0225

Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements

A. Laenen, H. Geys, T. Vangeneugden, and G. Molenbergh

PSYCHOMETRIKA—VOL. 1, NO. 1, 1–0
MARCH 1936

APPLYING LINEAR MIXED MODELS TO ESTIMATE RELIABILITY IN CLINICAL TRIAL DATA WITH REPEATED MEASUREMENTS

Annouschka Laenen and Helena Geys

limburgs universitair centrum

Tony Vangeneugden

janssen research foundation

Geert Molenberghs

limburgs universitair centrum

The first author was supported in part by the Minimal Psychiatric Data Registry of the Belgian Federal Ministry of Public Health and Social Affairs. The second author was supported by the Institute for the Promotion of Innovation by Science and Technology (IWT) in Flanders, Belgium. In addition, the authors are grateful to the Janssen Research Foundation for kind permission to use their data.

Requests for reprints should be sent to Geert Molenberghs, Limburgs Universitair Centrum, Center for Statistics, Universitaire Campus, B-3590 Diepenbeek, Belgium.

Abstract

Repeated measures are exploited to study reliability in the context of psychiatric health sciences. It is shown how test-retest reliability can be derived using linear mixed models when the scale is continuous or quasi-continuous and using generalized linear mixed models when the scale is dichotomous. The advantage of this approach is that the full modeling power of mixed models can be used. Repeated measures with a different mean structure can be used to usefully study reliablity, correction for covariate effects is possible, and a complicated variance-covariance structure between measurements is allowed. In case the variance structure reduces to a random intercept (compound symmetry), classical methods are recovered. With more complex variance structures (e.g., including random slopes of time and/or serial correlation), time-dependent reliability functions are obtained. The methodology is motivated by and applied to data from five double-blind randomized clinical trials, comparing the effects of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia.

Key words: Reliability; Linear Mixed Model; Repeated Measurements; Psychiatry; Rating Scale.

1. Introduction

Measurement in psychiatric health sciences seldomly relies on objective criteria. The subjective nature of the information to be gathered renders the development of scales in this area far from easy. One difficulty is that external conditions can influence the response that is given on such a scale, like for example the person who administers the test or the time of measurement. Therefore, whenever a mental health measurement scale is developed, its psychometric properties are typically checked. Two important properties in this respect are *reliability* and *validity*.

Reliability reflects the amount of error inherent in any measurement and hence, in a general sense, how replication of the administration would give a different result (Streiner and Norman 1995). The validity of an instrument is defined as the degree to which it measures what it purports to measure. In this paper we will concentrate on the former and propose a flexible way to calculate the reliability of psychiatric symptom scales, measured repeatedly over time.

Two main classical approaches are *internal consistency* and *reproducibility* of the instrument. Internal consistency is the extent to which individual items are consistent with each other and reflect a single underlying construct. In operational terms, internal consistency represents the average of the correlations among all items in the instrument. Several measures that are often used to measure internal consistency are: Cronbach's alpha coefficient (Cronbach 1951), Kuder-Richardson (Kuder and Richardson 1953) and factorial analyses. However, Streiner and Norman (1995) argue that these measures of internal consistency as measures for the reliability should be interpreted with great caution. They are based on performance observed in a single sitting, but there are many sources of variance which occur from day to day or between observers which do not enter into the calculation.

Reproducibility of an instrument is measured as either *inter-observer reliability* or *intra-observer reliability*. The former is only applicable for interviewer-administered questionnaires and is the degree to which a measurement yields stable scores when administered by different interviewers, rating the same subjects. The calculation of an intraclass correlation coefficient (Deyo, Dierh and Patrick 1991) is one of the most commonly used methods. Intra-observer reliability or test-retest reliability, is the degree to which a measure yields stable scores at different points in time for subjects who are assumed not to have changed on the domains being assessed. Also here, the calculation of the intraclass correlation coefficient is one of the most commonly used methods. A major difficulty in this approach is to select the appropriate time interval. If it is too long, the assumption of no change is not realistic. If it is too short, raters might remember the previous answer, thereby influencing the ratings.

Reliability is not a fixed property of a certain instrument. Measures of reliability depend on the population that has been evaluated using this instrument, they can as well depend on the raters and furthermore the measures can change over time. The reliability is typically higher if the instrument is used in an heterogenous population, and the measure can be increased by training or practice of the raters.

Wiley and Wiley (1970) presented a method to disentangle the effects of lack of stability (change over time) from the effects of poor instrument precision in repeated measurements. In their model it is postulated that a subject's true score at the second testing is linearly related to, but not necessarily the same as, the true score at the first testing. Applying these assumptions in the calculations of reliability, they come to a time-function of reliability. Also Tisak and Tisak (1996) present a method to incorporate the aspect of time in the calculation of the reliability and validity of an instrument.

Dunn (1989) describes how the reliability of an instrument can be derived using analysis of variance techniques. Furthermore, he extends this technique to mixed models that incorporate random effects such as, for example, a rater effect. Also in generalizability theory (Chronbach et al. 1963), the emphasis is on multiple sources of measurement error. An important goal is to recognize different sources of error variance and to measure and reduce the influence of these sources on the measurement.

Section 2 introduces data from a meta-analysis of five clinical trials comparing antipsychotic agents for the treatment of chronic schizophrenia. Section 3 reviews the concept of reliability, introduces a new and flexible way to calculate the reliability of continuous measurement scales, measured repeatedly over time, and extends this approach to the case of binary rating scales. Section 4 applies the methods, introduced in sections 3.2 and 3.3 on the data described in section 2. Finally, section 5 contains some concluding remarks.

2. Motivating Study

In this section we introduce individual patient data from five double-blind randomized clinical trials, comparing the effects of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia. Schizophrenia has long been recognised as a heterogeneous disorder with patients suffering from both "negative" and "positive" symptoms. Negative symptoms are characterized by deficits in social functions such as, for example, poverty of speech, apathy and emotional withdrawal. Positive symptoms entail more florid symptoms such as delusions, hallucinations, and disorganized thinking, which are superimposed on the mental status (Kay, Fiszbein and Opler, 1987).

Several measures can be considered to assess a patient's global condition. The *Positive and Negative Syndrome Scale* (PANSS) (Kay, Opler and Lindenmayer 1988) consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia (Kay, Fiszbein and

Opler 1987). The Clinician's Global Impression (CGI) is generally accepted as a subjective clinical measure of change. Here, we will consider the CGI overall change versus baseline. This is a 7-grade scale used by the treating physician to characterize how well a subject has improved since baseline.

Since the label in most countries recommend that risperidone is most effective in schizophrenia at doses ranging from 4 to 6 mg/day, we include only patients in our analyses that received either these doses of risperidone or an active control like haloperidol, levomepromazine, perphenazine and zuclopenthixol. Depending on the trial, treatment was administered for a duration of 4 to 8 weeks. For example in the international trials (by Peuskens and the Risperidone Study Group 1995, Chounard, Jones and Remington 1993, Marder and Meibach 1994, and Hoyberg et al. 1993) patients received treatment for 8 weeks; in the study by Blin, Azorin and Bouhours (1996) patients received treatment for 4 weeks, while in the study by Huttunen et al. (1995) patients were treated over a period of 6 weeks.

3. Methodology

First, we give a general outline of the concept of reliability. Thereafter, we will introduce two model families that will further be used to approach this quantity in a longitudinal setting. Which of these two families is used, depends on the type of outcome, that can be continous or binary.

