A family of measures to evaluate scale reliability in a longitudinal setting

Peer-reviewed author version

LAENEN, Annouschka; ALONSO ABAD, Ariel; MOLENBERGHS, Geert & VANGENEUGDEN, Tony (2009) A family of measures to evaluate scale reliability in a longitudinal setting. In: JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES A-STATISTICS IN SOCIETY, 172. p. 237-253.

DOI: 10.1111/j.1467-985X.2008.00554.x
Handle: http://hdl.handle.net/1942/9180

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 concept of reliability denotes one of the most important psychometric properties of a measurement scale. Reliability refers to the capacity of the scale to discriminate between subjects in a given population. In classical test theory, it is often estimated using the intraclass correlation coefficient based on two replicate measurements. However, the modelling framework used in this theory is often too narrow when applied in practical situations. Generalizability theory has extended reliability theory to a much broader a framework, but is confronted with some limitations when applied in a longitudinal setting. In this paper, we explore how the definition of reliability can be generalized to a setting where subjects are measured repeatedly over time. Based on four defining properties for the concept of reliability, we propose a 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 and that the reliability of the instrument can depend on the way it is used. The methodology is motivated by and illustrated on data from a clinical study on schizophrenia. Based on this study, we estimate and compare the reliabilities of two different rating scales to evaluate the severity of the disorder.

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

1 Introduction

Rating scales are mainly used when the trait of interest cannot be observed directly, such as in the measurement of depression, anxiety, and quality of life. They play an important role in many scientific fields and are often used within a longitudinal framework to evaluate a patient's evolution over time. For instance, in psychiatry and psychology, longitudinal measurements of rating scales are commonly used to obtain precise diagnostics and to study the efficacy of new treatments or therapeutic procedures. In spite of their undoubted advantages, longitudinal studies also bring some methodological challenges, especially from a statistical modelling perspective. For example, in such studies, patients usually exhibit a systematic change or evolution over time on top of which an individualized evolution, characterized by correlated subject-specific effects, is present. Additionally, serial correlation and heterogenous variance components are frequently encountered as well (Verbeke and Molenberghs 2000).

Whenever a new rating scale is developed, its validity and reliability must be evaluated. However, reliability is not an intrinsic property of an instrument but rather changes depending on the population in which it is used. As a consequence, the reliability of a measurement scale should be evaluated not only when the scale is created but also every time it is introduced to a different population or translated into a new language.

The most widely used definition of reliability was given in classical test theory (CTT), where it was defined as "the ratio of the true score variance to the observed score variance" (Lord and Novick 1968). This concept of reliability was developed within a cross-sectional setting and based on a rather restrictive modelling framework. As a result, it is difficult to apply in longitudinal studies, with their peculiarities mentioned earlier. The simple modelling framework of CTT does not allow for taking into account these characteristics, a necessary feature though to avoid bias in the estimation of the parameters and in the

inferential procedures. Basically, measures of reliability are model-based quantities and their scope and applicability will never extend beyond the scope and applicability of the model they are based on.

Perhaps the most important attempt to extend the concept of reliability to a longitudinal scenario came from generalizability theory (G-theory), which was developed to explicitly model the multiple sources of variation present in a measurement system (Cronbach *et al* 1963, 1972, Brennan 2001). Undoubtedly, G-theory is one of the most relevant developments in the psychometric field and since its introduction it has been applied to many areas of psychology and education.

G-theory is based on a much more solid modelling framework than CTT. Essentially, it takes advantage of all flexibility given by the analysis-of-variance models with random effects. Of course, the usefulness of G-theory in evaluating reliability in longitudinal settings depends on the adequacy of these models to describe the specific characteristics of this type of data structure. Unfortunately, the modelling framework used in G-theory can be applied in a longitudinal setting only if strong and unrealistic assumptions are made. Some of these assumptions are: stability of the true scores over time, uncorrelated error structure, uncorrelated random effects, equal variance over time and, in its most classical formulation, it also requires a missing completely at random mechanism when data are incomplete. Note that all of these assumptions are quite restrictive for longitudinal studies, and they seriously limit the applicability of G-theory in longitudinal scenarios. Applying G-theory models in a setting where these assumptions are violated will lead to biased estimates of the variance components and, as a consequence, to biased estimates of the reliability parameters, or G coefficients (Diggle, Liang, and Zeger 1994, Verbeke and Molenberghs 2000, Smith and Luecht 1992, Bost 1995, Molenberghs and Kenward 2007).

In the present work, we extend the concept of reliability to a longitudinal framework

using hierarchical linear models. This type of linear mixed models (LMM) were also used by Vangeneugden et al. (2004) to extend the concept of reliability. Depending on the complexity of the model, these authors defined reliability either 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 (2007) also used hierarchical linear models to evaluate reliability in a longitudinal setting. Unlike the earlier authors, they approach reliability not from a correlation perspective but rather from an axiomatic point of view, and defined reliability through a set of four simple properties. 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 show that the R_T can be framed in a much broader setting. Actually, this measure is just a special case of a more general family of reliability measures. Interestingly, different members of this family seem to capture different sides of the reliability problem and have different interpretations.

