, D. R., Sauer, P. L., and Young, M. (1995). Composite reliability in structural equations modeling. *Educational and Psychological Measurement,* **55**, 394–406.

Barchard, K., and Hakstian, A. R. (1997). The effects of sampling model on inference with coefficient alpha. *Educational and Psychological Measurement,* **57**, 893–905.

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *The British Journal of Mathematical and Statistical Psychology,* **48**, 211–220.

Bedeian, A. G., Day, D. V., and Kelloway, E. K. (1997). Correcting for measurement error attenuation in structural equation models: some important reminders. *Educational and Psychological Measurement,* **57**, 785–799.

Benjamin, K. (1945). An I.B.M. technique for the computation of ΣX^2 and ΣXY. *Psychometrika,* **10**, 61–63.

Bentler, P. M. (1968). Alpha-maximized factor analysis (alphamax): its relation to alpha and canonical factor analysis. *Psychometrika,* **33**, 335–345.

Bentler, P. M. (1972). A lower bound method for the dimension-free measurement of internal consistency. *Social Science Research,* **1**, 343–357.

Bentler, P. M., and Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika,* **45**, 249–267.

Bentler, P. M., and Woodward, J. A. (1983). *The greatest lower bound to reliability.* In H. Wainer and S. Messick (eds.), Principals of modern psychological measurement. Erlbaum, New Jersey. 237–253.

Bentler, P. M., and Woodward, J. A. (1985). On the greatest lower bound to reliability. *Psychometrika,* **50**, 245–246.

Blinkhorn, S. F. (1997). Past imperfect, future conditional: Fifty years of test theory. *The British Journal of Mathematical and Statistical Psychology,* **50**, 175–185.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* John Wiley & Sons, New York.

Box, G. E. P., and Müller, M. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics,* **29**, 610–611.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology,* **3**, 296–322.

Brownell, W. A. (1933). On the accuracy with which reliability may be measured by correlating test halves. *Journal of Experimental Education,* **1**, 204–215.

Burt, C. (1955). Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology,* **8**, 103–118.

Callender, J. C., and Osburn, H. G. (1979). An empirical comparison of coefficient alpha, Guttman's Lambda-2, and MSPLIT maximized split-half reliability estimates. *Journal of Educational Measurement,* **16**, 89–99.

Cochran, W. G. (1977). *Sampling Techniques.* 3rd ed., John Wiley & Sons, New York.

Coolican, H. (1994). *Research Methods and Statistics in Psychology.* 2nd ed., Hodder & Stoughton, London.

Crocker, L., and Algina, J. (1986). *Introduction to Classical* & *Modern Test Theory.* Holt, Rinehart and Winston.

Cronbach, L. J. (1943). On estimates of test reliability. *The Journal of Educational Psychology,* **34**, 485–494.

Cronbach, L. J. (1946). A case study of the split-half reliability coefficient. *The Journal of Educational Psychology,* **37**, 473–480.

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika,* **12**, 1–16.

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

Cronbach, L. J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika,* **53**, 63–70.

Cronbach, L. J. (1990). *Essentials of psychological testing.* 5th ed., Harper & Row, New York.

Cronbach, L. J., and Azuma, H. (1962). Internal-consistency reliability formulas applied to randomly sampled single-factor tests: An empirical comparison. *Educational and Psychological Measurement,* **22**, 645–665.

Cronbach, L. J., and Hartmann, W. (1954). A note on negative reliabilities. *Educational and Psychological Measurement,* **14**, 342–346.

Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology*, **16**, 137–163.

Cronbach, L. J., Schönemann, P., and McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement,* **25**, 291–312.

Cronbach, L. J., and Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika,* **16**, 167–188.

Davis, F. B. (1945). The reliability of component scores. *Psychometrika,* **10**, 57–60.

Dressel, P. L. (1940). Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika,* **5**, 305–310.

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

Feldt, L. S. (1975). Estimation of the reliability of a test divided into two parts of unequal lengths. *Psychometrika,* **40**, 557–561.

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

Francis, I. (1974). Factor analysis: fact or fabrication. *Mathematical Chronicle,* **3**, 9–44.

Garnett, J. C. M. (1919). On certain independent factors in mental measurement. *Proceedings of The Royal Society of London,* Series A, **96**, 91–111.

Gilmer, J. S., and Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika,* **48**, 99–111.

Green, B. F. Jr. (1950). A note on the calculation of weights for maximum battery reliability. *Psychometrika,* **15**, 57–61.

Greene, V. L., and Carmines, E. G. (1979). *Assessing the reliability of linear composites.* In Schuessler, K. F. (ed.), Sociological Methodology. Jossey-Bass, San Francisco. 160–175.

Guilford, J. P. (1954). *Psychometric Methods.* 2nd ed., McGraw-Hill.

Guilford, J. P., and Fruchter, B. (1978). *Fundamental Statistics in Psychology and Education.* 6th ed., McGraw-Hill.

Gulliksen, H. (1950). *Theory of Mental Tests.* John Wiley & Sons, New York.

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

Handy, U., and Lentz, T. F. (1934). Item value and test reliability. *Journal of Educational Psychology,* **25**, 703–708.

Harman, H. H. (1967). *Modern Factor Analysis.* 2nd ed., University of Chicago Press.

Heise, D. R., and Bohrnstedt, G. W. (1970). *Validity, invalidity and reliability.* In Borgatta, E. F. and Bohrnstedt, G. W. (eds.), Sociological Methodology. Jossey-Bass, San Francisco. 104–129.

Holzinger, K. J. (1923). Note on the use of Spearman's prophecy formula for reliability. *Journal of Educational Psychology,* **14**, 302–305.

Horst, P. (1953). Correcting the Kuder-Richardson reliability for dispersion of item difficulties. *Psychological Bulletin,* **50**, 371–374.

Horst, P. (1966). *Psychological Measurement and Prediction.* Wadsworth.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology,* **24**, 417–441, 498–520.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika,* **6**, 153–160.

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

Jones, L. V., and Appelbaum, M. I. (1989). Psychometric methods. *Annual Review of Psychology,* **40**, 23–43.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika,* **32**, 443–482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika,* **34**, 183–202.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika,* **57**, 239–251.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika,* **36**, 109–133.

Jöreskog, K. G., and Lawley, D. N. (1968). New methods in maximum likelihood factor analysis. *The British Journal of Mathematical and Statistical Psychology,* **21**, 85–96.

Jöreskog, K. G., and Sörbom, D. (1983). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood.* International Educational Services, Chicago.

Kaiser, H. F. (1960). The application of electronic computers in factor analysis. *Educational and Psychological Measurement,* **20**, 141–151.

Kaiser, H. F., and Caffrey, J. (1965). Alpha factor analysis. *Psychometrika,* **30**, 1–14.

Kaiser, H. F., and Michael, W. B. (1977). Little Jiffy factor scores and domain validities. *Educational and Psychological Measurement,* **37**, 363–365.

Kelley, T. L. (1921). The reliability of test scores. *Journal of Educational Research,* **3**, 370–379.

Kelley, T. L. (1924). Note on the reliability of a test: a reply to Dr. Crum's criticism. *Journal of Educational Psychology,* **15**, 193–204.

Kelley, T. L. (1925). The applicability of the Spearman-Brown formula for the measurement of reliability. *Journal of Educational Psychology,* **16**, 300–303.

Kelley, T. L. (1942). The reliability coefficient. *Psychometrika,* **7**, 75–83.

Kernighan, B. W., and Ritchie, D. M. (1988). *The C programming language.* 2nd ed., Prentice Hall, New Jersey.

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

Kristof, W. (1972). On a statistic arising in testing correlation. *Psychometrika,* **37**, 377–384.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika,* **39**, 491–499.

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

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of The Royal Society of Edinburgh,* **60**, 64–82.

Lawley, D. N., and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method.* 2nd ed., Butterworths, London.

Ledermann, W. (1938). Shortened method for estimation of mental factors by regression. *Nature*, **141**, 650.

Lefkowitz, M. M., and Greene, H. E. (1962). Obtaining components essential to a number of statistical analyses by use of the IBM accounting machine. *Educational and Psychological Measurement,* **22**, 183–186.

Levy, P. (1995). Charles Spearman's contributions to test theory. *The British Journal of Mathematical and Statistical Psychology,* **48**, 221–235.

Liou, M. (1989). A note on reliability estimation for a test with components of unknown functional lengths. *Psychometrika,* **54**, 153–163.

Liukkonen, J., and Leskinen, E. (1999). The reliability and validity of scores from the children's version of the perception of success questionnaire. *Educational and Psychological Measurement,* **59**, 651–664.

Lord, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika,* **20**, 1–22.

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika,* **23**, 291–296.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley, London.

Lyerly, S. B. (1958). The Kuder-Richardson formula (21) as a split-half coefficient, and some remarks on its basic assumption. *Psychometrika,* **23**, 267–270.

Mosier, C. I. (1936). A note on item analysis and the criterion of internal consistency. *Psychometrika,* **1**, 275–282.

Mosier, C. I. (1941). A short cut in the estimation of split-halves coefficients. *Educational and Psychological Measurement,* **1**, 407–408.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika,* **8**, 161–168.

Mustonen, S. (1966). Symmetrinen transformaatioanalyysi [Symmetric transformation analysis]. *Alkoholipoliittisen tutkimuslaitoksen tutkimusseloste N:o* 24.

Mustonen, S. (1988). *Sucros in SURVO 84C.* SURVO 84C Contributions 2. Department of Statistics, University of Helsinki.

Mustonen, S. (1989). *Programming SURVO 84 in C.* SURVO 84C Contributions 3. Department of Statistics, University of Helsinki.

Mustonen, S. (1992). *Survo, An Integrated Environment for Statistical Computing and Related Areas.* Survo Systems, Helsinki.

Mustonen, S. (1995). *Tilastolliset monimuuttujamenetelmät [Multivariate statistical methods].* Survo Systems, Helsinki.

Mustonen, S. (1999). *Matrix computations in Survo.* Proceedings in the Eighth International Workshop on Matrices and Statistics, Department of Mathematical Sciences, University of Tampere. *http://www.helsinki.fi/survo/matrix99.html*

Mustonen, S., and Vehkalahti, K. (1997). Survo as an environment for statistical research and teaching. *Bulletin of ISI*, Istanbul, Proceedings 2, 69–72. *http://www.helsinki.fi/survo/isi97.html*

Novick, M. R., and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, **32**, 1–13.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory.* 3rd ed., McGraw-Hill.

