A Unified Approach to Multi-item Reliability

Peer-reviewed author version

ALONSO ABAD, Ariel; LAENEN, Annouschka; MOLENBERGHS, Geert; GEYS, Helena & VANGENEUGDEN, Tony (2010) A Unified Approach to Multi-item Reliability. In: BIOMETRICS, 66 (4). p. 1061-1068.

DOI: 10.1111/j.1541-0420.2009.01373.x
Handle: http://hdl.handle.net/1942/11540

A Unified Approach to Multi-item Reliability

Ariel Alonso[1], **Annouschka Laenen**[1], **Geert Molenberghs**[1,2]

Helena Geys[3,1], **Tony Vangeneugden**[4,1]

[1]Center for Statistics, Hasselt University, Agoralaan 1, B-3590 Diepenbeek, Belgium

[2]Biostatistical Centre, Katholieke Universiteit Leuven, Leuven, Belgium

[3]Johnson & Johnson Pharmaceutical Research and Development, Beerse, Belgium

[4]Tibotec, Johnson & Johnson, Mechelen, Belgium

Email: ariel.alonso@uhasselt.be

Abstract

The reliability of multi-item scales has received a lot of attention in the psychometric literature, where a myriad of measures like the Conbrach's α or the Spearman-Brown formula have been proposed. Most of these measures, however, are based on very restrictive models that account only for unidimensional instruments. In this paper we introduce two measures to quantify the reliability of multi-item scales based on a more general model. We show that they capture two different aspects of the reliability problem and satisfy a minimum set of intuitive properties. The relevance and complementary value of the measures is studied and earlier approaches are placed in a broader theoretical framework. Finally, we apply them to investigate the reliability of the *Positive And Negative Syndrome Scale*, a rating scale for the assessment of the severity of schizophrenia.

Keywords: Reliability, Multi-item Rating Scales, Factor Analysis.

1 Introduction

Rating scales play a prominent role in psychology and psychiatry where they are frequently used to make precise diagnostics and to evaluate the efficacy of new treatments

or therapeutic procedures. They are also prevalent in health-related quality of life studies and studies for the clinical evaluation of pain, among others.

The intrinsic uncertainties of measurements obtained with rating scales are related to the concepts of validity and reliability. In general, validity tells whether a measurement scale measures what it is supposed to measure in the context it is applied and, hence, validity should always be a primary concern when evaluating an instrument. Another important aspect is the precision of the measurements, that is, their reliability.

The study of the reliability of multi-item scales has received a large amount of attention in the psychometric literature. The Spearman-Brown formula, the Kuder-Richardson formulas, including the well-known KR-20, its slight and famous variation known as Cronbach's α, the five lower bounds introduced by Guttman, and the measure proposed by Mosier are some of the proposals to quantify reliability in this context (Tarkkonen and Vehkalahti 2005). It has been extensively shown, however, that these measures equal reliability only under rather stringent assumptions (Novick and Lewis 1967, Green and Yang 2009). When these assumptions are not met, the previous measures can not be considered a proper quantification of reliability but merely a lower bound for it (Novick and Lewis 1967, Green and Yang 2009, Sijtsma 2009). Therefore, they are nowadays mainly considered as measures for the *internal consistency* of an instrument, which indicates the homogeneity of the items, or, equivalently said, how much they measure a unidimensional underlying construct. Nevertheless, the appropriateness of these coefficients to evaluate internal consistency has also been questioned (Sijtsma 2009). Even though their suitability is severely restricted by the assumptions they need, some of the previous measures have been routinely used and misused in many practical situations (Sijtsma 2009).

Based on a more general modelling framework, we will introduce two measures to quantify the reliability of multi-item scales. These measures satisfy a pre-defined minimum set of intuitive and appealing properties and, when the necessary modelling assumptions are met, they permit to recover some of the previous proposals as special cases. Along this line we will also illustrate the relevance and complementary value of our approach and

place earlier approaches in a broader theoretical framework.

Section 2 explains the methodological framework and introduces two new measures for multi-item scale reliability. We study these measures under some specific unidimensional models in Section 3 as well as under more general, multidimensional models in Section 4. In this section we further elaborate on the difference between sum-scores and multivariate outcomes for reliability estimation. Sections 5 and 6 attend to the concepts of correlation and prediction and their relations to reliability. Finally, in Section 7 the previously introduced methods are illustrated on a real case study.

2 Methodology

Let us start by introducing the general measurement model on which all the measures will be based. We assume that we have a multi-item scale, formed by p items. Further, we assume that for the i^{th} subject the following measurement model holds

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ denotes the p-dimensional vector of observed scores, $\boldsymbol{\tau}_i = (\tau_{i1}, \tau_{i2}, \ldots, \tau_{iq})'$ denotes a q-dimensional vector of true scores with $q \leq p$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{ip})'$ is a p-dimensional vector of measurement errors, $\boldsymbol{B}$ is a $p \times q$ full column rank matrix that describes the functional relationship between the observed and true scores and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ is a vector describing the mean of the observed scores. Additionally, we assume that: i) $E(\boldsymbol{\varepsilon}_i) = \boldsymbol{0}$ with $\mathrm{Cov}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma}$, ii) $E(\boldsymbol{\tau}_i) = \boldsymbol{0}$ with $\mathrm{Cov}(\boldsymbol{\tau}_i) = \boldsymbol{D}$, and finally that iii) $\boldsymbol{\tau}_i$ and $\boldsymbol{\varepsilon}_i$ are independent.

Based on the previous assumptions, if we define $\boldsymbol{G} = \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'$, the variance-covariance matrix of the measured items $\boldsymbol{V} = \mathrm{Cov}(\boldsymbol{X}_i)$ can be written as

$$\boldsymbol{V} = \boldsymbol{G} + \boldsymbol{\Sigma}. \tag{2}$$

Model (1) comprises many model families. For instance, if one assumes that $\boldsymbol{D} = \boldsymbol{I}$ and $\boldsymbol{\Sigma}$ is a diagonal matrix, then it reduces to the classical orthogonal factor analytic model.

