Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Multivariate Analysis 96 (2005) 172 – 189

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

# Measurement errors in multivariate measurement scales

L. Tarkkonen, K. Vehkalahti*

*Department of Mathematics and Statistics, University of Helsinki, P.O. Box 54, FI-00014, Finland*

Received 25 October 2003
Available online 11 November 2004

**Abstract**

Our aim is to construct a general measurement framework for analyzing the effects of measurement errors in multivariate measurement scales. We define a measurement model, which forms the core of the framework. The measurement scales in turn are often produced by methods of multivariate statistical analysis. As a central element of the framework, we introduce a new, general method of estimating the reliability of measurement scales. It is more appropriate than the classical procedures, especially in the context of multivariate analyses. The framework provides methods for various topics related to the quality of measurement, such as assessing the structural validity of the measurement model, estimating the standard errors of measurement, and correcting the predictive validity of a measurement scale for attenuation. A proper estimate of reliability is a requisite in each task. We illustrate the idea of the measurement framework with an example based on real data.

*AMS 2000 subject classification:* 62H25; 62J99; 62H20; 91C05; 97C40

*Keywords:* Measurement error; Measurement scale; Factor analysis; Reliability; Validity

## 1. Introduction

In statistical research, we are often interested in estimating some population parameters based on a random sample. The uncertainty then comes from the sampling. Sometimes, however, our data include all records under study, and no sampling is then needed. It is also

* Corresponding author. Fax: +358 9 191 24 872.
*E-mail address:* Kimmo.Vehkalahti@helsinki.fi (K. Vehkalahti).

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.09.007

possible to work in the individual level, without a need to consider how the results would be generalized to a population level. In some circumstances, it might even be difficult to define the population.

The other source of uncertainty comes from the measurement, which is an essential concept in science. Measurement is needed regardless of the sampling procedure. As the conclusions of empirical studies are based on values measured on research objects, it is 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. In addition, the measurements should be reliable, in the sense that the researchers can rely on the precision of the measuring instrument. The precision is stated by *reliability*, the ratio of the true variance to the total variance of the measurement. The true variance excludes the variance caused by the random *measurement error*. Reliability defines the resolution of the measurement, and tells us how small differences we can talk about.

Traditional statistical models concentrate on the sampling variation, and often treat the measurement errors with neglect, e.g., by including them in the sampling variation, or simply by assuming that the subjects are measured without error. This is rather vague because sampling and measurement are clearly different procedures. In a given study, it might be useful to find out, which one is the main source of uncertainty. If the measurement is inaccurate, increasing the sample size will not improve it. Instead, the measurement errors should be taken into account by using suitable approaches of modeling.

In this paper, we construct a general framework for analyzing the effects of measurement errors in multivariate measurement scales. The central concepts of the framework are *measurement model* and *measurement scale*. The measurement model relates our framework to the factor analysis model [20] and its generalizations [6,10,18], but it is also a multidimensional generalization of the classical, one-dimensional true score model [22, Chapter 3]. Throughout this paper, we assume that the subjects are random but the items are fixed, i.e. we do not consider random sampling of items (see [21]).

In many application areas it is typical to create one-dimensional scales, e.g., preference scales or predictive regression scales. Nevertheless, the scales are constructed from measurements of several multidimensional attributes. Therefore, we term these scales *multivariate measurement scales*. They connect our measurement framework with various methods of multivariate statistical analysis, such as regression analysis, canonical correlations, or discriminant analysis.

It is essential to ensure that the scales are reliable, i.e. the variation caused by the measurement error is minimized. Therefore, as a central element of the framework, we introduce a new, general method of estimating the reliability of multivariate measurement scales. We establish our method on the classical definition of reliability, and show that the well-known Cronbach's $\alpha$ [12] is a restricted special case of our method.

### 1.1. Historical background

The concept of reliability is due to Charles Spearman already at the turn of the 20th century. In 1904, Spearman [27] proposed a formula for correcting the effects of measurement errors in order to find the true relation between two variables. This idea, together with his famous application, measuring the general intelligence [28], marks the introduction of factor analysis. Spearman's original theory of the general factor and the specific factor corresponds to the one-factor case in the modern terminology. Spearman [29] also introduced the term'reliability coefficient', when developing an enhanced form of his correction formula. The formula, derived independently by Brown [11], and thus called the Spearman–Brown formula, became a classical research method in behavioral and social sciences.

Factor analysis was generalized to its modern, multidimensional form in the 1930s by Thurstone [32], but the lack of adequate computing facilities restricted its usage for decades. As Bartholomew [4, pp. 216–217] writes, factor analysis was born before its time, and it had to mark time until the technology caught up.

Meanwhile, a variety of reliability coefficients were developed, most of them following the Spearman–Brown tradition. Especially the work of Kuder and Richardson [19] at the end of the 1930s had an impact on later studies. In 1951, their *formula* 20 was extended and renamed to *coefficient* $\alpha$ by Cronbach [12]. Since then, Cronbach's $\alpha$ has been used and studied extensively (see, e.g., [14,16,25,2,26,7,15,34,3,8]). It has become a universal procedure of estimating the reliability [9, p. 182], although it is only a lower bound [25,7], and may give negative estimates [13]. However, alternative approaches (see, e.g., [17,33]) have not been adopted as common methods.

For historical reasons, the research on reliability and factor analysis has been focused 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 natural sciences. However, this is only partially true, since examples from unreliable measurements could be drawn from all of science.

### 1.2. Basic concepts

Methods for assessing the quality of measurements have been developed especially in the area of *classical test theory* of psychometrics. From the point of view of this study, the concept of reliability is central. In order to formally define reliability, we establish some notation for the basic concepts.

Let $x$ be the observed variable and $\tau$ the latent true score [22, p.56]. Let $\varepsilon$ be the random measurement error. The fundamental equation of the classical true score model is

$$x = \tau + \varepsilon, \tag{1.1}$$

with $E(\varepsilon) = 0$ and $\rho_{\tau\varepsilon} = 0$, where $E$ denotes the expectation and $\rho$ denotes the correlation [22, p. 56]. This allows separating the true score variance from the measurement error variance. Thus the variance of $x$ can be written as

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_\varepsilon^2 \tag{1.2}$$

[22, p. 57]. Based on the concepts above, we can define reliability.