Pandey, T. N., and Hubert, L. (1975). An empirical comparison of several interval estimation procedures for coefficient alpha. *Psychometrika,* **40**, 169–181.

Payne, W. H., and Anderson, D. E. (1968). Significance levels for the Kuder-Richardson twenty: an automated sampling experiment approach. *Educational and Psychological Measurement,* **28**, 23–39.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **6**, 559–572.

Peel, E. A. (1947). Prediction of a complex criterion and battery reliability. *British Journal of Psychology, Statistical Section,* **1**, 84–94.

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

Raju, N. S. (1979). Note on two generalizations of coefficient alpha. *Psychometrika,* **44**, 347–349.

Raju, N. S. (1982). On test homogeneity and maximum KR-20. *Educational and Psychological Measurement,* **42**, 145–152.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika,* **20**, 93–111.

Reinhardt, B. (1996). *Factors affecting coefficient alpha: a mini Monte Carlo study.* In Thompson, B. (ed.), Advances in Social Science Methodology. Texas A&M University. **4**, 3–20.

Reuterberg, S., and Gustafsson, J.-E. (1992). Confirmatory factor analysis and reliability: testing measurement model assumptions. *Educational and Psychological Measurement,* **52**, 795–811.

Richardson, M. W. (1936). Notes on the rationale of item analysis. *Psychometrika,* **1**, 69–76.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review,* **9**, 99–103.

Seber, G. A. F. (1984). *Multivariate Observations.* John Wiley & Sons, New York.

Sedere, M. V., and Feldt, L. S. (1977). The sampling distributions of the Kristof reliability coefficient, the Feldt coefficient, and Guttman's Lambda-2. *Journal of Educational Measurement,* **14**, 53–62.

Shapiro, A. (1985). A note on the asymptotic distribution of the greatest lower bound to reliability. *Psychometrika,* **50**, 243–244.

Sharp, S. (1997). "Much more at home with 3.999 pupils than with four": The contributions to psychometrics of Sir Godfrey Thomson. *The British Journal of Mathematical and Statistical Psychology,* **50**, 163–174.

Smith, K. W. (1974). On estimating the reliability of composite indexes through factor analysis. *Sociological methods & research,* **2**, 485–510.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology,* **15**, 72–101.

Spearman, C. (1904b). General intelligence objectively determined and measured. *American Journal of Psychology,* **15**, 201–293.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology,* **3**, 271–295.

Tarkkonen, L. (1987). *On Reliability of Composite Scales.* Statistical studies 7. Finnish Statistical Society.

ten Berge, J. M. F., Snijders, T. A. B., and 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., and Hofstee, W. K. B. (1999). Coefficient alpha and reliabilities of unrotated and rotated components. *Psychometrika,* **64**, 83–90.

Tezuka, S., and L'Ecuyer, P. (1991). Efficient and portable combined Tausworthe random number generators. *ACM Transactions on Modelling and Computer Simulation,* **1**, 99–112.

Thomson, G. H. (1939). *The Factorial Analysis of Human Ability.* University of London Press.

Thomson, G. H. (1940). Weighting for battery reliability and prediction. *British Journal of Psychology,* **30**, 357–366.

Thurstone, L. L. (1931). Multiple Factor Analysis. *Psychological Review,* **38**, 406–427.

Thurstone, L. L. (1935). *Vectors of Mind.* The University of Chicago Press, Chicago.

Thurstone, L. L. (1947). *Multiple-Factor Analysis.* The University of Chicago Press, Chicago.

Tucker, L. R. (1949). A note on the estimation of test reliability by the Kuder-Richardson formula (20). *Psychometrika,* **14**, 117–119.

Tucker, L. R., and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika,* **38**, 1–10.

Weiss, D. J., and Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology,* **32**, 629–658.

Werts, C. E., Rock, R. D., Linn, R. L., and Jöreskog, K. G. (1978). A general method of estimating the reliability of a composite. *Educational and Psychological Measurement,* **38**, 933–938.

Wherry, R. J., and Gaylord, R. H. (1943). The concept of test and item reliability in relation to factor pattern. *Psychometrika,* **8**, 247–264.

Wilcox, R. R. (1992). Robust generalizations of classical test reliability and Cronbach's alpha. *The British Journal of Mathematical and Statistical Psychology,* **45**, 239–254.

Woodhouse, B., and Jackson, P. W. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous types: II. A search procedure to locate the greatest lower bound. *Psychometrika,* **42**, 579–592.

Woodward, J. A., and Bentler, P. M. (1978). A statistical lower bound to population reliability. *Psychological Bulletin,* **85**, 1323–1326.

Woodward, J. A., and Bentler, P. M. (1979). Application of optimal sign-vectors to reliability and cluster analysis. *Psychometrika,* **44**, 337–341.

Zimmerman, D. W., Zumbo, B. D., and Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement,* **53**, 33–49.

Zubin, J. (1934). The method of internal consistency for selecting test items. *Journal of Educational Psychology,* **25**, 345–356.

Appendix A

Programs for computations and simulations

This appendix includes source codes of the programs made by the author for computing the reliabilities and implementing the Monte Carlo simulation experiments described in chapter 6. The appendix consists of four sections:

A.1 Survo module RELIAB for computing the reliabilities
A.2 Sucro program for reliability simulations
A.3 Implementation of the simulation experiments
A.4 Sucro program for computing the deviation statistics

A.1 Survo module RELIAB for computing the reliabilities

Survo is an open system, where new modules may be programmed in the C language, applying ready-made library functions for various tasks (Mustonen 1989). The following Survo module RELIAB has been programmed by the author for computing the reliabilities of different measurement scales, according to the input provided by the user. The module relies heavily on the general routines of Survo, but it is not among the most typical modules, as it does not utilize any data files. Instead, it operates with the matrix files. The input to the module consists of various matrices, such as correlation and factor matrices.

The words given in **boldface** refer to the functions, global variables and other definitions provided by the C libraries of Survo (Mustonen 1989). The *emphasized* functions are defined within the RELIAB module. Other functions belong to the standard C libraries (see Kernighan and Ritchie 1988, for example).

The modularity of Survo allows using large number of global variables within the modules. The other modules are not disturbed in any way, because the modules are independent of each other. They communicate only with the main process (the Survo editor), one at a time, when the corresponding command is activated. At runtime, the modules share a variety of global variables with the editor process and they may modify them as well. Knowing these simple rules, the usage of global variables shortens the function calls and makes the code simpler.