It is also related to the modeling framework used in Generalizability Theory (Cronbach, Gleser, Nanda and Rajaratnam 1972) and contains as a special cases three models that have played a prominent role in the quantification of reliability. These can be defined as: (i) *Parallel tests* obtained when μ and τ_i are scalars, $\boldsymbol{B} = \boldsymbol{\beta} = \mathbf{1} = (1, 1, \ldots, 1)'$, and $\text{Cov}(\boldsymbol{\varepsilon}_i) = \sigma^2 \boldsymbol{I}$; (ii) *Essentially tau-equivalent tests* obtained when $\boldsymbol{B} = \boldsymbol{\beta} = \mathbf{1}$, τ_i is a scalar and $\text{Cov}(\boldsymbol{\varepsilon}_i) = \text{diag}(\sigma_j^2)$, with $j = 1, \ldots, p$; and (iii) *Congeneric tests* obtained when $\boldsymbol{B} = \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$, τ_i is a scalar and $\text{Cov}(\boldsymbol{\varepsilon}_i) = \text{diag}(\sigma_j^2)$. Importantly, the parallel test model forms the basis for the definition of reliability in the Classical Test Theory (CTT) (Lord and Novick 1968).

While these models all assume a unidimensional true score, Model (1) allows a multidimensional vector of random effects and correlated error components. Stemming from identifiability issues, some restrictions may be needed to estimate the parameters, however, in what follows we will work with Model (1) in its most general form.

In order to extend the concept of reliability to the more general scenario implied by Model (1), we will introduce a minimum set of properties one would expect any meaningful measure of reliability should satisfy. Essentially, if R denotes a measure of reliability then: 1) $0 \leq R \leq 1$; 2) if $R = 0$ then $\boldsymbol{X}_i$ does not convey any information about the true scores $\boldsymbol{\tau}_i$, i.e, $\boldsymbol{B} = \mathbf{0}$; 3) if $R = 1$ then there exist linear functions ψ_1 and ψ_2 so that $P\left(\psi_1(\boldsymbol{X}_i) = \psi_2(\boldsymbol{\tau}_i)\right) = 1$, i.e, the true and observed scores are deterministically related; and 4) under parallel tests $R = \dfrac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2}$.

A new measure of reliability that fulfills properties (1)–(4) is given by

$$R_T = 1 - \frac{\text{tr}(\boldsymbol{\Sigma})}{\text{tr}(\boldsymbol{V})}. \tag{3}$$

Note that the previous expression closely resembles the formula of reliability used in CTT, but now one summarizes the variability of the observed scores and the error terms using the trace of the variance-covariance matrices $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$, respectively. The R_T coefficient can be rewritten as

$$R_T = \sum_{j=1}^{p} \nu_j R_j,$$

where $\nu_j = V_{jj}/\Sigma_{k=1}^{p} V_{kk}$ is a weight associated with the j^{th} item and $R_j = G_{jj}/V_{jj}$ denotes its reliability. The previous expression clearly shows that R_T is just a weighted average of the items' reliability, where the weight associated with item j is the proportion of the total variability of the scale this item accounts for. The rationale behind this set of weights becomes clear from the fact that, in general, variability is information. Essentially, an item with no variability will stay constant over all subjects in the population and will be useless. However, variability alone is not enough to evaluate the utility of an item. In fact, an item with a high variability can also be useless if most of this variability is due to measurement error. Therefore, the R_T coefficient looks for a compromise between these two important factors, i.e., the informational value of the item and the quality of that information captured by its reliability. The term $\nu_j R_j$ could then be interpreted as the relevant information conveyed by item j and the R_T coefficient as the relevant information conveyed by the entire scale.

Further, it is possible to show that R_T forms part of a more general family, based on the generalized eigenvalues associated with the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$. Indeed, $R_T \in \Omega$, where Ω is defined as

$$\Omega = \left\{ \theta : \theta = 1 - \sum_{j=1}^{p} w_j \lambda_j, \quad w_j > 0 \quad \text{and} \quad \sum_{j=1}^{p} w_j = 1 \right\}, \tag{4}$$

and the λ_js are the roots of $q(\lambda) = |\boldsymbol{\Sigma} - \lambda \boldsymbol{V}| = 0$. Associated with the generalized eigenvalue λ_j we have the so-called generalized eigenvectors $\boldsymbol{c}_j$s, which are the non-zero solutions of the linear equations $(\boldsymbol{\Sigma} - \lambda_j \boldsymbol{V})\boldsymbol{c}_j = \boldsymbol{0}$. It is useful to note that the λ_js could be equivalently defined as the eigenvalues of the matrix $\boldsymbol{H} = \boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2}$, where $\boldsymbol{V}^{1/2}$ denotes the symmetric square root of $\boldsymbol{V}$. Note that $\boldsymbol{H}$ is a symmetric matrix and, therefore, it can can be rewritten as $\boldsymbol{H} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$ where $\boldsymbol{P}$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \text{diag}(\lambda_j)$. If we call $\boldsymbol{M} = \boldsymbol{P}'\boldsymbol{V}\boldsymbol{P}$ then it can be shown that R_T is obtained by setting $w_j = \dfrac{m_{jj}}{\text{tr}(\boldsymbol{M})}$, where m_{jj} denotes the j^{th} element in the diagonal of $\boldsymbol{M}$. All members of Ω satisfy properties (1)–(4) and we refer the reader to the web appendix for a proof.

In multivariate analysis, the generalized variance of a random vector can be defined

using either the trace or the determinant of the corresponding variance-covariance matrix. Replacing the trace in the definition of R_T by the determinant leads to

$$R_\Lambda = 1 - \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{V}|} = 1 - |\boldsymbol{\Sigma}\boldsymbol{V}^{-1}|. \tag{5}$$

This measure also satisfies properties (1)–(4) and it is possible to show that $R_\Lambda = 1 - \prod_{j=1}^{p} \lambda_j$, i.e, this new coefficient is also based on the generalized eigenvalues used to construct the Ω family. In the following sections we will further argue that, unlike R_T which is a measure of average item reliability, R_Λ quantifies the reliability of the entire vector of items.

As previously stated, parallel, essentially tau-equivalent and congeneric tests have played a prominent role in the evaluation of reliability of multi-item scales. In the next section, we will apply the two newly introduced reliability coefficients in the scenarios defined by these models. It is important to point out that these measures are valid in more general settings than those defined by (i)–(iii). However, their performance in these special cases will help to increase our understanding of their properties and interpretation.

3 Reliability with Unidimensional True-score Models

Let us start by considering the simplest of the three special cases: the parallel test. In this setting, the decomposition of the variance-covariance matrix given in (2) takes the form: $\boldsymbol{V} = \sigma_\tau^2 \mathbf{1}\mathbf{1}' + \sigma^2 \boldsymbol{I}$ and from this expression easily follows that $R_T = \sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$. Basically, if we assume that the items of a scale form parallel tests, then each single item satisfies the model used in CTT, i.e., $X_{ij} = \mu_j + \tau_i + \varepsilon_{ij}$, and the reliability of all the items equals $\rho_{xx} = \sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$. Earlier, we have shown that R_T is a weighted average of the items' reliability. It then follows logically that, under the assumptions of parallelism, R_T equals the common item reliability.

