TECHNICAL
REPORT

0655

A FAMILY OF MEASURES TO EVALUATE SCALE RELIABILITY IN A LONGITUDINAL SETTING

LAENEN, A., ALONSO, A., MOLENBERGHS, G. and T. VANGENEUGDEN

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

A Family of Measures to Evaluate Scale Reliability in a Longitudinal Setting

Annouschka Laenen, Ariel Alonso, Geert Molenberghs

Center for Statistics, Hasselt University,
Agoralaan 1, B-3590 Diepenbeek, Belgium
email: annouschka.laenen@uhasselt.be

Tony Vangeneugden

Tibotec, Johnson & Johnson, 2800 Mechelen, Belgium

Summary. The reliability of a measurement scale is one of its most important psychometric properties. Reliability not only has a considerable clinical impact but is also crucial in empirical research owing to its direct influence on the statistical analysis of the measurement results. Reliability refers to the reproducibility of the measurement outcome and, in the classical setting, is defined as one minus the ratio between the error variance and the total variance. Frequently, reliability is estimated using the intraclass correlation coefficient based on two replicate measurements. In this paper, we explore how the definition of reliability can be generalized, keeping the spirit of the original concept, to the more realistic setting where repeated measurements are available. Based on four defining properties for the concept of reliability, we propose an uncountable family of reliability measures, which circumscribes the area in which reliability measures should be sought for. It is shown how different members assess different aspects of the problem. The methodology is motivated by and illustrated on data from a clinical study on schizophrenia.

Key words: Reliability, Longitudinal data, Clinical trials, Hierarchical models, Rating scales.

1 Introduction

Frequently, measurements in research and medical practice are based on rating scales, especially in fields like psychology and psychiatry. When using a rating scale, the study of its psychometric properties such as validity and reliability is of utmost importance.

Validity refers to the extent to which the instrument properly measures the underlying trait of interest. Reliability refers to the extent to which the measurement is reproducible, or the degree to which it is influenced by measurement error.

The properties of a statistical analysis of scale measurements depend directly on the scale's reliability. For example, the correlation between two variables is a direct function of the coefficient of reliability. In regression models, the reliability of a covariate influences the effect size. Sensitivity and specificity are affected when a scale with low reliability is used for classification or prediction, and there is a direct relationship between the reliability of the measurement and the power of a study. Additionally, the reliability of a measurement also defines an upper bound for the validity of this measurement (Fleiss 1986; Lachin 2004).

The calculation of a reliability coefficient arises from classical test theory (Lord and Novick 1968), where the outcome of a test for subjects $i = 1, \ldots, n$ is modeled as $Y_i = T_i + \varepsilon_i$, where Y_i represents the observed variable, T_i is the true score and ε_i the corresponding measurement error. One rarely thinks of T_i as an actual true score; rather it is defined as the expected value of Y_i if the subject were re-measured an infinite number of times. It is assumed that the measurement errors are mutually uncorrelated, as well as independent of the true scores. Given these assumptions, we have $\mathrm{Var}(Y_i) = \mathrm{Var}(T_i) + \mathrm{Var}(\varepsilon_i)$, and the reliability of a measuring instrument can be defined as the ratio of the true score variance to the observed score variance:

$$R = \frac{\mathrm{Var}(T_i)}{\mathrm{Var}(Y_i)} = \frac{\mathrm{Var}(T_i)}{\mathrm{Var}(T_i) + \mathrm{Var}(\varepsilon_i)}. \tag{1}$$

It can easily be shown that (1) equals the correlation between two measurements, assuming the so-called steady state condition, i.e., both measurements have equal means (true scores) and error variances. Therefore, reliability estimation is classically based on the correlation of two replicate measurements. However, in medical research, a steady state condition in patients is often doubtful and can flaw reliability research. Vangeneugden et al. (2004) have shown that this assumption can be relaxed, by using linear mixed models. The change in the condition of the patient can then be modeled within the fixed-effects structure and estimated simultaneously with the covariance parameters necessary for the calculation of the intraclass correlation (ICC) or a generalization thereof. Depending on the complexity of the model, these authors define reliability as a single correlation, a correlation that depends on the time lag between two measurements, or an entire correlation matrix for any pair of measurements. Laenen, Alonso, and Molenberghs (2006) also use linear mixed models in this setting. However, they approach reliability not from a correlation perspective but starting from its basic definition as "the ratio between the true score variance and the observed score variance". Further, they provide a single yet meaningful measure of reliability, the so-called R_T, which is independent of the structure of the model used to fit the data and hence facilitates interpretation and applicability.

In this paper, we position this measure R_T in a broader framework. We show that, starting from four defining properties, any measure of reliability should be sought for within a family of which all members fulfill this restricted set of criteria.

Section 2 describes the case study data. Section 3 elaborates the methodology for finding a measure for reliability, and Section 4 investigates the properties of some of such measures, based on simulations. Section 5 applies the methodology to the case study introduced in Section 2.

