PSYCHOMETRIKA—VOL. 74, NO. 1, 49–64
MARCH 2009
DOI: 10.1007/S11336-008-9079-7

RELIABILITY OF A LONGITUDINAL SEQUENCE OF SCALE RATINGS

ANNOUSCHKA LAENEN, ARIEL ALONSO, AND GEERT MOLENBERGHS

HASSELT UNIVERSITY

TONY VANGENEUGDEN

TIBOTEC, JOHNSON & JOHNSON

Reliability captures the influence of error on a measurement and, in the classical setting, is defined as one minus the ratio of the error variance to the total variance. Laenen, Alonso, and Molenberghs (Psychometrika 73:443–448, 2007) proposed an axiomatic definition of reliability and introduced the R_T coefficient, a measure of reliability extending the classical approach to a more general longitudinal scenario. The R_T coefficient can be interpreted as the average reliability over different time points and can also be calculated for each time point separately. In this paper, we introduce a new and complementary measure, the so-called R_Λ, which implies a new way of thinking about reliability. In a longitudinal context, each measurement brings additional knowledge and leads to more reliable information. The R_Λ captures this intuitive idea and expresses the reliability of the entire longitudinal sequence, in contrast to an average or occasion-specific measure. We study the measure's properties using both theoretical arguments and simulations, establish its connections with previous proposals, and elucidate its performance in a real case study.

Key words: reliability, linear mixed model, longitudinal data, psychiatry, rating scale.

1. Introduction

Frequently, in clinical practice and clinical trials, patients are measured repeatedly over time. For instance, in psychiatry and psychology, this type of longitudinal evaluations constitutes a very powerful tool to obtain precise diagnostics as well as to evaluate the efficacy of new treatments or therapeutic procedures. However, longitudinal studies also bring some methodological challenges, especially from a statistical modeling perspective. Indeed, in such studies, patients usually exhibit a systematic change or evolution over time in addition to an individualized evolution that is characterized by correlated subject-specific effects. Moreover, serial correlation and heterogenous variance components are frequently present (Verbeke & Molenberghs, 2000). A longitudinal modeling framework should be able to address the special characteristics of this type of data in order to avoid estimation bias.

Rating scales play an important role in many scientific areas and are mainly used when the trait of interest cannot be observed directly, such as the measurement of depression, anxiety, and quality of life. Very often, these instruments are used within a longitudinal framework in view of studying patients' time course. Nevertheless, whenever a new measurement scale is developed, its validity and reliability ought to be evaluated. Reliability is not an intrinsic property of an instrument, but rather changes depending on the population in which it is used. Therefore, the reliability of a measurement scale should be evaluated every time the scale is introduced to a different population.

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

Requests for reprints should be sent to Annouschka Laenen, Hasselt University, Hasselt, Belgium. E-mail: annouschka.laenen@uhasselt.be

The conventional concept of reliability was conceived within a cross-sectional scenario, i.e., a scenario where every patient is measured on only one occasion. In this setting, the classical test theory (CTT) defines reliability as "the ratio of the true score variance to the observed score variance" (Lord & Novick, 1968). In spite of being intuitively appealing, this classical definition can be difficult to apply in longitudinal studies, mainly due to the restrictive modeling framework within which classical test theory is cast. In fact, any extension of the concept of reliability to a longitudinal scenario should take into account all the specific characteristics of this data type and the model used in CTT simply is not suitable to achieve that. Essentially, measures of reliability are model based quantities and their scope and applicability will never go beyond the scope and applicability of the model they are based on.

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. An important attempt to extend the concept of reliability to a longitudinal setting was done using generalizability theory (from now on referred to as G-theory). G-theory was developed by Cronbach, Gleser, Nanda, Rajaratnam (1972) to explicitly model the multiple sources of variation present in a measurement system. It is undoubtedly one of the most relevant developments in the psychometric field over the last 40 years and since its introduction, G-theory has grown steadily in popularity and has been applied to virtually all areas of psychology and education. The basic mathematical model on which G-theory is based, the analysis-of-variance model with random effects, is quite solid. However, the utility of G-theory to evaluate reliability in longitudinal studies depends on the adequacy of this model to describe the specific data structure encountered. Unfortunately, the G-theory modeling framework can be applied to a longitudinal setting only if strong and unrealistic assumptions are made. For instance, we will need to assume stability of the true scores over time, uncorrelated error structure, uncorrelated random effects, equal variance over time and, in its most classical formulation, it will also require a missing completely at random mechanism when data are incomplete. All the previous assumptions are quite restrictive for longitudinal studies, limiting the applicability of G-theory. 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 G coefficients (Diggle, Liang, & Zeger, 1994; Verbeke & Molenberghs, 2000; Smith & Luecht, 1992; Bost, 1995; Molenberghs & Kenward, 2007).

Hence, the main objective of our work is to extend the classical concept of reliability to a very general modeling framework that takes into account the specifics of longitudinal data. In this, we build further on earlier work by Laenen et al. (2007). These authors introduced the R_T coefficient, expressing the average reliability over the time points in a longitudinal framework. However, the same methodology allows to obtain occasion-specific reliability estimates. Where the latter can be useful for studying reliability over time, a single reliability estimate has the benefit of simplicity, mainly when a large number of measurements is taken, or in case two rating scales are being compared. In this paper, we introduce the R_Λ coefficient, a new measure for reliability which is complementary to the R_T coefficient. In a longitudinal context, each measurement brings additional information. One could argue that the total information ought to be more reliable than the information obtained at separate time points. The R_Λ coefficient corresponds to this idea and expresses the reliability of the entire longitudinal sequence, in contrast to an average, or occasion-specific measure. One could say that this measure gives shape to the clinical premise stating that the longer we study and observe a patient, the more reliable our conclusions relative to that patient will be.

Section 2, starts by citing the underlying methodological framework and summarizing previous work in this context and continues with the elaboration of a new proposal for measuring reliability. In Sect. 3, established and novel measures are studied and compared by means of a simulation study. In Sect. 4, the previously developed methods are applied to a case study on schizophrenia.