When applied to this specific setting, R_Λ takes the form $R_\Lambda = p\rho_{xx}/\{(p-1)\rho_{xx} + 1\}$.

Interestingly, under these assumptions, R_Λ equals the Spearman-Brown formula. The expression indicates that the reliability of the instrument is an increasing function of the number of items, which is an intuitive and appealing result. In fact, every new item added to the scale will bring certain level of information about the true score τ_i, even if this information is contaminated by measurement error. As a consequence, the expanded scale will always contain more or at least the same amount information about τ_i than the original scale. Intuitively, the reliability of a scale is the amount of information on the true scores that the scale conveys. Therefore, it is reasonable that adding new items to the instrument can only increase the reliability of the conclusions derived from it.

It is important to recall at this point that R_Λ quantifies the reliability of the entire scale, i.e., the multivariate vector $\boldsymbol{X}_i$. However, the Spearman-Brown formula was originally obtained as the reliability of the scale $Y_i(\mathbf{1}) = \mathbf{1}'\boldsymbol{X}_i$ under the parallel test assumptions. We thus find that, under these assumptions, the reliability of the entire scale $\boldsymbol{X}_i$ equals the reliability of the simple sum score. Nevertheless, as we will illustrate later, in more general settings $Y_i(\mathbf{1})$ no longer has the same reliability as the entire scale $\boldsymbol{X}_i$ but, as expected, the reliability of a summary statistic like $Y_i(\mathbf{1})$ is usually smaller than the one of the entire instrument.

Essentially tau-equivalent tests relax the assumptions of parallel tests by allowing item-specific error variances so that $\boldsymbol{V} = \sigma_\tau^2 \mathbf{1}\mathbf{1}' + \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = \text{diag}(\sigma_j^2)$. Under these assumptions, R_T takes the form $R_T = \sigma_\tau^2/(\sigma_\tau^2 + S_{E1})$, where $S_{E1} = (\sum_j \sigma_j^2)/p$. Note that R_T is a decreasing function of S_{E1} and, therefore, if a new item $(p+1)$ is added to a scale, then

$$R_T(p) \leq R_T(p+1) \quad \text{if and only if} \quad \sigma_{p+1}^2 \leq \frac{\sum_j \sigma_j^2}{p}.$$

This implies that the expanded instrument will have a higher average reliability if and only if the error variance of the new item is smaller than the average error variance of the other items of the scale. Hence, the R_T coefficient can either increase or decrease when a new item is added, depending on the "quality" of such an item.

Turning to R_Λ, we first need to compute the determinant of $\boldsymbol{V}$. It is easy to show that if $\boldsymbol{V} = \sigma_\tau^2 \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Sigma}$ then

$$|\boldsymbol{V}| = (1 + \sigma_\tau^2 \boldsymbol{\beta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}) \cdot |\boldsymbol{\Sigma}|. \tag{6}$$

For essentially tau-equivalent tests $\boldsymbol{\beta} = \mathbf{1}$, $\boldsymbol{\Sigma} = \text{diag}(\sigma_j^2)$, and from (6) we obtain $R_\Lambda = S_{E2}/(1 + S_{E2})$, which is an increasing function of S_{E2}, with $S_{E2} = \sum_{j=1}^{p} \sigma_\tau^2/\sigma_j^2$. Obviously, adding a new item to the scale can only increase the value of S_{E2} and, therefore, R_Λ is always an increasing function of the number of items. Note however that, if the new item comes contaminated with a lot of measurement error then $\sigma_\tau^2/\sigma_{p+1}^2$ will be negligible and R_Λ will remain nearly constant.

Finally, congeneric convey the most general model among the three special cases. In this scenario the variance-covariance matrix takes the more general form $\boldsymbol{V} = \sigma_\tau^2 \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Sigma}$ and $R_T = \sigma_\tau^2/(\sigma_\tau^2 + S_{C1})$, with $S_{C1} = \sum_j \sigma_j^2 / \sum_j \beta_j^2$. Like before, adding a new item can increase or decrease the value of R_T depending on the impact of the new item on S_{C1}. Moreover, from (6) easily follows that $R_\Lambda = \sigma_\tau^2 S_{C2}/(1 + \sigma_\tau^2 S_{C2})$, with $S_{C2} = \sum_j \beta_j^2/\sigma_j^2$, and like for tau-equivalent tests, R_Λ can only increase its value when a new item is added.

The above reflections are a useful aid in understanding the meaning and the complementarity of the two new measures. Whereas R_T provides us with information on the quality of the items in a scale, regardless of their number, the R_Λ coefficient informs us on the amount of information the total package of items contain on the underlying traits.

However, due to the strong assumptions on which they are based, the applicability of the modelling frameworks analyzed in this section is very limited (Green and Yang 2009). In the next section we will apply the new measures in the more general scenario defined by Model (1). The weaker assumptions that this model requires enhance its practical value and, as a consequence, the newly proposed measures will also allow us to approach the reliability problem in more general settings.

4 Reliability with Multidimensional True-score Models

In models (i)–(iii), τ_i is a scalar, which means that unidimensionality of the instrument is assumed. Although ideally scales aim to assess primarily a single construct for interpretability (McDonald 1981, Hattie 1985), they frequently include additional factors and are not unidimensional (Reise, Waller, and Comrey 2000). Actually, many psychometricians argue that it is preferable that the structure underlying the items is more consistent with a complex hierarchical model than with a unidimensional model (e.g., McDonald 1999, Reise *et al* 2000).

Werts *et al* (1978) extended the measurement models (i)–(iii) by assuming a factor model for the true scores, thence allowing multiple dimensions in the measurement instrument. The specific factors in their model are considered as part of the true scores, so that the model contains specific factors as well as an error component. Such a model might, however, lead to identifiability problems. In their data example, Werts *et al* (1978) assume the specific factors to be zero. Tarkkonen and Vehkalahti (2005) suggested considering the specific factors as measurement errors.

