Coping With Memory Effect and Serial Correlation When Estimating Reliability in a Longitudinal Framework

Peer-reviewed author version

LAENEN, Annouschka; ALONSO ABAD, Ariel; MOLENBERGHS, Geert; VANGENEUGDEN, Tony & Mallinckrodt, Craig H. (2010) Coping With Memory Effect and Serial Correlation When Estimating Reliability in a Longitudinal Framework. In: APPLIED PSYCHOLOGICAL MEASUREMENT, 34 (4). p. 255-266.

DOI: 10.1177/0146621609349494
Handle: http://hdl.handle.net/1942/11064

Applied Psychological Measurement

Impact of ignoring serial correlation and memory effect on reliability estimates and plausible solutions.

Journal:	*Applied Psychological Measurement*
Manuscript ID:	APM-08-03-027.R2
Manuscript Type:	Manuscripts
Keywords:	Hierarchical models, Rating scales, Reliability, Memory effect, Serial correlation

Impact of ignoring serial correlation and memory effect on reliability estimates and plausible solutions.

Abstract

Longitudinal studies are currently permeating clinical trials in psychiatry. It is therefore of the utmost importance to study the psychometric properties of the rating scales, frequently used in these trials, within a longitudinal framework. However, intra-subject serial correlation and memory effects are problematic issues frequently encountered in longitudinal data. In the present work we study, via simulation, the impact of uncontrolled sources of serial correlation on newly proposed measures designed to evaluate reliability in a longitudinal scenario. We also address the relationship between serial correlation and memory effect. The simulations illustrate that ignoring serial correlation can have a severe impact on the estimates of reliability parameters and inferences related to them. We also argue that the underlying modeling framework allows correcting for this type of correlation and avoiding bias. Moreover, it can adjust for the presence of a memory effect. Nevertheless, to achieve that, a careful model building is required.

Keywords: Hierarchical Model, Memory effect, Rating Scales, Reliability, Serial correlation.

Introduction

Longitudinal clinical trials are becoming a standard tool for the evaluation of new psychiatric drugs. Moreover, in psychiatry, longitudinal data are also frequently encountered in clinical practice where patients are measured repeatedly over time in view of obtaining precise diagnostics as well as evaluating the effect of a treatment or a therapeutic intervention.

Such evaluations are typically carried out using rating scales, which are mainly used when the trait of interest cannot be observed directly, such as the measurement of depression, anxiety, or quality of life. Whenever a new measurement scale is developed, its validity and reliability ought to be evaluated. Reliability is, however, not an intrinsic property of an instrument but rather changes over time and with the population to which it is applied. Therefore, the reliability of a measurement scale should be evaluated every time the scale is introduced to a different population or translated into a different language.

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. Indeed, in such studies, patients usually exhibit a systematic change or evolution over time in addition to an individual-specific evolution that is characterized by correlated subject-specific effects. Moreover, serial correlation and heterogeneous variance components (i.e., variance functions changing over time and/or with covariate levels) are frequently present as well (Verbeke & Molenberghs, 2000).

In the so-called *classical test theory* (CTT), the reliability of a measurement was defined as the ratio of the true score variability and the total variability (Lord & Novick, 1968). In this scenario and under some additional assumptions, reliability equals the correlation between two measurements on the same subjects. These assumptions state that, for both measurements: (i) the true scores are equal; (ii) the error variances are equal; and (iii) the measurement errors are independent. In this framework, the reliability of a measurement can then be estimated by rating a group of subjects at two time-separated occasions and then calculating Pearson's correlation coefficient between both measurements. Note that a test-retest scheme is the simplest possible longitudinal design.

It is fair to say that test-retest reliability has always been controversial. A fundamental issue with the approach resides in finding the optimal length of the time interval between the first and the second measurement. Whenever measuring living organisms, it is clear that the characteristics being measured might change from one replication to another. The usual approach is therefore to take the time interval sufficiently short so that it would be safe to assume that the underlying process is unlikely to have changed in important ways. However, if both measurements are taken too close in time, it is quite probable that the rater will recall his/her previous ratings and, therefore, the new assessments will likely be influenced by them. Usually, the rater will give similar ratings in each of the replications. This effect of memory is not limited to the case where raters make subjective decisions but can also occur in a second attempt on a cognitive ability

test, or when filling out a questionnaire on political attitudes (Dunn, 1989; Streiner & Norman, 1995).