2. Methodology

2.1. *The Linear Mixed Model*

In the previous section, we described some of the limitations of the modeling framework used in G-theory when applied within a longitudinal framework. In the last decades, a substantial amount of work has been carried out to try and solve some of the modeling problems discussed previously. Many of these proposals are primarily based on path analysis or structural equations, and have been developed to estimate reliability in a longitudinal setting without assuming stability of the true scores (Heise, 1969; Jagodzinski & Kühnel, 1987; Werts, Breland, Grandy, & Rock, 1980; Wiley & Wiley, 1970). Nevertheless, to evade the requirement of true score stability when estimating reliability, these models often impose additional assumptions that may have questionable validity in this setting. For instance, it is usually assumed that the changes in the true scores across time follow a simplex pattern (Heise, 1969; Wiley & Wiley, 1970; Werts, Linn, & Jøreskog, 1977).

Additionally, some of these approaches also make strong assumptions regarding the pattern of measurement errors across time such as equal reliabilities across time (Heise, 1969), equal error variances over time (Wiley & Wiley, 1970) or uncorrelated error structures (Tisak & Tisak, 1996). Raykov (2000) criticizes the equal-reliability assumption of Heise (1969) and proposed a model that circumvents this limitation. Nevertheless, his model still assumed uncorrelated error terms. Many other authors have focused on the advantages and disadvantages of using a first-order autoregressive structure to describe within-subject evolution over time (Kenny & Zautra, 1995; Hertzog & Nesselroade, 1987; Cole, Martin, & 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 linear mixed model formulation), a term representing the state which is equivalent to the serial correlation component in linear mixed model (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 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.

Against the background of Laenen, Alonso, and Molenberghs (2007) and Vangeneugden, Laenen, Geys, Renard, and Molenberghs (2004), we will set our proposals for quantifying reliability within a linear mixed models framework. These allow 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. Being able to account 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 effect structure of the model in a very flexible way using, for example, fractional polynomials (Royston & Atman, 1994) or nonparametric approaches such as splines (Verbyla, Cullis, Kenward, & Welham, 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, and m-dependent structure among others. More general serial correlations have been accommodated by Verbeke and Molenberghs (2000). These authors have described different techniques for exploring serial correlation within the LMM framework and have offered flexible tools to model serial correlation based on fractional polynomials. The assumption of equal error variance over time can also be relaxed easily and fully general variance functions can be considered. A linear mixed-effects model can generally be written as

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

where $\boldsymbol{Y}_i$ is the p_i dimensional vector of responses for subject i, $1 \le i \le n$, with n denoting 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, $\boldsymbol{\varepsilon}_{(2)_i}$ is a p_i-dimensional vector of components of serial correlation, and $\boldsymbol{\varepsilon}_{(1)_i}$ 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 et al. (2007) to quantify reliability in this very general scenario. Further, we will argue for the utility of having a measure that is able to capture the reliability of the entire sequence of observations and finally, we will introduce and study such a measure. Even though it is not strictly necessary from a mathematical point of view, from now on we will only consider the single-trial setting with a balanced design, and we will assume that $\Sigma_i = \Sigma$, $\Sigma_{Di} = \Sigma_D$, and $V_i = V$ for all i. This will considerably simplify notation and facilitate interpretation. Note further that these assumptions are usually met in clinical trials, the premier environment we are working in.

2.2. *Summarizing the Concept of R_T*

Using the just-introduced modeling framework, Laenen et al. (2007) extended the classical concept of reliability through a set of defining properties. According to these authors, any measure of reliability θ, defined in a general setting, should satisfy: (i) $0 \le \theta \le 1$; (ii) $\theta = 0$ if and only if there is only measurement error: $V = \Sigma$; (iii) $\theta = 1$ if and only if there is no measurement error: $\Sigma = 0$; and (iv) in the cross-sectional setting, θ should equal the true-score variance to observed variance ratio, used in classical test theory. 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 probability density function.

Further, these authors propose the so-called R_T coefficient, a measure of reliability that satisfies the previous properties. The R_T coefficient is defined as $R_T = 1 - \frac{\mathrm{tr}(\Sigma)}{\mathrm{tr}(V)}$. Additionally, Laenen, Alonso, Molenberghs, and Vangeneugden (2009) showed that the R_T coefficient can be incorporated into a more general framework, based on the generalized eigenvalues related to the matrices V and Σ. Indeed, these authors introduced the following family of reliability parameters:

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

where the elements w_j are weights, $\rho_j^2 = 1 - \lambda_j$ and λ_j are solutions of the equation

$$q(\lambda) = |\Sigma - \lambda V| = 0. \tag{3}$$

All members of Ω satisfy the defining properties (i)–(iv) and one can further show that the R_T coefficient belongs to this family. Note that even though it is not an explicit requirement of the properties (i)–(iv), the R_T coefficient mimics the general functional form of the classical definition of reliability. Indeed, the trace of the variance-covariance matrix is usually regarded in multivariate analysis as a plausible generalization of the univariate concept of variance. From this perspective, it is easy to see that the functional form of the R_T coefficient is very similar to the one used in CTT. In the next section, we will summarize the variability in this multivariate setting, using another plausible generalization of the concept of variance: the determinant of the variance-covariance matrix. Remarkably enough, such a change leads to a completely new measure of reliability, with different mathematical properties and interpretation.

2.3. R_Λ: *Measuring the Reliability of an Entire Sequence*

As stated before, in multivariate analysis, the generalized variance of a random vector can be defined using either the trace or the determinant of the corresponding variance-covariance matrix. Replacing the traces in the definition of the R_T coefficient by the determinant of the variance-covariance matrix leads to the following expression for reliability:

$$R_\Lambda = 1 - \left|\Sigma V^{-1}\right|. \tag{4}$$

Note that R_Λ is closely related to the Wilks' Lambda statistic (Johnson & Wichern, 1998), well known in multivariate analysis. It can be proven that the parameter R_Λ satisfies the following set of properties: (a) $0 \le R_\Lambda \le 1$; (b) $R_\Lambda = 0$ if and only if there is only measurement error: $V = \Sigma$; (c) $R_\Lambda = 1$ if and only if $|\Sigma| = 0$; and (d) in the cross-sectional setting, the true score variance to observed variance ratio is obtained. Properties (a), (b), and (d) are identical to properties (i), (ii), and (iv). However, property (c) is slightly different from (iii). Only when $\Sigma = \sigma^2 I$, (c) and (iii) are equivalent. In a sense, (a)–(d) contain (i)–(iv) in that if θ satisfies (i)–(iv), then it will also satisfy (a)–(d) and, therefore, the later provides a more flexible defining set of properties for reliability.