**Definition 1.** (Lord and Novick [22, p. 61]) Reliability of $x$ under the true score model (1.1) is the squared correlation between $x$ and $\tau$, denoted by

$$\rho_{x\tau}^2 = \frac{\sigma_\tau^2}{\sigma_x^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_x^2} = \frac{1}{1 + \frac{\sigma_\varepsilon^2}{\sigma_\tau^2}}.$$

The definition gives three equivalent forms of reliability, expressed with the components of Eq. (1.2). The first form says that reliability is the ratio of the true score variance $\sigma_\tau^2$ to the total variance $\sigma_x^2$. The second one expresses this using the measurement error variance $\sigma_\varepsilon^2$. The last form of the definition, which does not contain $\sigma_x^2$ explicitly, follows by dividing one by the inverse of the first form. The last form is also related to the *signal-to-noise ratio* of the measurement [22, p. 119].

In order to estimate the reliability, some further assumptions are needed, since $x$ is the only observable quantity in Eq. (1.1). The classical approach is based on the principle of *parallel measurements* [22, p. 48]. Two measurements $x_1 = \tau + \varepsilon_1$ and $x_2 = \tau + \varepsilon_2$ are said to be parallel if we assume that $\sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2$ and $\rho_{\varepsilon_1\varepsilon_2} = 0$. From the assumptions it then follows that $\sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma_x^2$. Omitting the subscripts from $x_1$ and $x_2$ we obtain their correlation, denoted by

$$\rho_{xx} = \frac{\sigma_\tau^2}{\sigma_x^2} = \rho_{x\tau}^2$$

[22, p. 61]. Hence, the correlation between the parallel measurements is a way to estimate the reliability, but it requires quite rigorous assumptions on the measurement errors. In addition, the true score is assumed to be strictly one dimensional.

Most reliability estimators in the classical test theory are based on the parallel model and its variants. However, it is important that reliability may be defined without using the concept of parallel measurements [22, p. 61]. Yet notations of type $\rho_{xx}$ are commonly used for reliability estimators.

Using the above definition, the measurement error variance can be given in the form

$$\sigma_\varepsilon^2 = \sigma_x^2(1 - \rho_{xx}).$$

By taking square roots, we obtain

$$\sigma_\varepsilon = \sigma_x\sqrt{1 - \rho_{xx}}, \tag{1.3}$$

which is known as the *standard error of measurement* [22, p. 67]. It indicates the accuracy of discriminating between the observations. The standard error of measurement is useful because it is expressed in the units of the measurement.

The measurement framework we are constructing provides a method for estimating and using the reliability in more general circumstances, without serious restrictions. Nevertheless, our method is established on the same definition of reliability.

## 2. Measurement framework

In order to assess the quality of measurements in multidimensional situations, it is essential to distinguish between two central concepts: (1) the measurement model, which specifies the structure of the measurement, and discriminates it from the use of the items, and (2) the measurement scale, which is a combination of the measured items, and represents a realization of the theoretical notions. In this paper, we focus on linear measurement scales.

### *2.1. Measurement model*

The structure of the measurement is a relationship between the observed variable and the true score. In the general case, we have $p$ observed variables or measured items $x_1, x_2, \ldots, x_p$, and $k$ true scores $\tau_1, \tau_2, \ldots, \tau_k$ $(k < p)$. To analyze their relationship we define a measurement model

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{B\tau} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ is defined as the expectation of $\boldsymbol{x} = (x_1, \ldots, x_p)'$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)'$ is the true score, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$ is the measurement error, and the pattern matrix $\boldsymbol{B} \in \mathbb{R}^{p\times k}$ specifies the relationship between $\boldsymbol{x}$ and $\boldsymbol{\tau}$.

The basic assumptions of model (2.1) are $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $Cov(\boldsymbol{\tau}, \boldsymbol{\varepsilon}) = \boldsymbol{0}$. From (2.1), the definition $E(\boldsymbol{x}) = \boldsymbol{\mu}$, and the assumptions it follows that $E(\boldsymbol{\tau}) = \boldsymbol{0}$. We also assume that $Cov(\boldsymbol{\tau}) = \boldsymbol{\Phi}$ and $Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$. Under these assumptions we can present the covariance matrix $\boldsymbol{\Sigma}$ between the measured items $x_1, x_2, \ldots, x_p$ in a form

$$\boldsymbol{\Sigma} = E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'] = E[(\boldsymbol{B\tau} + \boldsymbol{\varepsilon})(\boldsymbol{B\tau} + \boldsymbol{\varepsilon})'] = \boldsymbol{B\Phi B}' + \boldsymbol{\Psi}, \tag{2.2}$$

where the true score variation is separated from the measurement error variation, as in (1.2). In addition to the above assumptions, we assume that $\boldsymbol{B}$ has full column rank, denoted by $r(\boldsymbol{B}) = k$, and that the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$ are positive definite.

The measurement model (2.1) is related to several different approaches of modeling: the classical true score model (1.1), the approach of the generalizability theory [14], and the factor analysis model [20].

The essential difference between the measurement model (2.1) and the classical true score model (1.1) is the dimensionality of the true score. In (2.1) the true score is truly multidimensional, while in (1.1) it is strictly one dimensional.

In generalizability theory [14], which is based on procedures from analysis of variance, number of specific variance components, associated with different features of the measurement process, are included in the model as sources of errors. However, according to the fundamental equation (1.1), measurement error is purely random noise. Thus any specific sources of variation related to the measurement process should be modeled rather as true scores, but this is not possible in generalizability theory because the true score remains one-dimensional.

In factor analysis model [20, p. 6], the dimensionality is not a problem, although the original model [27] was one dimensional in the modern terminology. However, the concept of specific factor has often been unclear. It is typical to interpret the specific factors as measurement errors [1, p. 570], but it is also possible to interpret them as true scores, assuming that we can identify them. Hence, the measurement model (2.1) corresponds to the factor analysis model, when the common factors are seen as true scores, and the specific factors are either interpreted as measurement errors or true scores, depending on the identification.

When multinormality is assumed, the parameters of the measurement model (2.1) can be estimated from a sample covariance matrix by the maximum likelihood method. However, in the general form of the model, there are too many parameters to be estimated. Since not all of them can be identified at the same time, it is necessary to impose some restrictions. In this sense, the model is analogous to various generalizations of the factor analysis model [6,10,18].

Model (2.1) conforms to the orthogonal factor analysis model, if we assume that the measurement errors do not correlate with each other, i.e. $\boldsymbol{\Psi}$ is diagonal, and the true scores are uncorrelated and standardized, i.e. $\boldsymbol{\Phi} = \boldsymbol{I}_k$, an identity matrix of order $k$. In an exploratory approach, this model, with an appropriate factor rotation, is often sufficient to specify the structure of the measurement. It is also useful to examine the residuals of the model, i.e. the estimated variances and covariances of the measurement errors. Fine-tuning the assumptions, by estimating some of the covariances of the measurement errors, or fixing some elements of the pattern matrix $\boldsymbol{B}$, takes the approach to a more confirmatory direction.