The problem of memory reappears whenever we want to study reliability in a more general longitudinal study, i.e., where subjects are measured at more than two occasions using the same rating scale. When a memory effect emerges in such a setting it implies that observations closer to each other in time are more alike than observations further apart. Basically, this is the same effect produced by a so-called *serial correlation component*, a term used to capture exactly this type of effect in the association structure (Verbeke & Molenberghs, 2000).

Ignoring serial correlation, originating from memory effects or other sources, can have a serious impact on the estimated reliability coefficients. In the present work we study via simulations the bias produced by such uncontrolled sources of serial correlation when employing recently proposed reliability coefficients. Our study complements previous research that has reported the effect of ignoring intra-subject serial correlation on the G-coefficients within a generalizability theory (G-theory) framework (Cronbach, Gleser, Nanda & Rajaratnam, 1972). Further, we argue in favor of hierarchical linear models as valuable tools to accommodate intra-subject serial correlation and to avoid bias in the variance components and the reliability coefficients derived therefrom. Importantly, we claim that this type of model can aid when adjusting for a possible memory effect.

In the following section, a summary is given of previous studies evaluating the impact of serial correlation on the estimation of reliability in a longitudinal framework when G-theory is used. Thereafter, linear mixed models are

introduced and we elaborate on how reliability can be studied within this frame. Some newly introduced measures, designed to evaluate reliability in a longitudinal scenario, are presented. Further, we analyze the impact of uncontrolled serial correlation on these new reliability measures via simulations. Finally, the methodology is illustrated by means of a case study.

Ignoring Intra-subject Serial Correlation

An important attempt to extend the concept of reliability to a longitudinal setting was done using generalizability theory, developed by Cronbach et al. (1972) to explicitly model the multiple sources of variation present in a measurement system. G-theory has played a prominent role in the psychometric field over the last 40 years. The basic mathematical model on which it is based is solidly rooted in analysis of variance with random effects. 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 (DeShon, Ployhart & Sacco, 1998). One such assumption is the presence of an uncorrelated and homoscedastic error structure. However, correlated error structures are frequent in longitudinal studies. Usually, observations close in time exhibit a stronger association than observations with more time separation. Ignoring this correlation will introduce bias in the variance-component estimates and, as a result, in the generalizability coefficients. This has been documented in the literature. For example, Smith and Luecht (1992) investigated the effect of ignoring correlated errors in a longitudinal

framework. Their results showed that not taking into account this correlation will lead to an overestimation of the subject-specific parameters' variance and to overestimating the generalizability coefficient. In their simulations, these authors considered a stationary correlated error structure, i.e., the error terms were allowed correlated but still with equal variances over time. Bost (1995) studied this issue further by examining the effect of both stationary and non-stationary auto-regressive error variance-covariance matrices. His results showed that, in the presence of non-stationary auto-regressive error, the G-coefficients were usually underestimated and the magnitude of the bias increased with the number of observations. Clearly, these results indicate that variance components estimates and the resulting generalizability coefficients can be severely biased when longitudinal data are analyzed under the assumption of independent errors across time. Incorrectly assuming a stationary variance for the error structure also results in bias. Unfortunately, the classic modeling paradigm used in G-theory is not designed to capture this type of associations and assumes uncorrelated error terms with equal variance over time.

Laenen, Alonso and Molenberghs (2007) and Laenen, Alonso, Molenberghs and Vangeneugden (2009) proposed an extension of the concept of reliability to a longitudinal framework, based on a simple set of defining properties. Additionally, these authors introduced two measures of reliability that satisfied these defining properties, the so-called R_T and R_Λ coefficients. They based their proposals on hierarchical linear models. Here, we will study the impact of ignored sources of serial correlation on R_T and R_Λ, and we will stress the importance of

a careful model building exercise. We will argue in favor of linear mixed models as a valuable tool to account for many different sources of serial correlation including the one emanating from a possible memory effect.