The source code includes certain additional features for the simulation purposes of this study. They are highlighted in the text by black on grey. Those features are used by the sucro program /RLBSIMUL, which is described in section A.2.

```
/* reliab.c (c) K.Vehkalahti 1993-2000 */
#include <stdio.h>
#include <stdlib.h>
#include <stdarg.h>
#include <ctype.h>
#include <malloc.h>
#include <math.h>
#include "survo.h"
#include "survoext.h"
extern char **spa, **spb; /* from survodat.h */

int results_line, model, alpha, orthogonal, errorneous;
double *CORR, *FACT, *RFACT, *COEFF, *COEFF2, *MSN, *COV;
double *FT, *TMP, *RCOV, *RCOVd, *FC, *COEF, *COVd;
int mX, nF, nW, nW2;
char *clabR, *clabF, *clabW, *rlabW, *clabW2, *rlabW2, *clab, *rlab;
int lr, lc, type;
char expr[129], weight[LNAME];
int wdim, w2dim;

int simul;
FILE *outfile;
double LSum;

#define NORMAL12  (model==1 || model==2)
#define WEIGHTED  (model>=3)
#define NO_RFACT  (model==1 || model==3 || model==5)
#define WEIGHTED2 (model>4)

void usage (void);
int check_parameters (void);
int check_specifications (void);
void no_memory (void);
int allocate_memory (void);
int compute_weights (void);
void free_spaces (void);
int reliabilities (void);
void print_line (void);
char *trim1 (char *);
void trim2 (char *);
void save_resmats (void);
void write_line (void);
void add_loadms (void);
void do_cov_matrix (void);
double compute_alpha (void);
void s_err (char *, ...);
```

```
void main (int argc, char *argv[])
{
      int i;
      if (argc==1) {
            printf("Use this program as a Survo module only!\n");
            return;
      }
      s_init (argv[1]);
      spec_init (r1+r-1);
      i=check_parameters ();   if (i<0) return;
      i=check_specifications ();   if (i<0) return;
      i=allocate_memory ();   if (i<0) return;
      i=compute_weights ();   if (i<0) return;
      reliabilities ();
      free_spaces ();
}

void usage (void) { s_err ("See: RELIAB?"); }

void s_err (char *fmt, ...)
{
      char buffer[LLENGTH];
      va_list ap;

      va_start (ap, fmt);
      vsprintf (buffer, fmt, ap);
      va_end (ap);

      sur_print ("\n");
      sur_print (buffer);
      WAIT;
}

int check_parameters (void)
{
      int i, j, k;

      if (g<3) {
            usage ();
            return -1;
      }
      results_line=0; model=1; alpha=0;
      i=matrix_load (word[1], &CORR, &mX, &j, &rlab, &clabR, &lr, &lc, &type, expr);
      if (i<0) return -1;
      if (j!=mX) {
            s_err ("%s is not a proper correlation matrix!", word[1]);
            return -1;
      }
      i=matrix_load (word[2], &FACT, &j, &nF, &rlab, &clabF, &lr, &lc, &type, expr);
      if (i<0) return -1;
      if (j!=mX) {
            s_err ("Correlation matrix and factor matrix are incompatible!");
            return -1;
      }
```

```
    if (g>3) {
        i=edline2 (word[3], 1, 0);
        if (i==0) {
            i=matrix_load (word[3], &RFACT, &j, &k, &rlab, &clab, &lr, &lc, &type, expr);
            if (i<0) return -1;
            model=2;
            if (j!=k) {
                s_err ("%s is not a proper factor correlation matrix!", word[3]);
                return -1;
            }
            if (j!=nF) {
                s_err ("Factor correlation matrix and factor matrix are incompatible!");
                return -1;
            }
        } else {
            results_line=i;
        }
    }
    if (g>4) {
        if (results_line) {
            s_err ("Wrong number of parameters!");
            usage ();
            return -1;
        }
        i=edline2 (word[4], 1, 0);
        if (i==0) {
            s_err ("Invalid edit line: %s!", word[4]);
            return -1;
        }
        results_line=i;
    }
    return 1;
}

int check_specifications (void)
{
    int i, j, k, n;

    j=spfind ("WEIGHT");
    if (j>=0 && strcmp (spb[j], "0")) {
        strcpy(weight, spb[j]);
        i=matrix_load (weight, &COEFF, &k, &nW, &rlabW, &clabW, &lr, &lc, &type, expr);
        if (i<0) return -1;
        model+=2;
        if (k-1!=mX && k!=mX ) {
            s_err ("Incompatible dimensions in coefficient matrix %s!", weight);
            return -1;
        }
        wdim=k;
```

```
        j=spfind ("WEIGHT2");
        if (j>=0) {
            strcpy (sbuf, spb[j]);
            strcat (weight, "*"); strcat (weight, sbuf);
            i=matrix_load (sbuf, &COEFF2, &k, &nW2, &rlab, &clabW2, &lr, &lc, &type, expr);
            if (i<0) return -1;
            model+=2;
            if (k-1!=nW && k!=nW ) {
                s_err ("Incompatible dimensions in coefficient matrix %s!", spb[j]);
                return -1;
            }
            w2dim=k;
        }
    }
    j=spfind ("RESULTS");
    if (j>=0) results=atoi (spb[j]);
    j=spfind ("MSN");
    if (j>=0) {
        alpha=1;
        if (*spb[j]=='*') {
            MSN=(double *)malloc (mX*2*sizeof (double));
            if (MSN==NULL) { no_memory (); return -1; }
            for (i=0; i<mX; i++) for (j=0; j<2; j++)
                MSN[i+mX*j]=(double)j;
        } else {
            i=matrix_load (spb[j], &MSN, &k, &n, &rlab, &clab, &lr, &lc, &type, expr);
            if (i<0) return -1;
            if (n<2 || n>3) {
                s_err ("%s is not a proper MSN-matrix!", spb[j]);
                return -1;
            }
            if (k!=mX) {
                s_err ("Incompatible dimensions in MSN-matrix %s!", spb[j]);
                return -1;
            }
        }
        i=matrix_load (word[1], &COV, &mX, &n, &rlab, &clab, &lr, &lc, &type, expr);
        if (i<0) return -1;
    }
    simul=0;
    j=spfind ("OUTFILE");
    if (j>=0) {
        outfile=fopen (spb[j],"a");
        if (outfile==NULL) {
            s_err ("Can not open output file %s!", spb[j]);
            return -1;
        }
        LSum=0.0;
        simul=1;
    }
    return 1;
}

void no_memory (void) { s_err ("Not enough memory!"); }
```

```
int allocate_memory (void)
{
    int i, j;
    double chsum;

    orthogonal=1;
    if (NO_RFACT) {
        RFACT=(double *)malloc (nF*nF*sizeof (double));
        if (RFACT==NULL) { no_memory (); return -1; }
        for (i=0; i<nF; i++) for (j=0; j<nF; j++)
            if (i==j) RFACT[i+nF*j]=1.0; else RFACT[i+nF*j]=0.0;
    } else {
        chsum=0.0;
        for (i=0; i<nF; i++) for (j=0; j<nF; j++)
            chsum+=RFACT[i+nF*j];
        if (chsum!=nF) orthogonal=0;
    }
    if (orthogonal) {
        for (i=0; i<mX; i++) for (j=0; j<nF; j++)
            if (FACT[i+mX*j] > 1.0) {
                s_err ("Error: factor loading > 1.0 in element (%d,%d) of factor matrix!",
                    i+1, j+1);
                return -1;
            }
    }
    FT=(double *)malloc (mX*nF*sizeof (double));
    if (FT==NULL) { no_memory (); return -1; }
    TMP=(double *)malloc (mX*nF*sizeof (double));
    if (TMP==NULL) { no_memory(); return -1; }
    RCOV=(double *)malloc (mX*mX*sizeof (double));
    if (RCOV==NULL) { no_memory (); return -1; }
    RCOVd=(double *)malloc (mX*1*sizeof (double));
    if (RCOVd==NULL) { no_memory (); return -1; }
    FC=(double *)malloc (mX*1*sizeof (double));
    if (FC==NULL) { no_memory (); return -1; }
    if (alpha) {
        COVd=(double *)malloc (mX*1*sizeof (double));
        if (COVd==NULL) { no_memory (); return -1; }
    }
    return 1;
}

int compute_weights (void)
{
    int i, j, k, l;
    double a;

    if (!WEIGHTED) return 1;
    if (WEIGHTED2) {
        COEF=(double *)malloc ((mX+1)*nW2*sizeof (double));
        if (COEF==NULL) { no_memory (); return -1; }
        for (i=0; i<mX+1; i++) for (j=0; j<nW2; j++)
            COEF[i+(mX+1)*j]=0.0;
        for (i=1; i<mX+1; i++) {
            for (j=0; j<nW2; j++) {
                a=0.0;
```

```
                if (wdim==mX) {
                    if (w2dim==nW) {
                        for (k=0, l=0; k<nW && l<nW; k++, l++)
                            a+=COEFF[(i-1)+mX*k]*COEFF2[(l+0)+nW*j];
                    } else {
                        for (k=0, l=0; k<nW && l<nW; k++, l++)
                            a+=COEFF[(i-1)+mX*k]*COEFF2[(l+1)+(nW+1)*j];
                    }
                } else {
                    if (w2dim==nW) {
                        for (k=0, l=0; k<nW && l<nW; k++, l++)
                            a+=COEFF[i+(mX+1)*k]*COEFF2[(l+0)+nW*j];
                    } else {
                        for (k=0, l=0; k<nW && l<nW; k++, l++)
                            a+=COEFF[i+(mX+1)*k]*COEFF2[(l+1)+(nW+1)*j];
                    }
                }
                COEF[i+(mX+1)*j]=a;
            }
        }
        if (wdim==mX) {
            rlabW2=(char *)malloc ((mX+1)*8*sizeof (char));
            if (rlabW2==NULL) { no_memory (); return -1; }
            strcpy (rlabW2, "Constant"); strcat (rlabW2, rlabW);
        } else {
            rlabW2=rlabW;
        }
        sprintf (sbuf, "Second_order_scale_coefficients_(%s)", weight);
        matrix_save ("WEIGHT2.M", COEF, mX+1, nW2, rlabW2, clabW2, lr, lc, -1, sbuf, 0, 0);
        free (rlabW2);
    } else {
        COEF=(double *)malloc ((mX+1)*nW*sizeof (double));
        if (COEF==NULL) { no_memory (); return -1; }
        for (i=0; i<mX+1; i++) {
            for (j=0; j<nW; j++) {
                if (wdim==mX) {
                    if (i==0) COEF[i+(mX+1)*j]=0.0;
                    else COEF[i+(mX+1)*j]=COEFF[(i-1)+mX*j];
                } else {
                    COEF[i+(mX+1)*j]=COEFF[i+(mX+1)*j];
                }
            }
        }
    }
    return 1;
}

void free_spaces (void)
{
    if (WEIGHTED) free (COEF);
    free (FC); free (RCOVd); free (TMP); free (FT);
    if (NO_RFACT) free (RFACT);
    if (WEIGHTED2) free (COEFF2);
    if (WEIGHTED) free (COEFF);
```

```
    if (!NO_RFACT) free (RFACT);
    free (FACT); free (CORR);
}

int reliabilities (void)
{
    int h, i, j, k, l, q, up_lim;
    double a, b, rxx;
    char mod[LNAME], lab[LNAME], reli[LNAME], alph[LNAME];

    mat_transp (FT, FACT, mX, nF);
    mat_mlt (TMP, FACT, RFACT, mX, nF, nF);
    mat_mlt (RCOV, TMP, FT, mX, nF, mX);
    for (i=0; i<mX; i++) for (j=0; j<mX; j++)
        RCOV[i+mX*j]=CORR[i+mX*j]-RCOV[i+mX*j];
    errorneous=0;
    for (i=0, j=0; i<mX; i++, j++) {
        RCOVd[i]=RCOV[i+mX*j];
        if (RCOVd[i] < 0.0) errorneous=1;
    }
    for (i=0; i<mX; i++) for (j=0; j<mX; j++)
        if (RCOVd[i]<=0.0) CORR[i+mX*j]=0.0;
        else CORR[i+mX*j]=RCOV[i+mX*j]/sqrt(RCOVd[i]);
    for (i=0; i<mX; i++) for (j=0; j<mX; j++)
        if (RCOVd[i]<=0.0) CORR[i*mX+j]=0.0;
        else CORR[i*mX+j]/=sqrt(RCOVd[i]);
    save_resmats ();
    output_open (eout);
    strcpy (sbuf, "Reliabilities according to models E2 and E3:");
    if (WEIGHTED) { sprintf (mod, " (weighted by %s)", weight); strcat (sbuf, mod); }
    print_line ();
    strcpy (sbuf, "E2: errors do not correlate; E3: errors may correlate.");
    print_line ();
    if (alpha) do_cov_matrix ();
    if (WEIGHTED)  up_lim=nW;
    if (WEIGHTED2) up_lim=nW2;
    if (NORMAL12)  up_lim=nF;
    for (k=0; k<=up_lim; k++) {
        if (k<up_lim) {
            if (WEIGHTED) {
                for (i=0; i<mX; i++) FC[i]=COEF[(i+1)+(mX+1)*k];
                if (WEIGHTED2) strncpy (lab, &clabW2[lc*k], lc);
                else strncpy (lab, &clabW[lc*k], lc);
                lab[lc]='\0'; trim2 (lab);
            }
            if (NORMAL12) {
                for (i=0; i<mX; i++) FC[i]=FACT[i+mX*k];
                strncpy (lab, &clabF[lc*k], lc);
                lab[lc]='\0'; trim2 (lab);
```

```
                if (simul) {
                    a=0.0;
                    for (i=0; i<mX; i++) a+=FC[i]*FC[i];
                    fprintf (outfile, "L\\%s %.10f\n", lab, a);
                    LSum+=a;
                }
            }
        } else {
            if (WEIGHTED) break;
            for (i=0; i<mX; i++) FC[i]=1.0;
            strcpy (lab, "Sum");
            if (simul) fprintf (outfile, "L\\%s %.10f\n", lab, LSum);
        }
        if (alpha) {
            rxx=compute_alpha ();
            if (simul) fprintf (outfile, "a\\%s %.10f\n", lab, rxx);
            fnconv (rxx, accuracy, alph);
            trim1 (alph); trim2 (alph);
        }
        mat_mlt (TMP, FACT, RFACT, mX, nF, nF);
        mat_transp (FT, FACT, mX, nF);
        mat_mlt (CORR, TMP, FT, mX, nF, mX);
        mat_mlt (TMP, FC, CORR, 1, mX, mX);
        mat_mlt (&a, TMP, FC, 1, mX, 1);
        for (l=0; l<2; l++) {
            if (l==0) {
                for (i=0; i<mX; i++) TMP[i]=FC[i]*RCOVd[i];
                sprintf (mod, "%s\\E2", lab);
            } else {
                mat_mlt (TMP, FC, RCOV, 1, mX, mX);
                sprintf (mod, "%s\\E3", lab);
            }
            mat_mlt (&b, TMP, FC, 1, mX, 1);
            rxx=1.0/(1.0+b/a);
            if (simul) fprintf (outfile, "%s %.10f\n", mod, rxx);
            fnconv (rxx, accuracy, reli);
            if (l==0) {
                sprintf (sbuf, "%s=%s  ", mod, trim1 (reli));
            } else {
                strcat (sbuf, mod); strcat (sbuf, "=");
                strcat (sbuf, trim1 (reli));
                if (alpha) {
                    strcat (sbuf, "   (Cronbach's alpha:");
                    strcat (sbuf, alph); strcat (sbuf, ")");
                }
                print_line ();
            }
        }
    }
    output_close (eout);
    add_loadms ();
    return 1;
}
```

```
void print_line (void)
{
    int res_line;
    if (results<31) res_line=0; else res_line=results_line;
    if (results>=0) output_line (sbuf, eout, res_line);
    if (res_line) ++results_line;
}

char *trim1 (char *s)
{
    while (*s==' ') ++s;
    return (s);
}

void trim2 (char *s)
{
    while (*s) s++; s--;
    while(*s==' ') s--; s++;
    *s='\0';
}

void save_resmats (void)
{
    strcpy (sbuf, "Residual_covariances");
    matrix_save ("RCOV.M", RCOV, mX, mX, clabR, clabR, lr, lc, 10, sbuf, 0, 0);
    strcpy (sbuf, "Residual_correlations");
    matrix_save ("RCORR.M", CORR, mX, mX, clabR, clabR, lr, lc, 10, sbuf, 0, 0);
}

void add_loadms (void)
{
    if (results_line==0) return;
    if (results<31) return;
    if (errorneous) {
        strcpy (sbuf, "At least one communality was greater than 1, which implies");
        write_line ();
        strcpy (sbuf, "that there might be error in the data or in the model!");
        write_line ();
    } else {
        strcpy (sbuf, "LOADM RCOV.M,##.###,END+2 / ");
        strcat (sbuf, "Residual covariance matrix");
        write_line ();
        strcpy (sbuf, "LOADM RCORR.M,##.###,END+2 / ");
        strcat (sbuf, "Residual correlation matrix");
        write_line ();
        strcpy (sbuf, "LIMITS=-0.9,-0.2,-0.1,0.1,0.2,0.9,1 ");
        strcat (sbuf, "SHADOWS=7,8,1,0,1,8,7 ");
        write_line ();
    }
}
```

```
void write_line (void)
{
    edwrite (space, results_line, 1);
    edwrite (sbuf, results_line, 1); results_line++;
}

void do_cov_matrix (void)
{
    int i, j;
    double a;

    for (i=0; i<mX; i++) {
        a=MSN[i+mX]; COVd[i]=a*a;
        for (j=0; j<mX; j++) {
            COV[i+mX*j]*=a; COV[i*mX+j]*=a;
        }
    }
}

double compute_alpha (void)
{
    int i, j;
    double a, b;

    a=0.0; b=0.0;
    for (i=0; i<mX; i++) a+=FC[i]*COVd[i]*FC[i];
    mat_mlt (TMP, FC, COV, 1, mX, mX);
    mat_mlt (&b, TMP, FC, 1, mX, 1);
    return ((double)mX/((double)mX-1.0)*(1.0-a/b));
}
```