To acquire a better insight into the R_T and R_Λ coefficients, as well as to better understand their relationship, we will study their behavior in an important special case, the random intercept model. Let us then start by assuming that Model (1) holds with $b_i \sim N(0, \sigma_b^2)$ and $\varepsilon_{(1)i} + \varepsilon_{(2)i} = \varepsilon_i \sim N(0, \sigma^2 I)$. It is easy to prove that in this setting,

$$R_T = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}, \tag{5}$$

i.e., the ratio of true-score variance to total variance. It is important to point out that the usual approach followed when estimating reliability in a longitudinal framework is based on the calculation of the reliability at each time point separately (Tisak & Tisak, 1996; Wiley & Wiley, 1970; Raykov, 2000). This typically leads to a function of reliability that changes over time. Note further that both the R_T and the R_Λ coefficients can also be calculated at each time point, leading again to a general function of reliabilities across time. However, they also offer a global measure of reliability that nicely complements their time functions. We believe this is an important issue because having a global measure of reliability, valid under such a general scenario, can substantially aid when comparing and interpreting the results obtained from two or more scales and can facilitate the understanding of their psychometric properties. It is intuitively clear that a single meaningful measure is much easier to analyze, understand, and interpret than several functions of changing reliabilities over time.

To understand how the R_T coefficient summarizes these reliabilities over time, let us note that under the assumed model, subsequently applying the classical definition of reliability at each time point leads to (5). Therefore, for this model, reliability is stable across time. The R_T coefficient can thus be seen as an average of the time-point reliabilities. If we now calculate R_Λ under the same model, it can be shown that

$$R_\Lambda = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma^2}{p}}. \tag{6}$$

This expression is very interesting from a theoretical as well as practical point of view. First, let us note that (6) is similar in spirit to the Spearman–Brown prediction formula (Spearman, 1910; Brown, 1910) in the sense that reliability increases with an increasing number of observations. A second important issue is that R_Λ goes to one as the number of time points goes to infinity. This shows that unlike the R_T coefficient, the R_Λ coefficient does not capture the average reliability but rather the reliability of the sequence as a whole. Increasing the numbers of time points, we also increase the amount of useful information about the patient, even if it is contaminated by measurement error. Actually, (6) confirms a clinical truth: the longer we follow a patient, the more reliable will be our conclusions about this patient. The practical implications of this result are also appealing. We can study the number of repeated measurements needed to obtain a certain level of reliability R_Λ, the number can be derived as $p = \frac{\sigma^2}{\sigma_b^2}\frac{R_\Lambda}{1-R_\Lambda}$. 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 prespecified value for R_Λ. We have used the random-intercept model to gain more insight into the meaning of the R_Λ coefficient. However, the assumptions on which this model is based will be too restrictive in many real applications. The following theorem extends the previous result to a totally general scenario and confirms our interpretation for this measure.

Theorem 1. *Let us assume that Model (1) holds for a balanced study design in which p time points have been considered. Further, let us denote by $R_{\Lambda(p)}$ the corresponding value of the R_Λ coefficient in this setting. If q additional observations are taken, then the new value of the R_Λ coefficient for the $p+q$ time points sequence satisfies $R_{\Lambda(p+q)} \geq R_{\Lambda(p)}$.*

The proof of the theorem can be found at the authors' web page, www.uhasselt.be/censtat. Theorem 1 proves in a very general setting that increasing our information about the patients can only increase the reliability of our conclusions, a very plausible and appealing result. We believe that in a longitudinal framework this measure is more attractive than the classical reliability functions previously proposed. Indeed, the main objective of longitudinal studies is to get information from the entire profile and not to analyze each time point separately. The R_Λ coefficient quantifies precisely this, i.e., the reliability of the whole profile we have at hand.

2.4. The Relationship between R_Λ and Ω

We have seen that every member of the Ω family (2) is a weighted sum of the $\rho_j^2 = 1 - \lambda_j$. Also, R_Λ can be written as a function of these elements. In fact, it is possible to show that:

$$R_\Lambda = 1 - \prod_{j=1}^{p}\left(1 - \rho_j^2\right).$$

R_Λ can further be seen as an upper bound for Ω. Indeed, if $w_j > 0$ and $\sum w_j = 1$ then:

$$\sum_{j=1}^{p} w_j\lambda_j \geq \prod_{j=1}^{p} \lambda_j^{w_j} \geq \prod_{j=1}^{p} \lambda_j.$$

Note that the first part of the inequality is the general form of the well-known relationship between the arithmetic and geometric means, whereas the second part comes from the fact that if $0 \leq w_j \leq 1$ then $\lambda_j^{w_j} \geq \lambda_j$. From this expression, we have:

$$\theta = 1 - \sum_{j=1}^{p} w_j\lambda_j \leq 1 - \prod_{j=1}^{p} \lambda_j = R_\Lambda.$$

This final inequality shows that $\theta \leq R_\Lambda$ for all $\theta \in \Omega$ and, therefore, the R_Λ coefficient can be interpreted as an upper bound for the family. This result totally coincides with our interpretation of the R_Λ coefficient as a measure of reliability for the entire sequence and our interpretation of the Ω family as summary measures of "average" reliability.

2.5. Reliability as a Measure of Association between True and Observed Scores

As pointed out in the previous section, the Ω family and R_Λ are built based on the same basic elements. Nevertheless, the practical interpretation of the ρ_j^2 is not totally clear. In the present section, we approach reliability from an alternative point of view that will help us to clarify the role and interpretation of the ρ_j^2. In classical test theory, Lord and Novick (1968) have proven that reliability equals the squared correlation between the observed score Y_i and the true score T_i, i.e., $R = \text{Corr}(Y_i, T_i)^2$. It would then be natural to explore whether such a connection also holds in the more general scenario considered here. Therefore, in what follows, we will study the relationship between the measures of reliability previously introduced and the squared association between $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$. Let us start by defining $\boldsymbol{S}_i = (\boldsymbol{Y}_i\ \boldsymbol{b}_i)'$. The following theorem will allow us to quantify this association.