In Section 2, the case study is introduced. Section 3 discusses the modelling framework used in the present work, as well as other approaches used in the psychometric literature. Additionally, a family of reliability measures is proposed and its properties are analyzed. Section 4 investigates the properties of some such measures, based on simulations and explores the relationship between the new proposals and the G coefficients. Finally, Section 5 applies the methodology to the case study.

2 Case Study

Schizophrenia is one of the most disabling and emotionally devastating illnesses known to man. It is characterized by a constellation of distinctive and predictable symptoms. The symptoms that are most commonly associated with the disease are called positive symptoms, that denote the presence of grossly abnormal behavior. These include thought disorder, delusions, and hallucinations. Less obvious than the positive symptoms but equally serious are the deficit or negative symptoms that represent the absence of normal behavior. These include flat or blunted affect (i.e. lack of emotional expression), apathy, and social withdrawal.

Several instruments can be considered to assess a patient's condition. The Brief Psychiatric Rating Scale (BPRS) is an 18-item scale that has been successfully used since 1967 to evaluate schizophrenic patients and to demonstrate the efficacy of antidepressant, antianxiety and antipsychotic drugs. It has also been used in epidemiological studies, gero-psychiatric research, and to compare diagnostic concepts between countries (Overall and Gorham 1988).

Another highly useful scale in the assessment of schizophrenia is the Positive and Negative Syndrome Scale (PANSS) (Kay, Fizbein, and Opler 1987). PANSS is a 30-item rating instrument evaluating the presence and severity of positive, negative and general psychopathology of schizophrenia. The scale was developed based on the BPRS and the Psychopathology Rating Scale and was conceived as an operationalized, drug-sensitive instrument that provides a balanced representation of positive and negative symptoms and gauges their relationship to one another and to global psychopathology. PANSS was designed as an advance on BPRS, addressing broader psychopathology and therefore it is expected to achieve greater reliability.

The case study is a randomized clinical trial, investigating the effect of risperidone as

compared to an active control for the treatment of chronic schizophrenia. A total of 453 patients were evaluated using both rating scales, PANSS and BPRS, at baseline and after 1, 2, 4, 6, and 8 weeks, respectively. The upper part of Figure 1 shows the evolution of the individual profiles for both instruments over time. Note that in this study the instruments were used longitudinally to evaluate the efficacy of a new drug. The main objective of the present work is to developed tools that could allow us to study the reliability of these or other scales in such a longitudinal scenario, likely the most frequently encountered scenario in practice.

3 Methodology

In general, each data structure presents unique problems for the estimation of reliability, but longitudinal data, with their different sources of variation and correlation, present some of the most challenging problems for defining and estimating reliability. In Section 1, we described some of the limitations of the modelling framework used in CTT and G-theory when applied within a longitudinal setting.

Many proposals have appeared over the last decades to solve some of these modelling limitations. They are frequently based on path analysis or structural equations, and have been developed to estimate reliability in a longitudinal setting dropping the assumption of stability for the true scores (Heise 1969, Jagodzinski and Kühnel 1987, Werts *et al* 1980, Wiley and Wiley 1970). In any event, to dodge the requirement of true score stability when estimating reliability, these models often impose additional assumptions that may also have questionable validity in a longitudinal setting. For example, it is usually assumed that the changes in the true scores across time follow a simplex pattern (Heise 1969, Wiley and Wiley 1970, Werts, Linn, and Joreskøg 1977).

Some of these approaches also make strong assumptions regarding the pattern of measurement errors across time, for instance, they assume equal reliabilities over time (Heise 1969), equal error variances over time (Wiley and Wiley 1970) or uncorrelated error structures (Tisak and Tisak 1996). Raykov (2000) criticizes the equal-reliability assumption of Heise (1969) and proposed a model that circumvents this limitation. However, his model still assumed uncorrelated error terms, another doubtful assumption in several longitudinal studies. Many other authors have discussed the merits and disadvantages of using a first-order autoregressive structure to describe within-subject evolution over time (Kenny and Zautra 1995, Hertzog and Nesselroade 1987, Cole, Martin and Steiger 2005). The model discussed by Kenny and Zautra (1995) decomposes the observed scores as an overall constant that is allowed to change over time but does not depend on any covariate, a trait, or subject-specific parameter (equivalent to a random intercept in the LMM formulation), a term representing the state which is equivalent to the serial correlation component in LMM and a random error equivalent to the random error also present in LMM. Unlike the LMM, the model assumes that the variance explained by each source is the same for all time points. Another important difference with the LMM is that the so-called trait-state-error model (TSE) imposes a first-order autoregressive structure for the state factor. Hertzog and Nesselroade (1987) criticized the first-order autoregressive assumption and claim it is not flexible enough to be applied to some data structures.