Methodology

Linear mixed models (LMM) allow incorporating many of the previously discussed longitudinal features, including varying true scores, correlated random effects, heteroscedastic error components, and correlated error terms. Being able to account for all of these complexities within the same modeling paradigm is extremely important to guarantee unbiased results when estimating reliability. Essentially, one would like to consider the following general model

$$Y_{ij} = \mu_{ij} + \tau_{ij} + \xi_{ij}, \quad (1)$$

where Y_{ij} denotes the observed score of subject i at time point j, μ_{ij} is a general mean that can vary over time, τ_{ij} is the true score of subject i at time point j and ξ_{ij} is the corresponding error component. Note that in the previous model the τ_{ij} s are subject-specific random variables that can vary over time, i.e., we are not assuming that the true scores are constant over time. Within the linear mixed model framework one can explicitly model the true scores as linear functions of time by considering $\tau_{ij} = \mathbf{z}_j \mathbf{b}_i$ where $\mathbf{z}_j$ is a row-vector that may depend on time and $\mathbf{b}_i$ is a vector of subject specific coefficients. Similarly to classical linear regression, this function is linear in the subject-specific parameters $\mathbf{b}_i$ but it does

not need to be linear in time. For example, one could consider the following expression to model the true scores as a function of time $\tau_{ij} = b_{i0} + b_{i1} t_j \log(t_j)$. In the previous formula $\mathbf{z}_j = (1 \;\; t_j \log(t_j))$ and $\mathbf{b}_i = \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix}$. One of the biggest advantages of this approach is that the time evolution of the subject-specific true scores τ_{ij} is now entirely characterized by a vector of subject-specific coefficients $\mathbf{b}_i$ that does not vary over time. Basically, the τ_{ij} s and $\mathbf{b}_i$ are equivalent quantities and, therefore, $\mathbf{b}_i$ can be treated as a vector of true scores itself.

Model 1 is a special case of the more general family of linear mixed models. Indeed, assuming a balanced study design, in the sense that all patients are evaluated at a common set of measurement occasions, the general linear mixed model can be written as

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}\mathbf{b}_i + \boldsymbol{\xi}_{(1)i} + \boldsymbol{\xi}_{(2)i} \qquad (2)$$

$\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\xi}_{(1)\mathrm{i}} \sim \mathrm{N}(\mathbf{0}, \mathbf{R})$, $\boldsymbol{\xi}_{(2)\mathrm{i}} \sim \mathrm{N}(\mathbf{0}, \tau^2 \mathbf{H})$,

$\mathbf{b}_i$, $\boldsymbol{\xi}_{(1)i}$ and $\boldsymbol{\xi}_{(2)i}$ are independent,

where $\mathbf{Y}_i$ is a p-dimensional vector of repeated measurements on certain trait for subject i and i takes values from 1 to n. Further, $\mathbf{X}_i$ and $\mathbf{Z}$ are fixed (p × q) and (p × r) dimensional matrices of known covariates, $\boldsymbol{\beta}$ is a q-dimensional vector of fixed effects, $\mathbf{b}_i$ is a r-dimensional vector containing the random effects,

$\xi_{(2)i}$ is a p-dimensional vector of components of serial correlation, and $\xi_{(1)i}$ is a p-dimensional vector of residual errors. Additionally, $\mathbf{D}$ is a general (r × r) covariance matrix, $\mathbf{H}$ is a (p × p) correlation matrix, τ^2 is a variance parameter, and $\mathbf{R}$ is an (p × p) covariance matrix.

Another reason why this model is interesting in reliability research is the fact that it allows to simultaneously model mean (i.e., fixed-effects), random-effects, and residual variability structures. The systematic evolution over time can then be modeled as part of the fixed-effects structure and one can also explicitly model the time evolution of the true scores, making the steady-state assumption unnecessary (Vangeneugden, Laenen, Geys, Renard, & Molenberghs, 2004). On the other hand, the models' ability to distinguish between different sources of variability (Laird & Ware, 1982; Verbeke & Molenberghs, 2000) makes it especially suitable for reliability estimation. Model 2 implies the marginal model $\mathbf{Y}_i \sim \mathrm{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V})$ where $\mathbf{V} = \mathbf{ZDZ}'+\boldsymbol{\Sigma}$ ($\mathbf{A}'$ denoting the transpose of matrix $\mathbf{A}$) with $\boldsymbol{\Sigma} = \tau^2\mathbf{H} + \mathbf{R}$. Note that the total variability is decomposed into two parts: the first one $\mathbf{ZDZ}'$, accounts for the variability of the subject-specific parameters or true scores, whereas the second one, $\boldsymbol{\Sigma}$, includes all remaining sources of variability. Exactly these two sources of variability constitute the main reliability ingredients.