2 Case Study

The study is concerned with individual patient data from a randomized clinical trial, investigating the effect of risperidone as compared to an active control for the treatment of chronic schizophrenia. Schizophrenic patients suffer from both 'positive' and 'negative' symptoms. Positive symptoms generally imply occurrences beyond normal experience whereas negative symptoms bear the connotation of diminished experience.

A total of 453 patients were evaluated, using two different rating scales at baseline and after 1, 2, 4, 6, and 8 weeks, respectively. The Positive an Negative Syndrome Scale (PANSS) consists of 30 items and is highly useful in the assessment of schizophrenia (Kay, Fizbein, and Opler 1987). The Brief Psychiatric Rating Scale (BPRS) is an 18-item scale, essentially a shorter version of the PANSS.

3 Methodology

We start by briefly reviewing the linear mixed model on which the present approach is based. Thereafter, we recall the minimum set of defining properties, introduced by Laenen, Alonso, and Molenberghs (2006). Finally, we introduce a family of parameters fulfilling this set of properties and study some of its elements.

3.1 The Linear Mixed Model

A linear mixed-effects model allows repeated measurements to be modeled in terms of their means, variances and covariances within a normal distribution based framework (Laird and Ware 1982; Verbeke and Molenberghs 2000). Three components of variability can be distinguished. Random effects capture part of the variability coming from heterogeneity between individual subjects. A serial correlation component formalizes the fact that pairs of measurements with shorter time lags are generally more strongly correlated than measurements with larger time lags between them. A third component is the measurement error. A linear mixed-effects model can generally be written as

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon}_{(1)\boldsymbol{i}} + \boldsymbol{\varepsilon}_{(2)\boldsymbol{i}}, \tag{2}$$

where $\boldsymbol{Y}_i$ is the p_i dimensional vector of responses for subject i, $1 \leq i \leq n$ with n the number of subjects, and p_i the number of measurements for subject i. X_i and Z_i are fixed $(p_i \times q)$ and $(p_i \times r)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is the q-dimensional vector of fixed effects, $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$ is the r-dimensional vector containing the random effects, $\boldsymbol{\varepsilon}_{(2)\boldsymbol{i}} \sim N(\mathbf{0}, \tau^2 H_i)$ is a p_i-dimensional vector of components of serial correlation, and $\boldsymbol{\varepsilon}_{(1)\boldsymbol{i}} \sim N(\mathbf{0}, \Sigma_{Ri})$ is a p_i-dimensional vector of residual errors. Additionally, D is a general $(r \times r)$ covariance matrix, H_i is a $(p_i \times p_i)$ correlation matrix, τ^2 is a variance parameter, and Σ_{Ri} is a $(p_i \times p_i)$ covariance matrix. Note that H_i and Σ_{Ri} depend on i only through their dimension p_i, i.e., the set of unknown parameters will not depend upon i. The random terms $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_{(1)1}, \ldots, \boldsymbol{\varepsilon}_{(1)N}, \boldsymbol{\varepsilon}_{(2)1}, \ldots, \boldsymbol{\varepsilon}_{(2)N}$ are assumed to be independent. The implied marginal model is:

$$\mathbf{Y_i} \sim N(X_i\boldsymbol{\beta}, V_i),$$

where $V_i = \Sigma_{D_i} + \Sigma_i$, $\Sigma_{D_i} = Z_i D Z_i'$, and $\Sigma_i = \tau^2 H_i + \Sigma_{Ri}$.

3.2 Properties of a Measure of Reliability

Based on the concept of reliability proposed in the early literature (Lord and Novick 1968), Laenen, Alonso, and Molenberghs (2006) asserted that any meaningful measure of reliability R should satisfy: (i) $0 \leq R \leq 1$, (ii) $R = 0$ if and only if there is only measurement error: $V_i = \Sigma_i$, (iii) $R = 1$ if and only if there is no measurement error: $\Sigma_i = 0$, and (iv) in the cross-sectional setting the classical expression for reliability (1) is recovered. Further, these authors proposed the following parameter, for quantifying reliability, that satisfies this set of properties:

$$R_T = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{tr}(\Sigma_i)}{\mathrm{tr}(V_i)}.$$

For a single-trial with a balanced design it simplifies to:

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

Note that the variability of the repeated measurements on the scale is summarized by the trace of its variance-covariance matrix. In a similar way, the error variabilities are summarized by the trace of the variance-covariance matrix associated with the error vectors $\boldsymbol{\varepsilon}_{(1)\boldsymbol{i}}$ and $\boldsymbol{\varepsilon}_{(2)\boldsymbol{i}}$.

In the next section, we elaborate on the reliability concept in this general setting, and propose a family of which all members satisfy the four properties introduced above. In doing so we embed the measure R_T in a broader framework. Actually, it will be shown that R_T is merely a special member of this general family.