In general, these authors do not directly study the reliability of the multivariate instrument $\boldsymbol{X}_i$ but the reliability of a new scale $Y_i(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{X}_i$ formed as a weighted sum of the items' score. When Model (1) holds and $\boldsymbol{a} \in \mathbb{R}^p$, $Y_i(\boldsymbol{a})$ can be written as

$$Y_i(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{X}_i = \boldsymbol{a}'\boldsymbol{\mu} + \boldsymbol{a}'\boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{a}'\boldsymbol{\varepsilon}_i.$$

If $\sigma_Y^2 = \mathrm{Var}\,[Y_i(\boldsymbol{a})]$, then (1) implies $\sigma_Y^2 = \boldsymbol{a}'\boldsymbol{G}\boldsymbol{a} + \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}$. Tarkkonen and Vehkalahti (2005) proposed to quantify the reliability of $Y_i(\boldsymbol{a})$ as

$$\rho(\boldsymbol{a}) = \frac{\boldsymbol{a}'\boldsymbol{G}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = 1 - \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}}. \tag{7}$$

Notice that expression (7) is just the classical definition (CTT) of reliability applied to the measure $Y_i(\boldsymbol{a})$. In Section 2 we introduced the matrix $\boldsymbol{H} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, with $\boldsymbol{P}$ an orthogonal matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_j)$. Using this decomposition of $\boldsymbol{H}$ one can show that

$\boldsymbol{\Sigma} = \boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q}$, and $\boldsymbol{V} = \boldsymbol{Q}'\boldsymbol{Q}$, with $\boldsymbol{Q} = \boldsymbol{P}'\boldsymbol{V}^{1/2}$. Finally, we can rewrite $\rho(\boldsymbol{a})$ as

$$\rho(\boldsymbol{a}) = 1 - \frac{\boldsymbol{a}'\boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{a}}. \tag{8}$$

This new expression for $\rho(\boldsymbol{a})$ will play an important role in subsequent developments.

Werts *et al* (1978) proposed a quantification of reliability very similar to (7), actually, their proposal equals (7) when the specific factors, included in their model, are assumed to be zero. Tarkkonen and Vehkalahti (2005) further proved that $\rho(\mathbf{1}) \geq \alpha$, where α denotes the value of Cronbach's alpha coefficient, with equality if and only if $\boldsymbol{G} = \sigma_\tau^2 \mathbf{1}\mathbf{1}'$ ($\sigma_\tau^2 > 0$) and $\boldsymbol{\Sigma}$ diagonal, i.e., exactly the conditions defined by (ii).

Interestingly, if we apply the measures R_T and R_Λ to quantify the reliability of the previous weighted sum $Y_i(\boldsymbol{a})$, we find that $R_T(\boldsymbol{a}) = R_\Lambda(\boldsymbol{a}) = \rho(\boldsymbol{a})$. Obviously, in this univariate scenario, the average and total reliability coincide and, as a consequence, R_T and R_Λ are equal.

In the remainder of this section we will explore the relationship between the reliability of the unidimensional scale $Y_i(\boldsymbol{a})$ and the reliability of the original instrument $\boldsymbol{X}_i$.

4.1 The Ω Family

As stated before, a considerable part of the psychometric literature has focussed on studying the reliability of the family of scales $\Psi^* = \{Y_i(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{X}_i : \boldsymbol{a} \in \mathbb{R}^p\}$. Moving from a high-dimensional instrument $\boldsymbol{X}_i$ to a univariate version $Y_i(\boldsymbol{a})$ can considerably facilitate the practical use of the scale and the clinical interpretation of its results. Therefore, in clinical practice, psychiatrists and psychologists frequently work with weighted sums of multivariate scales.

Also from a psychometric perspective working with the univariate version $Y_i(\boldsymbol{a})$ represents an important simplification. Basically, such a reduction in the dimensionality of the instrument allows the direct application of the classical definition of reliability, as it was shown in (7). A relevant question that then arises is to which extend the reliability of the new scale $Y_i(\boldsymbol{a})$ reflects the reliability of the original instrument $\boldsymbol{X}_i$. We will try to

address this issue by studying the relationship between the scales in Ψ^* and the reliability measures contained in Ω.

Recall that, in Section 2, Ω was introduced as general family of plausible reliability measures for $\boldsymbol{X}_i$. Different measures were formed by assigning different weights to the generalized eigenvalues $\lambda_1, \ldots, \lambda_p$ in (4). The following two theorems will shed light on the relationship between the family of scales Ψ^* and the family of measures Ω.

Theorem 1 *If Model (1) holds and $\theta \in \Omega$ then there exists a vector $\boldsymbol{a} \in \mathbb{R}^p$ so that the reliability of the weighted scale $Y_i(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{X}_i$ equals θ, i.e., $\rho(\boldsymbol{a}) = \theta$.*

Given a member of the Ω family $\theta = 1 - \sum_j w_j \lambda_j$, one can construct the vector $\boldsymbol{\delta}$ which j^{th} component equals $\delta_j = \sqrt{w_j}$. The result then immediately follows from expression (8) with $\boldsymbol{a} = \boldsymbol{Q}^{-1}\boldsymbol{\delta}$. Theorem 1 shows that any reliability measure for $\boldsymbol{X}_i$, contained in Ω, can also be interpreted as the reliability of certain univariate scale in Ψ^*. The reverse relationship is also of interest, i.e., one would like to know whether for each $\boldsymbol{a} \in \mathbb{R}^p$ we can find a corresponding measure θ within Ω so that $\rho(\boldsymbol{a}) = \theta$. The following theorem will address this question, but let us first define the $p \times p$ matrix $\boldsymbol{\Gamma} = (\gamma_{ij})$, where $\gamma_{ij} = 1$ if $\lambda_i = \lambda_j$ and zero otherwise. Further, let us define the set $C = \{\boldsymbol{a} \in \mathbb{R}^p : T(\boldsymbol{a})_j \neq 0 \quad \forall j\}$ where $T(\boldsymbol{a}) = \boldsymbol{\Gamma}\,(\boldsymbol{\delta} \circ \boldsymbol{\delta})$, $\boldsymbol{\delta} = \boldsymbol{Q}\boldsymbol{a}$ and $\circ$ denotes the Hadamard product (entrywise). Note that if all the generalized eigenvalues are different then $\boldsymbol{\Gamma} = \boldsymbol{I}$ and $T(\boldsymbol{a})_j \neq 0$ is equivalent to the simpler equation $(\boldsymbol{Q}\boldsymbol{a})_j \neq 0$.