### 2.2. Measurement scale

Multivariate measurement scales are constructed from measurements of several multidimensional attributes. The most common way is to compute linear combinations of the measured items $x_1, x_2, \ldots, x_p$. In general, we have $m$ scales $\boldsymbol{u} = (u_1, \ldots, u_m)'$ as

$$\boldsymbol{u} = \boldsymbol{A}'\boldsymbol{x} = \boldsymbol{A}'\boldsymbol{\mu} + \boldsymbol{A}'\boldsymbol{B}\boldsymbol{\tau} + \boldsymbol{A}'\boldsymbol{\varepsilon}, \tag{2.3}$$

where $\boldsymbol{A} \in \mathbb{R}^{p\times m}$ is the matrix of the weights. We assume that $r(\boldsymbol{A}) = m$ and $\boldsymbol{B}'\boldsymbol{a}_i \neq \boldsymbol{0}, i = 1, \ldots, m$, where $\boldsymbol{a}_i$ is the $i$th column vector of $\boldsymbol{A}$. The case of one scale is denoted by $u = \boldsymbol{a}'\boldsymbol{x}$.

Using (2.3) and the properties of the measurement model (2.1), we can write the expectation

$$E(\boldsymbol{u}) = E(\boldsymbol{A}'\boldsymbol{\mu} + \boldsymbol{A}'\boldsymbol{B}\boldsymbol{\tau} + \boldsymbol{A}'\boldsymbol{\varepsilon}) = \boldsymbol{A}'\boldsymbol{\mu} \tag{2.4}$$

and by (2.2) we obtain the covariance matrix

$$Cov(\boldsymbol{u}) = \boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{A}'\boldsymbol{B}\boldsymbol{\Phi}\boldsymbol{B}'\boldsymbol{A} + \boldsymbol{A}'\boldsymbol{\Psi}\boldsymbol{A}, \tag{2.5}$$

where the total variation of $\boldsymbol{u}$ is split in two parts: (1) the variation generated by the true scores and (2) the variation generated by the measurement errors. This will be needed for the reliability of $\boldsymbol{u}$.

The matrix $\boldsymbol{A}$ can be completely given by operational definitions. The scale could be a known test or an index, where the weights are predetermined, possibly based on a certain theory or previous experience and knowledge. A simple example is the unweighted sum, where all items are equally weighted.

Another possibility is that the correlation between the scale and an external criterion variable should be maximized. In that case, the scale would be the regression estimate, the weights being the regression coefficients. If the criterion is multidimensional, the scale could be, e.g., a vector of canonical variables or discriminant functions.

The matrix $\boldsymbol{A}$ could also be chosen freely, e.g., to maximize the reliability of the measurement scale by proper and optimal use of the items. The idea of maximizing reliability has been discussed by many authors (see, e.g., [6,31,29]).

Other typical measurement scales include, e.g., psychological test scales and factor scores. As a new construction, we introduce *factor images*, which will be needed in assessing the structural validity of the measurement model. For the sake of clarity, we here prefer the term “factor” to “true score”.

Let us denote the factor images by $\boldsymbol{f} = (f_1, \dots, f_m)'$ and the factor scores by $\boldsymbol{s} = (s_1, \dots, s_m)'$. In both scales it is assumed that $m = k$, i.e. the number of scales equals the number of factors in the measurement model.

**Definition 2.** Factor images $\boldsymbol{f} = \boldsymbol{A}'\boldsymbol{x}$, where $\boldsymbol{A} = \boldsymbol{B\Phi}$, are measurement scales corresponding to the factors $\boldsymbol{\tau}$ in the measurement model.

Factor images could be illustrated as distorted mirror images of the factors, where the distortions are caused by the measurement errors in the observed variables. The matrix $\boldsymbol{B\Phi} = Cov(\boldsymbol{x}, \boldsymbol{\tau})$ is usually called the structure matrix. In the case of uncorrelated and standardized factors ($\boldsymbol{\Phi} = \boldsymbol{I}_k$), the structure matrix is equal to the pattern matrix $\boldsymbol{B}$.

**Definition 3.** Factor scores (by regression method [30]) $\boldsymbol{s} = \boldsymbol{A}'\boldsymbol{x}$, where $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}\boldsymbol{B\Phi}$, are measurement scales which give the optimal predictors for the factors $\boldsymbol{\tau}$ in the least squares sense [20, p. 107].

The expectations of factor images and factor scores follow immediately from (2.4): $E(\boldsymbol{f}) = \boldsymbol{\Phi B}'\boldsymbol{\mu}$ and $E(\boldsymbol{s}) = \boldsymbol{\Phi B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Similarly, the covariance matrices are $Cov(\boldsymbol{f}) = \boldsymbol{\Phi B}'\boldsymbol{\Sigma B\Phi}$ and $Cov(\boldsymbol{s}) = \boldsymbol{\Phi B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B\Phi}$. However, for the reliabilities of the scales we need the explicit separation of the variation as in (2.5). Thus, the covariance matrices of factor images and factor scores become

$$Cov(\boldsymbol{f}) = \boldsymbol{\Phi B}'\boldsymbol{B\Phi B}'\boldsymbol{B\Phi} + \boldsymbol{\Phi B}'\boldsymbol{\Psi B\Phi} \tag{2.6}$$

and

$$Cov(\boldsymbol{s}) = \boldsymbol{\Phi B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B\Phi B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B\Phi} + \boldsymbol{\Phi B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi\Sigma}^{-1}\boldsymbol{B\Phi}. \tag{2.7}$$

All measurement scales mentioned above could be called *first-order scales*. They are typically produced by various methods of multivariate statistical analysis. Sometimes it is useful to create linear combinations of the first-order scales for further analysis. These *second-order scales* $\boldsymbol{z} = (z_1, \dots, z_s)'$ would then be

$$\boldsymbol{z} = \boldsymbol{W}'\boldsymbol{u} = \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{x} = \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{\mu} + \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{B}\boldsymbol{\tau} + \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{\varepsilon}, \tag{2.8}$$

where $\boldsymbol{W} \in \mathbb{R}^{m \times s}$ is the matrix of the second-order weights. Equivalently with (2.4), we have $E(\boldsymbol{z}) = \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{\mu}$, and by (2.5) we obtain

$$Cov(\boldsymbol{z}) = \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{B\Phi B}'\boldsymbol{A}\,\boldsymbol{W} + \boldsymbol{W}'\boldsymbol{A}'\boldsymbol{\Psi A}\,\boldsymbol{W}. \tag{2.9}$$

The idea of the second-order scales is to take advantage of the original measurement model and other information on the structure of the measurement, when creating additional scales, e.g., for prediction, or classification of observations.