3.3 A Family of Parameters for Reliability

Alonso, Laenen, and Molenberghs (2004) introduced a family of parameters to evaluate criterion validity of psychiatric symptom scales, based on canonical correlations. In the evaluation of criterion validity, a new scale is compared to a criterion scale, with known performance. In this setting, canonical correlations are a useful tool to quantify the amount of information shared between both instruments. In the context of reliability, we study the reproducibility of a single scale, which implies that canonical correlations are no longer applicable. Nevertheless, we will show that the role played by canonical correlations in the validity research, is in the reliability context assumed by the relative eigenvalues associated with specific variance-covariance matrices. Let us start by introducing the following result.

Theorem 1 *Given the function* $q(\lambda) = |\Sigma - \lambda V|$, *if model (2) holds then: (i) all roots of* $q(\lambda) = 0$ *are real, and (ii) if* λ_j *is a root of* $q(\lambda) = 0$ *then* $0 \leq \lambda_j \leq 1$.

An outline of the proof can be found in Appendix A. Based on this theorem we define the family:

$$\Omega = \left\{ \theta : \theta = \sum_{j=1}^{p} w_j \rho_j^2 \quad \text{with} \quad w_j > 0 \quad \sum_{j=1}^{p} w_j = 1 \right\}. \tag{4}$$

The elements w_j are weights assigned to the parameters ρ_j^2, where $\rho_j^2 = 1 - \lambda_j$ with λ_j the roots of the equation $q(\lambda) = 0$, or equivalently, the eigenvalues of the matrix ΣV^{-1}. Further, it is easy to prove, using Theorem 1, that all elements of Ω satisfy the properties (i)–(iv), given in Section 3.2.

This family is structurally similar to the family introduced by Alonso et al. (2004) in the validity framework. The main difference is that here the ρ_j^2 are not the canonical correlations associated with the new and criterion scales but rather the relative eigenvalues associated with the total and error variance covariance matrices. Actually, family (4) is closer to the eigenvalue-based, root statistics used in multivariate analysis of variance models (Pillai's Trace, Wilks's Lambda, Hotelling-Lawley's Trace) than to the idea of canonical correlations.

Note also that, even though the Ω family is uncountable, it clearly delineates our search for reliability measures. We will now study some specific, important members.

3.3.1 R_T as Member of the Ω Family

From Theorem 2 in Appendix A, we know that there exists a non-singular matrix Q so that $\Sigma = (Q')^{-1} D_0 Q^{-1}$ and $V = (Q')^{-1} Q^{-1}$, where D_0 is a diagonal matrix whose diagonal elements are the roots of the polynomial equation $q(\lambda) = 0$. Plugging the previous expression into (3), we obtain

$$R_T = 1 - \frac{\mathrm{tr}((Q')^{-1} D_0 Q^{-1})}{\mathrm{tr}((Q')^{-1} Q^{-1})} = 1 - \frac{\mathrm{tr}(Q^{-1}(Q')^{-1} D_0)}{\mathrm{tr}(Q^{-1}(Q')^{-1})}.$$

Further, if we call $S = Q^{-1}(Q')^{-1} = (Q^{-1})(Q^{-1})'$, we have:

$$R_T = 1 - \frac{\mathrm{tr}(SD_0)}{\mathrm{tr}(S)} = 1 - \mathrm{tr}\left(\frac{S}{\mathrm{tr}(S)} D_0\right) = 1 - \sum_{j=1}^{p} w_j \lambda_j,$$

with $w_j = \dfrac{s_{jj}}{\mathrm{tr}(S)}$ and s_{jj} the jth element in the diagonal of S. Note that $s_{jj} \geq 0$ for all j and that

$$\sum_{j=1}^{p} w_j = \sum_{i=j}^{p} \frac{s_{jj}}{\mathrm{tr}(S)} = \frac{1}{\mathrm{tr}(S)} \sum_{j=1}^{p} s_{jj} = 1.$$

The rationale of these derivations is that R_T is an element of Ω, since

$$R_T = \sum_{j=1}^{p} w_j(1 - \lambda_j) = \sum_{j=1}^{p} w_j \rho_j^2 \quad \text{with} \quad w_j > 0 \quad \text{and} \quad \sum_{j=1}^{p} w_j = 1.$$

3.3.2 Other Members of the Ω Family

The uncountable nature of the Ω family implies that the choice of some special members to be scrutinized further must be based on pragmatic considerations. Retaining R_T is evident. Another intuitive choice is to set all weights equal to $w_j = 1/p$. We then have that

$$R_p \;=\; \sum_{j=1}^{p} \frac{1}{p} \rho_j^2 = \sum_{j=1}^{p} \frac{1}{p}(1 - \lambda_j) = 1 - \frac{1}{p} \sum_{j=1}^{p} \lambda_j = 1 - \frac{1}{p} \mathrm{tr}(\Sigma V^{-1}).$$