Theorem 2. *If Model* (1) *holds, then* $\boldsymbol{S}_i \sim N(\boldsymbol{\mu}_{0i}, \Sigma_{0i})$, *where*

$$\boldsymbol{\mu}_{0i} = \begin{pmatrix} X_i\boldsymbol{\beta} \\ \boldsymbol{0} \end{pmatrix} \quad \text{and} \quad \Sigma_{0i} = \begin{pmatrix} V_i & Z_i D \\ (Z_i D)' & D \end{pmatrix}.$$

The proof of this result is posted at the authors' web pages. Since $\boldsymbol{S}_i$ is multivariate normally distributed, canonical correlations are a natural way to quantify the association between $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$ (Johnson & Wichern, 1998). As before, we consider the case of a balanced clinical trial, where $\Sigma_{0i} = \Sigma_0$, $V_i = V$, and $Z_i = Z$, where $V = ZDZ' + \Sigma$. The canonical correlations associated with $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$ are then the eigenvalues of the matrix: $V^{-1}ZDD^{-1}DZ' = V^{-1}ZDZ' = V^{-1}(V - \Sigma) = I - V^{-1}\Sigma$. Moreover, it is possible to prove that if λ is a solution of (3), then it is also a root of the equation $|I - V^{-1}\Sigma - (1-\lambda)I)| = 0$. Thus, $1 - \lambda$ is an eigenvalue of $I - V^{-1}\Sigma$. This result shows that ρ_j^2 are just the canonical correlations between $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$. It is appealing to see that two equivalent classical definitions of reliability also concur in this extended setting. However, a high-dimensional vector of canonical correlations may be difficult to interpret and difficult to use when comparing two scales regarding their reliabilities. Therefore, aiming at easy interpretation, we have summarized the information about the reliability, contained in the canonical correlation vector, using meaningful functions of its elements. All the previous results clearly illustrate that like in classical test theory, the newly introduced reliability measures can also be interpreted as quantifications of the association between the true and observed scores.

2.6. *Relationship between the New Proposals and the G Coefficients*

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. Now, we will study the relationship between the R_T, the R_Λ, and the G coefficients when the assumptions of the G-theory modeling framework are met.

Let us then consider that the following model, used in generalizability theory holds:

$$Y_{ij} = \mu + b_i + \tau_j + \varepsilon_{ij}, \tag{7}$$

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 $\mathrm{Var}(Y_{ij}) = \sigma_b^2 + \sigma_\tau^2 + \sigma^2$ and the index of dependability for absolute decisions is:

$$\Phi = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_\tau^2}{p'} + \frac{\sigma^2}{p'}},$$

with p' the number of time points considered in the D-study. Note that using vector notation, Model (7) can be rewritten as:

$$\boldsymbol{Y}_i = \mathbf{1}_p \mu + \mathbf{1}_p b_i + \boldsymbol{\tau} + \boldsymbol{\varepsilon}_i, \tag{8}$$

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. This model can be seen as a special case of the linear mixed model we considered. Indeed, Model (8) 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 variably. As we stated before, it is important to differentiate the variability emanating from the subject-specific random effects and the one coming from other sources. In this case, 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

$$V = \mathrm{Var}(Y_i) = J_p \sigma_b^2 + I_p\bigl(\sigma_\tau^2 + \sigma^2\bigr), \tag{9}$$

where $J_p = \mathbf{1}_p \mathbf{1}_p'$ 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

$$R_T = 1 - \frac{\mathrm{tr}(\Sigma)}{\mathrm{tr}(V)} = 1 - \frac{p(\sigma_\tau^2 + \sigma^2)}{p(\sigma_b^2 + \sigma_\tau^2 + \sigma^2)} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\tau^2 + \sigma^2} = \Phi$$

for $p' = 1$. Further, we can calculate the value of R_Λ for what we will need the following result

$|aI_p + bJ_p| = a^{p-1}(a + pb)$ (Searle, 1982). We now can write

$$R_\Lambda = 1 - \frac{|\Sigma|}{|V|} = 1 - \frac{(\sigma_\tau^2 + \sigma^2)^p}{(\sigma_\tau^2 + \sigma^2)^{p-1}(p\sigma_b^2 + \sigma_\tau^2 + \sigma^2)}$$

$$= 1 - \frac{\sigma_\tau^2 + \sigma^2}{p\sigma_b^2 + \sigma_\tau^2 + \sigma^2} = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_\tau^2}{p} + \frac{\sigma^2}{p}} = \Phi$$

for $p' = p$. The previous result illustrates that, when the assumptions underlying Model (7) are met, the R_T and the R_Λ coefficients equal the index of dependability for 1 and p time points correspondingly. Note that this is in full agreement with the previous interpretations of the R_T and R_Λ coefficients as the average reliability and the reliability of the entire sequence, respectively. A very similar proof can be constructed to illustrate that after conditioning on the time points, the R_T and R_Λ equal the generalizability coefficients for relative decisions

$$E\rho^2 = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma^2}{p'}}.$$

These are a very appealing results. Indeed, the previous derivations show that G-theory coefficients also satisfy our defining properties and given the seminal success of G-theory in many applications, these results increase our confidence in the newly proposed reliability definition and coefficients.

3. A Simulation Study

A simulation study was set up to investigate the performance of both, the R_Λ and R_T coefficients. We considered 36 different simulation settings. In a first stage, the data were generated based on the following linear mixed model with random intercept:

$$Y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 Z_i + b_i + \varepsilon_{ij}, \qquad (10)$$

where Y_{ij} refers to an observation for subject i at time t_j, and Z_i is the treatment indicator variable. Further, $b_i \sim N(0, \sigma_b^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2 I)$, $\sigma_b^2 = 300$, $\sigma^2 = 30$, 300, or 3,000 and the sample size was set to $n = 50$ or 150. These choices for σ_b^2 and σ^2 allow us to study the performance of both measures R_T and R_Λ when the error variance is 9%, 50%, and 90% of the total variance, respectively. These settings correspond to high, medium, and low reliability. In a second stage, data were generated based on a linear mixed model with random intercept and random slope for time:

$$Y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 Z_i + b_{1i} + b_{2i} t_j + \varepsilon_{ij}, \qquad (11)$$

where $(b_{1i}, b_{2i})' \sim N(0, D)$, and $\varepsilon_{ij} \sim N(0, \sigma^2 I)$. D contains $\sigma_{b1}^2 = 300$, $\sigma_{b2}^2 = 5$, and $\sigma_{b1b2} = -1$. The same choices for σ^2 and n are made as before. The norm $||D||$ was used as an indication of the "size" of the random-effects variance and based on this, the values of the error variance account again for 9%, 50%, and 90% of the total variance.