For example, we could first perform a factor analysis of, say, fifty items and five factors, based on a measurement model with appropriate assumptions. As a result, we would save the factor scores as new variables in the data. Next, we could perform a regression analysis, using the factor scores as explanatory variables. The regression coefficients would then be the weights of the second-order scale. In each scale, the variation generated by the measurement errors is separated from the variation generated by the factors, which facilitates assessing reliability and validity of the scales.

### 2.3. Reliability

As a central element of our measurement framework, we introduce a new, general method of estimating the reliability of multivariate measurement scales. We establish our method on the classical definition of reliability, given in Section 1.2 as Definition 1. The building blocks of the method are the central concepts of the framework, namely measurement model and measurement scale.

The true score variance and the measurement error variance of the classical definition need to be generalized to the variance generated by the true scores, and the variance generated by the measurement errors, respectively. This was already completed in (2.5), where the total variation of the measurement scale $\boldsymbol{u}$ was split in two parts according to the measurement model (2.1). Now, it is sufficient to consider the variances, i.e. the diagonal elements $diag(\boldsymbol{A}'\boldsymbol{\Sigma A}) = diag(\boldsymbol{a}_1'\boldsymbol{\Sigma a}_1, \ldots, \boldsymbol{a}_m'\boldsymbol{\Sigma a}_m)$. Dealing similarly with $\boldsymbol{A}'\boldsymbol{B\Phi B}'\boldsymbol{A}$ and $\boldsymbol{A}'\boldsymbol{\Psi A}$, we have, in accordance with (2.5), a separation of the variances

$$diag(\boldsymbol{A}'\boldsymbol{\Sigma A}) = diag(\boldsymbol{A}'\boldsymbol{B\Phi B}'\boldsymbol{A}) + diag(\boldsymbol{A}'\boldsymbol{\Psi A}). \tag{2.10}$$

The method of estimating the reliability then follows, simply by dividing the proper variance expressions by another. In the classical notation, $\rho_{uu}$ stands for the reliability of $u$. Analogously, we denote the reliability of $\boldsymbol{u}$ by $\boldsymbol{\rho_u}$, which is a diagonal matrix of order $m$. Proceeding from (2.10), we obtain

$$\boldsymbol{\rho_u} = diag(\boldsymbol{A}'\boldsymbol{B\Phi B}'\boldsymbol{A}) \times [diag(\boldsymbol{A}'\boldsymbol{\Sigma A})]^{-1} \tag{2.11}$$

or in alternative forms (see Definition 1)

$$\boldsymbol{\rho_u} = \boldsymbol{I}_m - diag(\boldsymbol{A}'\boldsymbol{\Psi A}) \times [diag(\boldsymbol{A}'\boldsymbol{\Sigma A})]^{-1} \tag{2.12}$$

or

$$\boldsymbol{\rho_u} = \{\boldsymbol{I}_m + diag(\boldsymbol{A}'\boldsymbol{\Psi A}) \times [diag(\boldsymbol{A}'\boldsymbol{B\Phi B}'\boldsymbol{A})]^{-1}\}^{-1}, \tag{2.13}$$

where $\boldsymbol{I}_m$ is an identity matrix of order $m$. In the case of one scale, $u = \boldsymbol{a}'\boldsymbol{x}$, Eqs. (2.11)–(2.13) reduce to

$$\rho_{uu} = \frac{\boldsymbol{a}'\boldsymbol{B\Phi B}'\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma a}}, \tag{2.14}$$

$$\rho_{uu} = 1 - \frac{\boldsymbol{a}'\boldsymbol{\Psi a}}{\boldsymbol{a}'\boldsymbol{\Sigma a}} \tag{2.15}$$

and

$$\rho_{uu} = \frac{1}{1 + \frac{\boldsymbol{a}'\boldsymbol{\Psi a}}{\boldsymbol{a}'\boldsymbol{B\Phi B}'\boldsymbol{a}}}, \tag{2.16}$$

respectively.

Obviously, $0 \leqslant \rho_{uu} \leqslant 1$, but since our assumptions related to the matrices $\boldsymbol{B}$, $\boldsymbol{A}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Phi}$ imply that the variances of the scales are positive, we can infer that $0 < \rho_{uu} \leqslant 1$. The same applies to the diagonal elements of $\boldsymbol{\rho_u}$. The reliability of 1 is reached only if the measurement errors do not generate any variance to the corresponding scale. It is unlikely, however, since measurements practically always contain measurement errors.

Multiplying Eq. (2.12) from right by $diag(\boldsymbol{A}'\boldsymbol{\Sigma A})$ and rearranging the terms gives

$$diag(\boldsymbol{A}'\boldsymbol{\Psi A}) = (\boldsymbol{I}_m - \boldsymbol{\rho_u}) \times diag(\boldsymbol{A}'\boldsymbol{\Sigma A}), \tag{2.17}$$

which emphasizes the role of $\boldsymbol{\rho_u}$. The scales with a high reliability contain less variation generated by the measurement errors, and vice versa. The square root of (2.17) is the standard error of measurement, a generalization of Eq. (1.3). It indicates the accuracy of discriminating between the observations by different measurement scales. Since the standard error of measurement is expressed in the units of the measurement, it often provides a concrete picture of the measurement accuracy.

In Section 2.2, we presented three special measurement scales, namely the factor images $\boldsymbol{f}$, the factor scores $\boldsymbol{s}$, and the second-order scales $\boldsymbol{z}$. The reliabilities of these scales are denoted by $\boldsymbol{\rho_f}$, $\boldsymbol{\rho_s}$, and $\boldsymbol{\rho_z}$, and they follow immediately by substituting the covariance matrices of (2.6), (2.7), and (2.9) in (2.11). Hence, we have

$$\boldsymbol{\rho_f} = diag(\boldsymbol{\Phi B}'\boldsymbol{B\Phi B}'\boldsymbol{B\Phi}) \times [diag(\boldsymbol{\Phi B}'\boldsymbol{\Sigma B\Phi})]^{-1}, \tag{2.18}$$

$$\boldsymbol{\rho_s} = diag(\boldsymbol{\Phi B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B\Phi B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B\Phi}) \times [diag(\boldsymbol{\Phi B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B\Phi})]^{-1} \tag{2.19}$$

and

$$\boldsymbol{\rho_z} = diag(\boldsymbol{W}'\boldsymbol{A}'\boldsymbol{B\Phi B}'\boldsymbol{A\,W}) \times [diag(\boldsymbol{W}'\boldsymbol{A}'\boldsymbol{\Sigma A\,W})]^{-1}, \tag{2.20}$$

which may be written in any of the alternative forms given above.