It would be appealing to consider the elements of Ω corresponding to the largest and smallest eigenvalue of ΣV^{-1}, i.e., $\widetilde{\theta}_{\max} = \rho^2_{(p)}$ and $\widetilde{\theta}_{\min} = \rho^2_{(1)}$, where $\rho^2_{(j)}$ is the jth largest eigenvalue. However, the restrictions placed on the weights ($w_j > 0$) make $\widetilde{\theta}_{\max}$ and $\widetilde{\theta}_{\min}$ invalid choices. Nevertheless, we could define $\theta_{\max}$ and $\theta_{\min}$ in the following alternative way:

$$\begin{aligned} \theta_{\max} &= \sum_{j=1}^{p} w_j \rho_j^2 \quad \text{with} \quad w_p >> w_j \quad \text{for } j \neq p, \\ \theta_{\min} &= \sum_{j=1}^{p} w_j \rho_j^2 \quad \text{with} \quad w_1 \gg w_j \quad \text{for } j \neq 1. \end{aligned}$$

Note that, if the weights w_j are carefully chosen, we can be rather confident that for any arbitrary element of Ω: $\theta_{\min} \leq \theta \leq \theta_{\max}$. Indeed, for any given scale and independently of the element of Ω that one may use in the analysis, the reliability of the instrument will lie always in the interval $[\theta_{\min},\theta_{\max}]$. In the following section, we will investigate the performance of the previously defined elements of Ω via simulation.

4 Simulation Study

4.1 Design of the Simulation Study

Let us consider 12 different simulation settings. In a first stage, the data are generated based on the following linear mixed model with random intercept:

$$Y_{ij} \;=\; \beta_0 + \beta_1 t_j + \beta_2 Z_i + b_i + \varepsilon_{ij},$$

where Y_{ij} refers to an observation for subject i at time t_j, and Z_i is the treatment indicator variable. Further, $b_i \sim N(0, \sigma_b^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2 I)$, with $\sigma_b^2 = 300$. The error variability takes values $\sigma^2 = 30$, 300, or 3000, and the sample size was set to either $n = 50$ or 150. These choices for σ_b^2 and σ^2 allow us to study the performance of the elements of the Ω family when the error variance is 9%, 50%, and 90% of the total variance, respectively. These settings correspond to high, medium, and low reliability.

In a second stage, data are generated based on a linear mixed model with random intercept and random slope for time:

$$Y_{ij} \;=\; \beta_0 + \beta_1 t_j + \beta_2 Z_i + b_{1i} + b_{2i} t_j + \varepsilon_{ij}$$

where $(b_{1i}, b_{2i})' \sim N(0, D)$, $\varepsilon_{ij} \sim N(0, \sigma^2 I)$, and

$$D = \begin{pmatrix} 300 & -1 \\ -1 & 5 \end{pmatrix}.$$

The same choices for σ^2 and n are made.

In both stages, the mean parameters are fixed at $\beta_0 = 85$, $\beta_1 = 2.5$, $\beta_2 = 3$ to generate the data. These values are based on the results obtained when the previous models were fitted using the case study data. We consider $p = 5$ time points in all scenarios and, for each setting, 250 data sets are simulated.

The parameters $\theta_{\min}$ and $\theta_{\max}$ are specified in the following way:
$\theta_{\min} = \sum_{j=1}^{p} w_j \rho^2_{(j)}$ where $w_j = 0.999$ for $j = 1$ and $w_j = \dfrac{0.001}{p-1}$ otherwise, and
$\theta_{\max} = \sum_{j=1}^{p} w_j \rho^2_{(j)}$ where $w_j = 0.999$ for $j = p$ and $w_j = \dfrac{0.001}{p-1}$ otherwise.

Using restricted maximum likelihood, we calculate the point estimates, the confidence intervals, and the coverage percentage (CP) of the confidence intervals. A confidence interval, based on the delta method, can be derived for all members of the Ω family, assuming the weights are known constants. Details on the derivation of these confidence intervals can be found in Appendix B. This assumption is not fulfilled for R_T. Confidence intervals for R_T are calculated as described in Laenen, Alonso, and Molenberghs (2006). To avoid that confidence limits take values beyond the $[0, 1]$ range, a logit transformation is applied.

4.2 Results of the Simulation Study

Point estimates, true values, average confidence intervals, and coverage percentages are given in Tables 1–3 for R_T, R_p, and $\theta_{\max}$, respectively, showing that accurate point estimates for all parameters can be obtained with a relative small sample size of 50 patients. A larger sample size, as expected, produces narrower confidence intervals. Furthermore, the coverage probabilities for all the asymptotic confidence intervals are generally around the pre-specified 95% level.

Considering the values of the point estimates, the measure R_T produces results in line with intuition. We obtain values close to 1 when the error variance is small compared to the model variance, we settle for values in the neighborhood of 0.50 in case the error variance and model variance are of a similar magnitude, and values are close to 0 when error variances are large.