As stated before, in the present work we will outline our proposals for quantifying reliability within a linear mixed models framework. This modelling paradigm will allow us to incorporate many of the previously discussed features, such as varying true scores, correlated error terms, including different types of serial correlation, heteroscedastic error components, and correlated random effects, in a very natural way. Accounting for all of these complexities within the same modeling paradigm is of the utmost importance to guarantee unbiased results when estimating reliability. For instance, we can incorporate the systematic variability of the true scores into the fixed-effects structure of the model

in a very flexible manner using, for example, fractional polynomials (Royston and Altman 1994) or non-parametric approaches such as splines (Verbyla *et al* 1999). Unlike in the model of Kenny and Zautra (1995), we could incorporate many different structures to account for serial correlation like Gaussian, first-order autoregressive, exponential, m-dependent structures, among others. The assumption of equal error variance over time can also be dropped and fully general variance functions can be considered. A linear mixed-effects model can generally be written as

$$\begin{aligned} &\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon}_{(1)\boldsymbol{i}} + \boldsymbol{\varepsilon}_{(2)\boldsymbol{i}}, \\ &\boldsymbol{b_i} \sim N(\mathbf{0}, D), \qquad \boldsymbol{\varepsilon}_{(1)\boldsymbol{i}} \sim N(\mathbf{0}, \Sigma_{Ri}), \qquad \boldsymbol{\varepsilon}_{(2)\boldsymbol{i}} \sim N(\mathbf{0}, \tau^2 H_i), \\ &\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_{(1)1}, \ldots, \boldsymbol{\varepsilon}_{(1)N}, \boldsymbol{\varepsilon}_{(2)1}, \ldots, \boldsymbol{\varepsilon}_{(2)N} \text{ independent}, \end{aligned} \tag{1}$$

where $\boldsymbol{Y}_i$ is the p_i dimensional vector of responses for subject i, $1 \leq i \leq n$, n denotes 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, $\boldsymbol{b_i}$ is the r-dimensional vector containing the random effects, $\boldsymbol{\varepsilon}_{(2)\boldsymbol{i}}$ is a p_i-dimensional vector of components of serial correlation, and $\boldsymbol{\varepsilon}_{(1)\boldsymbol{i}}$ is a p_i-dimensional vector of residual errors. Additionally, D is a general $(r \times r)$ covariance matrix, associated with the subject-specific random effects, 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. Furthermore, H_i and Σ_{Ri} depend on i only through their dimension p_i.

Model (1) implies the marginal model $\boldsymbol{Y_i} \sim N(X_i\boldsymbol{\beta}, V_i)$, where $V_i = \Sigma_{D_i} + \Sigma_i$ with $\Sigma_{D_i} = Z_i D Z_i'$ and $\Sigma_i = \tau^2 H_i + \Sigma_{Ri}$. Note that the total variability is decomposed into a component stemming from the subject-specific random effects and a residual variability component. The remaining variability is the sum of a serial correlation part and an error part, but we will generically refer to it as the error variability.

In what follows, we will discuss a proposal by Laenen, Alonso, and Molenberghs (2007)

to quantify reliability in this very general scenario. Further, we will introduce a general family of reliability measures that contained this proposal as a special case.

3.1 Properties of a Reliability Measure

Laenen, Alonso, and Molenberghs (2007) extended the concept of reliability to a longitudinal scenario using a simple set of four defining properties. Essentially, these authors 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 true-score variance to observed variance ratio, used in classical test theory, should be recovered. This type of *axiomatic* definitions have been successfully applied in many different areas, so as to extend concepts, originally defined in a simple setting, to more general scenarios. For instance, the same approach was used in mathematics to define the concept of distance or in probability and statistics to define the concept of a probability density function.

Further, these authors proposed the so-called R_T, a parameter that satisfies the previous set of properties, to quantify reliability. Assuming a balanced design where $\Sigma_i = \Sigma$ and $V_i = V$ for all i, the R_T takes the form:

$$R_T = 1 - \frac{\text{tr}(\Sigma)}{\text{tr}(V)}. \tag{2}$$

Note that, in the previous expression, 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. Importantly, it will be shown that R_T is merely a special member of this general family.

3.2 A Family of Parameters for Reliability

Alonso et al. (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 generalized 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 (1) holds then: (i) all roots of* $q(\lambda) = 0$*, the so-called generalized eigenvalues, are real, and (ii) if* λ_j *is a root of* $q(\lambda) = 0$ *then* $0 \leq \lambda_j \leq 1$*.*

A detailed proof of the previous result can be obtained from the authors. Based on this theorem, we can now define the family:

$$\Omega = \left\{ \theta : \theta = \sum_{j=1}^{p} w_j \rho_j^2 \quad \text{with} \quad w_j > 0 \quad \sum_{j=1}^{p} w_j = 1 \right\}. \tag{3}$$

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 3.1.