Finally, we briefly examine the relationship between our measurement framework and the classical approach, especially concerning the way of estimating the reliability. Indeed, special cases of our method include some well-known procedures, in particular Cronbach's $\alpha$ [12], which "has been derived dozens of times from different theoretical starting points" [9, p. 182]. To show how Cronbach's $\alpha$ relates to our method, we derive it once more, applying the concepts of our measurement framework.

We begin from the measurement model (2.1). The first assumption behind Cronbach's $\alpha$ concerns the true score: it is assumed to be one dimensional. Hence, the model becomes

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{b}\tau + \boldsymbol{\varepsilon},$$

where $\boldsymbol{b} = (b_1, \dots, b_p)'$ and $\tau$ is the true score. In addition, the items are assumed to be equally good indicators of the same true score, and thus we can write $\boldsymbol{b} = \mathbf{1} = (1, \dots, 1)'$, which leads to a simpler model

$$\boldsymbol{x} = \boldsymbol{\mu} + \mathbf{1}\tau + \boldsymbol{\varepsilon}.$$

The basic assumptions of the model are identical with (2.1). In addition, the measurement errors are assumed to be uncorrelated with each other. Denoting $\sigma_\tau^2$ by $\varphi$ and following (2.2), we have a matrix of equal covariances

$$\boldsymbol{\Sigma} = \mathbf{1}\varphi\mathbf{1}' + diag(\boldsymbol{\Psi}).$$

The traditional measurement scale is the unweighted sum of the items, denoted here by $u = \mathbf{1}'\boldsymbol{x}$. Its variance is obtained by applying (2.5):

$$\sigma_u^2 = \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} = \mathbf{1}'\mathbf{1}\varphi\mathbf{1}'\mathbf{1} + \mathbf{1}'diag(\boldsymbol{\Psi})\mathbf{1} = p^2\varphi + \mathrm{tr}(\boldsymbol{\Psi}),$$

where tr denotes the trace. Thus, by (2.14), the reliability of $u$ becomes

$$\rho_{uu} = \frac{p^2\varphi}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}} = \frac{p}{p-1}\left(\frac{p(p-1)\varphi}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right) = \frac{p}{p-1}\left(\frac{p^2\varphi - p\varphi}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right),$$

but since $p^2\varphi = \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1} - \mathrm{tr}(\boldsymbol{\Psi})$ and $\mathrm{tr}(\boldsymbol{\Psi}) + p\varphi = \mathrm{tr}(\boldsymbol{\Sigma})$, we end up with

$$\rho_{uu} = \frac{p}{p-1}\left(1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}\right) = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^p \sigma_{x_i}^2}{\sigma_u^2}\right),$$

which is the original form of Cronbach's $\alpha$ [12, p. 299].

Hence, Cronbach's $\alpha$ is a special case of $\boldsymbol{\rho_u}$ (2.11), but only when the assumptions on both the measurement model and the measurement scale are extremely strict: the true score is one dimensional, the items have equal covariances, and the scale is an unweighted sum. It is notable that these assumptions are not visible in the appearances of Cronbach's $\alpha$. It seems that all that is needed is the covariance matrix of the items. However, the assumptions affect the conclusions nevertheless. Biased or even meaningless values are obtained, if the assumptions are not valid. A common problem is to obtain negative values with Cronbach's $\alpha$, although it is essentially a ratio of two variances, i.e. nonnegative parameters.

Most empirical problems are multidimensional. It is difficult to develop items that measure only one true score. However, the strict assumption of the one-dimensional true score is the essence in Cronbach's $\alpha$, or likewise in the classical true score model (1.1) and in the generalizability theory [14].

In our general approach, we have multiple true scores, which may be correlated if desired. Even the measurement errors may correlate with each other, if the other parts of the model are acceptably specified. The measurement model and measurement scale are defined without any serious restrictions. The assumptions related to the matrices $\boldsymbol{B}, \boldsymbol{A}, \boldsymbol{\Sigma}$, and $\boldsymbol{\Phi}$ are merely technical. They ensure that the scales have positive variances, i.e. they are not constants. Our method of estimating the reliability of measurement scales is a general method, as it does not suffer from any strict assumptions.

### 2.4. Validity

The concept of validity includes several aspects that have been analyzed in the literature of psychological measurement (see, e.g., [5, p. 20]). Within our framework, two aspects can be addressed, namely *structural validity* and *predictive validity*. The structural validity is a property of the measurement model, and the predictive validity is a property of the measurement scale. In addition to the statistical approach, the knowledge of the theory and practice of the application is necessary.

Since the measurement model forms the core of the measurement framework, and affects the quality of the measurement scales, the question of structural validity is essential. It should be examined based on the estimation of the measurement model (2.1). A definite hypothesis on the dimension of the true score and the effects of the true scores on the observed variables may then be tested, and the lack of structural validity revealed. The true score or factor images and the residuals of the model are useful in this task. An appropriate factor rotation also contributes to the structural validity of the measurement model.

If an external criterion variable $y$ is available, the predictive validity of a measurement scale $u$ can be assessed by $\rho_{uy}$, the correlation between the scale and the criterion. However, this correlation is attenuated by the measurement errors in the original measurements [22, p. 71]. A correction for attenuation, introduced already by Spearman [27], takes advantage of the reliabilities of $u$ and $y$ to provide an estimate of the correlation between the true value of the scale and the criterion.

Using the measurement model (2.1), we denote the true value of the scale $u = \boldsymbol{a}'\boldsymbol{x}$ by

$$v = u - \boldsymbol{a}'\boldsymbol{\varepsilon} = \boldsymbol{a}'(\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\tau}). \tag{2.21}$$

Then, we can correct the predictive validity $\rho_{uy}$ for attenuation by

$$\rho_{vy} = \frac{\rho_{uy}}{\sqrt{\rho_{uu}\rho_{yy}}}, \tag{2.22}$$

where $\rho_{uu}$ and $\rho_{yy}$ are the reliabilities of $u$ and $y$, respectively. If there is no specific information on the precision of the criterion, we must assume that $\rho_{yy} = 1$.

The attenuation formula (2.22) stresses the importance of the reliability in assessing the predictive validity. Underestimation of $\rho_{uu}$, which is a commonly known weakness of the classical procedures, leads to overestimation of $\rho_{vy}$ [22, p. 138]. Serious underestimation of reliability will simply explode the estimate of $\rho_{vy}$ and make it useless. Therefore, a proper estimate of reliability is a requisite in assessing both the precision and the predictive validity of measurement scales.

## 3. Example of application

In the following, we consider predicting the performance in mathematics by a psychological test. Our data originate from a study conducted for 115 pupils of secondary school in Finland [23]. We note that more than a psychological data analysis, our example serves as a technical demonstration of the measurement framework.