A.2 Sucro program for reliability simulations

The operations of Survo are combined with its macro language as sucros (Mustonen 1988). The following sucro program /RLBSIMUL has been programmed by the author for studying the reliability measures by Monte Carlo simulation. The sucro code consists of nearly 700 lines. A great part of that is comments, which document the features and the control flow of the sucro. Much of the actual sucro code keeps track of the administrative things like result files and random number seeds, thus glueing the operations together in a reasonable way. The way of using /RLBSIMUL is demonstrated in chapter 6.2.4.

The most essential points of the sucro are marked with **white on black**. They are:

"Start" (on line 161): Up to this point, the sucro has checked that enough parameters are given, and displayed short instructions, if needed. Then, it checks that the given parameters are valid, and creates the frames for the simulation.

"main" (on line 438): The simulation takes place in a loop. The results are saved to various files for further processing.

"Return" (on line 605): The sucro ends its job and terminates.

```
   1  1 SURVO 98  Wed Mar 01 21:33:33 2000              D:\V\    1000   100 0
   1 *
   2 *TUTSAVE RLBSIMUL
   3 / RLBSIMUL.TUT (c) K.Vehkalahti 1994-2000
   4 / ***********************************************************************
   5 / [Original structure based on SM’s sucros for transformation analysis.]
   6 / ***********************************************************************
   7 /
   8 / Set the fastest speed and make sure that the environment is "normal",
   9 / i.e. the insert mode is not on, the reference point is cancelled,
  10 / the first column is set to the point of return when ENTER is pressed,
  11 / no blocks or shadow lines are being defined, etc.
  12 /
  13 *{tempo -1}{init}
  14 /
  15 / If the first parameter is "ERR", then a severe error has occurred,
  16 / and a special error handler is automatically launched. This sucro
  17 / includes its own error handler (labeled "Error_handler"), which is
  18 / then called. It is important to check this situation in the first
  19 / place before continuing further.
  20 /
  21 - if W1 ’=’ ERR then goto Error_handler
  22 /
  23 / Before altering anything in the edit field, the sucro must make sure
  24 / that it can perfectly restore the situation after it has ended.
  25 / A ready-made sucro SUR-SAVE is used as a subroutine to save the
  26 / current setup (the edit field and the position of the cursor).
  27 / The setup is saved in to files in the directory of temporary files
  28 / using the name RLBSIMUL as the common part. That name is passed on
  29 / in the sucro memory as the item W1.
  30 / Before SUR-SAVE is called, the stack-like sucro memory is saved,
  31 / because all sucros use the same sucro memory.
  32 /
  33 *{save stack}{W1=RLBSIMUL}{call SUR-SAVE}
  34 /
  35 / The sucro memory which includes the original parameters given by
  36 / the user, is restored. Controlling the speed and breaking the
  37 / execution of the sucro is prevented by critical sucros like SUR-SAVE.
  38 / The control is given back to the user by {break on}.
  39 /
  40 *{del stack}{load stack}{break on}
  41 /
  42 / If the user has given command
  43 /    /RLBSIMUL ?
  44 / the sucro displays instructions of its usage.
  45 / If the sixth parameter (random number seed) is greater than zero
  46 / (which also implies that the first five parameters were given),
  47 / the sucro continues execution in the label "Start".
  48 / Otherwise, not enough parameters were given, so the instructions
  49 / are displayed.
  50 /
  51 - if W1 ’=’ ? then goto Usage
  52 - if W6 > 0 then goto Start
  53 /
  54 / Short instructions of the sucro are displayed in the edit field.
```

```
  1  1 SURVO 98  Wed Mar 01 21:33:33 2000                D:\V\   1000  100 0
/ First, the current edit field is cleared by putting the cursor to
/ the first position in the field, erasing everything on that line
/ and issuing a SCRATCH command. {R} means pressing ENTER and {act}
/ activating an operation by pressing ESC.
/
+ Usage: {jump 1,1,1,1}{erase}{erase}SCRATCH{act}{R}
/
/ Instructions are written to the edit field.
/
*/RLBSIMUL A,<rotation>,N,<res_file>,<n>,<rand>[,<exit>]          {R}
*simulates factor analysis by repeating <n> times generation of   {R}
*sample of size N according to a factor model X=AF+U, where A is  {R}
*p*r matrix of orthogonal factors F. The theoretical factor matrix{R}
*A is rotated by the ROTATE operation with ROTATION=<rotation>    {R}
*(either VARIMAX, COS, OBLIMIN or '-' (none)) to A0.              {R}
*   For each sample a ML factor solution with r factors is        {R}
*computed giving an estimated factor matrix A1, which is compared {R}
*to A0 by a symmetric transformation analysis. That gives a trans-{R}
*formation matrix L0, which is in an orthogonal case (VARIMAX     {R}
*rotation) used as a rotation matrix for A1. In oblique cases (COS{R}
*or OBLIMIN rotation) L0 is used as an initial matrix for further {R}
*transformation analysis (done by a hybrid Survo module TRAN1).   {R}
*   The reliabilities of models E2 and E3 (see: RELIAB?), the sums{R}
*of squares of the factor loadings and Cronbach's alpha are the   {R}
*variables in <res_file>.SVO , which is automatically created.    {R}
*   The samples are generated by using the rand() function with   {R}
*the original seed number <rand> which is incremented by 1 before {R}
*each sample. The seed appears as a CASE variable in <res_file> . {R}
/
/ The user is asked to press ENTER after reading the instructions.
/ The execution then continues at label "End1" near the end of the
/ sucro code.
/
*{message}        Press ENTER!@{R}
- on key
-     key ENTER: continue
-    wait 3000
*{goto End1}
/
/ **********************************************************************
/
/ For the sake of clarity, the items of the sucro memory (W1, W2, etc.)
/ are renamed. First the parameters of the sucro:
/
/ def WA=W1                Factor matrix (pxr)
/ def Wrotation=W2         Rotation method (VARIMAX, COS, OBLIMIN, -)
/ def WNobs=W3             # of observations in sample
/ def Wfile=W4             Data file for the results
/ def WN=W5                Max # of replicates
/ def Wrand=W6             Seed number for first rand()
/ def Wexit=W7             1: exit Survo after experiments done
/
/ Then the other definitions (variables of the sucro):
/
/ def Wr=W8                # of factors (r)
/ def Worthogonal=W9       1: orthogonal rotation, 0: oblique
/ def Wn=W10               # of current replicate
/ def Wout=W11             Original output file
/ def Wtmp=W12             Temporary usage
/ def Wmnsimul=W13         MNSIMUL indicator
/                            (0=first call, 1=subsequent calls)
/ def Wseedfile=W14        Filename for the seeds (INSEED-OUTSEED)
/ def Wrelfile=W15         Output file for RELIAB results
/ def Wseed1=W16           Seed #1 from seedfile
/ def Wseed2=W17           Seed #2 from seedfile
/ def Wi=W18               Loop counter variable
/ def Wtransmtx=W19        MTX file for symmetric transformation analysis
/ def Wrotmtx=W20          MTX file for factor rotation
/ def Wscoremtx=W21        MTX file for factor scores
/ def WHeywood=W22         Heywood error indicator (FA)
/
/ **********************************************************************
```

```
 1  1 SURVO 98  Wed Mar 01 21:33:33 2000                D:\V\   1000  100 0
/
/ An error handler takes care of severe or fatal situations which
/ might occur during sucro run. An example would be a sucro trying
/ to open a data file which for some reason does not exist.
/ Usually there is no need to replace the default error handler of
/ the system, which just halts the sucro and displays the appropriate
/ cause of the error. Here, however, a special error handler is defined
/ for some rare cases where the correlation matrix is so singular
/ that the maximum likelihood factor solution is not found. To ensure
/ that the execution is not halted, the sucro enables its own error
/ handler during the factor analysis phase.
/ If an error occurs, the execution continues from here. First the
/ default error handler is restored. When the error has occurred, the
/ factor analysis module has replaced the sucro memory with the
/ following information:
/    ERR@<error_no.>@<name_of_operation>@<error_message>@
/ The first item, "ERR" was already checked earlier. Now the second
/ item is investigated. It is printed to the edit field before the
/ sucro memory is restored and the loop counter decremented (as the
/ current replicate failed). In any case, the execution is continued
/ at label "main". However, if the <error_no.> was 27, indicator
/ WHeywood is set. The value 27 stands for <error_message>
/    "Not a positive definite input matrix!"
/
+ Error_handler: {error handler SURVOERR}{line start}{erase}{erase}