3.1. Reliability

In the classical test theory, the outcome of a test is modeled as

$$X = \tau + \varepsilon, \tag{1}$$

where X represents an observation or measurement, τ is the true score and ε the corresponding measurement error. It is assumed that the measurement errors are mutually uncorrelated as well as with the true scores. If this assumption is correct, we obtain

$$\mathrm{Var}(X) = \mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon). \tag{2}$$

The reliability of a measuring instrument is defined as the ratio of the true score variance to the observed score variance, i.e.,

$$R = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(X)} \tag{3}$$

or

$$R = \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon)}. \tag{4}$$

In the case of two parallel measurements, we have $X_1 = \tau + \varepsilon_1$ and $X_2 = \tau + \varepsilon_2$, with $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \mathrm{Var}(X)$ and $\mathrm{Var}(\varepsilon_1) = \mathrm{Var}(\varepsilon_2) = \mathrm{Var}(\varepsilon)$. Further, the covariance of the two measurements equals

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(\tau + \varepsilon_1, \tau + \varepsilon_2) = \mathrm{Var}(\tau) \tag{5}$$

and the correlation between the two measurements can be written as

$$\begin{aligned}\mathrm{Corr}(X_1, X_2) &= \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(\boldsymbol{X}_1)}\sqrt{\mathrm{Var}(\boldsymbol{X}_2)}} \\ &= \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon)} = R. \end{aligned} \tag{6}$$

The outcomes X_1 and X_2 can, for example, be two subscores of a test, in which case we are also talking about *split-halve reliability*. If the scores are two measurements of the same instrument, measured at different moments in time, then we are dealing with *test-retest reliability*. When the scores are obtained by two different raters, at one moment in time, then the measure is called *inter-rater reliability*.

Next, we will see how advantage can be made of linear and generalized linear mixed models in case repeated or longitudinal measures are taken, rather than a single measure or a pair of measures.

3.2. Linear Mixed Models

Methods for continuous data form the best developed and most advanced body of research, while the same is true for software implementation. This is natural, since the special status and the elegant properties of the multivariate normal distribution simplify model building and ease software development. It is in this area that the linear mixed model is situated (Laird and Ware 1982, Verbeke and Molenberghs 2000). Gaussian data can be modeled entirely in terms of their means, variances and covariances. The parameters of the mean model are referred to as *fixed-effects* parameters, and the parameters of the variance-covariance model as *covariance parameters*. The fixed-effects parameters capture the influence of explanatory variables on the mean structure, exactly as in the standard linear model. However, the occurence of random effects and a structured covariance matrix distinguishes the linear mixed model from the standard linear model. The need for covariance modeling arises quite frequently in applications such as when repeated measurements are taken on the same experimental unit, with spatially correlated data, or when experimental units can be grouped into clusters and data from a cluster are correlated. One can distinguish between three components of variability. Part of the covariance structure arises from so-called *random effects*, i.e., additional covariate effects with random parameters. These are effects which arise from the characteristics of individual subjects. The variances of the random-effects parameters are commonly referred to as *variance components* (Searle, Casella, and McCulloch 1992). Another component of the variability is the serial correlation which captures that measurements taken close together in time are typically more strongly correlated than those taken further apart in time. On a sufficiently small time-scale, this kind of structure is almost inevitable. The last component is the measurement error: when the measurement process involves fuzzy determinations, the results may show substantial variation even when two measurements are taken at the same time from the same subject.

The standard linear model is one of the most commonly used statistical models and is represented by:

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \tag{7}$$

where $\boldsymbol{Y}_i$ is the n_i dimensional vector of responses for the ith subject, $\boldsymbol{\beta}$ are the fixed-effects parameters, X_i is the design matrix and $\boldsymbol{\varepsilon}_i$ is the unknown random error vector. The linear mixed-effects model for repeated measurements generalizes the standard linear model as follows (Laird and Ware 1982):

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{8}$$

where X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is the p dimensional vector containing the fixed effects, $\boldsymbol{b}_i$ is the q dimensional vector containing the random effects, and $\boldsymbol{\varepsilon}_i$ is an n_i dimensional vector of residual components whose elements are no longer required to be independent and homogeneous. Explicitly, one assumes:

$$\boldsymbol{b}_i \sim N(\mathbf{0}, D),$$

$$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Sigma_i),$$

$$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \text{ independent,}$$

where D is a general symmetric $(q \times q)$ covariance matrix and Σ_i is an$(n_i \times n_i)$ covariance matrix which depends on i only through its dimension n_i, i.e., the set of unknown parameters in Σ_i will not depend upon i. Note that when $\Sigma_i = \sigma^2 I_{n_i}$ and $Z_i = \mathbf{0}$, the mixed model reduces to the standard linear model.

From (8) it follows that, conditional on the random effect $\boldsymbol{b}_i$, $\boldsymbol{Y}_i$ is normally distributed with mean vector $X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i$ and with covariance matrix Σ_i. Further, $\boldsymbol{b}_i$ is assumed to be normally distributed with mean vector $\mathbf{0}$ and covariance matrix D. Let $f(\boldsymbol{y}_i|\boldsymbol{b}_i)$ and $f(\boldsymbol{b}_i)$ be the corresponding density functions. We then have that the marginal density function of $\boldsymbol{Y}_i$ is calculated by

$$f(\boldsymbol{y}_i) = \int f(\boldsymbol{y}_i|\boldsymbol{b}_i)f(\boldsymbol{b}_i)d\boldsymbol{b}_i$$

which can easily be shown to be the density function of a n_i dimensional normal distribution with mean vector $X_i\boldsymbol{\beta}$ and with covariance matrix

$$V_i = Z_iDZ_i' + \Sigma_i.$$

Hence, V_i can be modeled by specifying the structure of Z_i, D and Σ_i. Z_i is set up in a similar fashion as X_i, the design matrix for the fixed effects parameters. Typically, parameters are estimated using maximum likelihood or restricted maximum likelihood (Verbeke and Molenberghs 2000).

In Section 4, we will show, in different settings, how we can easily derive the reliability of psychiatric symptom scales from such models, thereby generalizing the classical developments as outlined in Section 3.1.

3.3. Generalized Linear Mixed Models

The linear mixed model, described in the previous section, assumes that response variables are directly equated to a linear combination of fixed and random effects and that the error terms are normally distributed (Littell et al. 1996). If one (or both) of these assumptions is violated one can resort to a generalized linear mixed model (GLMM). Fixed-effects generalized linear models (GLM) have been studied extensively in the literature (Nelder and Wedderburn 1972, McCullagh and Nelder 1989). Let us assume that all measures from all subjects are stacked into a vector $Y = (Y_1^T, ..., Y_n^T)^T$, with similar conventions for other vectors and matrices. This implies the index i drops from notation.

The basic form of a generalized linear model is:

$$\boldsymbol{\eta} = X\boldsymbol{\beta}, \tag{9}$$

where $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, $\boldsymbol{\mu} = E(\boldsymbol{Y})$ and g is an appropriate link function. Nelder and Wedderburn (1972) showed that maximum likelihood estimates for β can be obtained by iteratively solving

$$X'WX\beta = X'W\boldsymbol{y}^*, \tag{10}$$

where $W = D\Sigma^{-1}D$, $\boldsymbol{y}^* = \hat{\eta} + (\boldsymbol{y} - \hat{\mu})D^{-1}$, $D = (\partial\mu/\partial\eta)$, and $\Sigma = \text{Var}(\boldsymbol{\varepsilon}) = \Sigma_\mu^{1/2}A\Sigma_\mu^{1/2}$.