The central question is the dimensionality of the problem. It should be answered primarily based on the theory of the application. Here we follow the researcher's conception and assume that the true score is two dimensional, and consists of verbal ability and deductive ability. The abilities are measured with a psychological test constructed of nine items, which are described in Table 1. The mathematical performance is assessed by two separate criteria: (1) the result in a national mathematical examination and (2) the school grade in mathematics.

### 3.1. Measurement model

The assessment of the quality of the measurement begins by the specification and estimation of the measurement model. Our approach of the study is exploratory. We assume the measurement errors uncorrelated with each other, and the true scores uncorrelated and standardized. This leads to a measurement model, which conforms to the orthogonal factor analysis model. In practice, that model is often sufficient to specify the structure of the measurement, although the framework also allows making more general assumptions.

Hence, the model is written as

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

which is equal to model (2.1) and its assumptions, except that we now assume that $Cov(\boldsymbol{\tau}) = \boldsymbol{I}_2$, and $Cov(\boldsymbol{\varepsilon}) = diag(\boldsymbol{\Psi})$. Under these assumptions the covariance structure becomes

$$\boldsymbol{\Sigma} = \boldsymbol{B}\,\boldsymbol{B}' + diag(\boldsymbol{\Psi}), \tag{3.2}$$

but, since the items may have different scales of measurement, it is preferable to standardize them, i.e. use correlations instead of covariances. Then, $\boldsymbol{\Sigma}$ denotes the correlation matrix given in Table 2. The highest correlation is 0.708 between the verbal items V13 and V5.

The elements of the matrix $\boldsymbol{B}$, which are now the factor loadings, can be estimated by any factor analysis method. Similarly, we could apply confirmatory factor analysis, if we had more support from the psychological theory. According to our assumption of the dimensionality, $k = 2$, and hence we extract a two-factor solution using the maximum likelihood (ML) method, and apply varimax rotation for an easier interpretation. The loadings are presented in Table 3, together with the communalities.

### 3.2. Structural validity

Let us first consider different aspects of the structural validity of the measurement model. An important part of the structure is an appropriate factor rotation. Here, we accept the varimax rotation without discussing it in detail, and in the sequel, we simply denote the rotated factor matrix by $\boldsymbol{B}$.

Table 1
Descriptions of the observed variables

| | |
|---|---|
| Psychological test items: | |
| V13 | Verbal fluency, completion of a sentence |
| R4/0 | Deduction, continuing a list of numbers |
| D | Deductive reasoning |
| V5 | Verbal problem, finding synonyms |
| S1 | Spatial comprehension, unfolded figures |
| I | Inductive reasoning, series of numbers |
| Vz | Visual problem, inverted figures |
| N3 | Numerical ability, arithmetics |
| P1 | Perceptual recognition, numbers and letters |
| Criterion variables: | |
| Exam | National mathematical examination |
| Grade | School grade in mathematics |

Table 2
Correlations of the psychological test items

| Item | V13 | R4/0 | D | V5 | S1 | I | Vz | N3 | P1 |
|---|---|---|---|---|---|---|---|---|---|
| V13 | 1.000 | | | | | | | | |
| R4/0 | 0.349 | 1.000 | | | | | | | |
| D | 0.394 | 0.597 | 1.000 | | | | | | |
| V5 | 0.708 | 0.404 | 0.410 | 1.000 | | | | | |
| S1 | −0.108 | 0.274 | 0.050 | 0.077 | 1.000 | | | | |
| I | 0.406 | 0.671 | 0.576 | 0.476 | 0.205 | 1.000 | | | |
| Vz | −0.016 | 0.275 | 0.261 | −0.049 | 0.390 | 0.296 | 1.000 | | |
| N3 | 0.411 | 0.572 | 0.520 | 0.368 | 0.112 | 0.641 | 0.253 | 1.000 | |
| P1 | 0.542 | 0.508 | 0.426 | 0.371 | 0.128 | 0.508 | 0.018 | 0.427 | 1.000 |

Table 3
Factor loadings and communalities, two factors

| Item | ML solution | | Varimax rotation | | Communality |
|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 1 | Factor 2 | |
| V13 | 0.997 | −0.017 | **0.991** | −0.110 | 0.995 |
| R4/0 | 0.363 | 0.736 | 0.430 | **0.699** | 0.673 |
| D | 0.406 | 0.575 | 0.457 | 0.535 | 0.495 |
| V5 | 0.713 | 0.187 | **0.727** | 0.120 | 0.543 |
| S1 | −0.102 | 0.369 | −0.067 | 0.377 | 0.146 |
| I | 0.420 | 0.730 | 0.486 | **0.688** | 0.709 |
| Vz | −0.010 | 0.404 | 0.027 | 0.404 | 0.164 |
| N3 | 0.422 | 0.592 | 0.475 | 0.550 | 0.529 |
| P1 | 0.550 | 0.366 | 0.581 | 0.313 | 0.436 |

The columns of the matrix $\boldsymbol{B}$ given in Table 3 are the factor images $\boldsymbol{f}$ (see Definition 2). They seem to support our assumption of the dimensionality: the first factor image corresponds to the verbal ability, the highest loadings being on items V13 and V5, while the

Table 4
Factor loadings and communalities, three factors

| Item | ML solution | | | Varimax rotation | | | Communality |
|---|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 3 | Factor 1 | Factor 2 | Factor 3 | |
| V13 | 0.641 | −0.017 | 0.764 | **0.964** | 0.219 | 0.131 | 0.995 |
| R4/0 | 0.479 | 0.658 | 0.070 | 0.226 | **0.745** | −0.246 | 0.667 |
| D | 0.338 | 0.596 | 0.246 | 0.258 | 0.680 | −0.023 | 0.529 |
| V5 | 0.577 | 0.136 | 0.445 | **0.671** | 0.311 | −0.056 | 0.550 |
| S1 | 0.690 | −0.009 | −0.720 | 0.015 | 0.034 | −**0.997** | 0.995 |
| I | 0.468 | 0.682 | 0.155 | 0.268 | **0.778** | −0.175 | 0.709 |
| Vz | 0.294 | 0.268 | −0.263 | −0.030 | 0.284 | −0.382 | 0.228 |
| N3 | 0.398 | 0.581 | 0.217 | 0.285 | 0.674 | −0.085 | 0.543 |
| P1 | 0.499 | 0.316 | 0.297 | 0.474 | 0.449 | −0.105 | 0.438 |

second factor image corresponds to the deductive ability, with the highest loadings on items R4/0 and I.