The mean parameters were fixed at $\beta_0 = 85$, $\beta_1 = 2.5$, and $\beta_2 = 3$. These values are in line with the results obtained in the analysis of the case study. We considered $p = 3$, 6, and 9 time points of measurement and 500 data sets were simulated in each setting.

Tables 1 and 2 show the results of the simulation study for the random intercept model (RI) and random intercept and slope model (RIS), respectively. Confidence intervals were calculated using matrix differential calculus, combined with the delta method. More information on the calculations can be found in Laenen et al. (2007) for R_T and in Appendix for R_Λ.

TABLE 1.
Simulation Results for RI model: R_T and R_Λ average point estimates and 95% confidence intervals for different error percentages (%), number of time points (p), and sample sizes (50, 150).

%	p	R_T		R_Λ	
		50	150	50	150
9	3	0.91 [0.86; 0.94]	0.91 [0.89; 0.93]	0.97 [0.95; 0.98]	0.97 [0.96; 0.98]
9	6	0.90 [0.86; 0.93]	0.91 [0.89; 0.93]	0.98 [0.97; 0.99]	0.98 [0.98; 0.99]
9	9	0.90 [0.86; 0.93]	0.91 [0.89; 0.93]	0.99 [0.98; 0.99]	0.99 [0.99; 0.99]
50	3	0.50 [0.37; 0.64]	0.50 [0.43; 0.58]	0.75 [0.61; 0.85]	0.75 [0.68; 0.81]
50	6	0.49 [0.38; 0.61]	0.50 [0.43; 0.57]	0.85 [0.77; 0.90]	0.86 [0.82; 0.89]
50	9	0.49 [0.38; 0.60]	0.50 [0.44; 0.56]	0.89 [0.84; 0.93]	0.90 [0.87; 0.92]
90	3	0.10 [0.04; 0.52]	0.10 [0.04; 0.30]	0.24 [0.09; 0.74]	0.24 [0.10; 0.55]
90	6	0.09 [0.04; 0.28]	0.09 [0.05; 0.16]	0.35 [0.17; 0.69]	0.38 [0.24; 0.54]
90	9	0.09 [0.04; 0.20]	0.09 [0.06; 0.14]	0.45 [0.26; 0.69]	0.47 [0.35; 0.60]

TABLE 2.
Simulation Results for RIS model: R_T and R_Λ average point estimates and 95% confidence intervals for different error percentages (%), number of time points (p), and sample sizes (50, 150).

%	p	R_T		R_Λ	
		50	150	50	150
9	3	0.91 [0.87; 0.94]	0.92 [0.90; 0.93]	0.98 [0.96; 0.99]	0.99 [0.98; 0.99]
9	6	0.94 [0.91; 0.96]	0.94 [0.93; 0.95]	1.00 [1.00; 1.00]	1.00 [1.00; 1.00]
9	9	0.96 [0.94; 0.97]	0.96 [0.95; 0.97]	1.00 [1.00; 1.00]	1.00 [1.00; 1.00]
50	3	0.54 [0.40; 0.67]	0.53 [0.45; 0.62]	0.79 [0.59; 0.91]	0.79 [0.67; 0.87]
50	6	0.61 [0.50; 0.70]	0.61 [0.55; 0.67]	0.94 [0.88; 0.97]	0.94 [0.91; 0.96]
50	9	0.70 [0.62; 0.78]	0.71 [0.66; 0.75]	0.98 [0.96; 0.99]	0.98 [0.97; 0.99]
90	3	0.16 [0.06; 0.53]	0.14 [0.06; 0.34]	0.33 [0.11; 0.82]	0.31 [0.13; 0.65]
90	6	0.15 [0.07; 0.29]	0.14 [0.09; 0.22]	0.50 [0.26; 0.76]	0.50 [0.34; 0.67]
90	9	0.20 [0.12; 0.30]	0.20 [0.12; 0.30]	0.70 [0.51; 0.84]	0.71 [0.51; 0.84]

Table 1 clearly illustrates that the values for R_T, based on a random intercept model with homogeneous error variances do not depend on the number of time points, as shown in (5). Not surprisingly, confidence intervals tend to be narrower with larger sample sizes and with increasing number of time points.

The values for R_Λ, under the same model, increase with the number of time points. When the error variability is relatively small compared to the total variability (9%), we get high values for both the R_T and the R_Λ coefficients. However, when the error variability is 50% of the total variability, unlike with the R_T coefficient, we still obtain very high values for R_Λ in case 6 or 9 measurements are taken. This means that when there is a lot of measurement error, still very reliable information can be obtained over the whole sequence of measurements when the measurement is repeated a sufficient number of times. Even repeating the measurement three times, the combined information could be considered as reliable ($R_\Lambda = 0.79$). A similar result is found for R_Λ under the random intercept and random slope model (Table 2).

Under this model, however, we find an interesting result for the R_T coefficient. Table 2 shows an increase of R_T for an increasing number of time points. In what follows, we will further explore the relationship between R_T and p.