*{print W2}{load stack RLBERR.STK}{Wn=Wn-1}{line start}{save word Wtmp}
- if Wtmp <> 27 then goto main
*{WHeywood=1}{goto main}
/
/ ************************************************************************
/ Normal execution begins here.
/ ************************************************************************
/ The edit field is initialized and the reference point #1 is set.
/
+ Start: {line start}{erase}{erase}INIT 400,100,30{act}{R}{ref set 1}
/
/ Save the current file name of the output (results) file. Reset it
/ so that no output is saved to an external text file by any of the
/ statistical operations. This saves time and disk space.
/
*OUTPUT{act}{r}{save word Wout}{R}
*OUTPUT -{act}{home}{erase}
/
/ Check the <rotation> parameter and set the Worthogonal indicator
/ according to it. If an invalid rotation method is given, display
/ an error message and exit the sucro through the label "End2".
/
*{Worthogonal=1}
- switch Wrotation
-   case       -: goto Rotate_A
-   case VARIMAX: goto Rotate_A
-   case OBLIMIN: goto Oblique
-   case     COS: goto Oblique
-        default: continue
*{message} <rotation> one of: -, VARIMAX, COS, OBLIMIN. Press ENTER!@
- on key
-    key ENTER: goto End2
-   wait 300
*{goto End2}
+ Oblique: {Worthogonal=0}
+ Rotate_A: {ref jump 1}SCRATCH {act}{home}
/
/ Find the number of columns in the theoretical factor matrix (WA),
/ and save it to Wr.
/
*MAT DIM {print WA}{act}
*{line end} col{print WA}={act}{ins} {ins}{save word Wr}{home}{erase}
/
/ Rotate the theoretical factor matrix according to <rotation>
/ (unless there is only one factor or no rotation wanted).
/
- if Wrotation '=' - then goto rot0
```

```
  1  1 SURVO 98  Wed Mar 01 21:33:33 2000             D:\V\   1000  100 0
- if Wr > 1 then goto rot
/
/ No rotation, just copy the factor matrix to %A0.
/
+ rot0: MAT %A0={print WA}{act}{R}{goto R&MSN}
/
/ Rotation by ROTATE operation, which gives the rotated matrix as
/ AFACT.M. Copy it to %A0.
/
+ rot: ROTATE {print WA},{print Wr} / ROTATION={print Wrotation}{act}{R}
*MAT %A0=AFACT.M{act}{R}
/
/ Create theoretical correlation matrix %R and MSN-matrix %MSN
/ from the theoretical factor matrix. Matrix names beginning with
/ '&' are for temporary usage and they will be removed later.
/
+ R&MSN: {ref jump 1}SCRATCH {act}{home}
*MAT DIM {print WA}{R}
*MAT &R=MMT({print WA}){R}
*MAT &D=VD(&R){R}
*MAT &D=DV(&D){R}
*MAT &I=IDN(row{print WA},row{print WA}){R}
*MAT &PSI!=&I-&D{R}
*MAT %R=&R+&PSI{R}
*MAT &C=CON(row{print WA},1){R}
*MAT %MSN=ZER(row{print WA},2){R}
*MAT %MSN(1,2)=&C{R}
*MAT RLABELS FROM {print WA} TO %MSN{R}
/
/ Use continuous activation of consecutive lines; it is especially
/ fast with matrix operations.
/
*{ref jump 1}{pre}{act}
/
/ Create the data file for the results. If it already exists, do not
/ overwrite. CHECK gives "OK" if file exists, "NOT FOUND!" if not.
/ Double check is needed, because the user might give the whole path
/ of the file. The default path is the current data path which is
/ retrieved by using a sucro memory operation {save datapath}.
/
*{ref jump 1}SCRATCH {act}{home}{save datapath Wtmp}
*CHECK {print Wtmp}{print Wfile}.SVO{act}
*{next word}{save word Wtmp}{home}{erase}
- if Wtmp '=' OK then goto PrepareLoop
*CHECK {print Wfile}.SVO{act}
*{next word}{save word Wtmp}{home}{erase}
- if Wtmp '=' OK then goto PrepareLoop
/
/ The data file did not exist, so the sucro creates it according to
/ the dimensions of the theoretical factor matrix using the FILE
/ CREATE operation. The date and the sucro call are written as
/ comments to the header of the file.
/
*FILE CREATE {print Wfile}{R}
*Reliabilities by simulation{R}
*DATE{act}{R}
*/RLBSIMUL {print WA},{print Wrotation},{print WNobs},{print Wfile},
*{print WN},{print Wrand}{R}
*{R}
/
/ Write the types, lengths, names and formats of the fields for
/
/   id number of the observation (Wrand)           CASE
/   seeds of the random number generator           SEED1, SEED2
/   reliabilities of r factor images (model E2)    F1\E2, F2\E2, ...
/   reliability of an unweighted sum (model E2)    Sum\E2
/   reliabilities of r factor images (model E3)    F1\E3, F2\E3, ...
/   reliability of an unweighted sum (model E3)    Sum\E3
/   Cronbach's alphas of r factors                 a\F1, a\F2, ...
/   Cronbach's alpha (the usual one)               a\Sum
/   efficiencies of r factors (sums of squares)    L\F1, L\F2, ...
/   sum of the sums of squares of the loadings     L\Sum
```

```
  1  1 SURVO 98  Wed Mar 01 21:33:33 2000              D:\V\    1000  100 0
/   reliabilities of r factor scores (model E2)    %1\E2, %2\E2, ...
/   reliabilities of r factor scores (model E3)    %1\E3, %2\E3, ...
/   Cronbach's alphas of r factor scores           a\%1, a\%2, ...
/
*FIELDS:{R}
/
/ This COUNT command will be activated after the fields are written
/ to give numbers 1,2,3,... for them. "11" is a mask for the COUNT.
/ Reference point #2 is set for returning to this line later.
/
*COUNT CUR+2,END-1,CUR+1{ref set 2}{R}
* 11{R}
/
/ Constant fields for identifying the experiment afterwards:
/
*    S  8 CASE     (########){R}
*    N  8 SEED1    (############){R}
*    N  8 SEED2    (############){R}
/
/      Reliabilities of factor images (model E2):
*{Wi=0}
+ E2: {Wi=Wi+1}
- if Wi > Wr then goto E2e
*    N  8 F{print Wi}\E2    (##.#####){R}{goto E2}
+ E2e:     N  8 Sum\E2   (##.#####){R}
/
/      Reliabilities of factor images (model E3):
*{Wi=0}
+ E3: {Wi=Wi+1}
- if Wi > Wr then goto E3e
*    N  8 F{print Wi}\E3    (##.#####){R}{goto E3}
+ E3e:     N  8 Sum\E3   (##.#####){R}
/
/      Cronbach's alphas:
*{Wi=0}
+ a: {Wi=Wi+1}
- if Wi > Wr then goto ae
*    N  8 a\F{print Wi}     (##.#####){R}{goto a}
+ ae:     N  8 a\Sum    (##.#####){R}
/
/      Efficiencies of factors (and their sum):
*{Wi=0}
+ L: {Wi=Wi+1}
- if Wi > Wr then goto Le
*    N  8 L\F{print Wi}     (##.#####){R}{goto L}
+ Le:     N  8 L\Sum    (##.#####){R}
/
/      Reliabilities of factor scores (model E2):
*{Wi=0}
+ sE2: {Wi=Wi+1}
- if Wi > Wr then goto sE2e
*    N  8 %{print Wi}\E2    (##.#####){R}{goto sE2}
+ sE2e:
/      Reliabilities of factor scores (model E3):
*{Wi=0}
+ sE3: {Wi=Wi+1}
- if Wi > Wr then goto sE3e
*    N  8 %{print Wi}\E3    (##.#####){R}{goto sE3}
+ sE3e:
/      Cronbach's alphas of factor scores:
*{Wi=0}
+ sa: {Wi=Wi+1}
- if Wi > Wr then goto sae
*    N  8 a\%{print Wi}     (##.#####){R}{goto sa}
+ sae: END
/
/ Now activate the COUNT command and remove the two lines which were
/ needed for it.
/
*{ref jump 2}{act}{del line}{del line}
/
/ The description is complete, so activate FILE CREATE and clear the
```

```
  1  1 SURVO 98  Wed Mar 01 21:33:33 2000                D:\V\   1000  100 0
/ field again. The data file for the results is now created.
/
*{ref jump 1}{act}{line start}{erase}SCRATCH{act}{home}
/
/ Some preparations must be done before entering the main loop:
/   The loop counter (Wn) is initialized to 0.
/   The MNSIMUL indicator (Wmnsimul) is set to 0 (for the first round).
/   The Heywood indicator (WHeywood) is set to 0 (default state).
/
+ PrepareLoop: {Wn=0}{Wmnsimul=0}{WHeywood=0}
/
/   Two files are declared for later use:
/     1. Wseedfile will be used for the seeds of the random number
/        generator (INSEED-OUTSEED technique).
/     2. Wrelfile will be used for the output of RELIAB module.
/   Both files will be created to the directory of temporary files
/   (given by system parameter 'tempdisk').
/
*{save tempdisk Wseedfile}{Wseedfile=Wseedfile&RELSEEDS.DAT}
*{save tempdisk Wrelfile}{Wrelfile=Wrelfile&RELOUT}
/
/ To maximize speed, the essential matrix operations are not done
/ inline, but instead using matrix chain files by MATRUN command.
/ These series of matrix operations will be carried out on every
/ round of the main loop.
/ The chain files are also created to the tempdisk and used from
/ there. "MTX" is the traditional file type of the chain files.
/ The files are plain text files.
/
*{save tempdisk Wtransmtx}{Wtransmtx=Wtransmtx&SYMTRANS.MTX}
*{save tempdisk Wrotmtx}{Wrotmtx=Wrotmtx&ROTATION.MTX}
*{save tempdisk Wscoremtx}{Wscoremtx=Wscoremtx&FSCORES.MTX}
/
/ Create a matrix chain file for symmetric transformation analysis:
/
*SAVEP {print Wtransmtx}{R}
*MAT &C12=MTM2(FACT.M,%A0){R}
*MAT SINGULAR_VALUE DECOMPOSITION OF &C12 TO &U,&D,&V{R}
*MAT &L=MMT2(&U,&V){R}
*{ref jump 1}{act}SCRATCH {act}{home}
/
/ Create a matrix chain file for factor rotation:
/
*SAVEP {print Wrotmtx}{R}
*MAT &FACT=FACT.M*&L{R}
*MAT &RFACT=MTM(&L){R}
*MAT &RFACT=inv(&RFACT){R}
*{ref jump 1}{act}SCRATCH {act}{home}
/
/ Create a matrix chain file for computing the factor scores by
/ regression method:
/
*SAVEP {print Wscoremtx}{R}
*MAT DIM &FACT{R}
*MAT &A=&FACT'{R}
*MAT &A=NRM(&A){R}
*MAT &P=NORM'{R}
*MAT TRANSFORM &P BY X#*X#{R}
*MAT &A=CON(row&FACT,1){R}
*MAT &P=&P-&A{R}
*MAT &P=(-1)*&P{R}
*MAT TRANSFORM &P BY 1/X#{R}
*MAT &P!=DV(&P){R}
*MAT &A=&FACT'{R}
*MAT &A=&A*&P{R}
*MAT &B=&A*&FACT{R}
*MAT &B=&RFACT+&B{R}
*MAT &B=INV(&B){R}
*MAT &B=&B*&A{R}
*MAT &B=&B'{R}
*MAT &A=MSN.M(*,2){R}