The reliabilities of the factor images are obtained by applying Eq. (2.18). Since now $\boldsymbol{\Phi} = \boldsymbol{I}_2$, we have

$$\rho_f = diag[(\boldsymbol{B}'\boldsymbol{B})^2] \times [diag(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})]^{-1},$$

which gives a reliability of 0.920 for the verbal ability factor image, and 0.861 for the deductive ability factor image. It is not uncommon that in a multidimensional context, some dimensions are measured with greater precision than others. Here, the verbal ability seems to be slightly better measured.

Table 3 indicates that the items S1 and Vz have very low communalities. Instead of rejecting the items, we could consider the option that we are missing one dimension. In some circumstances, a new common factor may clear the structure, at least if there is something in common with the weak items. We could think that the items S1 and Vz were representing some sort of a spatial concept. We remind that this example is mainly a technical demonstration of the measurement framework. To show how a new factor affects the estimates and the structural validity, we extract a three-factor solution similarly as above, and examine its properties.

The rotated three-factor solution in Table 4 is consistent with the two-factor solution, but it also reveals the weakness of the spatial concept. The only item that has a considerable loading on the third factor is S1, the spatial comprehension. Adding an extra dimension will increase the estimates of the reliabilities, since the measurement errors are estimated based on the residuals of the model, and increasing the number of factors decreases the residuals. Here, the reliabilities of the first two factor images are 0.925 and 0.894. The reliability of the new spatial concept factor image is also 0.925. However, the dimensionality of the measurement model should be primarily decided according to the theory of the application.

We decide to keep all items and continue with two factors, according to the original assumption of the researcher. Two factors are enough to make the point in this technical

Table 5
Estimated error variances and covariances from two-factor solution

| Item | V13 | R4/0 | D | V5 | S1 | I | Vz | N3 | P1 |
|---|---|---|---|---|---|---|---|---|---|
| V13 | 0.005 | | | | | | | | |
| R4/0 | −0.000 | 0.327 | | | | | | | |
| D | −0.000 | 0.027 | 0.505 | | | | | | |
| V5 | 0.000 | 0.008 | 0.014 | 0.457 | | | | | |
| S1 | −0.000 | 0.039 | −0.121 | 0.081 | 0.854 | | | | |
| I | −0.000 | −0.018 | −0.014 | 0.040 | −0.021 | 0.291 | | | |
| Vz | 0.002 | −0.019 | 0.033 | −0.117 | **0.240** | 0.005 | 0.836 | | |
| N3 | 0.000 | −0.017 | 0.008 | −0.044 | −0.063 | 0.032 | 0.018 | 0.471 | |
| P1 | 0.001 | 0.039 | −0.007 | −0.089 | 0.049 | 0.010 | −0.124 | −0.021 | 0.564 |

demonstration, and probably three factors of nine items would be exaggeration. Our conclusion is that the two-dimensional structure can be considered valid, although certain items seem to be poor. In further studies, better items would be needed. Then it could be possible to improve the analysis in three dimensions.

Hence, we revert to the two-factor solution of Table 3. Yet another way to assess the structural validity, especially the dimensionality, is to look at the residuals, i.e. the estimates of the measurement error variances and covariances given in Table 5. According to (3.2), this matrix should be diagonal. Larger deviations from zero in the covariance elements imply that we should reconsider the dimension of the true score, or perhaps the assumption of the correlations of the measurement errors.

The residuals do not seem to indicate any hidden variation, except the largest residual 0.240, which is the estimated covariance between the measurement errors of S1 and Vz, the spatial and visual items. Those items also have the lowest communalities in the two-factor model. If we were creating a scale of spatial and visual skills, we would obviously need more items to strengthen the scale. If the spatial concept was essential from the theoretical point of view, we should perhaps make a new iteration of the analysis with better items.

### *3.3. Measurement scales*

We now change the focus from the measurement model to the measurement scales. It is worthwhile to note that the same items can be used to create scales according to varying criteria by weighting them differently. As stated earlier, there are two separate criteria for predicting the performance in mathematics: the national examination and the school grade. We denote these criteria by $y_1$ and $y_2$, respectively. For both criteria, we have to find the scales that have the highest predictive validity.

The best linear predictors for $y_1$ and $y_2$ are found by linear regression analyses, using the test items as explanatory variables (see Table 6). The ‘best’ items for predicting the performance in the national examination are D and I, the deductive and inductive reasoning. The corresponding *t*-values are 1.937 and 2.578, respectively. The school grade of mathematics is ‘best’ predicted by S1 and N3, the spatial comprehension and numerical ability. Their *t*-values are 2.479 and 2.126, respectively. The multicollinearity may of course affect the interpretation of the regression coefficients.

Table 6
Regression coefficients for criterion variables

| Criterion | V13 | R4/0 | D | V5 | S1 | I | Vz | N3 | P1 |
|---|---|---|---|---|---|---|---|---|---|
| Exam ($y_1$) | 0.067 | 0.197 | **0.205** | −0.075 | 0.107 | **0.318** | −0.025 | 0.017 | −0.046 |
| Grade ($y_2$) | −0.029 | 0.125 | −0.014 | 0.083 | **0.258** | 0.113 | −0.172 | **0.256** | −0.142 |

Let us denote the scales predicting $y_1$ and $y_2$ by $u_1 = \boldsymbol{a}_1'\boldsymbol{x}$ and $u_2 = \boldsymbol{a}_2'\boldsymbol{x}$, where $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are the vectors of the regression coefficients given in Table 6. The estimate of the predictive validity is given by the correlation between the scale and the criterion. The figures are not very high: $\rho_{u_1y_1} = 0.623$ and $\rho_{u_2y_2} = 0.464$. The effect of the measurement errors reduces the correlations. To employ the correction for attenuation, we need the reliabilities of the scales and the criteria. As we do not have any information on the latter, we have to assume that $\rho_{y_1y_1} = \rho_{y_2y_2} = 1$. Using Eq. (2.14), we obtain the reliabilities of the scales:

$$\rho_{u_1u_1} = \frac{\boldsymbol{a}_1'\boldsymbol{B}\,\boldsymbol{B}'\boldsymbol{a}_1}{\boldsymbol{a}_1'\boldsymbol{\Sigma}\boldsymbol{a}_1} = 0.808$$

and

$$\rho_{u_2u_2} = \frac{\boldsymbol{a}_2'\boldsymbol{B}\,\boldsymbol{B}'\boldsymbol{a}_2}{\boldsymbol{a}_2'\boldsymbol{\Sigma}\boldsymbol{a}_2} = 0.447$$

and hence we can correct the estimates of the predictive validity for attenuation, by applying the attenuation formula (2.22). Following (2.21), we denote the true values of the scales by $v_1 = \boldsymbol{a}_1'(\boldsymbol{\mu} + \boldsymbol{B\tau})$ and $v_2 = \boldsymbol{a}_2'(\boldsymbol{\mu} + \boldsymbol{B\tau})$ to obtain

$$\rho_{v_1y_1} = \frac{\rho_{u_1y_1}}{\sqrt{\rho_{u_1u_1}\rho_{y_1y_1}}} = 0.693$$

and

$$\rho_{v_2y_2} = \frac{\rho_{u_2y_2}}{\sqrt{\rho_{u_2u_2}\rho_{y_2y_2}}} = 0.694. \tag{3.3}$$

Although the figures are practically equal, it seems that the national examination might be a simpler and more reliable indicator of the mathematical ability than the school grade.