Generalized linear mixed models are a useful extension of GLMs, involving the addition of random effects and correlated errors. The generalized linear mixed model replaces (9) with

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zb} \tag{11}$$

where $\boldsymbol{\beta}$ is a vector of unknown fixed effects with known model matrix $\boldsymbol{X}$, $\boldsymbol{b}$ is a vector of unknown random effects with known model matrix $\boldsymbol{Z}$ and $\boldsymbol{\eta}$ is the link function $g(\boldsymbol{\mu})$ and $\boldsymbol{\mu}$ is the conditional mean of the observations $\boldsymbol{y}$, given the random effects $\boldsymbol{b}$. As in the conventional mixed model, the random effects $\boldsymbol{b}$ are in most mainstream papers assumed to follow a Gaussian distributed. However, this assumption is not crucial and has been relaxed (Lee and Nelder 1996). Likelihood inference for generalized linear mixed models requires evaluation of integrals (Breslow and Clayton 1993), where the integral's dimension is equal to the number of random effects. Zeger and Karim (1991) avoid the need for numerical integration by casting the generalized linear random-effects model in a Bayesian framework and by resorting to the Gibbs sampler. Wolfinger and O'Connell (1993) circumvent numerical integration by using pseudo-likelihood (and restricted pseudo-likelihood) procedures. The latter approach is implemented in the SAS macro GLIMMIX and is essentially a random-effects extension of (10). The GLIMMIX macro is known to have some drawbacks such as, for example, downward biases in fixed-effects and covariance parameters. In contrast, the MLWIN software, the MIXOR software package (Hedeker and Gibbons 1994) and the SAS procedure NLMIXED use either better approximations or numerical integration and are known to have better properties.

It might therefore seem sensible to avoid the use of the GLIMMIX macro. However, this procedure is the only one that allows for random-effects *and* residual (serial) correlation in the context of GLMMs. For this reason, we continue to use the GLIMMIX macro.

4. Data Analyses

Let us now apply the previously developed methodology on the pooled data described in Section 2. We will assess the reliability for both the PANSS and CGI scales, using the SAS procedure MIXED. The SAS codes for fitting the subsequent models and their respective reliabilities can be found in the Appendix.

4.1. The PANSS Scale

As mentioned earlier, the PANSS scale is a continuous response with 30 items. For this response we considered in turn three different models and calculated the corresponding reliability measures.

4.1.1. Model 1

First, we assume a linear mixed model with a random intercept. In that case, the repeated measurements of the PANSS for subject i satisfy:

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i b_i + \boldsymbol{\varepsilon}_i \tag{12}$$

with X_i the design matrix for the fixed effects which includes an intercept term, time, treatment and the interaction between time and treatment, Z_i a n_i dimensional vector of ones, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 I)$ and $b_i \sim N(\mathbf{0}, d^2)$. The fitted components are $\widehat{d^2} = 311.00$ and $\widehat{\sigma}^2 = 125.14$. To show how the reliability can be derived from these data we first rewrite the model as:

$$Y_{ijk} = \mu_{jk} + b_i + \varepsilon_{ijk}$$

where Y_{ijk} is the measure at time point j for subject i under treatment k; μ_{jk} groups the fixed-effects structure, b_i is still the random intercept and ε_{ijk} is the measurement error. As we have seen in Section 3.1, the reliability reduces to the correlation between two parallel measurements. For measurements at time points s and t we then have:

$$\begin{aligned} R &= \text{Corr}(Y_{isk}, Y_{itk}) \\ &= \frac{\text{Cov}(\mu_{sk} + b_i + \varepsilon_{isk}, \mu_{tk} + b_i + \varepsilon_{itk})}{\sqrt{\text{Var}(b_i + \varepsilon_{isk})}\sqrt{\text{Var}(b_i + \varepsilon_{itk})}} \\ &= \frac{\text{Cov}(b_i, b_i)}{d^2 + \sigma^2} \\ &= \frac{d^2}{d^2 + \sigma^2}. \end{aligned} \tag{13}$$

The reliability expresses the ratio of the variance explained by the model to the total observed variance. The link of (13) with the intuitive definition of reliability as we have expressed in (4) is obvious. For data containing two measurements per subject, this value equals the test-retest reliability of the instrument. For any series of repeated measurements, this value gives a global measure of the correlation between the measurements within subjects. For the PANSS data this global reliability measure yields a value of $\widehat{R} = 0.713$ (s.e. 0.012). The standard error is calculated using the delta method.

Table 1.

PANSS. Analysis of variance tables (weeks 6 and 8). (k=2 is the number of measurements per patient)

Source of variation	Df	Sum of Sq.	Mean Sum of Sq.	Exp. Sum of Sq.
Between patient	485	404812.1	834.664 (BMS)	$\sigma_e^2 + k\sigma_p^2$
Within patients (error)	486	22247.5	45.777 (WMS)	σ_e^2
Total	971	427059.6		

4.1.2. Comparison Between Model 1 and the Classical Approach

Here, we will show the analogy of the presented method for estimating reliability to the classical approach (Bartko 1966, Fleiss 1989, Dunn 1989) through comparing results of both methods. In the classical approach, typically two measurements are compared (test-retest reliability), assuming them to be parallel (equal true scores and equal error variances). First, we apply Model 1 to a reduced set of data with only two measurements in the case that this assumption is valid. These results will be compared to the classical approach. Second, we discuss both methods in the case that the assumption of parallel measures is violated.

We assume parallel measures for the subset of patients who both have week 6 and week 8 PANSS measurements. This assumption is viable since the means are 69.24 and 68.84, respectively, with standard deviations equal to 17.7+ and 21.5, respectively.

In the classical approach, reliability is estimated by the intraclass correlation coefficient wich can be derived from a one way analysis of variance with patient as factor (Table 1). From the table, we derive that the estimates for the variance components are $\widehat{\sigma}_e^2 = 45.78$ and $\widehat{\sigma}_p^2 = 394.44$. The estimate for the intraclass correlation coefficient of reliability (Bartko 1966) then is:

$$\widehat{R}_c = \frac{\widehat{\sigma}_p^2}{\widehat{\sigma}_p^2 + \widehat{\sigma}_e^2} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (k-1)\text{WMS}} = 0.896. \tag{14}$$

The reliability, estimated using the same set of data but based on linear mixed Model 1 is based on $\widehat{d}^2 = 394.77$ and $\widehat{\sigma}^2 = 45.72$ and therefore

$$\widehat{R} = \frac{394.77}{394.77 + 45.72} = 0.896.$$

The results of both types of analyses almost coincide.

Let us now consider a different set of time points (baseline, i.e., week 0 *versus* week 8). The means then are 89.9 and 68.8, respectively, with standard deviations equal to 20.5 and 21.5, respectively. Here, the assumption of equal true scores is violated.

The classical intraclass correlation coeficient as derived via the one-way ANOVA table is equivalent to a Pearson product-moment correlation coefficient between the pairs of measurements, in which each pair would enter the calculation twice, i.e., both the pair (Y_{isk}, Y_{itk}) would enter as well as the pair (Y_{itk}, Y_{isk}) (Dunn 1989). Therefore, the reliability, calculated in the classical way, will be biased by the change in true score. This bias, however is removed if the reliability is estimated via Model 1. The proposed model follows a hierarchical approach where the variance components are determined while fixed effects (time effect, treatment effect, and the interaction thereof) are taken into account. As a result, the corresponding reliability is then also fixed-effects corrected. This procedure could be mimicked within the classical paradigm by replacing the true scores with residual scores, i.e., where the time by treatment group mean values are subtracted from the true scores. Hence, we are left with three possible approaches, summarized in Table 2.

Table 2 clearly shows the classical method, without adjustment, would lead to incorrect results. This is to be expected since the assumptions required for the method to be valid (no mean change, stable variance between both measurements) are not satisfied.

TABLE 2.