Theorem 2 *Let us assume that Model (1) holds. If $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{a} \neq \boldsymbol{0}$, and there exists a $\theta \in \Omega$ so that $\rho(\boldsymbol{a}) = \theta$ then $\boldsymbol{a} \in C$. Similarly, if $\boldsymbol{a} \in C$ then there exists a $\theta \in \Omega$ so that $\rho(\boldsymbol{a}) = \theta$.*

A proof of this result can be found in the web appendix. Theorem 2 shows that not all the scales in Ψ^* will properly reflect the reliability of $\boldsymbol{X}_i$, at least not in the sense the members of Ω do it. Essentially, Ω is not equivalent to Ψ^* but to the family $\Psi = \{Y_i(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{X}_i : \boldsymbol{a} \in C\}$. Formally, Ψ^* will be equivalent to a more general family Ω^* which can

be defined as

$$\Omega^* = \left\{ \theta : \theta = 1 - \sum_{j=1}^{p} w_j \lambda_j, \quad w_j \geq 0 \quad \text{and} \quad \sum_{j=1}^{p} w_j = 1 \right\}.$$

Note, however, that the elements of Ω^* do not necessarily satisfy properties (1)–(4). Additionally, the reliability of some members of Ψ^* can dramatically differ from the reliability of the original instrument $\boldsymbol{X}_i$ and this difference can have an important impact on the conclusions these scales produce. To illustrate this, let us denote by $C(\boldsymbol{B})$ the column space associated with $\boldsymbol{B}$. Further, we will consider $\boldsymbol{a} \in C(\boldsymbol{B})^{\perp}$, where $C(\boldsymbol{B})^{\perp}$ denotes the orthogonal complement of $C(\boldsymbol{B})$. Obviously, $Y_i(\boldsymbol{a}) \in \Psi^*$ and assuming that Model (1) holds, we have

$$Y_i(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{X}_i = \boldsymbol{a}'(\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_i) = \boldsymbol{a}'\boldsymbol{\mu} + \boldsymbol{a}'\boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{a}'\boldsymbol{\varepsilon}_i = \boldsymbol{a}'\boldsymbol{\mu} + \boldsymbol{a}'\boldsymbol{\varepsilon}_i.$$

Clearly, this scale does not contain any information about the true scores $\boldsymbol{\tau}_i$ and $\rho(\boldsymbol{a}) = 0$ irrespectively of the reliability of the original scale $\boldsymbol{X}_i$. It is possible to show that this scale does not belong to Ψ. Indeed, denoting $\boldsymbol{\delta} = \boldsymbol{Q}\boldsymbol{a}$, from (2) we have that $\boldsymbol{Q}\boldsymbol{a} = \boldsymbol{P}'\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{a}$. Moreover, from $\boldsymbol{H} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$ one gets that $\boldsymbol{P}'\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{a} = \boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{a}$ and this implies $\boldsymbol{\delta} = \boldsymbol{\Lambda}\boldsymbol{\delta}$ or, equivalently, $\delta_j = \lambda_j \delta_j$ for all j. If $\boldsymbol{V} \neq \boldsymbol{\Sigma}$ there will be at least a k so that $\lambda_k \neq 1$. It then easily follows that $T(\boldsymbol{a})_k = 0$. Only when $\boldsymbol{V} = \boldsymbol{\Sigma}$, i.e, when the reliability of $\boldsymbol{X}_i$ is equal to zero, $Y_i(\boldsymbol{a})$ will have the reliability of the original instrument. As stated at the beginning of this section, working with unidimensional versions of $\boldsymbol{X}_i$, like $Y_i(\boldsymbol{a})$, can considerably simplify the clinical use and interpretation of the scale. Frequently, one would like the new scale $Y_i(\boldsymbol{a})$ to reflect as much as possible the characteristics of the original and more complex instrument $\boldsymbol{X}_i$. However, the previous results show that these unidimensional versions do not always mimic the psychometric properties of the original scale. This can be specially important if $Y_i(\boldsymbol{a})$ is used to gain information about the performance of $\boldsymbol{X}_i$. The previous finding can be a useful guideline to construct a unidimensional version of $\boldsymbol{X}_i$ that preserves its reliability.

4.2 Weighted Score versus Multivariate Score

In Section 3 we found that, for parallel tests, the reliability of the simple sum score equals the reliability of the entire scale $\boldsymbol{X}_i$. Since parallel tests contain only one latent true-score, it is intuitively logical that a single well-chosen linear combination can fully capture all the information in the data. The following theorem states that such a linear combination can only be found when the true scores are unidimensional, otherwise, the reliability of a weighted sum will always be smaller than or equal to the reliability of the entire scale.

Theorem 3 *Let us assume that Model (1) holds and* $\boldsymbol{V} \neq \boldsymbol{\Sigma}$ *with* $\boldsymbol{\Sigma}$ *not singular. If* $\boldsymbol{a} \in \mathbb{R}^p$*, then the reliability of the scale* $Y_i(\boldsymbol{a}) = \boldsymbol{a}'\boldsymbol{X}_i$ *is always smaller than or equal to the reliability of* $\boldsymbol{X}_i$*, i.e.,* $\rho(\boldsymbol{a}) \leq R_\Lambda$*. The equality is obtained if and only if rank*$(\boldsymbol{B}) = 1$ *and* $\boldsymbol{a}$ *is proportional to* $\boldsymbol{c}_{(1)}$*, a generalized eigenvector associated with the smallest generalized eigenvalue* $\lambda_{(1)}$*.*

For a detailed proof we refer the reader to the web appendix. Note that if $\text{rank}(\boldsymbol{B}) = 1$ then the full column rank assumption implies that $q = 1$ and, therefore, $\boldsymbol{\tau}_i$ is a scalar. The theorem also establishes R_Λ as an upper bound for the reliability of an entire family of instruments constructed from the original set of items. Notice that if the value of this measure is low, then any instrument derived as a weighted sum of the original items will have an even lower reliability and will be basically useless.

Furthermore, in the special cases considered in Section 3, we showed that R_Λ is an increasing function of the number of items. The following theorem extends this result to the more general scenario implied by Model (1).

Theorem 4 *Let us assume that Model (1) holds and* $\boldsymbol{\Sigma}$ *is not singular. Further, denote by* $R_\Lambda(\boldsymbol{X}_p)$ *the value of* R_Λ *for the p-dimensional scale* $\boldsymbol{X}_p$*. If r additional items are added to* $\boldsymbol{X}_p$*, then the value of* R_Λ *for this new* $(p+r)$*-dimensional scale* $\boldsymbol{X}_{p+r}$ *satisfies* $R_\Lambda(\boldsymbol{X}_{p+r}) \geq R_\Lambda(\boldsymbol{X}_p)$*.*

A detailed proof can also been seen in the web appendix. All the previous findings indicate that the R_Λ coefficient can be interpreted as a generalization of the Spearman-Brown formula. Finally, in the following two sections we will study the relationship between these newly proposed measures and two other important concepts: correlation and prediction.