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 a function of the generalized eigenvalues associated with the total and error variance covariance matrices.

Note also that, even though the Ω family is uncountable, it clearly delineates our search for reliability measures. In general this is not a new situation. In other fields, concepts like the mathematical concept of distance, are defined through a minimum set of properties that lead to many specific instances. Having many elements to quantify a concept is not always undesirable. Indeed, it could allow us to approach a wide variety of problems in a very flexible way. For example, the Mahalanobis distance has been successfully used in cluster analysis and classification analysis in multivariate statistics, whereas the distance based on the uniform norm is the basic concept underlying the Kolmogorov-Smirnov test. In what follows, we will study some specific, important members of the Ω family in more detail and we will try to shed light on their specific meaning and interpretation.

3.2.1 R_T as Member of the Ω Family

It is possible to show, using the results of Graybill (1983, chapter 12), that there exists a non-singular matrix Q so that $\Sigma = (Q')^{-1}D_0Q^{-1}$ 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 (2), we obtain

$$R_T = 1 - \frac{\text{tr}[(Q')^{-1}D_0Q^{-1}]}{\text{tr}[(Q')^{-1}Q^{-1}]} = 1 - \frac{\text{tr}[Q^{-1}(Q')^{-1}D_0]}{\text{tr}[Q^{-1}(Q')^{-1}]}.$$

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

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

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

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

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

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

3.2.2 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

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

It would also 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 asymptotic confidence intervals constructed for the elements of Ω via simulation. We also try to explore whether the different members lead to intuitively plausible results in some special settings and try to clarify their interpretation.

4 Simulation Study

4.1 Design of the Simulation Study

We consider 12 different simulation settings. In a first stage, 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, $b_i \sim N(0, \sigma_b^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, 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 intuitively 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 $(b_{1i}, b_{2i})' \sim N(0, D)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, 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 (Verbeke and Molenberghs 2000), 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. This assumption is not fulfilled for R_T. Confidence intervals for R_T are calculated as described in Laenen, Alonso, and Molenberghs (2007). 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. Only when a large amount of measurement error is present and a limited number of patients is available, the asymptotic confidence intervals fail

to reach the pre-specified level of confidence. However, the problem is solved when the sample size increases.

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 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 $V = \sigma_b^2 J + \sigma^2 I$. It can be shown that in this scenario:

$$\rho^2_{(p)} = \frac{p\sigma_b^2}{p\sigma_b^2 + \sigma^2}. \tag{4}$$

From (4), it can be seen that this measure increases with the number of time points. Note that (4) fully resembles the Spearman-Brown prediction formula, where the role of the number of items is now played by the number of measurements. 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. Indeed, increasing the numbers of time points, we also increase the amount of useful information about the patient, even if it comes contaminated with measurement error. Another important implication of (4) 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. This is a very encouraging result. Given the subjective nature of psychiatric and psychological research we should expect that in many situations rating scales will be affected by considerable measurement error. The previous finding hints on the fact that such an instrument can still be very valuable if it

is used in a proper way. We could also calculate the necessary number of time points to reach a specified level of reliability $\theta_{\max}$. Indeed, in general

$$p \approx \frac{\sigma^2}{\sigma_b^2} \frac{\theta_{\max}}{1 - \theta_{\max}}.$$

Note that if we aim at a reliability of 1, p will go to infinity. The equation further shows that, as long as $\sigma_b^2 \neq 0$, it will always be possible to achieve convergence: there always will be a certain number of measurements p that results in a pre-specified value for $\theta_{\max}$.

Turning to the third measure, R_p, we observe again a totally different pattern. This measure generally gives 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 $\sigma^2 \neq 0$, $R_p = \sigma_b^2/(p\sigma_b^2 + \sigma^2)$. 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 $\theta_{\max}$ are functionally related. It can be shown that $R_p = \frac{\rho_{(p)}^2}{p} \approx \frac{\theta_{\max}}{p}$. R_p can therefore be interpreted as the average contribution per measurement to the total reliability of the whole sequence. Where large values of $\theta_{\max}$ 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 global reliability $\theta_{\max}$.

The parameter $\theta_{\min}$ 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, $\theta_{\max}$ gives the reliability of the entire sequence of measurements. Further, R_p gives the average contribution to $\theta_{\max}$ 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. Arguably, in some cases, it will be of interest to consider a few measures simultaneously. SAS-macros to obtain the different measures can be obtained from the authors.

As stated in the introduction, one of the most important attempts to estimate reliability in a longitudinal framework was based on G-theory and the use of the G coefficients. In the next section, we will study the relationship between some members of the Ω family and these G coefficients.