Interestingly, $\theta_{\max}$ takes higher values in all the settings. With 50% of the variability originating from error, it takes values above 0.80. To gain intuition about this behavior let us recall that $\theta_{\max} \approx \rho^2_{(p)}$ and consider the random intercept model, where $\Sigma = \sigma^2 I$ and $V = \sigma^2_b J + \sigma^2 I$. It can be shown that in this scenario:

$$\rho^2_{(p)} = \frac{p\sigma^2_b}{p\sigma^2_b + \sigma^2}. \tag{5}$$

From (5) it can be seen that this measure increases with the number of time points. Actually, $\theta_{\max}$ seems to quantify the reliability of the entire series of measurements, in contrast to R_T, which gives an average reliability. Note that, from this perspective, $\theta_{\max}$ is in total agreement with clinical intuition: the longer a patient is followed, the more reliable our conclusions about that patient will be. Another important implication of (5) is that we can obtain reliable information from an instrument that produces a lot of measurement error, as long as we take a sufficiently high number of measurements.

Looking at the third measure, R_p, we observe again a totally different pattern. This measure gives generally low values. Even when the error variance is small compared to the model variance, R_p reaches values far below 1. Studying R_p under the random intercept model, it can easily be shown that, if $\sigma^2 \neq 0$, $R_p = \sigma^2_b/(p\sigma^2_b + \sigma^2)$. Note that, unlike $\theta_{\max}$, R_p is a decreasing function of the number of time points. The expression further shows that, even when the error variance is very small, the measure R_p can never exceed $1/p$. Additionally, R_p is not a continuous function of σ^2 for $\sigma^2 = 0$. Indeed,

$\lim_{\sigma^2 \to 0} R_p = \frac{1}{p} \neq 1 = R_p(\sigma^2 = 0)$. In spite of their differences, R_p and θ_{max} are functionally related. It can be shown that $R_p = \frac{\rho^2_{(p)}}{p} \approx \frac{\theta_{\text{max}}}{p}$. R_p can therefore be interpreted as the average contribution per measurement to the total reliability of the whole sequence. Where large values of θ_{max} can, in principle, always be obtained by increasing the number of repeated measurements, R_p is more a measure of efficiency. It shows us at what 'cost' we obtain a large θ_{max}.

The parameter θ_{min} gives the lowest estimates of all members of the Ω family. The simulation study shows that the measure takes values close to 0 under all circumstances considered. The informative value of this measure is therefore very limited.

Comparing the different parameters in the present simulation study has made clear that different measures can lead to rather divergent messages. While the R_T should be interpreted as the average reliability, θ_{max} gives the reliability of the entire sequence of measurements. Further, R_p gives the average contribution to θ_{max} at each time point, and can be seen as a measure of efficiency.

Which measure is preferred will depend on the circumstances of the research and the scientific question one wants to address. The R_T is closest to the intuition behind the classical concept of reliability and might therefore be preferred in some settings. However, other members of Ω might bring valuable information as well. A parallel can be drawn with the concept of distance in mathematics that, based on no more than three properties, has several operationalizations. Arguably, in some cases, it will be of interest to consider a few measures simultaneously. In the next section, we will further explore the performance of these proposals using a real case study.

5 Analysis of the Case Study

In this section, we will study the parameters introduced for the schizophrenia data, introduced in Section 2. All reliability estimates are obtained from the estimated covariance parameters resulting from fitting a linear mixed model to the data. A model building step is therefore crucial to find the best fitting model for the data at hand. The model selection was based on the AIC and restricted maximum likelihood was used for parameter estimation (Verbeke and Molenberghs 2000). For all three scales, the final model has the general form:

$$Y_{ij} = \mu_{ij} + b_{i0} + b_{i1}t_j + \varepsilon_{ij}$$

where Y_{ij} denotes the score (either PANSS or BPRS) for subject i at time point t_j, μ_{ij} summarizes the fixed-effects structure, encompassing treatment, categorical time, and their interaction, $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$ with D a 2×2 unstructured variance-covariance matrix, $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma)$, and $\Sigma = \text{diag}(\sigma_j^2)$.

Table 4 presents the reliability estimates for the different parameters and for both scales, together with the 95% confidence interval. Clearly, both scales have high average reliabilities characterized by large estimates of R_T and the value of $\theta_{\max}$ indicates that highly reliable results can be achieved with six measurements per subject.

It has been shown earlier (Laenen, Alonso, and Molenberghs 2006) that R_T is slightly higher for PANSS than for BPRS. However the two confidence intervals overlap. Also for R_p and $\theta_{\max}$ the point estimates for PANSS and BPRS are almost identical.

PANSS, with 30 items, is conceived as a more complete extension of BPRS, having 18 items. Nevertheless, the previous results illustrate that this additional complexity does not bring a considerable gain in reliability. Similar results have been found by Alonso et al. (2002) when studying criterion validity. These authors also obtained very similar values of trial-level validity and individual-level validity for these two scales. Finally, we should not that the choice between different instruments usually is not only based on statistical aspects and clinical considerations must be taken into account as well.

6 Discussion

The reliability of a measurement is not only relevant from a clinical point of view but it directly affects the results of a statistical analysis based upon it. Therefore, reliability is a concept of the utmost importance in the evaluation of a rating scale to be used in clinical trials.