5 Reliability and Correlation

In CTT reliability equals the squared correlation between the observed and true scores. Actually, in CTT one can alternatively define reliability by using the previous property as its definition. In this section, we will study the relationship between the measures of reliability previously introduced and the squared association between $\boldsymbol{X}_i$ and $\boldsymbol{\tau}_i$. Let us start by denoting $\boldsymbol{S}_i = \begin{pmatrix} \boldsymbol{X}_i \\ \boldsymbol{\tau}_i \end{pmatrix}$. If one further assumes that the true scores and the error terms are normally distributed then the following result follows

Theorem 5 *If Model (1) holds,* $\boldsymbol{\tau}_i \sim N(\mathbf{0}, \boldsymbol{D})$ *and* $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ *then:* $\boldsymbol{S}_i \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ *where*

$$\boldsymbol{\mu}_0 = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix} \quad \textit{and} \quad \boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{V} & \boldsymbol{BD} \\ (\boldsymbol{BD})' & \boldsymbol{D} \end{pmatrix}.$$

A natural way to quantify the association between $\boldsymbol{X}_i$ and $\boldsymbol{\tau}_i$ in this context is through the use of canonical correlations. The squared canonical correlations of $\boldsymbol{S}_i$ are then the eigenvalues of the matrix $\boldsymbol{V}^{-1/2}\boldsymbol{Z}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{Z}'\boldsymbol{V}^{-1/2} = \boldsymbol{I} - \boldsymbol{H}$. It is easy to show that if λ is an eigenvalue of the matrix $\boldsymbol{H}$ then $1-\lambda$ is an eigenvalue of the matrix $\boldsymbol{I} - \boldsymbol{H}$. The implications of these results are very appealing. In fact, they show that if $\theta \in \Omega$ then $\theta = \sum_{j=1}^{p} w_j \rho_j^2$ where the elements $\rho_j^2 = 1 - \lambda_j$ are just the squared canonical correlations of the observed and true scores. Similarly, it is easy to show that $R_\Lambda = 1 - \prod_{j=1}^{p}(1-\rho_j^2)$. It is appealing to see that two equivalent classical definitions of reliability also concur in this extended setting. Notice that any extension of the classical definition of reliability should necessarily be based on the ρ_j^2, if it wants to retain its interpretation as the

squared correlation between the observed and true scores. However, a high-dimensional vector of squared canonical correlations may be difficult to interpret and difficult to use when comparing two scales regarding their reliabilities. Therefore, aiming at an easier interpretation, the new measures summarize the information about the reliability, contained in the vector of squared canonical correlations, by using meaningful functions of its elements.

6 Reliability and Prediction of the True Scores

The prediction of the true scores has been at the center of many developments in CTT and item response theory (IRT). In the present section we will study the relationship between this problem and the concept of reliability.

Theorem 5 implies that $\boldsymbol{\tau}_i|\boldsymbol{X}_i \sim N\left(\boldsymbol{\mu}(\boldsymbol{X}_i), \boldsymbol{\Sigma}^p\right)$, where $\boldsymbol{\mu}(\boldsymbol{X}_i) = \boldsymbol{D}\boldsymbol{B}'\boldsymbol{V}^{-1}\left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)$ and $\boldsymbol{\Sigma}^p = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{B}'\boldsymbol{V}^{-1}\boldsymbol{B}\boldsymbol{D}$. If one uses $\boldsymbol{\mu}(\boldsymbol{X}_i)$ as a predictor for $\boldsymbol{\tau}_i$ then it is possible to show that the generalized variance of this prediction satisfies $|\boldsymbol{\Sigma}^p| = |\boldsymbol{D}|(1 - R_\Lambda)$. The previous expression illustrates that the generalized variance of the prediction is directly proportional to the variability of the true scores and inversely proportional to the reliability of the scale. As a consequence, more reliable scales will produce more accurate predictions of the true scores, a very intuitive and appealing result. Using the trace instead of the determinant to quantify the variability of the prediction, leads to the analogous expression $\mathrm{tr}(\boldsymbol{\Sigma}^p) = \mathrm{tr}(\boldsymbol{D})(1 - R_{PT})$ where

$$R_{PT} = \frac{\mathrm{tr}(\boldsymbol{D}\boldsymbol{B}'\boldsymbol{V}^{-1}\boldsymbol{B}\boldsymbol{D})}{\mathrm{tr}(\boldsymbol{D})}.$$

It can be easily shown that R_{PT} satisfies the properties (1)–(4) introduced in Section 2. Observe also that $\mathrm{Cov}\left(\boldsymbol{\mu}(\boldsymbol{X}_i)\right) = \boldsymbol{D}\boldsymbol{B}'\boldsymbol{V}^{-1}\boldsymbol{B}\boldsymbol{D}$ and, therefore, R_{PT} can be interpreted as the proportion of the total variability of $\boldsymbol{\tau}_i$ that $\boldsymbol{X}_i$ explains. If we assume that $\boldsymbol{D} = \boldsymbol{I}$ then $R_{PT} = (p/q)R_p$ where $R_p = 1 - \dfrac{1}{p}\mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{V}^{-1})$ is a member of the Ω family. For parallel, congeneric and essentially tau equivalent tests R_{PT} coincides with R_Λ and, after some algebraic transformations, one can show that, in general, $R_\Lambda \approx 1 - e^{-qR_{PT}}$. Obviously,

more research is necessary to fully understand the connection between reliability and the prediction of the true scores and between the measures of reliability introduced in the previous sections and the predictive measure of reliability introduced here.

7 A Case Study in Schizophrenia

Schizophrenia is a complex and heterogeneous disorder with variable symptoms. To improve research clarifying the diversity in the disorder, Kay, Fiszbein and Opler (1987) developed a standardized instrument; the Positive And Negative Syndrome Scale (PANSS). The instrument contains 30 items (symptoms), which are all scored on a 7-grade scale ranging from "absent" to "extreme." As reflected by the name of the scale, schizophrenia is often described in terms of positive and negative symptoms. Positive symptoms include hallucinations and delusions and are typically regarded as manifestations of psychosis. Negative symptoms are so-named because they are considered to be the loss or absence of normal traits or abilities, and include features such as blunted affect, apathy, and social withdrawal. Besides these two dimensions, general psychopathology was included as a third a priori factor in the PANSS (Kay *et al* 1987). However, empirical research suggests the existence of five factors, which can be described as: negative syndrome, positive syndrome, excitement, depressive symptoms, and cognitive dysfunction (Lindemayer *et al* 1995). Many other studies have confirmed a five-factor structure for this scale (e.g. Van der Gaag *et al* 2006a).