Let us assume that Model (1) holds and that information from p time points is available. We can then write R_T as

$$R_{T(p)} = 1 - \frac{\mathrm{tr}(\Sigma_p)}{\mathrm{tr}(\Sigma_{Dp}) + \mathrm{tr}(\Sigma_p)} = 1 - \frac{\frac{\mathrm{tr}(\Sigma_p)}{\mathrm{tr}(\Sigma_{Dp})}}{1 + \frac{\mathrm{tr}(\Sigma_p)}{\mathrm{tr}(\Sigma_{Dp})}} = 1 - \frac{x_p}{1 + x_p},$$

with $x_p = \mathrm{tr}(\Sigma_p)/[\mathrm{tr}(\Sigma_{Dp})]$. If we define $f(x_p) = x_p/(1 + x_p)$, then $f(x_p)$ is an increasing function of x_p, and $R_{Tp} = 1 - f(x_p)$ decreases when x_p increases. If information on an additional time point $p + 1$ becomes available, then

$$x_{p+1} = \frac{\mathrm{tr}(\Sigma_p) + \sigma^2_{p+1}}{\mathrm{tr}(\Sigma_{Dp}) + z_{p+1} D z'_{p+1}}.$$

It is easy to show that $x_p > x_{p+1}$, and thus $R_{T(p)} < R_{T(p+1)}$ if and only if

$$\frac{\mathrm{tr}(\Sigma_p)}{\mathrm{tr}(\Sigma_{Dp})} > \frac{\sigma^2_{p+1}}{z_{p+1} D z'_{p+1}}.$$

We thus see that R_T can both increase and decrease when adding an extra measurement, the direction depending on the ratio between the error and true score variability in such an extra measurement. This is an intuitively logical result since R_T can be seen as an "average" reliability over a whole sequence of measurements. Adding an extra measurement which error-true score variance ratio is "worse" than the same ratio for the previous p measurements will make the average reliability go down. Adding a measurement of which the error-true score variance ratio is better than the same ratio for the previous p measurements will result in the opposite effect. We will illustrate this finding with an additional simulation study.

We revisit the results of the simulations following the RI model (10) for $p = 3$, as presented in Table 1. We further generated data with an extra time point ($p = 4$), in such a way that the extra measurement satisfies

$$\frac{\mathrm{tr}(\Sigma_p)}{\mathrm{tr}(\Sigma_{Dp})} < \frac{\sigma^2_{p+1}}{z_{p+1} D z'_{p+1}},$$

or equivalently, under the present model

$$\frac{\sum_{j=1}^{p} \sigma_j^2}{p} < \sigma^2_{p+1},$$

thereby expecting R_T to decrease compared to the results displayed in Table 1. Precisely, the data were generated based on Model (10), where $b_i \sim N(0, \sigma_b^2)$, $\varepsilon_{ij} \sim N(0, \Sigma)$, with $\sigma_b^2 = 300$, and with Σ a diagonal matrix with the first three diagonal elements equal to σ^2 and the fourth diagonal element equal to $2\sigma^2$, $\sigma^2 = 30$, 300, and 3,000. Figure 1 summarizes our findings. We indeed observe that the values for R_T decrease with a larger number of time points. However, the values of R_Λ increase.

This simulation study illustrates once more that the R_T and R_Λ coefficients should be interpreted in different ways. R_T is an average reliability taken over a number of measurements. Adding time points with "low" reliability will pull the average down, adding "reliable" measurements will lift the average up. Unlike R_T, the R_Λ coefficient quantifies the reliability of the whole sequence of measurements. Adding more measurements to the sequence will never decrease our total information about the true scores. Obviously, the magnitude of the increase will depend on

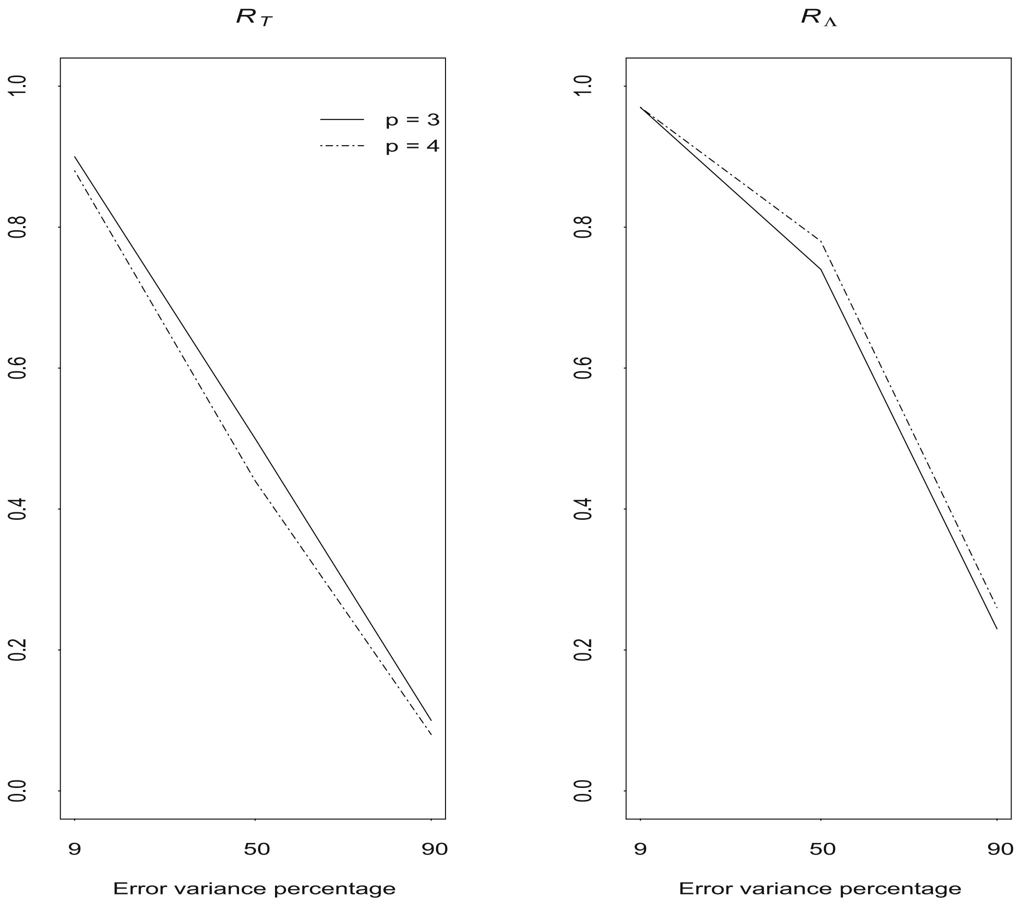

FIGURE 1.
Simulation study. R_T and R_Λ for additional time point with large error variability.

the amount of measurement error that contaminates the new observations. Adding measurements with little measurement error will lead to a faster increase of R_Λ than what measurements with a lot of measurement error would do.