A test-retest reliability study essentially consists of taking two replicate measurements. However, in clinical studies it is common practice to measure a patient's condition repeatedly over time. It is therefore good practice to take advantage of the available longitudinal data when estimating test-retest reliability. Vangeneugden et al. (2004) already showed how linear mixed models can correct for evolutions in the patient's condition while estimating reliability. These authors discussed merits and advantages of such an approach. Obviously, the price to pay is the need to make modeling assumptions. Laenen, Alonso, and Molenberghs (2006) introduced, with the parameter R_T, an alternative for the intraclass correlation coefficient as a measure of reliability, based on the original definition of reliability as the ratio of the true score variance and the total variance. The same authors introduce a basic set of properties that should be fulfilled by any parameter for reliability.

In this paper, we have defined an entire family of which all members satisfy these four properties. I doing so, we have established that any measure of reliability should be built from the relative eigenvalues related to error and total variance-covariance matrices. Different weights assigned to these eigenvalues lead to different members of the family. A few key members of this family were scrutinized further, the R_T being one of them.

A simulation study demonstrates that there are clear and important differences in the meaning of the different members. Since different measures give different messages, they cannot be compared on objective criteria when selecting one as the 'best' measure. The measure to be used will depend on the circumstances of the study. It might be of interest to consider more than one measure simultaneously. In a similar fashion, a family of parameters has been introduced to evaluate the criterion validity of psychiatric symptom scales (Alonso et al. 2004). It is appealing to see that the two most important psychometric characteristics of a scale can be investigated using similar methodologies.

Appendix A

Proof of Theorem 1. In the proof we will use the following theorem (Graybill 1983).

Theorem 2: Let A and B be symmetric matrices of order $p \times p$, then:

1. If A is positive definite, then there exists a nonsingular matrix Q such that $Q'AQ = I$ and $Q'BQ = D_0$, where D_0 is a diagonal matrix whose diagonal elements are the roots of the polynomial equation $q(\lambda) = |B - \lambda A| = 0$.
2. If A and B are positive semidefinite, then there exists a nonsingular matrix P such that $P'AP = D_1$ and $P'BP = D_2$, where D_1 and D_2 are diagonal matrices (Newcomb 1960).

Applying now the first part of Theorem 2 with $A = V$ and $B = \Sigma$ we obtain that there exists a nonsingular matrix Q such that:

$$Q'VQ = I \quad \Rightarrow \quad V = (Q')^{-1}(Q)^{-1} \tag{6}$$

$$Q'\Sigma Q = D_0 \quad \Rightarrow \quad \Sigma = (Q')^{-1}D_0(Q)^{-1} \tag{7}$$

with D_0 diagonal and the elements of the diagonal are the roots of the equation $q(\lambda) = |\Sigma - \lambda V| = 0$. From (6)- (7) it follows that: $\Sigma V^{-1} = (Q')^{-1}D_0(Q)^{-1}(Q)(Q') = (Q')^{-1}D_0Q'$, so that

$$\mathrm{tr}(\Sigma V^{-1}) = \mathrm{tr}[(Q')^{-1}D_0Q'] = \mathrm{tr}(D_0) = \sum_{j=1}^{p}\lambda_j \quad \text{where} \quad D_0 = \mathrm{diag}(\lambda_j)_{j=\overline{1,p}}.$$

Further we have from the second part of Theorem 2 that $P'VP = D_1$ and $P'\Sigma P = D_2$, where P is nonsingular and D_1, D_2 are diagonals. It then follows that $V = (P')^{-1}D_1P^{-1}$ and $\Sigma = (P')^{-1}D_2P^{-1}$, and thus

$$\begin{aligned} |\Sigma - \lambda V| &= |(P')^{-1}D_2P^{-1} - (P')^{-1}(\lambda D_1)P^{-1}| \quad = \quad |P'|^{-1}|P|^{-1} \quad |D_2 - \lambda D_1| \\ &= \frac{|D_2 - \lambda D_1|}{|P|^2}. \end{aligned}$$

In case that $q(\lambda) = \dfrac{D_2 - \lambda D_1}{|P|^2} = 0$ then $\lambda_j = \dfrac{d_{2j}}{d_{1j}}$ with $D_1 = \mathrm{diag}(d_{1j})$ and $D_2 = \mathrm{diag}(d_{2j}), \quad j = 1, ..., p$. This proves the first part of Theorem 1.