Even though the five-factor model is confirmed by several studies, differences are often found in the exact allocation of the items to the factors. Such differences might be related to the use of different statistical techniques or model assumptions, but also to differences in the investigated populations. Dolfus and Petit (1995), for example, did not observe a depression dimension in an acute population while it was observed in a chronic population. A plausible explanation is that depressive symptoms cannot be expressed when positive symptoms are very severe. In many of the studies investigating the factor

structure of PANSS, models have been developed where each item loads only on one factor. The underlying aim is to divide the scale in separate sub-scales composed of clearly distinguished sets of items. Van der Gaag *et al* (2006b) showed, by means of a cross-validation study, that allowing some items to load on more than one factor leads to a better model fit.

7.1 Data Analysis

The case study data consist of clinical trial baseline measurements taken from 520 patients with a diagnosis of chronic schizophrenia after a single-blind placebo washout period (Chouinard *et al* 1993, Marder and Meibach 1994).

The first step is to find a well fitting model that provides us with the parameter estimates necessary for the estimation of reliability. As expressed in (2), variability in the observations comes from two sources, the latent variables (random effects) and the measurement errors. Since both are unobserved, model restrictions are inevitable to avoid identifiability problems. A factor-analytic approach was applied to fit and compare different models.

As many studies of this scale have suggested a five-factor structure, we fitted an exploratory 5-factor model. Note that this model is a special case of Model (1), when $\boldsymbol{D} = \boldsymbol{I}$, $\boldsymbol{\Sigma}$ is a diagonal matrix, and $\boldsymbol{B}$ unstructured. As a sensitivity analysis we also considered an exploratory 7 factor model. Two confirmatory 5-factor models were also fitted in which restrictions were mainly laid on the $\boldsymbol{B}$ matrix by allowing the items to load only on pre-defined factors. In the first one, each item loaded on one factor only. Basically, the model followed the five sub-scales proposed by Marder, Dabis and Chouinard (1997) with $\boldsymbol{D}$ an unstructured correlation matrix and $\boldsymbol{\Sigma}$ a diagonal matrix. The second confirmatory model was the one proposed by Van der Gaag (2006b). In this model, several items can load on more than one factor. Further, some factors were assumed to be correlated and also some pre-specified measurement errors could be correlated.

All models were fitted using maximum likelihood and compared using the Consistent Akaike's Information Criterion (CAIC) and the Schwarz's Bayesian Criterion (SBC).

Both criteria showed that the model introduced by Van der Gaag (2006b) was the one producing the best fit, as can be seen in Table 1.

Table 2 presents the reliability estimates for the four fitted models. Interestingly, all these models produced relatively similar estimates suggesting certain degree of "robustness" with respect to the model assumptions. Only for the poorest fitting model (CFA Marder) the R_T estimate is somewhat lower. The previous results show that while finding the 'best' model can be hard, it is sufficient to find a good fitting model in order to estimate reliability.

The R_T coefficient indicates the average item reliability and estimates are close to 0.50, which is, for a single item, certainly an acceptable level. As stated before, R_Λ quantifies the information content when all items are considered jointly, i.e., it expresses the reliability of the entire multivariate scale. The fact that individual items already achieve a decent reliability level and that PANSS contains no less than 30 items, explains why we obtain values for R_Λ very close to one.

In practice, the sum score of the PANSS items is mostly used for clinical evaluation and data analysis. We have already shown that working with the sum of the item scores always leads to a certain amount of information loss. Table 2, however, shows that the reliability of the sum score, expressed by $\rho(\mathbf{1})$, although lower than R_Λ still has a very high value. The results thus illustrate that summing the PANSS items leads to a relatively small loss of information. It is important to point out here that these two reliability measures are valid at two different levels. Indeed, the R_Λ quantifies the amount of information shared by the vector of observed scores and the vector of true scores, whereas $\rho(\mathbf{1})$ quantifies the information shared by a well-chosen linear combination of the observed scores and a corresponding linear combination of the true scores. At any rate, the high reliability of the sum score obtained for this scale suggests that working with the sum for clinical evaluation and data analysis may be a sensible strategy given the substantial simplification that it brings.

Interest may also lie in estimating the patients' scores on the PANSS sub-scales. For

example, Marder *et al* (1997) investigated drug-effectiveness on the different dimensions of schizophrenia. Reliability estimates for the separate sub-scales can be obtained by replacing the full matrices in (3), (5), and (7) by sub-matrices related to the variances and covariances between the items in the sub-scale. Table 3 presents the point estimates of the three reliability measures, for each of the five sub-scales. The estimates are based on the five-factor exploratory factor-analytic model. The items in each of the different sub-scales provide sufficient information on the respective underlying dimensions as indicated by R_Λs greater than 0.85. Additionally, the sum-score reliabilities are all above 0.75. Interestingly, the negative sub-scale clearly has a higher average (R_T) and sum-score [$\rho(\mathbf{1})$] reliability than the positive sub-scale, however the R_Λs are similar. This owes to the fact that the positive sub-scale has 8 items whereas the negative sub-scale has 7. We can also see that for the positive sub-scale, about 15% of information is lost due to summing the item scores, for the negative sub-scale only 5% is lost. Finally, it is clearly illustrated that both R_Λ and $\rho(\mathbf{1})$ are affected by the number of items, resulting in lower estimates for the subscales compared to the total scale. This is clearly not the case for R_T. The latter is therefore more informative on the quality of the items, irrespective of the size of the scale. Note that item-specific estimates could also be obtained.

Finally we also estimated the predictive measure of reliability R_{PT} introduced in Section 6. For the 5-factor EFA model $R_{PT} = 0.829$. We have shown in Section 6 that R_{PT}, like R_Λ, is linked to the prediction of the true scores. The high values obtained for these two measures hint then on the possibility of accurately predicting the true scores, using the information conveyed by the observed scores. Remarkably, the approximation $R_\Lambda \approx 1 - e^{-qR_{PT}}$ obtained under the assumption $\boldsymbol{D} = \boldsymbol{I}$ seems to hold under more general models like the 5-factor CFA model proposed by Van der Gaag for which we found $\widehat{R}_{PT} = 0.861$.

8 Discussion

In this paper we have introduced two measures of reliability, the so-called R_T and R_Λ coefficients. Unlike many previously proposed measures, for example, the Spearman-Brown formula or the Cronbach's α, the R_T and R_Λ coefficients are valid under a more general modelling framework that can include multivariate true-scores. As a consequence, these coefficients will allow the assessment of reliability in practical situations where the previous measures are not valid or can merely be considered as a lower bound for it.