PANSS. Three methods to calculate reliability (weeks 0 and 8).

Type of analysis	Between variance	Within variance	Reliability coefficient
Classical (raw data)	57.0	441.6	0.114
Model 1 (raw data)	169.2	218.1	0.437
Classical (residuals)	169.1	217.2	0.438

However, the modeling approach presented here is still able to make use of these data since it allows the correction for mean level and for heterogeneous variances. These are important advantages of the mixed modeling approach. Further, there are three additional advantages: the mixed model approach can be applied when (1) there are more than two measurement occasions, (2) not all subjects have the same number of measurements (due to missingness or irregularly spaced measurement times) and (3) more complicated variance-covariance structures within subjects exist. To study these advantages further additional, more elaborate models will be considered in subsequent sections.

4.1.3. Model 2

The use of random effects in the assessment of reliability dates back to Bartko (1966) and has been described by Dunn (1989). Model 1 builds upon this work. In addition, we will introduce serial correlation and then generalize the calculation of reliability to this situation. Explicitly, the second model combines a random intercept with serial correlation. Typical choices for such serial correlation structures are based on exponentially or Gaussian decaying processes. These are standard available in the SAS procedure MIXED (Littell et al. 1996). In order to choose the covariance structure that is best fitting the data, an

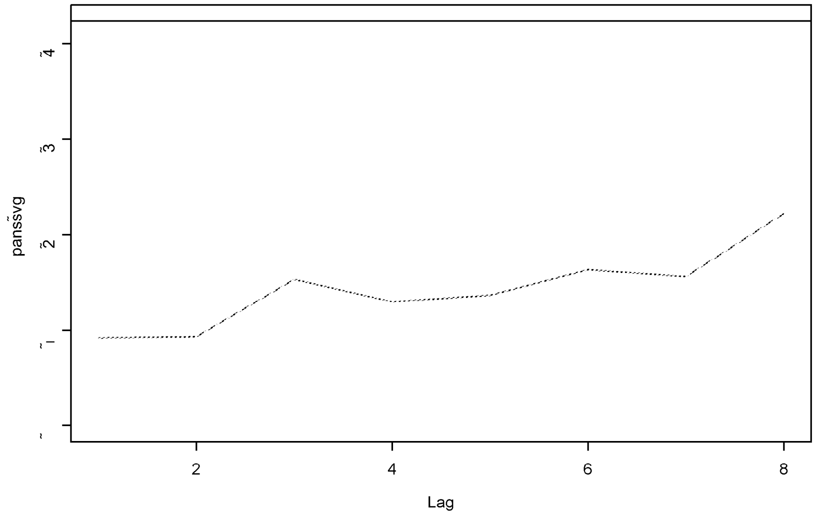

Figure 1.

Empirical variogram of the PANSS data.

empirical variogram was created which is shown in Figure 1. For a formal introduction to the variogram in the context of longitudinal data, we refer to Diggle, Liang and Zeger (1994) or Verbeke and Molenberghs (2000). The value of the variogram at time lag zero is an indication for the relative importance of the measurement error, the discrepancy between the variogram at the largest time lag and the process variance (represented as a level straight line at the top of the plot) is an indication for the importance of the random intercept. The shape of the variogram describes the serial correlation process. The strength of the process is indicated by the amount of increase between zero and maximum time lags, while the shape of the curve is indicative for the shape of the process of serial decay. In this case, we opt for a Gaussian serial process. Precisely, we retain Model (12), solely replacing the variance-covariance matrix Σ of $\boldsymbol{\varepsilon}_i$ by a matrix with elements

$$\Sigma_{ss} = \tau^2 + \sigma^2,$$

$$\Sigma_{st} = \tau^2 \exp(-u_{st}^2/\rho^2), \qquad s \neq t,$$

where σ^2 denotes the measurement error variance and the remaining part is the serial variance component with u_{st} the time-lag between measurents Y_{isk} and Y_{itk} for subject i and treatment k. The estimated covariance parameters of this model, applied to the PANSS data, are $\widehat{d}^2 = 103.21$, $\widehat{\tau}^2 = 274.97$, $\widehat{\rho} = 6.38$, and $\widehat{\sigma}^2 = 65.21$.

As with Model 1, to derive a measure of reliability, we rewrite the model as:

$$Y_{ijk} = \mu_{jk} + b_i + w_{ij} + \varepsilon_{ijk}, \tag{15}$$

where w_{ij} is now the correlated spatial Gaussian component of the residual variability structure with $w_{ij} \sim N(0, \tau^2)$ and ε_{ijk} merely the measurement error with variance σ^2. For time points s and t it follows that

$$\text{Var}(Y_{isk}) = \text{Var}(\mu_{sk} + b_i + w_{is} + \varepsilon_{isk}) = d^2 + \tau^2 + \sigma^2 = \text{Var}(Y_{itk})$$

and

$$\begin{aligned}\text{Cov}(Y_{isk}, Y_{itk}) &= \text{Cov}(\mu_{sk} + b_i + w_{is} + \varepsilon_{isk}, \mu_{tk} + b_i + w_{it} + \varepsilon_{itk}) \\ &= \text{Cov}(b_i, b_i) + \text{Cov}(w_{is}, w_{it}) \\ &= d^2 + \tau^2 \exp\left(\frac{-u_{st}^2}{\rho^2}\right)\end{aligned}$$

and therefore the reliability can be calculated as a function of time-lag u between two measurements

$$\begin{aligned}R(u) &= \text{Corr}(Y_{isk}, Y_{itk}) \\ &= \frac{\text{Cov}(Y_{isk}, Y_{itk})}{\sqrt{\text{Var}(Y_{isk})}\sqrt{\text{Var}(Y_{itk})}} \\ &= \frac{d^2 + \tau^2 \exp\left(\frac{-u_{st}^2}{\rho^2}\right)}{d^2 + \tau^2 + \sigma^2}.\end{aligned} \tag{16}$$

In spite of the correction for the fixed time effect, the covariance parameter estimates show a considerable remaining serial component in the PANSS data. As can be seen from formula

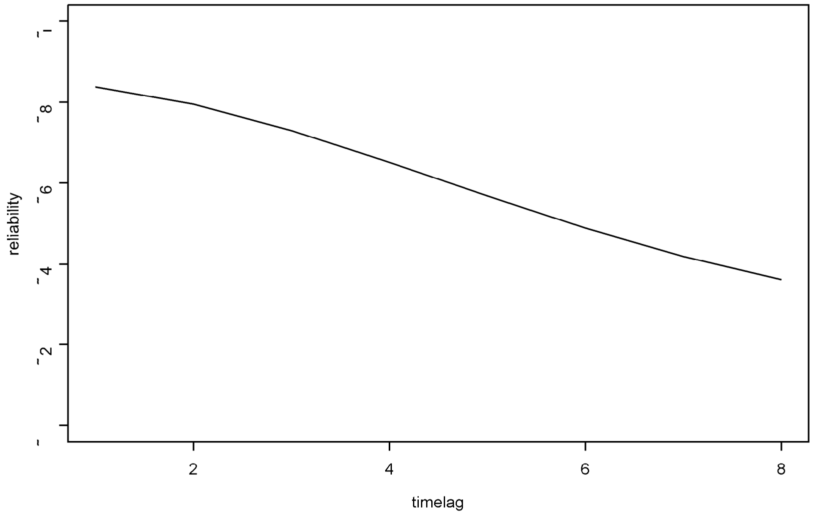

FIGURE 2.

PANSS. Reliability as a function of the time-lag u between any two measurements of a subject.

(16), a strong serial effect will lead to a fastly decreasing reliabilty for increasing time lags. This is shown in Figure 2.