*MAT TRANSFORM &A BY 1/X#{R}
```

```
  1  1 SURVO 98  Wed Mar 01 21:33:33 2000              D:\V\    1000  100 0
*MAT &A=DV(&A){R}
*MAT &B=&A*&B{R}
*MAT &C=MSN.M(*,1){R}
*MAT &C=&C'{R}
*MAT &C=&C*&B{R}
*MAT &C=(-1)*&C{R}
*MAT FCOEFF.M=ZER(row&FACT+1,col&FACT){R}
*MAT FCOEFF.M(1,1)=&C{R}
*MAT FCOEFF.M(2,1)=&B{R}
*MAT FCOEFF.M(1,0)="Constant"{R}
*MAT CLABELS "%" TO FCOEFF.M{R}
*MAT NAME FCOEFF.M AS FCOEFF{R}
*{ref jump 1}{act}SCRATCH {act}{home}
/
/ *********************************************************************
/ All preparations are finally completed. The main loop begins.
/ *********************************************************************
/
/ Increment the original seed number (Wrand), which is only used as an
/ id number after the INSEED-OUTSEED technique was added.
/ Also increment the counter variable (Wn). Check if enough replicates
/ were done already. In that case, exit through the "Return" label.
/
+ main: {Wrand=Wrand+1}{Wn=Wn+1}
- if Wn > WN then goto Return
/
/ Clear the field and write a simple message to the first line telling
/ the number of the current replicate.
/
*{ref jump 1}SCRATCH {act}{home}{u}{erase}{erase}
*{form}Experiment {write Wfile}: computing {write Wn}/{write WN}...{R}
/
/ OUTSEED is used every time to get the two seeds of the random number
/ generator to a text file. RND specification, however, is used only
/ once (on the first round), when Wmnsimul is 0. After the first round,
/ the INSEED specification will be used, instead. This ensures that we
/ get non-overlapping series of random numbers during the experiment.
/
*OUTSEED={print Wseedfile}
- if Wmnsimul = 1 then goto not1st
* RND=rand({print Wrand}){Wseed1=Wrand}{Wseed2=MISSING}
+ not1st: {R}
/
/ A sample of WNobs observations from multivariate normal distribution
/ is generated according to the given means, stddevs and correlations
/ in %MSN and %R. The sample will be saved in data file &TEST.SVO.
/
*MNSIMUL %R,%MSN,&TEST,{print WNobs},{print Wmnsimul} / TYPES=8{R}
/
/ On the first time, the MNSIMUL indicator (Wmnsimul) is 0, which means
/ that the principal components are computed and the necessary matrices
/ of coefficients are saved. The seed given by the user is used (by
/ the RND specification above), so the INSEED is skipped.
/
- if Wmnsimul = 0 then goto Activate_Simulation
/
/ Beginning from the second round, the seeds are loaded from the seed
/ file and saved to the sucro memory. INSEED specification is written.
/
*LOADP {print Wseedfile}{act}{R}
*{save word Wseed1}{next word}{save word Wseed2}{home}{erase}{u}{erase}
*INSEED={print Wseedfile}
/
/ Activate the MNSIMUL command written above.
/ After the generation of the sample, Wmnsimul is set to 1 to indicate
/ that MNSIMUL may use the saved coefficients and that INSEED-OUTSEED
/ technique is to be used with the random number generator.
/
+ Activate_Simulation: {u}{line start}{act}{Wmnsimul=1}{R}{R}
/
/ Compute correlations of the sample. Output: CORR.M (correlations)
/ and MSN.M (means and standard deviations). No output to the edit
```

```
  1  1 SURVO 98  Wed Mar 01 21:33:33 2000                D:\V\   1000  100 0
/ field and no temporary output to the screen.
/
*CORR &TEST / PRIND=0 FAST=1{act}{R}
/
/ Set the error handler and save the current sucro memory to a file.
/ (See "Error_handler" label above.)
/
*{error handler RLBSIMUL}{save stack RLBERR.STK}
/
/ Compute the factor matrix according to the maximum likelihood (ML)
/ principle. The number of factors to be computed is r (Wr). Output:
/ factor loadings (FACT.M).
/
*FACTA CORR.M,{print Wr}
/
/ Usually this is just activated, but if the correlation matrix was
/ previously not positive definite, the error handler (see above)
/ has set WHeywood to 1. In that case, a FEPS specification is used
/ to give more tolerance for the ML algorithm. WHeywood is reset to 0.
/
- if WHeywood = 0 then goto Activate_FA
* / FEPS=0.005 {WHeywood=0}
/
/ Straight after activation, the default error handler is restored.
/ If an error occurs before that, the error handler defined above
/ is responsible for restoring the default handler.
/
+ Activate_FA: {act}{error handler SURVOERR}{R}
/
/ Find the optimal transformation matrix for rotating the factor matrix
/ to the same position as the theoretical factor matrix. In orthogonal
/ cases, symmetric transformation analysis is enough.
/
*MATRUN {print Wtransmtx}{act}{R}
- if Worthogonal = 1 then goto Rotation
/
/ In oblique cases the matrix given by the symmetric transformation
/ analysis (&L.MAT) is not enough, but it is a good initial solution,
/ which can be improved iteratively. A hybrid module TRAN1 (SM 8/94) is
/ used for the iteration. Its output, L1.M is then copied to &L.MAT.
/
*TRAN1 FACT.M,%A0,&L / PENALTY=10000 {act}{R}
*MAT &L=L1.M{act}{R}
/
/ Now rotate the factor matrix (FACT.M) using the optimal transformation
/ matrix (&L.MAT). Output: rotated factor matrix (&FACT.MAT) and the
/ factor correlations (&RFACT.MAT). In orthogonal cases the latter is
/ an identity matrix (rxr).
/
+ Rotation: MATRUN {print Wrotmtx}{act}{R}
/
/ Compute the factor scores by regression method. Output: factor score
/ coefficients (FCOEFF.M).
/
*MATRUN {print Wscoremtx}{act}{R}
/
/ Delete the output file of RELIAB, because it always appends to it.
/ (It is an ordinary text file with no file type.)
/
*FILE DEL {print Wrelfile}.{act}{R}
/
/ Compute the reliabilities of the factor images and the unweighted sums
/ (both for models E2 and E3), Cronbach's alphas and the efficiencies
/ of the factors. They will appear in the output file (Wrelfile) as
/ consecutive lines.
/
*RELIAB CORR.M,&FACT,&RFACT / MSN=MSN.M OUTFILE={print Wrelfile}{act}{R}
/
/ Compute the reliabilities and Cronbach's alphas of the factor scores.
/ They will be appended to the output file (Wrelfile).
/
*RELIAB CORR.M,&FACT,&RFACT / WEIGHT=FCOEFF.M{act}{R}
```

```
  1  1 SURVO 98  Wed Mar 01 21:33:33 2000            D:\V\   1000  100 0
/
/ Delete the Survo data file which is used for the output manipulation.
/ (The type of the file is ".SVO" by default.)
/
*FILE DEL {print Wrelfile}{act}{R}
/
/ Save the contents of the output file to the Survo data file.
/
*FILE SAVE {print Wrelfile} TO {print Wrelfile}{act}{R}
/
/ Save the data file to a matrix file for additional manipulation.
/ (The type of the file is ".MAT" by default.)
/
*MAT SAVE DATA {print Wrelfile} TO {print Wrelfile}{act}{R}
/
/ Take the transpose to form an observation (record) of the values.
/ Save the id number of the observation (Wrand) to the matrix.
/
*MAT {print Wrelfile}={print Wrelfile}'{act}{R}
*MAT {print Wrelfile}(1,0)="{print Wrand}"{act}{R}
/
/ Delete the temporary data file again.
/
*FILE DEL {print Wrelfile}{act}{R}
/
/ Save the matrix file to a data file.
/
*FILE SAVE MAT {print Wrelfile} TO {print Wrelfile}{act}{R}
/
/ Copy the current seeds to new variables in the data file.
/
*SEED1={print Wseed1} SEED2={print Wseed2}{R}
*VAR SEED1,SEED2 TO {print Wrelfile}{act}{R}
/
/ Copy the record to the data file of the final results.
/ Repeat from the label "main".
/
*FILE COPY {print Wrelfile} TO {print Wfile}{act}{R}
*{goto main}
/
/ **********************************************************************
/
/ Return to the previous job after a normal execution. Remove any
/ messages from the message line. Delete all work matrices. Restore
/ the name of the results file which was saved earlier (Wout).
/
+ Return: {message}@{ref jump 1}{erase}
*MAT KILL &*{act}{line start}{erase}
*OUTPUT {print Wout}{act}{line start}{erase}
/
/ Update the description of the data file by adding the date and time
/ to the empty line reserved for this when the file was created.
/ Continue at "End2".
/
*FILE STATUS {print Wfile} / VARS= {act}{R}{R}{R}{R}
*DATE {act}{u4}UPDATE{act}{line start}{erase}{goto End2}
/
/ **********************************************************************
/
/ The end of the sucro: redefine W1 with no special name and exit
/ using different routes.
/
/ def W1=W1
/
/ "End1" is used after displaying the help text. First the previous
/ job (the edit field and the position of the cursor) is restored.
/ Then the syntax of the sucro is given as a new line after which
/ the sucro is is terminated at label "E".
/
+ End1: {W1=RLBSIMUL}{call SUR-RESTORE}{line start}{ins line}
*/RLBSIMUL A,<rotation>,N,<res_file>,<n>,<rand>
*{goto E}
```

```
   1  1 SURVO 98  Wed Mar 01 21:33:33 2000                D:\V\   1000  100 0
/
/ "End2" is used in all other situations. If <exit> parameter was used,
/ the sucro exits Survo by using the EXIT (F8) key (sucro code {exit})
/ and answering "Y" to the "Exit from Survo (Y/N)?" question. Otherwise
/ the previous job is restored and the sucro is terminated at label "E".
/
+ End2:
- if Wexit = 1 then goto ExitSurvo else goto NoExit
+ ExitSurvo: {exit}Y
+ NoExit: {W1=RLBSIMUL}{call SUR-RESTORE}
/
/ Remove any messages from the message line. Set the normal speed.
/ Terminate the sucro by the {end} code.
/
+ E: {message}@{tempo +1}{end}
*
```