Observe that, whenever serial correlation is present, whether caused by the rater's memory effects or by other sources, it can be incorporated within model 2. Frequently, this serial correlation component is modeled through an auto-

regressive structure with $\mathbf{H}_{jk} = \rho^{d_{jk}}$. In this case, $\mathbf{H}$ is then a correlation matrix, with ρ denoting the correlation between two measurements taken one unit of time apart, and d_{jk} the time lag between two measurements taken at times j and k.

As stated in the introduction, we argue that a memory effect will typically produce the same correlation pattern as a serial correlation component and, as a result, it could be absorbed into it. Clearly, other sources of association may also contribute to the presence of serial correlation and, therefore, we should not fully identify these two related but different concepts. In general, a strong serial correlation can be the reflection of a strong memory effect, a memory effect combined with other factors, or simply (a combination of) such other factors. Which of these scenarios is the true one is not relevant, but rather the fact that serial correlation is able to absorb each one of them. This is because one's primary interest is not in making inferences about serial correlation, but rather about reliability, with serial correlation treated as a nuisance characteristic.

The linear mixed model framework conveniently offers a large amount of flexibility for modeling serial correlation. For instance, Gaussian or exponential structures could replace an autoregressive structure when data points are not equally spaced, with heterogeneous versions further allowing for time- and covariate-dependent variance functions. Furthermore, on top of the serial correlation, additional measurement error variability can be superimposed.

In CTT, the reliability of a measurement is defined as the ratio of the true-score and total variability, or equivalently, as one minus the ratio of the error and

total variability. Considering all data at a fixed time point, say t_j, then one is back to the cross-sectional setting with true score $\tau_{ij} = \mathbf{z}_j \mathbf{b}_i$. Applying the classical definition of reliability at time point t_j then leads to the expression

$$R_{Tj} = 1 - \frac{\sigma^2_{\mathbf{\Sigma}jj}}{\sigma^2_{\mathbf{V}jj}} \tag{3}$$

with $\sigma^2_{\mathbf{\Sigma}jj}$ the jth diagonal element of the general error matrix $\mathbf{\Sigma}$, representing the error variability at time t_j, and $\sigma^2_{\mathbf{V}jj} = \mathbf{z}_j \mathbf{D} \mathbf{z}'_j + \sigma^2_{\mathbf{\Sigma}jj}$ the jth diagonal element of the matrix $\mathbf{V}$, representing the total variability at the same time point. Note also that the variance of the true scores at time t_j is then given by $Var(\tau_{ij}) = \mathbf{z}_j \mathbf{D} \mathbf{z}'_j$.

From expression 3 we can see that reliability is not necessarily constant over time, as frequently assumed in earlier approaches (Tisak & Tisak, 1996; Wiley & Wiley, 1970; Raykov, 2000). Two aspects may cause reliability to be different at different measurement occasions. First, if the variability of the measurement error decreases over time then the reliability will increase. Second, the reliability will also be affected if the true scores of subjects change over time, i.e, if the vector $\mathbf{z}_j$ depends on time (Heise, 1969; Jagodzinski & Kühnel, 1987; Werts et al., 1980). As previously stated, model 2 will allow the true scores to change over time as soon as a random slope for time is included in the random effect structure. A model including both, a random intercept term (b_{0i}) and a random

slope (b_{1i}), may lead, for example, to the following expression, introduced previously, for the true score of subject i at time t_j: $\tau_{ij} = b_{i0} + b_{i1} t_j \log(t_j)$.

Using expression 3 one can then calculate the reliability of the observed scores Y_{ij} at each time point, what naturally leads to a time varying function of reliability. However, interpretability may be substantially improved if one could have a meaningful summary measure. Such a measure is provided by the coefficient R_T that indicates the average reliability over the different time points

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

Actually, it is possible to show that R_T can be rewritten as

$$R_T = \sum_{j=1}^{p} w_j R_{Tj} \ ,$$

With weights $w_j = \sigma^2_{\mathbf{V}jj} \big/ tr(\mathbf{V})$. Basically, these weights quantify the proportion of the total variability that each time point accounts for. Notice that variability is information and, therefore, R_T establishes a compromise between the amount of information every time point conveys and the quality of that information, i.e., its reliability.