4.3 Relationship between the new proposals and the G coefficients

In order to quantify reliability in a longitudinal setting using the G coefficients we will assume that the following model, used in generalizability theory, holds:

$$Y_{ij} = \mu + b_i + \tau_j + \varepsilon_{ij}, \tag{5}$$

where Y_{ij} denotes the score for subject i ($i = 1 \ldots n$) at time point j ($j = 1 \ldots p$), μ denotes a constant general mean, $b_i \sim N(0, \sigma_b^2)$ is a subject-specific effect, $\tau_j \sim N(0, \sigma_\tau^2)$ denotes the time effect and the error terms are assumed independent with $\varepsilon_{ij} \sim N(0, \sigma^2)$. It is further assumed that b_i, τ_j, and ε_{ij} are independent. Under these assumptions

$\text{Var}(y_{ij}) = \sigma_b^2 + \sigma_\tau^2 + \sigma^2$ and the following G-theory coefficients, the so-called rho absolute error and rho relative error, are typically used to quantify reliability:

$$\begin{aligned} \rho_{ae} &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\tau^2 + \sigma^2}, \\ \rho_{re} &= \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_\tau^2}{p} + \frac{\sigma^2}{p}}. \end{aligned}$$

Note that, using vector notation, Model (5) can be rewritten as:

$$\boldsymbol{Y}_i = \mathbf{1}_p \mu + \mathbf{1}_p b_i + \boldsymbol{\tau} + \boldsymbol{\varepsilon}_i, \tag{6}$$

where $\boldsymbol{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$ denotes a column vector with all observations originating from subject i, $\mathbf{1}_p = (1, 1, \ldots, 1)'$ denotes a p-dimensional column vector, $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_p)'$ denotes a column vector with the time effects, and finally $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{ip})'$ denotes the column vector with all the error terms associated with subject i. Note that this model can be seen as a special case of the linear mixed model we considered. Indeed, Model (6) is a linear mixed model with only one subject-specific random effect and the error structure decomposed into a time component (which can be seen as a special type of serial correlation where the H_i matrix reduces to the identity), and a component that captures extra residual variability. Note further that in this scenario we have only one subject-specific random effect b_i and therefore for this model the variance-covariance matrix associated with the subject-specific random effects $D = \sigma_b^2$ is scalar. Using matrix notation, we can now write

$$V = \text{Var}(\boldsymbol{Y}_i) = J_p \sigma_b^2 + I_p(\sigma_\tau^2 + \sigma^2), \tag{7}$$

where $J_p = 1_p 1_p'$ and I_p is a $p \times p$ identity matrix. Employing notation previously introduced, we have $V = \Sigma_D + \Sigma$ with $\Sigma_D = J_p \sigma_b^2$ accounting for the variability coming from the subject-specific effect and $\Sigma = I_p(\sigma_\tau^2 + \sigma^2)$ accounting for the remaining variability. It now follows that

$$R_T = 1 - \frac{\text{tr}(\Sigma)}{\text{tr}(V)} = 1 - \frac{p(\sigma_\tau^2 + \sigma^2)}{p(\sigma_b^2 + \sigma_\tau^2 + \sigma^2)} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\tau^2 + \sigma^2} = \rho_{ae}.$$

In addition from (4) we can easily obtain

$$\theta_{\max} \approx \rho^2_{(p)} = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_\tau^2}{p} + \frac{\sigma^2}{p}} = \rho_{re}.$$

The previous results illustrate that, when the assumptions underlying Model (5) are met, the G-theory coefficients ρ_{ae} and ρ_{ar} can also be seen as special members of the Ω family. Similarly, it is possible to illustrate that, after integrating out the time effect, the average versions of ρ_{ae} and ρ_{ar} are also members of the Ω family.

This is a very appealing finding because the previous derivations show that both G-theory coefficients also satisfy our defining properties and can be seen as special cases of the Ω family. Given the seminal success of G-theory in many applications, these results increase our confidence in the newly proposed reliability definition and family.

5 Analysis of the Case Study

In this section, we will apply the previously introduced tools to the schizophrenia data described in Section 2. The idea is to evaluate the reliability of PANSS and BPRS when both scales are repeatedly measured over time. Notice that longitudinal measurements of both rating scales are frequently encountered in clinical trials as well as in common clinical practice.

As stated in Section 1, all reliability measures are based on a model that attempts to describe the data generating mechanism. Hence, a model building step is crucial to find the best fitting model for the data at hand. To this effect, model building guidelines, as laid out in, for example, Verbeke and Molenberghs (2000, Ch. 9) ought to be followed. We considered 15 different models. In all the cases a saturated fixed mean structure was used with one parameter for each treatment by time combination. Eventually, such a general structure for the fixed effects should help to guarantee unbiased estimates for the parame-

ters of the variance components, which are the building blocks of the reliability coefficients. We considered three different random effects, intercept, time, and time^2, allowing to flexibly model the individual evolutions over time. Furthermore, we analyzed five different structures for the error variance covariance matrix (Σ): three serial correlation structures including Gaussian, exponential, and power correlation matrices, a diagonal matrix with heterogeneous variances and a diagonal matrix with common variance.