4. Analysis of the Case Study

The case study data come from a clinical trial investigating the effect of risperidone compared to an active control for the treatment of chronic schizophrenia (Peuskens & the Risperidone Study Group, 1995). Three different rating scales were used as outcome measures; the Brief Psychiatric Rating Scale (BPRS) containing 18 items, the Positive and Negative Syndrome Scale (PANSS) with 30 items and including all items of the BPRS, and the Clinical Global Impression (CGI), which is a global rating of the change of the patient's condition compared to baseline measurement. A sample of 453 patients was evaluated at baseline and after 1, 2, 4, 6, and 8 weeks. In our analysis, both reliability coefficients, R_T and R_Λ, are calculated based on the covariance parameter estimates resulting from fitting a linear mixed model. A model building step is therefore crucial to find the best fitting model for the data at hand. Since interest primarily lies in the covariance structure, a complex fixed effects structure was adopted (Diggle, Liang, &

TABLE 3.
Schizophrenia Study: estimates and 95% confidence intervals for R_T and R_Λ.

	PANSS	BPRS	CGI
R_T	0.846 [0.825; 0.865]	0.821 [0.797; 0.842]	0.737 [0.700; 0.771]
R_Λ	0.994 [0.992; 0.996]	0.991 [0.988; 0.993]	0.977 [0.969; 0.983]

Zeger, 1994), containing time (categorically), treatment, and treatment by time interaction. The covariance structure that lead to the lowest AIC was selected. Restricted maximum likelihood was used for parameter estimation (Verbeke & Molenberghs, 2000). For all three scales, the final model takes the general form:

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

where Y_{ij} denotes the outcome for subject i at time point t_j, μ_{ij} summarizes the fixed effects, $\boldsymbol{b}_i \sim N(\mathbf{0}, D)$ with D a 2×2 unstructured variance-covariance matrix, and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \Sigma)$. For PANSS and BPRS, the best fitting covariance structure for the errors corresponds to $\Sigma = \text{diag}(\sigma_j^2)$. However, for CGI, $\Sigma = \tau^2 H$, with H corresponding to a spatial power serial correlation structure.

Table 3 presents the reliability estimates for both measures together with the 95% confidence interval. The R_T coefficient reveals good average reliabilities for the PANSS and BPRS scales, somewhat lower, but still acceptable is the reliability for CGI. This difference is to be expected since the former two scales are multi-item scales whereas CGI is a one-item instrument. Furthermore, it is interesting that the R_T values for both multi-item scales are very similar, even though BPRS is a subscale of PANSS and, therefore, simpler. This implies that the additional complexity of PANSS does not translate into a reliability gain. Note that similar findings were obtained by Alonso, Geys, Molenberghs, and Vangeneugden (2002) and Alonso, Geys, Molenberghs, and Kenward (2004) when investigating the criterion validity of the same scales.

When the entire sequence of repeated measurements is considered, very reliable information can be obtained from all three scales, with R_Λ values all close to 1. This is not unexpected since the average reliabilities for the scales (R_T) are already high and we have a fair number of time points. The R_Λ coefficient of CGI stays the lowest (however, still high), due to a lower average reliability and a smaller number of time points.

5. Discussion

In the present work, we have approached the problem of defining and estimating reliability within a longitudinal framework. Longitudinal studies are regularly used in clinical practice and research, especially in psychiatry and psychology, where they constitute a very powerful tool for diagnostic purpose and for evaluation. However, longitudinal data, with their various sources of variability and correlation, introduce many challenges into the estimation of reliability.

The classical definition of reliability, proposed in classical testing theory, is defined within a cross-sectional framework. It is based on a very simple model that cannot be used in a longitudinal scenario and, therefore, extensions are needed. We have proposed such extensions, making use of the linear mixed model framework. We have introduced a set of defining properties and further proposed a new measure, the so-called R_Λ coefficient, that fully captures the reliability of the entire sequence of observations.

We have mainly focused on studies where a balanced design was used. Indeed, balanced designs are very frequently used in clinical trials and even though missing data could break the balance originally designed, this will not represent a serious problem within the modeling framework used in this research. Indeed, linear mixed models are fitted using the marginal likelihood obtained after integrating out the random effects. It has been established that maximum likelihood and Bayesian methods are valid under a missing at random generating mechanism (MAR, Rubin, 1976) unlike frequentist methods such as least squares or generalized estimating equations (Liang & Zeger, 1986) that are only valid under the more restrictive missing completely at random mechanism (MCAR) (Rubin, 1976; Verbeke & Molenberghs, 2000; Molenberghs & Kenward, 2007).

Hence, using a likelihood approach, such as the LMM, will permit us to estimate in an unbiased fashion the Σ and V matrices in a balanced design with missing data, to the extent that the MAR assumption is plausible. This further implies that the R_Λ coefficient and all measures previously introduced can be unbiasedly estimated in this setting, too. It has been argued that MAR is a reasonable and plausible assumption in many clinical trials and, therefore, these measures will be applicable in many practical setting (Verbeke & Molenberghs, 2000; Molenberghs & Kenward, 2007).

In an unbalanced design, the concept of missing data becomes more diffuse because patients have a different number of observations taken at different time points and these measurements frequently do not follow a prespecified plan, making it more troublesome to precisely define what is meant by a missing observation. In such a setting, we could still estimate the necessary matrices, Σ_i and V_i, to calculate reliability. However, now they will vary across subjects and, as a consequence, an extra index i is needed. We could still define the R_T and R_Λ coefficients using the average of all subject contributions. Nevertheless, the interpretation of these quantities will become less clear in such a scenario.