4.1.4. Model 3

After adding serial correlation in Model 2 to the random-intercept Model 1, we now add random slope in time as well. The random-effects variance then equals

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}.$$

The estimated covariance parameters for the PANSS data are $\widehat{d}_{11} = 295.59$, $\widehat{d}_{12} = -4.3489$, $\widehat{d}_{22} = 5.9810$, $\widehat{\tau}^2 = 87.577$, $\widehat{\rho} = 0.8114$, and $\widehat{\sigma}^2 = 0.9443$.

The model can now be written as follows:

$$Y_{ijk} = \mu_{jk} + \begin{pmatrix} b_{i0} & b_{i1} \end{pmatrix} \begin{pmatrix} 1 \\ j \end{pmatrix} + w_{ij} + \varepsilon_{ijk}.$$

For time points s and t, we then have

$$\begin{aligned}\mathrm{Var}(Y_{isk}) &= \boldsymbol{z}_s D \boldsymbol{z}_s' + \tau^2 + \sigma^2,\\ \mathrm{Var}(Y_{itk}) &= \boldsymbol{z}_t D \boldsymbol{z}_t' + \tau^2 + \sigma^2,\\ \mathrm{Cov}(Y_{isk}, Y_{itk}) &= \boldsymbol{z}_s D \boldsymbol{z}_t' + \tau^2 \exp(-u_{st}^2/\rho^2).\end{aligned}$$

Here, $\boldsymbol{z}_s$ is the design row in Z corresponding to time s. From this we can derive, similar to the derivations done earlier, the reliability as a function of time point t and time-lag u:

$$R(t,u) = \mathrm{Corr}(Y_{isk}, Y_{itk}) = \frac{\boldsymbol{z}_s D \boldsymbol{z}_t' + \tau^2 \exp(\frac{-u_{st}^2}{\rho^2})}{\sqrt{\boldsymbol{z}_s D \boldsymbol{z}_s' + \tau^2 + \sigma^2}\sqrt{\boldsymbol{z}_t D \boldsymbol{z}_t' + \tau^2 + \sigma^2}}. \tag{17}$$

Table 3 presents a matrix of all the values of the test-retest reliability, depending on the point of measurement and the time lag between two measurements. Although the covariance parameter estimates show a considerably decreased serial effect compared to Model 2, the decrease in test-retest reliability when time lag increases remains, as can be seen from this table. It is not straigtforward to interpret this effect. In the introduction we mentioned two problems when choosing the time interval between two measurements for a test-retest reliability study. One is the possible change over time of the true score. For this problem, we introduce a solution by working with mixed models that allows correction for the fixed time effects. An important issue is the potential memory effect of raters when measurements are reasonably close in time. This may then explain the presence of a serial correlation, since the memory effect is likely to decrease with increasing time lags. What remains over a sufficiently long period of time is the random-effects structure, which may be seen as a long-term reliability. Figure 3 gives a graphical representation of the same information.

4.2. The CGI Scale

Next, we will consider the same three models for Clinician's Global Impression (CGI) overall change versus baseline, a scale going from 1 ('very much improved') to 7 ('very

Table 3.
PANSS. Test-retest reliability for Model 3, as a function of the measurement occasion and time lag between two measurements.

	time lag							
time point	1	2	3	4	5	6	7	8
0	0.811	0.741	0.710	0.673	0.631	0.588	0.545	0.504
1	0.813	0.748	0.723	0.692	0.659	0.624	0.590	.
2	0.820	0.762	0.743	0.719	0.693	0.666	.	.
3	0.832	0.781	0.767	0.750	0.730	.	.	.
4	0.846	0.802	0.793	0.781	.	.	.	.
5	0.862	0.824	0.818	0.692	0.631	.	.	.
6	0.877	0.845	0.767	0.719	0.659	0.588	.	.
7	0.892	0.824	0.793	0.750	0.693	0.624	0.545	.
8	0.892	0.845	0.818	0.781	0.730	0.666	0.590	0.504

much worsened') and used by the treating physician to characterize how well a subject has improved relative to baseline. We first consider CGI as a continuous response and apply the same methods that have been used for the PANSS scale. In the next subsection, we will apply generalized linear mixed methodology to accommodate the discrete nature of the scale.

The estimates for the variance components resulting from random-intercept Model 1, random-intercept and serial correlation Model 2, and random-intercept, random-slope and serial correlation Model 3 are presented in Table 4.

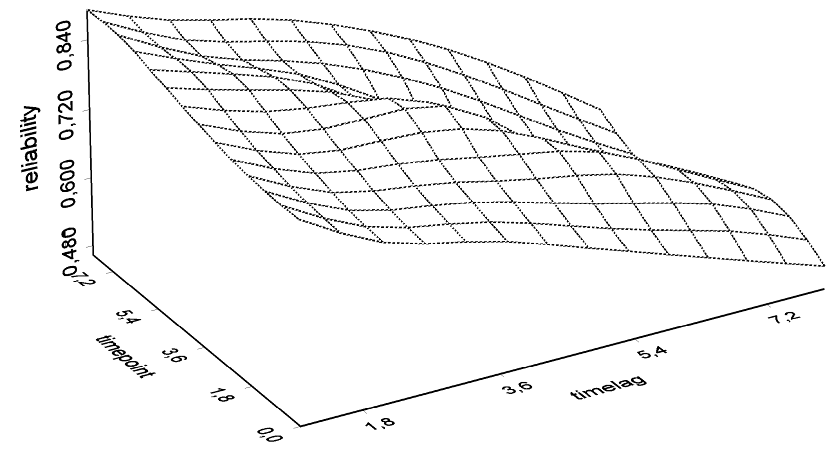

FIGURE 3.

PANSS. Model 3. Reliability as a function of the measurement time t and the time lag between two measurements u.

The reliability, derived under Model 1, calculated according to (13), equals $\widehat{R} = 0.570$ (s.e. 0.018). Let us consider the variogram for CGI, as presented in Figure 4. The variogram does not show an obvious structure, the spatial correlation appears to be rather weak. Let us confirm this using Model 2 (see Table 4). A comparison of the loglikelihood at maximum apparently does not underscore this message. However, no random effects have been added yet and therefore, the serial process may just capture omitted random-effects structure. For this model and as given by (16), reliability is again a function of time lag between measurements. Figure 5 gives a graphical display. The figure shows a gradual decrease in the reliability measurement when the time lag increases.

For Model 3 we add a random slope for time. The resulting covariance parameter etimates are represented as 3(a) in Table 4. The loglikelihood shows a further significant

Table 4.

CGI. Variance component estimates for Models 1–3.

		Estimates for Various Models				
Component	Par.	1	2	3 (a)	3 (b)	3 (c)
Var. rand. int.	d_{11}	0.7213	0.3501	0.0000	0.7001	0.6838
Cov. (rand. int., rand. slope)	d_{12}			0.0541	-0.0261	-0.0230
Var. rand. slope	d_{22}			0.0008	0.0216	0.0209
Serial process variance	τ^2		0.5701	0.5278	-1.6005	
Serial process corr. par.	ρ		5.3405	4.9901	0.4758	
Measurement error var.	σ^2	0.5449	0.3664	0.3591	1.9997	0.4061
-2 loglikelihood		8758.5	8578.4	8510.7	8546.2	8546.9

improvement. The default option with the MIXED procedure in SAS is to restrict all variance parameters to be nonnegative. However, a zero variance can be indicative for a negative variance component. Given the presence of measurement error and a serial correlation process, a negative variance component in the random-effects structure can still be compatible with an overal positive-definite variance-covariance structure. Therefore, we can relax the assumptions and allow for a general D matrix. For more details on negative variance components we refer to Verbeke and Molenberghs (2000). Practically, this is effected by including the 'lowerb' option in the PROC MIXED code. The corresponding estimates are labelled 3 (b) in Table 4. Now, a negative serial-correlation variance is obtained.