The model selection was based on the Akaike information criterion (AIC) and restricted maximum likelihood was used for parameter estimation (Verbeke and Molenberghs 2000). For the two scales, the final model takes the general form:

$$Y_{ij} = \mu_{ij} + b_{i0} + b_{i1}t_j + b_{i2}t_j^2 + \varepsilon_{ij},$$

where Y_{ij} denotes the score (either PANSS or BPRS) for subject i at time point t_j, μ_{ij} denotes the fixed-effects structure, encompassing a parameter for each treatment by time combination. For both scales the three random effects remained in the final model, so that $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$ with D a 3×3 unstructured variance-covariance matrix. Further, $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma)$, where for PANSS the best fitting covariance structure was a diagonal matrix with heterogeneous variances, $\Sigma = \text{diag}(\sigma_j^2)$, and for BPRS, a simple structure was selected $\Sigma = (\sigma^2 I)$. The lower part of Figure 1 shows the evolution over time of the individual residuals coming from the final models. No pattern can be detected in these plots what hints on the appropriateness of chosen the models. Additionally, Figure 2 displays the observed and fitted values for five randomly selected subjects for both scales. These plots also show a reasonable level of agreement between the final models and the data.

Figure 1 shows incomplete observations for some patients. Because the model fitting has a likelihood basis, the ensuing inferences are valid for both balanced as well as unbalanced data. Also, when the data are incompletely observed, the methodology remains statistically valid if the missing data mechanism is missing at random (Rubin 1976), in the sense

that missingness is allowed to depend on observed data but, given these, not further on unobserved data.

Having found suitable models, the next step is to calculate the different measures of reliability. Table 4 presents the estimated values of R_T, R_p, and $\theta_{\max}$ for both scales, together with the corresponding 95% confidence intervals. Clearly, both scales have very high average reliabilities, characterized by estimates of R_T that largely exceed 80%. This is not a surprising result. As stated in Section 2, these scales have been successfully used in clinical practice and research for many years and the large values of the R_T are in total agreement with that fact. Moreover, the estimates of $\theta_{\max}$ are larger than 96% for both scales. Note that these large estimates may be partially explained by the large values obtained for R_T. Now, $\theta_{\max}$ being a quantification of the reliability of the entire sequence, the previous results clearly illustrate that highly reliable results can be achieved when six measurements per subject are taken.

As expected, we observe that PANSS scores higher on all three reliability measures. PANSS, with 30 items, is conceived as an extension of BPRS, a scale with only 18 items. However, the differences are generally small. The left hand graph of Figure 3 plots R_T values per time point, which are calculated as $R_{Tj} = \dfrac{z_j D z_j'}{z_j D z_j' + \sigma_j^2}$ for time point j and express the reliability at each of the measurement occasions separately. It can be observed that BPRS is performing a little better in the beginning of the study, but is outperformed by PANSS at later observations. This scale exhibits a clear increase of reliability over time. Also, BPRS finds its reliability growing over time, but much less pronounced. We speculate that this increasing reliability over time could be the result of a learning effect of the rater. Such a learning effect could also explain the relative performance of both scales at the beginning of the study. Indeed, BPRS is not only simpler than PANSS, but generally more frequently used and therefore better known by clinicians. It is therefore not surprising that it leads to more reliable results than PANSS at the beginning of the

study. This effect is reversed once the rater gets more experience in the use of PANSS somewhere after the second measurement. It is important to point out that these are just plausible interpretations of the patterns we observed but of course they are speculative. The right hand graph of Figure 3 presents the $\theta_{\max}$ values cumulatively over time. At the first time point the values are given for the first observation only, at the second time point the values express the reliability of the joint observations at the first and the second measurement, and so on. The graph shows that around 10% of information is gained by taking a second measurement. It also indicates that reliabilities above 90% can be obtained with both instruments when three measurements are taken. Additional gain in information becomes smaller as more measurements are considered.

Essentially, these results illustrate that the additional complexity of PANSS over BPRS does not bring a considerable gain in reliability. This may suggest that in some practical situations the use of a simple scale like the BPRS could be more advisable. Similar results have been found by Alonso et al. (2002) when studying criterion validity. Indeed, these authors obtained very similar values of trial-level validity and individual-level validity for BPRS and PANSS. Nevertheless, we should point out 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 directly affects the results of a statistical analysis that is based thereupon (Fleiss 1986, Lachin 2004). Therefore, reliability is a key concept 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 research and practice it is common to measure a patient's condition repeatedly over time. It is therefore important to take advantage of the available longitudinal data when estimating reliability. Laenen, Alonso, and Molenberghs (2007) extended the concept of reliability to a more general longitudinal scenario using a basic set of four properties. Further, they introduced the parameter R_T which is based on a very general class of hierarchical linear models. Notice that using such a general modelling framework is of utmost important to avoid bias when dealing with such a complex data structure.