A summary measure can be very useful when a large number of repeated measurements are taken; or in case the researcher is interested in the general performance of the outcome scale over the entire longitudinal study. Also when two scales are to be compared, such a summary can simplify the interpretation and conclusions.

Besides this average reliability measure, Laenen et al. (2009) proposed a global reliability coefficient

$$R_{\Lambda} = 1 - \left|\boldsymbol{\Sigma}\mathbf{V}^{-1}\right| \quad (4)$$

It is clear from model 2 that in a longitudinal framework the variability is expressed by matrices rather than single variance coefficients. Two common ways of summarizing the variability from such a variance-covariance matrix are the so-called generalized variances, which are typically based on the concepts of trace and determinant, naturally leading to R_T and R_{Λ}, respectively. It can be noticed, however, that both measures still express one minus the proportion of the total variability that is due to measurement error, exactly as in the classical definition of reliability.

Remarkably, the R_{Λ} coefficient bears a different interpretation than R_T: R_{Λ} expresses the reliability of an entire longitudinal sequence. Indeed, as previously stated, R_T quantifies the average reliability over time, i.e., it is a summary measure of the cross-sectional reliabilities. On the other hand, R_{Λ} quantifies the reliability of the entire vector of observed scores with respect to the vector of subject-specific effects that describes the true scores evolution over time, i.e., the vector $\mathbf{b}_i$. Simply said, R_{Λ} quantifies the amount of information about $\mathbf{b}_i$ that $\mathbf{Y}_i$ conveys. Notice that the vector $\mathbf{b}_i$ is the element of the model that fully captures the longitudinal evolution of the true scores and can be considered itself a vector of true scores. In a longitudinal design, every new measurement for a

subject brings additional information on his/her vector of true scores $\mathbf{b}_i$. It is this increase in total information that is captured by R_Λ. As a consequence, the measure increases as the number of repeated measurements increases. Such a global measure can be very useful to analyze the total impact of measurement error on a longitudinal study. One frequently encounters that, although a single measurement can suffer from high measurement error (indicated by low R_{Tj}'s), the impact of that error is minimized when several repeated measurements are taken (indicated by high R_Λ). Essentially, what lies behind this behavior is that even though every Y_{ij} may not convey a lot of information about the corresponding τ_{ij}, the entire longitudinal set of observations $\mathbf{Y}_i$ may still convey a lot of information about the vector describing the time evolution of the τ_{ij}, i.e., the vector $\mathbf{b}_i$.

When the assumptions of CTT are met, both R_T and R_Λ reduce to the classical expression of reliability. Furthermore, when applied in a setting where G-theory assumptions are met, such as non-changing true scores or equal error variances at different time points, the two measures reduce to the index of dependability, and after conditioning on the time points they equal the generalizability coefficient. Both G-coefficients are commonly used to quantify reliability in a longitudinal scenario (Brennan, 2001; Laenen et al., in press). For a full discussion of the rationale and mathematical details behind R_T and R_Λ we refer the reader to the original papers (Laenen et al., 2007, 2009).

In the following section, we will design and carry out a simulation study to investigate the impact of the presence of serial correlation, whether caused by a memory effect or another source, on the two reliability coefficients introduced by Laenen et al. (2007, 2009). This allows for the evaluation of the robustness of R_T and R_Λ relative to potential misspecification of the error structure in model 2.

A Simulation Study

The design of the simulation study was a 2*3*2 complete factorial arrangement with: 2 types of subject-specific true scores, (1) true scores that were constant over time, i.e., a random intercept model, and (2) true scores that evolved linearly over time, i.e., a random intercept and random slope model; 3 levels of auto-regressive serial correlation were considered with values 0.1, 0.5, and 0.8; and two types of analyses were carried out (1) ignoring serial correlation and (2) fitting serial correlation.

The random-intercept model can be expressed as

$$Y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 d_i + \tau_{ij} + \xi_{ij} \quad (5)$$

where t_j denotes the time at which measurement j is taken, and d_i the treatment allocation for subject i. Further, $\tau_{ij} = b_{0i}$ indicates the true score for subject i which is not varying over time, with $b_{0i} \sim \mathrm{N}(0, \sigma_{b0}^2)$, and ξ_{ij} is the measurement error at time j for subject i, with $\boldsymbol{\xi}_\mathrm{i} \sim \mathrm{N}(\mathbf{0}, \tau^2\,\mathbf{H})$. We fix $\sigma_{\mathrm{b0}}^2 = 300$ and $\tau^2 = 100$, corresponding to a situation where the error variability accounts for one quarter of the total variability.