We have seen that the R_T coefficient expresses an average item reliability and it would easily allow to compare the quality of the items between two scales of different length. Additionally, we have shown that R_T can be framed into a more general family of measures, all of which satisfy a minimum set of intuitive properties.

The R_Λ coefficient captures the reliability of the entire multivariate set of items. The practical implications of such a measure are important. First, one has a tool that will alow the assessment of the reliability of the entire scale. If such an assessment shows a low value of reliability, then any simplified version of the instrument will be useless. Second, it also allows to evaluate the loss of reliability implied by using a simplified version of the instrument, constructed as a weighted sum of the item scores. This will be crucial, for instance, to evaluate the trade-off between the gain in interpretability and the loss in reliability implied by the use of this simpler univariate instrument.

We have also introduced a measure of predictive reliability, the so-called R_{PT}, this measure in particular and the entire connection between reliability and prediction in general, is a promising line of research that will certainly get attention in the future.

It is important to point out that the results obtained with general measures of reliability always need to be embedded into a broader analysis, in order to get a clearer idea of the properties and performance of the scale. For instance, the specific reliability of each item should be calculated. Additionally, one could also calculate the values of the general measures sequentially after adding one item at a time. This will help to understand the

particular contribution of each item to the global and average reliability. For simplicity, we have not carried out such a detailed analysis in our case study but it should certainly be part of any real application.

Finally, we would like to remark that similar measures can also be applied in a longitudinal framework (Laenen *et al* 2009a, 2009b). The fact that similar concepts can be applied to evaluate reliability in two different and important scenarios brings some degree of conceptual unity to the entire problem of estimating reliability.

Supplementary Materials

Web appendix is available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgements

The authors are grateful to J&J PRD for the data. We gratefully acknowledge support from Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data."

References

Chouinard, G., Jones, B., Remington, G., Bloom, D., Addington, D., MacEwan, G.W., Labelle, A., Beauclair, L., and Arnott, W. (1993). A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients *Journal of Clinical Psychopharmacology* **13**, 25–40.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: John Wiley.

Dolfus, S., and Petit, M. (1995). Principal-component analyses of PANSS and SANS-SAPS in schizophrenia: their stability in an acute phase. *European Psychiatry*, **10**,

97–106.

Green, S.B., and Yang, Y. (2009). Commentary on coefficient alpha: a cautionary tale. *Psychometrika*, **74**, 121–135.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of test and items. *Applied Psychological Measurement*, **9**, 139–164.

Kay, S.R., Fiszbein, A., and Opler, L.A. (1987). The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin*, **13**, 261–276.

Laenen, A., Alonso, A., Molenberghs, G., and Vangeneugden, T. (2009a). A family of parameters to investigate the reliability of a psychiatric symptom scale. *Journal of the Royal Statistical Society - Series A*, **172**, 1–17.

Laenen, A., Alonso, A., Molenberghs, G., and Vangeneugden, T. (2009b). Reliability of a longitudinal sequence of scale ratings. *Psychometrika.* **74**, 49–64.

Lindenmayer, J.P., Bernstein-Hyman, R., Grochowski, S., and Bark, N. (1995). Psychopathology of schizophrenia: initial validation of a 5-factor model. *Psychopathology*, **28**, 22–31.

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

Marder, S.R., and Meibach, R.C. (1994). Risperidone in the treatment of schizophrenia *American Journal of Psychiatry*, **151**, 825–835.

Marder, S.R., Dabis, J.M., and Chouinard, G. (1997). The effects of Risperidone on the five dimensions of schizophrenia derived by factor analysis: combined results of the North American trials. *Journal of Clinical Psychiatry*, **58**, 538–546.

McDonald, R.P. (1981). The dimensionality of test and items. *British Journal of Mathematical and Statistical Psychology*, **34**, 100–117.

McDonald, R.P. (1999). *Test Theory: A Unified Approach.* Hillsdale: Erlbaum.

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

Tarkkonen, L., and Vehkalahti, K. (2005). Measurement errors in multivariate measurement scales. *Journal of Multivariate Analysis* **96**, 172–189.

Reise, S.P., Waller, N.G., and Comrey, A.L. (2000). Factor analysis and scale revision. *Psychological Assessment*, **12**, 287–297.

Sijtsma, K. (2009). On the use, the misuse and the very limited usefulness of Cronbach's alpha. *Psychometrika*, **74**, 107–120.

Van der Gaag, M., Cuijpers A., Hoffman, T., Remijsen, M., Hijman, R., de Haan, L., van Meijel B., van Harten, P.N., Valmaggia, L., de Hert, M., and Wiersma, D. (2006a). The five-factor model of the Positive and Negative Syndrome Scale I: Confirmatory factor analysis fails to confirm 25 published five-factor solutions. *Schizophrenia Research,* **85**, 273–279.

Van der Gaag, M., Hoffman, T., Remijsen, M., Hijman, R., de Haan, L., van Meijel B., van Harten, P.N., Valmaggia, L., de Hert, M., Cuijpers A., and Wiersma, D. (2006b). The five-factor model of the Positive and Negative Syndrome Scale II: An ten-fold cross-validation of a revised model. *Schizophrenia Research,* **85**, 280–287.

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.

Table 1: *Fit statistics (CAIC and SBC) for two exploratory factor models (EFA) and two confirmatory factor models (CFA).*

model	CAIC	SBC
EFA 5 factors	-1240	-945
EFA 7 factors	-1213	-967
CFA Marder	-796	-401
CFA Van der Gaag	-1420	-1048

Table 2: *Point estimates [95% confidence intervals] for the reliability measures.*

model	R_T	R_Λ	$\rho(\mathbf{1})$
EFA 5 factors	0.479 [0.436; 0.522]	1.000 [0.999; 1.000]	0.911 [0.872; 0.939]
EFA 7 factors	0.521 [0.481; 0.562]	1.000 [1.000; 1.000]	0.918 [0.878; 0.946]
CFA Marder	0.414 [0.394; 0.435]	1.000 [0.999; 1.000]	0.895 [0.793; 0.951]
CFA Van der Gaag	0.446 [0.424; 0.468]	1.000 [1.000; 1.000]	0.888 [0.765; 0.952]

Table 3: *Reliability measures for selected sub-scales.*

	Positive	Negative	Cognitive	Excitement	Depression
R_T	0.401	0.571	0.436	0.590	0.466
R_Λ	0.949	0.942	0.914	0.902	0.858
$\rho(\mathbf{1})$	0.798	0.894	0.829	0.836	0.754