In this paper, we have discussed the relative advantages of the modelling paradigm used by Laenen, Alonso, and Molenberghs (2007) with respect to some of the proposals available in the psychometric literature to evaluate reliability in a longitudinal setting, like G-theory. We have also shown that, within this general modelling paradigm, the R_T introduced by these authors can be seen as a special case of a more general framework defined by an entire family of reliability measures, of which all members satisfy the four defining properties. In doing so, we have established that any measure of reliability should be built from the generalized eigenvalues related to the 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 answer different scientific questions, 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.

Interestingly, under some modelling assumptions, it is possible to show that the classically

used G coefficients to evaluate reliability in a longitudinal setting are also members of the Ω family. Therefore, they can also be seen as special cases of the general scenario introduced in the present work. This is a very relevant finding and the outstanding success of G-theory in many application in psychometric and education adds an extra value to it.

Finally, all the measures considered lead to the same conclusions about the two scales under study. Both PANSS and BPRS are very reliable scales. Using these instruments in a longitudinal fashion can increase the reliability to values close to 100% as the estimates of the $\theta_{\max}$ illustrate. This clearly hints on the advantages of using this type of instruments repeatedly over time. Indeed, many rating scales in related areas perform much less spectacularly than the two scales we have considered, and provide less reliable results at one measurement occasion. Such scales might profit most of the information gain that is obtained by taking additional measurements. Another interesting conclusion that emerged from the analysis is the similar performance of both scales in spite of the additional complexity present in PANSS. It is important to point out that even though such a result may lead to the advise of using a simpler equally reliable scale like BPRS over PANSS, clinical considerations are also of vital importance when taking such a decision.

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.

Bost, J.E. (1995). The effect of correlated errors on generalizability and dependability coefficients. *Applied Psychological Measurement*, **19** (2), 191–203.

Brennan, R.L. (2001). *Generalizability Theory.* New York: Springer.

Cole, D.A., Martin, N.C., & Steiger, J.H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, **10**, 3–20.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley.

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

Diggle, P.J., Liang, K.Y., & Zeger, S.L. (1994). *Analysis of longitudinal data.* Clarendon Press: Oxford.

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.

Heise, D.R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review* **34**, 93–101.

Hertzog, C., and Nesselroade, J.R. (1987). Beyond autoregressive models: some implications of the trait-state distinction for the structural modeling of developmental change. *Child Development*, **58**, 93–109.

Jagodzinski, W. and Kühnel, S.M. (1987). Estimation of reliability and stability in single-indicator multiple-wave models. *Sociological Methods and Research* **15**, 219–258.

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

Kenny, D.A., and Zautra A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, **63** (1), 52–59.

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

Laenen, A., Alonso, A., and Molenberghs, G. (2007). A measure for the reliability of a rating scale based on longitudinal clinical trial data. *Psychometrika*, **73**, 443–448.

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.

Molenberghs, G. & Kenward, M. (2007). *Missing Data in Clinical Studies.* Chichester: Wiley.

Raykov, T. (2000). A method for examining stability in reliability. *Multivariate Behavioral Research*, **35** (3), 289–305.

Royston, P., and Altman D.G. (1994). Regression using fractional polynomials of continuous covariates: parametric modelling. *Applied Statistics*, **43** (3), 429–467.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

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

Overall, J.E. and Gorham, D.R. (1988): The Brief Psychiatric Rating Scale (BPRS): Recent developments in ascertainment and scaling. *Psychopharmacology Bulletin* **24**, 97–99.

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.

Verbyla, A.P., Cullis, B.R., Kenward, M.G., & Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics* **48**, 269–311.

Werts, C. E., Breland, H.M., Grandy, L., and Rock, D. R. (1980). Using longitudinal data to estimate reliability in the presence of correlated measurement errors. *Educational and Psychological Measurement,* **40**, 19–29.

Werts, C.E., Linn C.E., and Jøreskog, K.G. (1977). A simplex model for analyzing academic growth. *Educational and Psychological Measurement*, **37** (3), 745–756.

Wiley, D.E., & Wiley, J.A. (1970). The estimation of measurement error in panel data. *American Sociological Review*, *35*, 112–117.