The model with random intercept and slope has the same general form as (5) but now $\tau_{ij} = b_{0i} + b_{1i}t_j$, i.e., the true scores are allowed to vary linearly over time, we further assume that

$$\mathbf{b}_i = \begin{pmatrix} b_{0\mathrm{i}} \\ b_{01} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{D}) \text{ with } \mathbf{D} = \begin{pmatrix} 300 & -1 \\ -1 & 5 \end{pmatrix}, \text{ and } \boldsymbol{\xi}_{\mathrm{i}} \sim \mathrm{N}(\mathbf{0}, \tau^2\, \mathbf{H}).$$

Values for the fixed effects were set to $\beta_0 = 85$, $\beta_1 = -2.5$, and $\beta_2 = 85$. Six equally spaced time points, at weeks 0, 2, 4, 6, 8, and 10, were considered, and the sample size was set equal to 250. Finally, a total of 250 data sets were generated for each of these six settings.

First we will illustrate the effect of instability (changing true scores) and serial correlation on ordinary reliability estimates, calculated as test-retest correlations. Table 1 presents Pearson correlations between the outcome at the first measurement (Y_{i0}) and the outcomes at later measurement occasions ($Y_{i2},\ldots,Y_{i10}$), for different strengths of serial correlation (ρ). For the random intercept model (RI) one can easily obtain the reliability of the scale as the ratio of the true score variability to the total variability

$$R = R_T = \frac{\sigma^2_{b0}}{\sigma^2_{b0} + \tau^2} = \frac{300}{300+100} = 0.75$$

Note that in this model the true score does not change over time and, therefore, it does not distort the pairwise correlations. Essentially, one can state that for the random intercept model the steady-state assumption is valid and all the misspecification is concentrated in the error structure. The upper half of Table 1

clearly shows that test-retest reliability can give a severely distorted image if serial correlation is present. Indeed, in case of small serial correlation, as expected, Pearson correlation coefficient can give stable and trustworthy results as an estimator of reliability, especially when using observations that are far apart. We must point out, however, that some overestimation can appear, even in this scenario, if the observations are close in time. Basically, this illustrates that correlation is a valid estimator for reliability, only when the serial correlation is very small or does not exist at all. However, with an increasing serial correlation the situation changes dramatically and reliability is usually strongly overestimated, especially for small time lags.

The classical definition of reliability does not apply to a model with random intercept and slope (RIS). We will then use the true value of R_T as a reference point. Using the parameter values chosen for the simulation study we obtain a value of 0.826. For this model, true scores change over time, i.e, different subjects can now evolve over time in different ways. The lower half of Table 1 shows that these changes in the true scores lower the correlations when time lag increases. This can lead to a severe underestimation of reliability if the two observations used to calculate the test-retest estimate are far apart. The serial correlation, on the other hand, produces the opposite effect, i.e, it increases the Pearson correlations. This clearly shows one of the most important problems associated with test-retest reliability: choosing two time points which are close enough in time to guarantee the steady-state assumption and, at the same time, far enough from each other to annul the effect of serial correlation. As the

simulation results clearly show, this optimal time point depends on the value of the unknown serial correlation and it can be extremely difficult to determine in practice. Notice also that even when such an optimal time point can be determined, this does not guarantee that bias will be fully avoided. As a summary, the results presented here illustrate that the classical approach to reliability is only justified when the necessary assumptions are fulfilled. Whenever a serial correlation is present or the true scores vary over time, this approach will not lead to correct estimates.

Let us now look at the effects of serial correlation on the R_T and R_Λ coefficients. We consider two different scenarios for analysis: (i) a correctly specified model that includes a serial correlation component with an auto-regressive structure and (ii) a miss-specified model that assumes an uncorrelated structure for the residual part, i.e., $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. Based on these models, we calculated the point estimates and confidence intervals for R_T and R_Λ.

Tables 2 and 3 present the true values for R_T and R_Λ, and the average of the estimated values over the 250 simulated data sets. The coverage probability (CP) indicates the percentage of the cases in which the true value lies within the estimated 95% confidence interval.