We will now show that $0 \leq \lambda_j \leq 1 \quad \forall j$. It is not difficult to see that $d_{1j} \geq 0$ and $d_{2j} \geq 0 \quad \forall j$, and therefore $\lambda_j \geq 0 \quad \forall j$. Additionally we have:
$V = \Sigma_D + \Sigma \quad \Leftrightarrow \quad P'VP = P'\Sigma_D P + P'\Sigma P$
so that $D_1 - D_2 = P'\Sigma_D P = RDR'$ with $R = P'Z$,
so $d_{j1} - d_{2j} = r_jDr_j' \geq 0$ where r_j is the jth row of R, and thus $0 \leq \lambda_j = \dfrac{d_{2j}}{d_{1j}} \leq 1$, completing the proof. □

Appendix B

Let θ be any member of Ω as defined in (4). Let ψ be the vector of the covariance parameters of a linear mixed-effects model. We know from ML theory that $\widehat{\psi} \sim N(\psi, \Sigma_P)$

where Σ_P is the variance-covariance matrix of $\widehat{\psi}$. Applying now the Delta method to $\widehat{\theta}$ we get: $\widehat{\theta} \sim N(\theta, \Delta\Sigma_P\Delta')$ where $\mathbf{\Delta} = \dfrac{\partial\theta}{\partial\psi}$. A $(1-\alpha)\%$ confidence interval for θ can then be given by

$$\left[\widehat{\theta} \pm z_{1-\frac{\alpha}{2}}\sqrt{\mathbf{\Delta}\Sigma_P\mathbf{\Delta'}}\right].$$

To avoid confidence limits to go beyond the $[0,1]$ range, a logit transformation is applied, with $l(\theta) = \log\left(\dfrac{\theta(\psi)}{1-\theta(\psi)}\right)$. A restricted $(1-\alpha)\%$ confidence interval for θ is then given by

$$\left[\frac{e^{l_1}}{1+e^{l_1}}, \frac{e^{l_2}}{1+e^{l_2}}\right],$$

with l_1 the lower limit and l_2 the upper limit of the confidence interval

$$\left[l(\widehat{\theta}) \pm \frac{z_{1-\frac{\alpha}{2}}}{\theta(1-\theta)}\sqrt{\mathbf{\Delta}\Sigma_P\mathbf{\Delta'}}\right].$$

Additionally, let us note that $\theta = 1 - \Sigma_j w_j \lambda_j$ where λ_j are the eigenvalues of ΣV^{-1}. This implies that there exists a nonsingular matrix P so that $\Lambda = P^{-1}\Sigma V^{-1}P$ with $\Lambda = \text{diag}(\lambda_j)$. On the other hand,

$$\begin{aligned} \theta &= 1 - \text{tr}(W\Lambda) \quad \text{where} \quad W = \text{diag}(w_j) \\ &= 1 - \text{tr}(WP^{-1}\Sigma V^{-1}P) \\ \theta &= 1 - \text{tr}(Q\Sigma V^{-1}) \quad \text{where} \quad Q = PWP^{-1}. \end{aligned}$$

Thus:

$$\frac{\partial\theta}{\partial z} = -\frac{\partial}{\partial z}\text{tr}(Q\Sigma V^{-1}) = -\text{tr}\left(\frac{\partial Q}{\partial z}\Sigma V^{-1}\right) - \text{tr}\left(Q\frac{\partial}{\partial z}(\Sigma V^{-1})\right) = -\text{tr}\left(Q\frac{\partial}{\partial z}(\Sigma V^{-1})\right).$$

From the product rule of differential calculus for matrices we get:

$$\frac{\partial}{\partial z}(\Sigma V^{-1}) = \frac{\partial\Sigma}{\partial z}V^{-1} + \Sigma\frac{\partial V^{-1}}{\partial z}$$

and therefore

$$\begin{aligned} \frac{\partial}{\partial z}(\Sigma V^{-1}) &= \left(\frac{\partial\Sigma}{\partial z} - \Sigma V^{-1}\frac{\partial V}{\partial z}\right)V^{-1} \\ \Rightarrow \frac{\partial\theta}{\partial z} &= \text{tr}\left[Q\left(\Sigma V^{-1}\frac{\partial V}{\partial z} - \frac{\partial\Sigma}{\partial z}\right)V^{-1}\right] \end{aligned}$$

where:

$$\begin{aligned}\frac{\partial \Sigma}{\partial z} &= \frac{\partial \tau^2}{\partial z}H + \tau^2 \frac{\partial H}{\partial z} + \frac{\partial \Sigma_R}{\partial z} \\ \frac{\partial V}{\partial z} &= Z\frac{\partial D}{\partial z}Z' + \frac{\partial \Sigma}{\partial z}\end{aligned}$$

So in general:

$$\begin{aligned}\frac{\partial \theta}{\partial z} &= \mathrm{tr}\left[V^{-1}Q\left(\Sigma V^{-1}\frac{\partial V}{\partial z} - \frac{\partial \Sigma}{\partial z}\right)\right] \\ Q &= PWP^{-1}, \quad W = \mathrm{diag}(w_j) \\ \frac{\partial \Sigma}{\partial z} &= \frac{\partial \tau^2}{\partial z}H + \tau^2 \frac{\partial H}{\partial z} + \frac{\partial \Sigma_R}{\partial z} \\ \frac{\partial V}{\partial z} &= Z\frac{\partial D}{\partial z}Z' + \frac{\partial \Sigma}{\partial z}\end{aligned}$$

And in particular:

1. $\dfrac{\partial \theta}{\partial d_D} = \mathrm{tr}\left[V^{-1}Q\Sigma V^{-1}\left(Z\dfrac{\partial D}{\partial d_D}Z'\right)\right]$
2. $\dfrac{\partial \theta}{\partial \tau^2} = \mathrm{tr}\left[V^{-1}Q\left(\Sigma V^{-1} - I\right)H\right]$
3. $\dfrac{\partial \theta}{\partial d_H} = \mathrm{tr}\left[V^{-1}Q\left(\Sigma V^{-1} - I\right)\tau^2\dfrac{\partial H}{\partial d_H}\right]$
4. $\dfrac{\partial \theta}{\partial d_{\Sigma_R}} = \mathrm{tr}\left[V^{-1}Q\left(\Sigma V^{-1} - I\right)\dfrac{\partial \Sigma_R}{\partial d_{\Sigma_R}}\right]$

Acknowledgments

The authors gratefully acknowledge support from Belgian IUAP/PAI network “Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data”.

References

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2002). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical Statistics* **12**, 161–179.

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. *Biometrics* **60**, 845–853.

Fleiss, J.L (1986). *Design and Analysis of Clinical Experiments.* New York: John Wiley.

Graybill, F.A. (1983). *Matrices with Applications in Statistics, 2nd ed.* Belmont, California: Wadsworth.

Kay, S.R., Fiszbein, A., and Opler, L.A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* **13**, 261–267.

Lachin, J.M. (2004). The role of measurement reliability in clinical trials. *Clinical Trials* **1**, 553–566.

Laenen, A., Alonso, A., and Molenberghs, G. (2006). A measure for the reliability of a rating scale based on longitudinal clinical trial data. *Submitted for publication.*

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

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

Newcomb, R.W. (1960). On the simultaneous digonalization of two semi-definite matrices. *Quarterly in Applied Mathematics* **19**, 144–146.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D. and Molenberghs G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, **25**, 13–30.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Table 1: *Simulation Results for R_T: true values, point estimates, average confidence intervals and coverage probabilities.*

		Random intercept model				Random intercept + slope model			
σ^2	n	true	est.	95% CI	CP	true	est.	95% CI	CP
30	50	0.91	0.90	[0.86; 0.93]	93	0.93	0.93	[0.90; 0.95]	94
30	150	0.91	0.91	[0.89; 0.93]	96	0.93	0.93	[0.92; 0.94]	91
300	50	0.50	0.50	[0.38; 0.61]	94	0.58	0.57	[0.47; 0.68]	95
300	150	0.50	0.50	[0.43; 0.57]	96	0.58	0.58	[0.51; 0.64]	93
3000	50	0.09	0.09	[0.04; 0.34]	90	0.12	0.14	[0.06; 0.33]	86
3000	150	0.09	0.09	[0.05; 0.18]	97	0.12	0.13	[0.07; 0.22]	94

Table 2: *Simulation Results for R_p: true values, point estimates, average confidence intervals and coverage probabilities.*

		Random intercept model				Random intercept + slope model			
σ^2	n	true	est.	95% CI	CP	true	est.	95% CI	CP
30	50	0.20	0.20	[0.19; 0.20]	95	0.36	0.36	[0.35; 0.38]	95
30	150	0.20	0.20	[0.20; 0.20]	97	0.36	0.36	[0.36; 0.37]	95
300	50	0.17	0.17	[0.15; 0.18]	96	0.24	0.23	[0.17; 0.30]	92
300	150	0.17	0.17	[0.16; 0.18]	98	0.24	0.24	[0.20; 0.28]	96
3000	50	0.07	0.06	[0.03; 0.22]	88	0.09	0.09	[0.04; 0.24]	92
3000	150	0.07	0.07	[0.04; 0.12]	96	0.09	0.09	[0.05; 0.16]	95

Table 3: *Simulation Results for $\theta_{\max}$: true values, point estimates, average confidence intervals and coverage probabilities.*

		Random intercept model				Random intercept + slope model			
σ^2	n	true	est.	95% CI	CP	true	est.	95% CI	CP
30	50	0.98	0.98	[0.97; 0.99]	96	0.98	0.98	[0.97; 0.99]	97
30	150	0.98	0.98	[0.97; 0.98]	98	0.98	0.98	[0.98; 0.99]	96
300	50	0.83	0.83	[0.74; 0.89]	96	0.86	0.86	[0.78; 0.91]	97
300	150	0.83	0.83	[0.78; 0.87]	97	0.86	0.86	[0.82; 0.89]	97
3000	50	0.33	0.32	[0.14; 0.70]	91	0.39	0.41	[0.21; 0.69]	93
3000	150	0.33	0.33	[0.19; 0.53]	98	0.39	0.40	[0.26; 0.56]	97

Table 4: *Schizophrenia Study: Three reliability parameters, applied to two scales: estimates and 95% confidence intervals.*

parameter	PANSS	BPRS
R_T	0.846 [0.825; 0.865]	0.821 [0.797; 0.842]
R_p	0.285 [0.277; 0.294]	0.280 [0.271; 0.289]
$\theta_{\max}$	0.976 [0.970; 0.980]	0.968 [0.962; 0.973]