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	0.890 [0.871; 0.907]	0.856 [0.839; 0.871]
R_p	0.414 [0.381; 0.448]	0.366 [0.347; 0.385]
$\theta_{\max}$	0.985 [0.968; 0.993]	0.968 [0.960; 0.975]

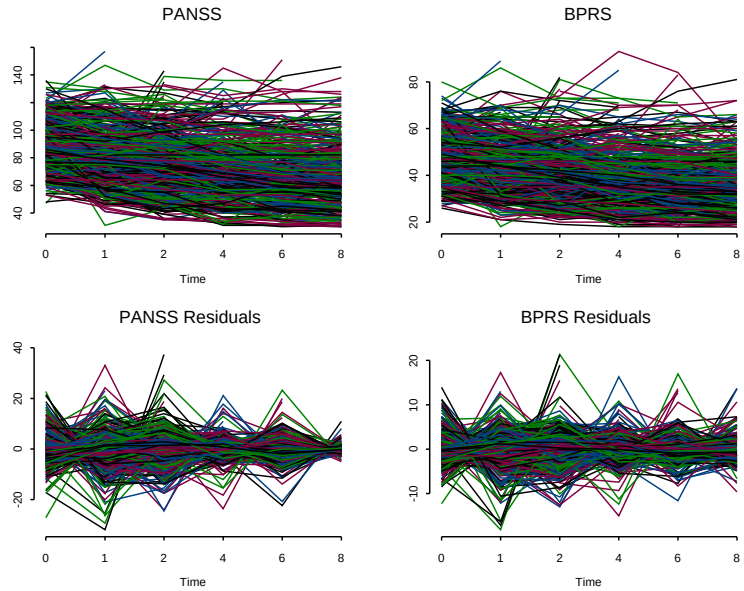

Figure 1: *Individual profiles (top) and residual profiles (bottom) for PANSS and BPRS.*

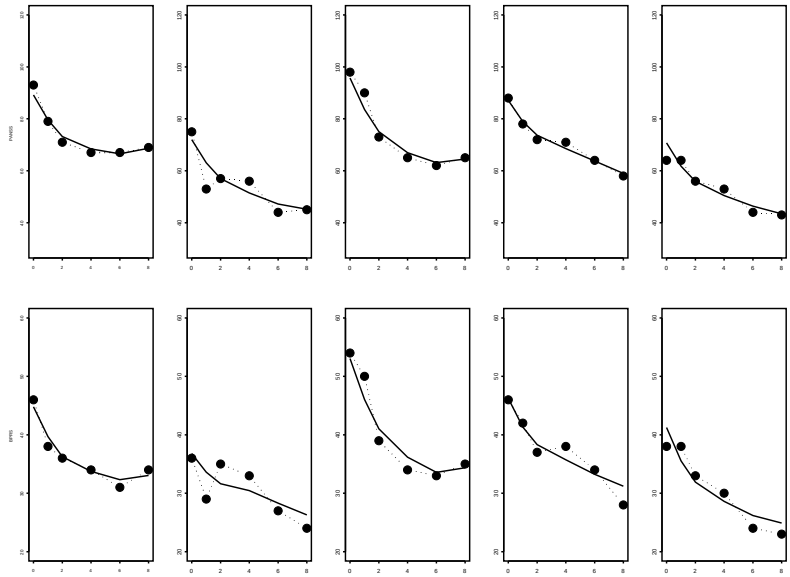

Figure 2: *Observed and fitted profiles for 5 randomly selected patients for PANSS (top) and BPRS (bottom).*

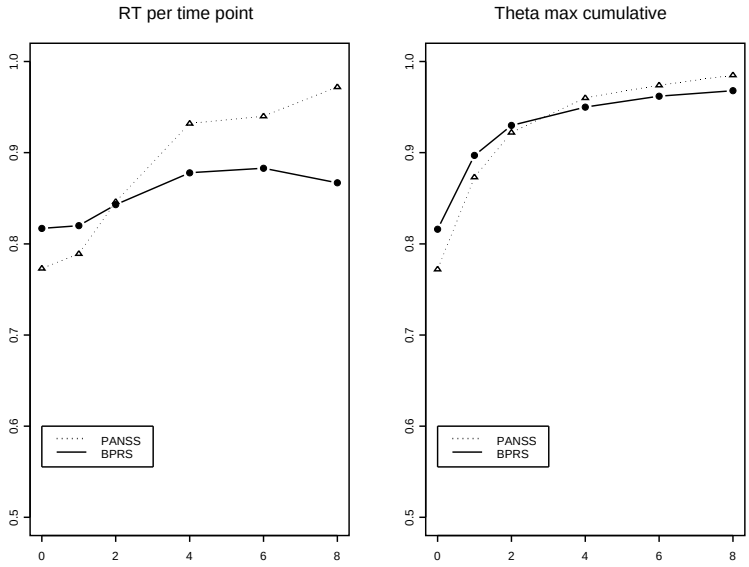

Figure 3: *R_T per time point (left) and $\theta_{\max}$ cumulative over time points (right).*