A.3 Implementation of the simulation experiments

The simulation experiments were implemented as sucros (Survo macros), using the sucro /RLBSIMUL (see section A.2) as a subroutine.

Following the logical structure of the five experimental settings (see chapter 6.3), we show how one of the settings was implemented. The rest of the settings are analogous with it. Let us consider the experimental setting V, with three factors and 35 items. As each setting includes eight different models, where the factor loadings vary, we will finally have eight independent tasks. Those tasks are given by eight batch files and collected as one batch file, S’51-58.BAT. This file (see Scheme A.1) will be executed from the operating system level.

```
   1  1 SURVO 98  Thu Mar 02 10:09:27 2000                D:\V\   1000  100 0
*
*SAVEP CUR+1,CUR+8,S’51-58.BAT
*CALL D:\V\S’51
*CALL D:\V\S’52
*CALL D:\V\S’53
*CALL D:\V\S’54
*CALL D:\V\S’55
*CALL D:\V\S’56
*CALL D:\V\S’57
*CALL D:\V\S’58
*
```

Scheme A.1. Saving the batch files for the models of the experimental setting V.

Each model in turn is simulated with 14 sample sizes, using different pseudo random numbers. This is accomplished by running separate Survo sessions with different start sucros. (Below, the main program of Survo, S.EXE, is assumed to reside in the directory C:\E.) Here, we consider the last call, namely the batch file S’58, which corresponds to the best model of this setting (see Scheme A.2).

```
  1  1 SURVO 98  Thu Mar 02 10:19:27 2000                D:\V\   1000  100 0
 73 *
 74 *SAVEP CUR+1,CUR+14,S'58.BAT
 75 *C:\E\S D:\V\S'58_040
 76 *C:\E\S D:\V\S'58_060
 77 *C:\E\S D:\V\S'58_080
 78 *C:\E\S D:\V\S'58_100
 79 *C:\E\S D:\V\S'58_120
 80 *C:\E\S D:\V\S'58_140
 81 *C:\E\S D:\V\S'58_160
 82 *C:\E\S D:\V\S'58_180
 83 *C:\E\S D:\V\S'58_200
 84 *C:\E\S D:\V\S'58_220
 85 *C:\E\S D:\V\S'58_240
 86 *C:\E\S D:\V\S'58_260
 87 *C:\E\S D:\V\S'58_280
 88 *C:\E\S D:\V\S'58_300
 89 *
```

Scheme A.2. Saving the batch file for the model 8 in the experimental setting V.

The start sucros select the appropriate parameters for a particular simulation experiment, call the /RLBSIMUL sucro and exit back to the operating system level.

The current Survo consists of the 32-bit SURVO 98, which is accessed through the 16-bit SURVO 84C. To call SURVO 98 from a batch file, this must be taken into account (see lines 52–59 in Scheme A.3).

```
  1  1 SURVO 98  Thu Mar 02 10:29:27 2000                D:\V\   1000  100 0
 48 A
 49 *TUTSAVE D:\V\S'58_???
 50 /  Monte Carlo simulation experiments                K.Vehkalahti 1999
 51 / *******************************************************************
 52 /     Check the type of the Survo running this sucro (1=84C, 3=98).
 53 *{save survotype W1}
 54 /     If we are already in SURVO 98, continue at "98".
 55 - if W1 = 3 then goto 98
 56 /     Otherwise, start SURVO 98 using this same sucro as a start sucro.
 57 *_S D:\V\S'58_???{act}
 58 /     After returning from SURVO 98, exit SURVO 84C:
 59 *{exit}Y{goto E}
 60 / *******************************************************************
 61 /     Select the appropriate working directory in SURVO 98:
 62 + 98: CD D:\V{act}
 63 /     Set the following parameters to the sucro memory:
 64 /     W1: Factor matrix    W2: Rotation    W3: N       W4: Results file
 65 /     W5: # of replicates  W6: Seed number W7: exit(1)
 66 *{W1=B58}{W2=-}{W3=???}{W4=S'58_???}{W5=500}{W6=2058???0}{W7=1}
 67 /     Call RLBSIMUL sucro as a subroutine:
 68 *{call RLBSIMUL}
 69 /     Exit SURVO 98: (to the SURVO 84C level)
 70 *{exit}Y
 71 /     The {end}:
 72 + E: {end}
 73 *
```

Scheme A.3. Defining the start sucros for the model 8 in the experimental setting V.

The bolded question in Scheme A.3 are replaced by the sample sizes, using a key sucro, which is defined on lines 23–29 in Scheme A.4.

```
  1  1 SURVO 98  Thu Mar 02 10:49:27 2000                D:\V\   1000  100 0
*
*TUTSAVE #S
/ Press F2-N-S on the first REPLACE below, to save the 14 sucros.
*{tempo 0}
+ a: {save word W1}
- if W1 '<>' REPLACE then goto e
*{pre}D{ref}{act}{search}TUTSAVE{R}{act}{ref}{R}{goto a}
+ e: {end}
*
*LINES=A+1,END
*REPLACE ??? 040 C
*REPLACE 040 060 C
*REPLACE 060 080 C
*REPLACE 080 100 C
*REPLACE 100 120 C
*REPLACE 120 140 C
*REPLACE 140 160 C
*REPLACE 160 180 C
*REPLACE 180 200 C
*REPLACE 200 220 C
*REPLACE 220 240 C
*REPLACE 240 260 C
*REPLACE 260 280 C
*REPLACE 280 300 C
*REPLACE 300 ??? C
*
A
```

Scheme A.4. Defining a key sucro for saving the start sucros for each sample size.

The commands for defining the structure of the experiment are in the top of the edit field (See Scheme A.5). When the first experiment has been defined, the other ones are obtained systematically, by changing names and numbers. The edit fields are saved, which makes it easy to go back and repeat the simulations, if necessary.

```
  1  1 SURVO 98  Thu Mar 02 10:53:27 2000                D:\V\   1000  100 0
*SAVE S'58
*
*REPLACE      B57       B58  C
*REPLACE      R57       R58  C
*REPLACE     S'57      S'58  C
*REPLACE "W6=2057" "W6=2058" C
*p=35 loading=0.95
*MAT B=ZER(p,3)
*MAT B1=CON(20,1,loading)
*MAT B(1,1)=B1
*MAT B1=CON(10,1,loading)
*MAT B(21,2)=B1
*MAT B1=CON(5,1,loading)
*MAT B(31,3)=B1
*MAT CLABELS F TO B
*MAT RLABELS Item TO B
*MAT R=B*B'+(IDN(p,p)-DIAG(B*B'))
*MAT NAME B AS Factor_matrix
*MAT NAME R AS Correlation_matrix
*MAT R58=R
*MAT B58=B
*
```

Scheme A.5. Definitions of the model structures in the experimental setting V.