As was done for the PANSS data, the test-retest reliability can be derived as a function of measurement time and time lag (Table 5). A graphical representation is given

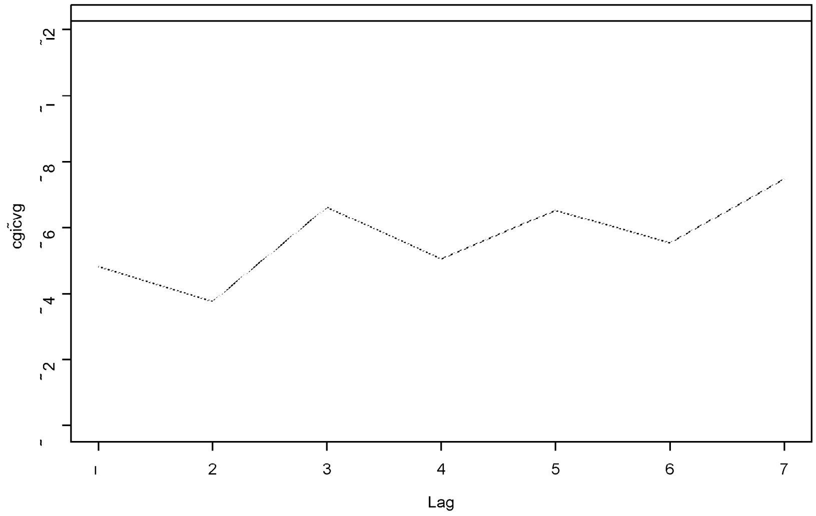

FIGURE 4.

Empirical variogram of the CGI dataset.

in Figure 6.

The variogram indicated the serial correlation process is likely not very important. Therefore, it is of interest to simplify Model 3 (b) by deleting the serial correlation process, retaining random intercepts and random slopes (Model 3 (c) in Tabel 4). Effectively, we then have $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 I)$. This can be derived from comparing the likelihoods as well. First, comparing the likelihoods of Models 1 and 2, there is some evidence the serial process would be needed, apparently contradicting this statement. However, secondly comparing Models 3 (b) and (c) there clearly is no need for the serial process. The disparity between both comparisons is because the first comparison is done in the absence of a random slope while the second one is conditional on the presence of a random slope. Given the latter, there is no further need for serial correlation. When the choice is between either a serial process (Model 2) or a random slope (Model 3 (c)), the likelihoods favor the random slope.

Comparing the likelihood of Model 3 (c) with Model 3 (b), also in the light of the

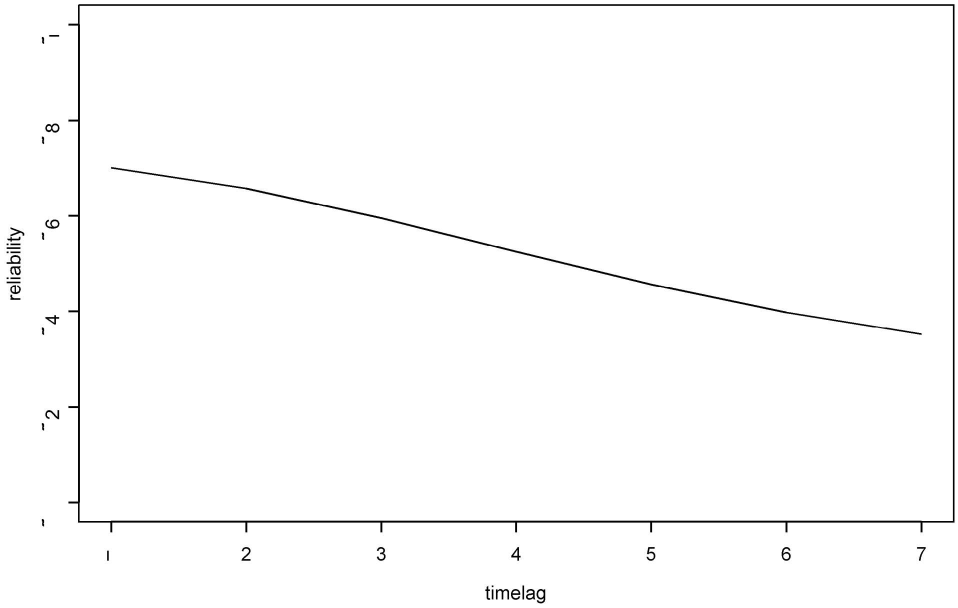

Figure 5.

CGI. Reliability as a function of the time-lag u between any two measurements of a subject.

sample size, is consistent with our conclusion that the serial process is not very important.

The model can now be written as

$$Y_{ijk} = \mu_{jk} + \left(\begin{array}{cc} b_{i0} & b_{i1} \end{array} \right) \begin{pmatrix} 1 \\ j \end{pmatrix} + \varepsilon_{ijk}$$

and

$$\mathrm{Var}(Y_{isk}) = z_s D z_s' + \sigma^2,$$

$$\mathrm{Var}(Y_{itk}) = z_t D z_t' + \sigma^2,$$

$$\mathrm{Cov}(Y_{isk}, Y_{itk}) = z_s D z_t'.$$

It is then straightforward to derive the reliability for measurements at times s and t:

$$R(s,t) = \frac{z_s D z_t'}{\sqrt{z_s D z_s' + \sigma^2}\sqrt{z_t D z_t' + \sigma^2}}. \qquad (18)$$

Table 5.
CGI. Test-retest reliability for Model 3 (b), as a function of the measurement occasion and time lag between two measurements.

	time lag						
time point	1	2	3	4	5	6	7
0	0.604	0.594	0.556	0.511	0.463	0.414	0.369
1	0.601	0.599	0.571	0.537	0.500	0.464	.
2	0.613	0.620	0.601	0.577	0.551	.	.
3	0.639	0.651	0.640	0.625	.	.	.
4	0.673	0.689	0.684	0.511	.	.	.
5	0.727	0.601	0.637	0.463	.	.	.
5	0.689	0.640	0.577	0.500	0.414	.	.
7	0.746	0.727	0.684	0.625	0.551	0.464	0.369

For the CGI dataset this leads to the values for reliability as presented in Table 6. The table shows an increase of reliability over succesive measurements for a constant lag between two measurements. In general, the values appear to decrease when the time lag between two measurements increase, however there are some exceptions for higher values for both time points.

4.3. Binary Responses on the CGI Scale

In this section we will indicate how the techniques developed in the previous sections for continuous data, can be applied when the outcome is binary, using the model proposed by Wolfinger and O'Connell (1993) as implemented in the SAS macro GLIMMIX and

Table 6.

CGI. Test-retest reliability for Model 3 (c), as a function of the measurement occasion and time lag between two measurements.

	time point 2							
time point 1	1	2	3	4	5	6	7	8
1		0.612	0.594	0.568	0.536	0.502	0.468	0.435
2			0.625	0.616	0.599	0.577	0.553	0.528
3				0.651	0.648	0.638	0.625	0.608
4					0.683	0.685	0.681	0.674
5						0.719	0.724	0.723
6							0.754	0.760
7								0.785
8								

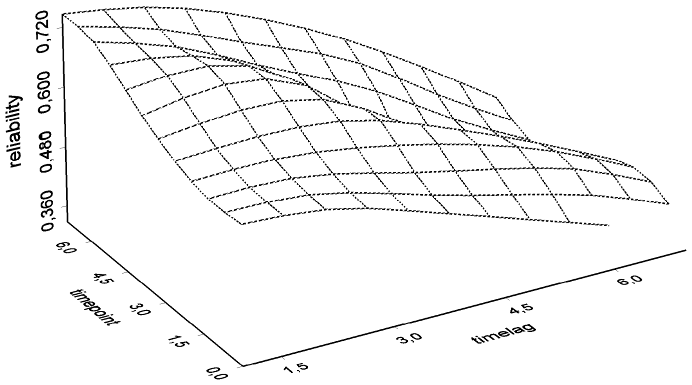

FIGURE 6.