Finally, we create scales for assessing the verbal and deductive abilities of the pupils. The best predictive scales in the least squares sense are the factor scores (see Definition 3) $\boldsymbol{s} = \boldsymbol{A}'\boldsymbol{x}$, where $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}\boldsymbol{B}$, since now $\boldsymbol{\Phi} = \boldsymbol{I}_2$. The coefficients $\boldsymbol{A}$ are given in Table 7. Since factor scores often represent the factors in further analyses, it is important to carefully examine the structural validity of the measurement model before computing the factor scores.

The reliabilities of the factor scores are obtained by applying Eq. (2.19) with $\boldsymbol{\Phi} = \boldsymbol{I}_2$:

$$\boldsymbol{\rho_s} = diag[(\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B})^2] \times [diag(\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B})]^{-1},$$

which gives a reliability of 0.994 for the verbal ability factor score and 0.850 for the deductive ability factor score. The corresponding standard errors of measurement, by (2.17) and taking

Table 7
Factor score coefficients, two factors

| Factor | V13 | R4/0 | D | V5 | S1 | I | Vz | N3 | P1 |
|---|---|---|---|---|---|---|---|---|---|
| Verbal | **0.928** | 0.037 | 0.020 | 0.013 | 0.005 | 0.042 | 0.007 | 0.022 | 0.014 |
| Deductive | −**0.614** | **0.338** | 0.171 | 0.061 | 0.065 | **0.377** | 0.073 | 0.189 | 0.097 |

square roots, are 0.077 and 0.357, respectively. The verbal ability factor score is practically identical to the best item V13, the verbal fluency, thus making the reliability of the scale exceptionally high. In practice, we should have more indicators of the ability dimensions.

In conclusion, it is essential to specify the structure of the measurement and assess its validity, but it is also important to create the proper scales. If any criteria are available, the weights for the items can be estimated by regression method. Otherwise, it is a good practice to compute the factor scores and weight them according to the theory. A proper way of estimating the reliability of measurement scales is a requisite in each phase of the analysis.

## Acknowledgments

We are grateful to *Dr. Simo Puntanen* for his thoughtful remarks and suggestions. Comments by the referees were also helpful, and improved the clarity.

## References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, 3rd ed., Wiley, New Jersey, 2003.
[2] D.J. Armor, Theta reliability and factor scaling, in: H.L. Costner (Ed.), Sociological Methodology, Jossey-Bass, San Francisco, 1974, pp. 17–50.
[3] K.A. Barchard, A.R. Hakstian, The effects of sampling model on inference with coefficient alpha, Educ. Psychol. Measure. 57 (1997) 893–905.
[4] D.J. Bartholomew, Spearman and the origin and development of factor analysis. Br. J. Math. Statist. Psychol. 48 (1995) 211–220.
[5] D.J. Bartholomew, The Statistical Approach to Social Measurement, Academic Press, London, 1996.
[6] D.J. Bartholomew, M. Knott, Latent Variable Models and Factor Analysis, Arnold, London, 1999.
[7] P.M. Bentler, J.A. Woodward, The greatest lower bound to reliability, in: H. Wainer, S. Messick (Eds.), Principals of Modern Psychological Measurement, Erlbaum, New Jersey, 1983, pp. 237–253.
[8] J.M.F. ten Berge, W.K.B. Hofstee, Coefficient alpha and reliabilities of unrotated and rotated components, Psychometrika 64 (1999) 83–90.
[9] S.F. Blinkhorn, Past imperfect, future conditional: fifty years of test theory, Br. J. Math. Statist. Psychol. 50 (1997) 175–185.
[10] K.A. Bollen, Structural Equations with Latent Variables, Wiley, New York, 1989.
[11] W. Brown, Some experimental results in the correlation of mental abilities, Br. J. Psychol. 3 (1910) 296–322.
[12] L.J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika 16 (1951) 297–334.
[13] L.J. Cronbach, W. Hartmann, A note on negative reliabilities, Educ. Psychol. Measure. 14 (1954) 342–346.
[14] L.J. Cronbach, N. Rajaratnam, G.C. Gleser, Theory of generalizability: a liberalization of reliability theory, Br. J. Statist. Psychol. 16 (1963) 137–163.
[15] L.J. Cronbach, Internal consistency of tests: analyses old and new, Psychometrika 53 (1988) 63–70.

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

[17] D.R. Heise, G.W. Bohrnstedt, Validity, invalidity and reliability. In: E.F. Borgatta, G.W. Bohrnstedt (Eds.), Sociological Methodology, Jossey-Bass, San Francisco, 1970, pp. 104–129.

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

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

[20] D.N. Lawley, A.E. Maxwell, Factor Analysis as a Statistical Method, 2nd ed., Butterworth, London, 1971.

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

[22] F.M. Lord, M.R. Novick, Statistical Theories of Mental Test Scores, Addison-Wesley, London, 1968.

[23] V. Malinen, Grounds for success in mathematics in secondary school, Unpublished Master's thesis (in Finnish), Department of Education, University of Helsinki, 1980.

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

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

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

[27] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101.

[28] C. Spearman, General intelligence objectively determined and measured, Am. J. Psychol. 15 (1904) 201–293.

[29] C. Spearman, Correlation calculated from faulty data, Br. J. Psychol. 3 (1910) 271–295.

[30] G.H. Thomson, The Factorial Analysis of Human Ability, University of London Press, 1939.

[31] G.H. Thomson, Weighting for battery reliability and prediction, Br. J. Psychol. 30 (1940) 357–366.

[32] L.L. Thurstone, Multiple factor analysis, Psychol. Rev. 38 (1931) 406–427.

[33] C.E. Werts, R.D. Rock, R.L. Linn, K.G. Jöreskog, A general method of estimating the reliability of a composite, Educ. Psychol. Meas. 38 (1978) 933–938.

[34] R.R. Wilcox, Robust generalizations of classical test reliability and Cronbach's alpha, Br. J. Math. Statist. Psychol. 45 (1992) 239–254.