A.4 Sucro program for computing the deviation statistics

The following sucro program /BIAS2MSE computes the squared bias and the mean squared error of the reliabilities, which are collected into separate Survo data files for each experimental setting.

```
1 SURVO 98  Thu Mar 02 23:05:55 2000           D:\V\   2000  100 0
*
*TUTSAVE BIAS2MSE
/  Computing the bias and MSE of the reliabilities   K.Vehkalahti 2000
/ *****************************************************************
/ def Wi=W1   Wj=W2   Wa=W3   Wb=W4   Wk=W5   We=W6   WE=W7   WM=W8
/ def We1=W11 We2=W12 We3=W13 We4=W14 We5=W15 We6=W16 We7=W17 We8=W18
/ def WE1=W21 WE2=W22 WE3=W23 WE4=W24 WE5=W25
/ def WM1=W31 WM2=W32 WM3=W33 WM4=W34 WM5=W35
/
/ Model numbers within experimental settings:
/
*{WE1=11 12 13 14 15 16 17 18}
*{WE2=21 22 23 24 25 26 27 28}
*{WE3=31 32 33 34 35 36 37 38}
*{WE4=41 42 43 44 45 46 47 48}
*{WE5=51 52 53 54 55 56 57 58}
/
/ Conditions for selecting the models from the files R'1,...,R'5:
/
*{We1=ORDER,00001,07000}
*{We2=ORDER,07001,14000}
*{We3=ORDER,14001,21000}
*{We4=ORDER,21001,28000}
*{We5=ORDER,28001,35000}
*{We6=ORDER,35001,42000}
*{We7=ORDER,42001,49000}
*{We8=ORDER,49001,56000}
/
/ Masks for selecting the variables from the files R'1,...,R'5:
/
*{WM1=}{WM1=WM1&--AAAAAA------}
*{WM2=}{WM2=WM2&--AAAAAA------}
*{WM3=}{WM3=WM3&--AAAAAA------}
*{WM4=}{WM4=WM4&--AAAAAAAAAA----------}
*{WM5=}{WM5=WM5&--AAAAAAAAAAAAAA-------------}
/
/ Outermost loop: (experimental settings)
/
*{WE=0}{ref set 8}
+ k: {WE=WE+1}
*{ref jump 8}
*{R}SCRATCH {act}{home}
- if WE > 5 then goto e
*{l4}a*b{u3}{r}
- switch WE
-   case 1: goto k1
-   case 2: goto k2
-   case 3: goto k3
-   case 4: goto k4
-   case 5: goto k5
+ k1: {print WE1}{WM=WM1}{goto kk}
+ k2: {print WE2}{WM=WM2}{goto kk}
+ k3: {print WE3}{WM=WM3}{goto kk}
+ k4: {print WE4}{WM=WM4}{goto kk}
+ k5: {print WE5}{WM=WM5}{goto kk}
+ kk: {R}{R}{R}
```

```
   1  1 SURVO 98  Thu Mar 02 23:05:55 2000              D:\V\    2000  100 0
/
/ Sample sizes:
/
*040 060 080 100 120 140 160 180 200 220 240 260 280 300{R}
*{R}{ref set 1}
/
/ Inner loop 1: (models)
/
*{Wi=0}
+ a: {Wi=Wi+1}
- if Wi > 8 then goto k
*{jump a,a,1}{save word Wa}{del3}{jump b,b,1}{copy}b+1{R}
- switch Wi
-   case 1: goto a1
-   case 2: goto a2
-   case 3: goto a3
-   case 4: goto a4
-   case 5: goto a5
-   case 6: goto a6
-   case 7: goto a7
-   case 8: goto a8
+ a1: {We=We1}{goto aa}
+ a2: {We=We2}{goto aa}
+ a3: {We=We3}{goto aa}
+ a4: {We=We4}{goto aa}
+ a5: {We=We5}{goto aa}
+ a6: {We=We6}{goto aa}
+ a7: {We=We7}{goto aa}
+ a8: {We=We8}{goto aa}
+ aa: {ref jump 1}{R}SCRATCH {act}{home}
/
/ Compute the true values of the reliabilities:
/
*CASE true{R}
*FILE DEL RELIAB.OUT{act}{R}
*COPY CUR-2,CUR-2 TO RELIAB.OUT{act}{R}
*MAT &F=INV(R{print Wa})*B{print Wa}{act}{R}
*MAT CLABELS "%" TO &F{act}{R}
*MSN=* OUTFILE=RELIAB.OUT{R}
*RELIAB R{print Wa},B{print Wa} / WEIGHT=0{act}{R}
*RELIAB R{print Wa},B{print Wa} / WEIGHT=&F{act}{R}
*FILE DEL &T{act}{R}
*FILE SAVE RELIAB.OUT TO &T{act}{R}
/
/ Inner loop 2: (sample sizes)
/
*{Wj=0}
+ b: {Wj=Wj+1}
- if Wj > 14 then goto a
*{jump b,b,1}{save word Wb}{del4}{ref jump 1}{R}SCRATCH {act}{home}
*FILE DEL &S{act}{R}
/
/ Combine the true values with the sample means (computed earlier with
/ the CORR operation and saved as matrix files like S'58_300.MAT):
/
*FILE SAVE MAT S'{print Wa}_{print Wb} TO &S{act}{R}
*VAR true:8=MISSING TO &S{act}{R}
*FILE COPY &T TO &S / MATCH=CASE MODE=2{act}{R}
*.....................................................................{R}
/
/ Form a matrix of the selected reliabilities:
/
*MASK={print WM}{R}
*IND={print We}{R}
*CASES=N,{print Wb}{R}
*MAT SAVE DATA R'{print WE} TO X / PRIND=0{act}{R}
*.....................................................................{R}
```

```
  1  1 SURVO 98  Thu Mar 02 23:05:55 2000             D:\V\   2000  100 0
/
/ Compute the MSE and (squared) bias of the reliabilities,
/ relative to the true values:
/
*n=500{R}
*MAT X=VD(X'*X){act}{R}
*MAT X(0,1)="sumsqr"{act}{R}
*MAT X=X/n{act}{R}
*FILE DEL &X{act}{R}
*FILE SAVE MAT X TO &X{act}{R}
*VAR mean=MISSING TO &X{act}{R}
*VAR true=MISSING TO &X{act}{R}
*FILE COPY &S TO &X / MATCH=CASE MODE=2 VARS=mean,true{act}{R}
*....................................................................{R}
*MAT SAVE DATA &X TO X{act}{R}
*{ref set 2}
*MAT Y=X(*,mean){R}
*MAT Z=X(*,true){R}
*MAT X=X(*,sumsqr){R}
*MAT W=Z{R}
*MAT TRANSFORM W BY Y AND X#*Y#{R}
*MAT TRANSFORM X BY W AND X#-2*Y#{R}
*MAT TRANSFORM X BY Z AND X#+Y#^2{R}
*MAT TRANSFORM X BY Z AND X#/Y#{R}
*MAT TRANSFORM Y BY Z AND (X#-Y#)^2/Y#{R}
*MAT X(0,1)="{print Wb}"{R}
*MAT X=X'{R}
*MAT NAME X AS "Relative_MSE"{R}
*MAT Y(0,1)="{print Wb}"{R}
*MAT Y=Y'{R}
*MAT NAME Y AS "Relative_bias^2"{R}
*FILE SAVE MAT X TO M'{print Wa} / TYPE=8{R}
*FILE SAVE MAT Y TO B'{print Wa} / TYPE=8{R}
*{ref jump 2}{pre}{act}{R}
*{goto b}
/
/ Cleanup:
/
+ e: {line start}{erase}{erase}
*MAT KILL &*{act}{home}{erase}
*MAT KILL X,Y,Z,W{act}{home}{erase}
*FILE DEL &T{act}{home}{erase}
*FILE DEL &S{act}{home}{erase}
*FILE DEL &X{act}{home}{erase}
+ End: {end}
*
```

STATISTICAL RESEARCH REPORTS

PUBLISHED

BY THE FINNISH STATISTICAL SOCIETY

1. PENTTI MANNINEN: Puolueiden kannatusosuuksien estimoinnin tarkkuus Demingin vyöhykepoiminnassa. (The Accuracy of Party Support Estimation in Deming Zone Selection), in Finnish, with English summary, Helsinki 1976.
2. TIMO HAKULINEN: On Competing Risks of Death, Helsinki 1977.
3. LARS-ERIK ÖLLER: Time Series Analysis of Finnish Foreign Trade, Helsinki 1978.
4. PEKKA LAIPPALA: The Empirical Bayes Two-Action Rules with Optimal Sample Size and Exponential Conditional Distributions, Helsinki 1980.
5. MARKKU NURMINEN: Some Developments in Quantitative Methods of Epidemiology, Helsinki 1982.
6. PENTTI SAIKKONEN: Comparing Asymptotic Properties of Some Tests Used in the Specification of Time Series Models, Helsinki 1985.
7. LAURI TARKKONEN: On Reliability of Composite Scales, Helsinki 1987.
8. JUNI PALMGREN: Models for Categorical Data with Errors of Observation, Helsinki 1987.
9. ARI VEIJANEN: On Estimation of Parameters of Partially Observed Random Fields and Mixing Processes, Helsinki 1989.
10. RITVA LUUKKONEN: On Linearity Testing and Model Estimation in Non-Linear Time Series Analysis, Helsinki 1990.
11. HELY SALOMAA: Factor Analysis of Dichotomous Data, Helsinki 1990.
12. KENNETH NORDSTRÖM: Contributions to the Comparison of Linear Models and to the Löwner-ordering Antitonicity of Generalized Inverses, Helsinki 1990.
13. SEPPO LAAKSONEN: Handling Household Survey Non-Response Data, Helsinki 1992.
14. MERVI EEROLA: On Predictive Causality in the Statistical Analysis of a Series of Events, Helsinki 1993.
15. MIKAEL LINDEN: Studies in Integrated and Cointegrated Economic Time Series, Helsinki 1995.
16. TADEUSZ DYBA: Precision of Cancer Incidence Predictions Based on Poisson Distributed Observations, Helsinki 2000.
17. KIMMO VEHKALAHTI: Reliability of Measurement Scales, Helsinki 2000.

ISBN 952-91-2818-5
ISSN 0356-3499
TUMMAVUOREN KIRJAPAINO OY
VANTAA 2000