CGI. Reliability as a function of measurement time t and time lag u.

described in Section 3.3. The binary version of CGI is defined as much or very much improved versus baseline (CGI equal to 1 or 2) versus the other categories.

In analogy with the continuous case of CGI (Table 4), we fit Models 1–3 (for the latter versions (a) and (c)) to the binary outcome. The corresponding variance components are given in Table 7.

Under Model 1, reliability is calculated as $\widehat{R} = 0.953$. We note that the value is much higher than the continuous outcome counterpart. While Streiner and Norman (1995) have found a reverse effect, we do believe our result is plausible since dichotomization can (but does not have to) lead to an increased concordance between measurements.

Under Model 2, the reliability is only a slowly varying function of time lag (Figure 7).

Under Model 3 (a), a zero serial process variance is obtained, and therefore we revert

Table 7.

CGI (binary version). Variance component estimates for Models 1–3.

		Estimates for Various Models			
Component	Par.	1	2	3 (a)	3 (c)
Var. rand. int.	d_{11}	8.2270	3.8484	45.9539	46.6942
Cov. (rand. int., rand. slope)	d_{12}			-7.9661	-8.0956
Var. rand. slope	d_{22}			2.7531	2.7778
Serial process variance	τ^2		0.2170	0.0000	
Serial process corr. par.	ρ		4.0560	1.9218	
Measurement error var.	σ^2	0.4008	0.4101	0.1296	0.1294

immediately to version (c) where the serial structure is removed from the model, leaving in random intercepts and random slopes. The reliability can be calculated according to (18). Results are shown in Table 8.

The results are rather peculiar. As was observed in the previous cases we see that for a constant time lag, the reliability increases over the weeks. And on the other hand the reliability decreases with increasing time lag. The rate at which this occurs is dramatic: for far apart time lags there is no residual reliability any more.

5. Discussion

A body of research exists on reliability, especially in psychology and educational sciences. In the past decades the topic is also entering the field of health sciences and especially the psychiatric health sciences because of the inherent subjectivity of the measures

Table 8.

CGI (binary version). Test-retest reliability for Model 3 (c), as a function of the measurement occasion and time lag between two measurements.

	time point 2							
time point 1	1	2	3	4	5	6	7	8
1		0.957	0.812	0.582	0.350	0.164	0.003	-0.070
2			0.939	0.782	0.593	0.428	0.302	0.207
3				0.941	0.824	0.701	0.598	0.517
4					0.961	0.893	0.825	0.766
5						0.978	0.943	0.907
6							0.988	0.970
7								0.994
8								

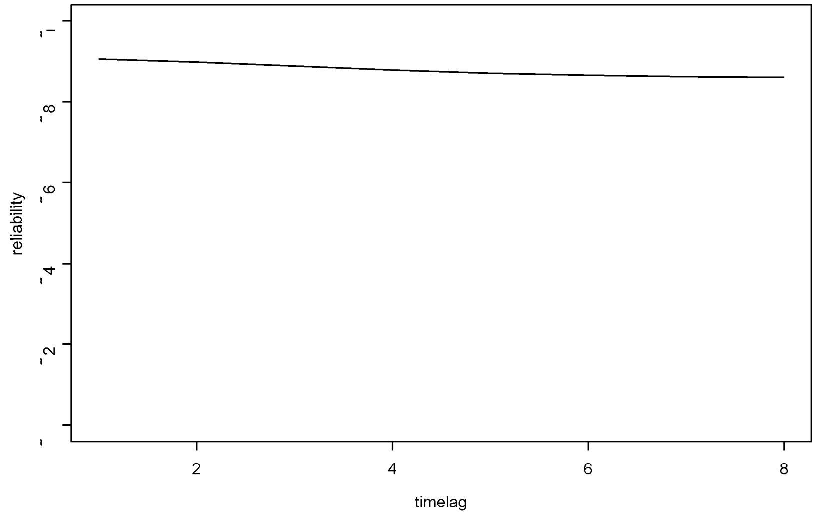

Figure 7.

CGI (binary version). Reliability as a function of the time-lag u between any two measurements of a subject.

employed in this field. Test-retest reliability as one of the classical approaches typically deals with the problem of time: how to disentangle the measurement error from real fluctuations in what you are measuring ?

Wiley and Wiley (1970) were among the first authors to deal with this problem by assuming a linear relationship between two adjacant measurements. In this way also reliability will have different values at both moments of measurement. Tisak and Tisak (1996) also stressed the fact that reliability is not a fixed property of an instrument but changes with time. They proposed a method to calculate a time-function of reliability. Dunn (1989) describes a method that uses components of variance in the calculation of reliability. He further extends this method to a mixed model to deal with rater effects by taking the rater into the model as a random effect. The mixed model methodology indeed allows to study variance components and fixed effects simultaneously. While, for this reliability study, we are primarily interested in the variance components, mixed-model methodology provides an interesting opportunity to model the fixed effects as well. We do not have to make the unrealistic assumption that there is no change in a patients situation over time or with treatment. Instead, such changes can be incorporated into the model.

When using repeated measurements a third source of variation can be taken into account when calculating reliability, the so-called serial correlation. In this work, a method has been proposed that allows for serial correlation in the calculation of test-retest reliability, as well as random effects and measurement error.

The method was applied to two psychiatric rating scales for schizophrenia: PANSS and CGI. For both scales, we observed a gradual decrease of reliability with increasing time lag between measurements. As mentioned earlier, there are different possible scenarios to explain such effects, such as memory effect of the raters or other covariates that are not taken into account in the model. For the PANSS scale we obtain reliability estimates from almost 0.90 to 0.50. Up to a time interval of five weeks, the reliability does not go below 0.60, which is considerable. The CGI scale shows less good results: up to a time lag of three weeks, the estimate of reliability remains above 0.55. Another result that occurs quite consistently is a slight increase in the reliability measure as time goes by, but for a fixed time lag. The reason for this is most likely a learning effect in the raters. In a different setting, one might also encounter learning effects in the study subjects. Of course, other perhaps complementary explanations cannot be excluded. We saw that dichotomising the CGI data to a binary outcome has the effect of increasing the reliability considerably. In order to calculate reliabilities in the context of repeated binary data, the generalized linear mixed model of Wolfinger and O'Connell (1993), as implemented in the SAS macro GLIMMIX, has been used. This methodology has the advantage of allowing for serial correlation together with random effects. The method does have some drawbacks, as mentioned earlier, and further research in this area is warranted.

The present method stresses once again the fact that reliability should not be perceived as a fixed quantity, but changes with circumstances. The quantity does not only change over time or with population, but also with the covariates incorporated into the model. The consequence is a more complicated picture, but arguably one that is also closer to the true nature of things. It also means that better and more specific conclusions can be drawn from a reliability study when an appropriate model is constructed, when it is known which sources of variability are under consideration. Modelling other sources of variation, like for example country or rater, is therefore an interesting topic for further research on the present data.