Appendix: A Confidence Interval for R_Λ

If $\hat{D}$, $\hat{\tau}$, $\hat{H}$, $\hat{\Sigma}_R$ denote the maximum likelihood estimators (MLE) for D, τ, H, and Σ_R, respectively, then the MLE for R_Λ is given by $\hat{R}_\Lambda = 1 - |\hat{\Sigma}\hat{V}^{-1}|$, where $\hat{V} = Z\hat{D}Z' + \hat{\tau}^2\hat{H} + \hat{\Sigma}_R$ and $\hat{\Sigma} = \hat{\tau}^2\hat{H} + \hat{\Sigma}_R$. According to the delta method, $\hat{R}_\Lambda \sim N(R_\Lambda, \mathbf{\Delta}\Sigma_P\mathbf{\Delta}')$, where Σ_P is the variance-covariance matrix of the parameter estimates and

$$\mathbf{\Delta}' = \left(\frac{\partial R_\Lambda}{\partial D}, \frac{\partial R_\Lambda}{\partial \tau^2}, \frac{\partial R_\Lambda}{\partial H}, \frac{\partial R_\Lambda}{\partial \Sigma_R}\right).$$

To avoid confidence limits beyond the [0, 1] range, a logit transformation is applied. In what follows, we will calculate the $\frac{\partial R_\Lambda}{\partial z}$ with z a scalar. Let us first note that

$$\left|\Sigma V^{-1}\right| = 1 - R_\Lambda \quad \Leftrightarrow \quad \ln|\Sigma| - \ln|V| = \ln(1 - R_\Lambda) = \gamma.$$

For simplicity, we will calculate $\frac{\partial \gamma}{\partial z}$. Notice that

$$\frac{\partial \gamma}{\partial z} = -\frac{1}{1 - R_\Lambda}\frac{\partial R_\Lambda}{\partial z}$$

and, therefore,

$$\frac{\partial R_\Lambda}{\partial z} = (R_\Lambda - 1)\frac{\partial \gamma}{\partial z}, \quad \text{where } \frac{\partial \gamma}{\partial z} = \frac{\partial}{\partial z}\ln|\Sigma| - \frac{\partial}{\partial z}\ln|V|.$$

If we call $\gamma_1 = \ln|\Sigma|$ and $\gamma_2 = \ln|V|$, then: $\frac{\partial \gamma}{\partial z} = \frac{\partial \gamma_1}{\partial z} + \frac{\partial \gamma_2}{\partial z}$.

We have:

$$\frac{\partial \gamma_1}{\partial z} = \operatorname{tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial z}\right),$$

but $\Sigma = \tau^2 H + \Sigma_R$ and, therefore,

$$\frac{\partial \Sigma}{\partial z} = \frac{\partial}{\partial z}\left(\tau^2 H + \Sigma_R\right) = \frac{\partial \tau}{\partial z}H + \tau^2\frac{\partial H}{\partial z} + \frac{\partial \Sigma_R}{\partial z}.$$

On the other hand,

$$\frac{\partial \gamma_2}{\partial z} = \operatorname{tr}\left(V^{-1}\frac{\partial V}{\partial z}\right),$$

but $V = ZDZ' + \Sigma$ and, therefore,

$$\frac{\partial V}{\partial z} = Z\frac{\partial D}{\partial z}Z' + \frac{\partial \Sigma}{\partial z},$$

implying that $\frac{\partial \gamma_2}{\partial z} = \operatorname{tr}[V^{-1}(Z\frac{\partial D}{\partial z}Z') + V^{-1}\frac{\partial \Sigma}{\partial z}]$.

Finally, the general formula is given by

$$\frac{\partial R_\Lambda}{\partial z} = (R_\Lambda - 1)\operatorname{tr}\left[V^{-1}\left(Z\frac{\partial D}{\partial z}Z'\right) + \left(V^{-1} + \Sigma^{-1}\right)\frac{\partial \Sigma}{\partial z}\right],$$

where $\frac{\partial \Sigma}{\partial z} = \frac{\partial \tau^2}{\partial z}H + \tau^2\frac{\partial H}{\partial z} + \frac{\partial \Sigma_R}{\partial z}$.

The formulae by cases can be summarized as follows:

$$\text{For } z \text{ element of } D\text{:} \quad \frac{\partial R_\Lambda}{\partial z} = (R_\Lambda - 1)\operatorname{tr}\left[V^{-1}\left(Z\frac{\partial D}{\partial z}Z'\right)\right]$$

$$\text{For } z = \tau^2\text{:} \quad \frac{\partial R_\Lambda}{\partial \tau^2} = (R_\Lambda - 1)\operatorname{tr}\left[\left(V^{-1} + \Sigma^{-1}\right)H\right]$$

$$\text{For } z \text{ element of } H\text{:} \quad \frac{\partial R_\Lambda}{\partial z} = (R_\Lambda - 1)\operatorname{tr}\left[\left(V^{-1} + \Sigma^{-1}\right)\tau^2\frac{\partial H}{\partial z}\right]$$

$$\text{For } z \text{ element of } \Sigma_R\text{:} \quad \frac{\partial R_\Lambda}{\partial z} = (R_\Lambda - 1)\operatorname{tr}\left[\left(V^{-1} + \Sigma^{-1}\right)\frac{\partial \Sigma_R}{\partial z}\right].$$

References

Alonso, A., Geys, H., Molenberghs, G., & 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., & Kenward, M.G. (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.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

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*(1), 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: Wiley.

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

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

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

Johnson, R.A., & Wichern, D.W. (1998). *Applied multivariate statistical analysis* (4th ed.). Englewood Cliffs: Prentice-Hall.

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

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

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

Liang, K.-Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Molenberghs, G., & Kenward, M.G. (2007). *Missing data in clinical studies*. Chichester: Wiley.

Peuskens, J., & the Risperidone Study Group (1995). Risperidone in the treatment of chronic schizophrenic patients: a multinational, multicentre, double-blind, parallel-group study versus haloperidol. *British Journal of Psychiatry*, *166*, 712–726.

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

Royston, P., & Atman, 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.

Searle, S.R. (1982). *Matrix algebra useful for statistics*. New York: Wiley.

Smith, P.L., & Luecht, R.M. (1992). Correlated effects in generalizability studies. *Applied Psychological Measurement*, *16*(3), 229–235.

Spearman, C. (1910). Correlation calculate from faulty data. *British Journal of Psychology*, *3*, 271–295.

Tisak, J., & Tisak, M.S. (1996). Longitudinal models of reliability and validity: a latent curve approach. *Applied Psychological Measurement*, *20*, 275–288.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & 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., & 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., Linn, C.E., & Jøreskog, K.G. (1977). A simplex model for analyzing academic growth. *Educational and Psychological Measurement*, *37*(3), 745–756.

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

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

Manuscript Received: 26 JAN 2007
Final Version Received: 4 JUL 2008
Published Online Date: 13 SEP 2008