A further important advantage of the present method is that it becomes possible to estimate trial-specific or population-specific reliability in clinical studies. This is especially true because even in studies, designed to assess reliability, it is difficult to exclude fluctuations in the true scores and furthermore these studies are often conducted with different populations and in different circumstances. Finally, when measurement sequences on a subset of respondents are incomplete, these data can still be used for analysis, unlike in the classical approaches.

Acknowledgements

Research supported by a PAI program P5/24 of the Belgian Federal Government (Federal Office for Scientific, Technical, and Cultural Affairs).

Appendix

```
MODEL 1
-------
proc mixed data=panss2 method=reml noclprint;
class treat id xtime_c;
model panss= xtime_c treat xtime_c*treat / outp=out;
random intercept/ type=un subject=id;
ods output covparms=cp;
run;

proc iml;
use cp;
read all into covpars;
close cp;
d=covpars[1];
sigma2=covpars[2];
relvec=d/(d+sigma2);
quit;

MODEL 2
-------
proc mixed data=panss2 method=reml noclprint;
class treat id xtime_c;
model panss= xtime_c treat xtime_c*treat / outp=out;
random intercept / type=un subject=id;
repeated xtime_c / type=sp(gau)(xtime) local subject=id;
ods output covparms=cp;
run;

proc iml;
use cp;
read all into covpars;
close cp;
d=covpars[1];
tau2=covpars[2];
rho=covpars[3];
sigma2=covpars[4];
u=t(1:8);
relvec=(d+tau2*exp(-(u##2/rho##2)))/(d+tau2+sigma2);
create rel var{u relvec};
append;
close rel;
quit;

MODEL 3
-------
proc mixed data=panss2 method=reml noclprint;
class treat id xtime_c;
model panss= xtime_c treat xtime_c*treat / outp=out;
random intercept xtime / type=un subject=id;
repeated xtime_c / type=sp(gau)(xtime) local subject=id;
ods output covparms=cp;
run;

proc iml;
```

```
use cp;
read all into covpars;
close cp;
d11=covpars[1];
d12=covpars[2];
d22=covpars[3];
dvec=d11||d12||d12||d22;
dmat=shape(dvec,2,2);
tau2=covpars[4];
rho=covpars[5];
sigma2=covpars[6];
tvar=0;
uvar=0;
relvar=0;
time=t(0:8);
relvec=j(9,8,0);
t=1;
do while (t<=nrow(time));
tvec=1||time[t];
vart=tvec*dmat*t(tvec)+tau2+sigma2;
 u=1;
 do while (u<=(8-time[t])|u<=time[t]);
 if time[t]+u<=8 then tuvec=1||(time[t]+u);
 if time[t]+u>8 then tuvec=1||(time[t]-u);
 vartu=tuvec*dmat*t(tuvec)+tau2+sigma2;
 covar=tvec*dmat*t(tuvec)+tau2*exp(-(u##2/rho##2));
 relvec[t,u]=covar/sqrt(vart*vartu);
 tvar=tvar//time[t];
 uvar=uvar//u;
 relvar=relvar//relvec[t,u];
 u=u+1;
 end;
t=t+1;
end;
tvar=tvar[2:nrow(tvar)];
uvar=uvar[2:nrow(uvar)];
relvar=relvar[2:nrow(relvar)];
create rel var{tvar uvar relvar};
append;
close rel;
quit;
```

References

Bartko JJ (1966) “The intraclass correlation coefficient as a measure of reliability,” *Psychological Reports*, **19**, 3–11.

Blin O, Azorin JM and Bouhours P (1996) “Antipsychotic and anxiolytic properties of risperidone, haloperidol and methotrimeprazine in schizophrenic patients,” *J. Clin. Psychopharmacol*, **16**, 38–44.

Breslow NE and Clayton DG (1993) “Approximate inference in generalized linear mixed models”, *J. Am. Statist. Assoc.*, **88**, 9–25.

Chounard G, Jones B and Remington G (1993) “A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients,” *J. Clin. Phychopharmacol*, **13**, 25–40.

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

Cronbach LJ, Rajaratnam N and Gleser GC (1963) “Theory of Generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, **16**, 137–163.

Deyo RA, Dierh P and Patrick D (1991). ”Reproducibility and Responsiveness of Health Status Measure Statistics and Strategies for Evaluation,” *Controlled Clinical Trials,***12**, 142-158.

Diggle PJ Liang K-Y and Zeger SL (1994) “Analysis of Longitudinal Data”, Clarendon Press: Oxford.

Dunn G (1989) “Design and Analysis of Reliability Studies: The statistical evaluation of measurement errors”, Oxford University Press: New York.

Dunn, G. (1989) “Design and Analysis of Reliability Studies,” Edward Arnold: London.

Hedeker D and Gibbons RD (1994) “A random-effects ordinal regression model for multilevel analysis”, *Biometrics*, **50**, 933–944.

Hoyberg OJ, Fensbo C, Remvig J, Lingjaerde OK, Slotei-Nielsen M and Salvesen I (1993)

"Risperidone versus perphenazine in the treatment of chronic schizophrenic patients with acute exacerbations," *Acta Psychiatr Scand*, **88**, 395–402, 1993.

Huttunen MO, Piepponen T, Rantanen H, Larmo I, Nyholm R and Raitasuo V (1995) "Risperidone versus zuclopenthixol in the treatment of acute schizophrenic episodes: a double-blind parallel-group trial," *Acta Psychiatr Scand*, **91**, 271–277.

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

Kay SR, Opler LA and Lindenmayer JP (1988) "Reliability and Validity of the Positive and Negative Syndrome Scale for Schizophrenics," *Psychiat. Res*, **23**, 99–110.

Kuder GF, Richardson MW (1951) "The Theory of Estimation of Test Reliability," *Psychometrika,* **2**, 151–160.

Laird NM and Ware JH (1982) "Random effects models for longitudinal data", *Biometrics*, **38**, 963–974.

Lee Y and Nelder JA (1996) "Hierarchical generalized linear models (with discussion)", *J. R. Statist. Soc. B*, **58**, 619–678.

Littell RC, Milliken GA, Stroup WW and Wolfinger RD (1996) "SAS System for Mixed Models", SAS Institute Inc., Cary, NC.

Marder SR and Meibach RC (1994) "Risperidone in the treatment of schizophrenia," *Am. J. Psychiatry*, **151**, 825–835.

McCullagh P and Nelder JA (1989) "Generalized Linear Models", Chapman & Hall: London.

Nelder JA and Wedderburn RWM (1972) "Generalized linear models", *J. R. Statist. Soc. B*, **135**, 370–384.

Peuskens J and the Risperidone Study Group (1995) "Risperidone in the treatment of chronic schizophrenic patients: a multi.tio.l, multicentre, double-blind, parallel-group study versus haloperidol," *Br J Psychiatry*, **166**, 712–726.

Searle SR, Casella G and McCulloch CE (1992) "Variance Components", John Wiley & Sons: New York.

Streiner DL and Norman GR (1995) "Health measurement scales", Oxford University Press: Oxford.

Tisak J and Tisak MS (1996) "Longitudinal models of Reliability and Validity: A Latent Curve Approach", *Applied Psychological Measurement*, **20/3**, 275–288.

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

Wiley DE and Wiley JA (1970) "The estimation of measurement error in panel data. *American Sociological Review*, **35**, 112–117.

Wolfinger R and O'Connell M (1993) "Generalized linear mixed models: a pseudo-likelihood approach", *J. Statist. Comp. Simul.*, **48**, 233–243.

Zeger SL and Karim MR (1991) "Generalized linear models with random effects: a Gibbs' sampling approach", *J. Am. Statist. Assoc.*, **86**, 79–95.