Let us first focus on the random-intercept setting. The first half of Table 2 illustrates that, when the model used to fit the data does not include a serial correlation component, both $\hat{R}_T$ and $\hat{R}_\Lambda$ overestimate the true values. As one

would expect, for the smallest values of ρ, the bias present in R_T is only minor and the misspecification seems to exert a weak impact only on the coverage probability of the corresponding confidence interval. However, a totally different image emerges when larger values of ρ are considered. In such scenarios, a large bias is observed in the point estimates of R_T and the coverage probability of the corresponding confidence interval is considerably smaller than the pre-specified 95% value.

Interestingly, R_Λ seems to be more sensitive to the misspecification. Indeed, even for the smallest values of ρ, a moderate bias appears in the point estimate of R_Λ and the coverage probability of the confidence intervals are also more seriously affected than the confidence intervals for R_T. Unsurprisingly but with important ramifications, the situation worsens considerably for larger values of serial correlation.

These findings fully coincide with the results reported by Smith and Luecht (1992) and Bost (1995) in their studies of the effect of ignoring a stationary correlated error structure on the estimation of the G-coefficients. Fortunately, unlike in the modeling framework used in G-theory, linear mixed models allow for the absorption of such a correlation structure. The second part of Table 2 shows the results obtained when the models fitted to the data included a serial correlation component. As one would expect, now there is bias in neither the R_T nor the R_Λ point estimates. Furthermore, the confidence intervals now enjoy coverage probabilities very close to their nominal level.

Interestingly, the true value of R_Λ decreases when the serial correlation increases, which is an entirely plausible feature. Indeed, it has been shown that R_Λ has the ability to increase with the number of time points, owing to the fact that every new observation purports additional information, even if it comes contaminated by measurement error (Laenen et al., 2009). Nevertheless, for a given number of time points, we have less information when different observations are strongly correlated, explaining lower R_Λ for larger values of ρ.

Table 3 displays the results obtained under the second setting, i.e., when the true scores vary linearly over time. The conclusions in this case are almost identical to our earlier ones. Note that, if the serial correlation is ignored, then the bias of the point estimates and the problem with the coverage probabilities of the confidence intervals seem to aggravate in this scenario, stemming from the more complicated random-effects structure. The second half of the table shows the results when the correct model was fitted to the data. Here again, there is no bias in the point estimate and the coverage probabilities are close to their nominal value. Only when the serial correlation was largest a moderate under-coverage was observed for the confidence intervals of both R_T and R_Λ. Nevertheless, some additional simulations (details not shown) proved that the problem completely disappears when the sample size is increased to 500 patients.

A Case Study

In this section, we will use R_T and R_Λ to evaluate the reliability of two widely

used rating scales in a longitudinal setting. The case study consists of two randomized double blind clinical trials that were set up to investigate treatment efficacy on major depressive disorder (Studies 5 and 6 in Mallinckrodt et al., 2003). The studies contain a total of 354 and 353 patients, respectively, randomly allocated to four treatment groups. The primary efficacy measure was the $HAMD_{17}$ total score, whereas the MADRS total score was used as secondary measure. The HAMD or Hamilton Rating Scale for Depression (Hamilton, 1960) was developed to assess the effectiveness of the first generations of anti-depressants. The scale quickly became the standard measure of depression severity for clinical trials of anti-depressants. It is until now the most commonly used measure for depression, even though several conceptual and psychometric problems have been described in the literature (Bagby, Ryder, Schuller & Marshall, 2004). The MADRS or Montgomery-Asberg Depression Rating Scale (Montgomery & Asberg, 1979) is a 10-item scale that was designed to address the limitations of the HAMD, and was supposed to capture contemporary definitions of depression and to be more sensitive to change. Ratings on both scales were taken at baseline and weeks 1, 2, 4, 6, 8, and 10.

Our simulations have clearly illustrated that a meticulous model building step is crucial to avoid bias in the variance-components and thence reliability estimates. In general, when estimating reliability, the main interest primarily lies in the covariance structure, and therefore, an elaborate fixed-effects structure was adopted, containing categorical time, treatment, investigator, and treatment by time interaction. This minimizes the risk of miss-specifying the mean structure

and hence of bias resulting there from (Diggle, Liang & Zeger, 2002).

The selection of the best fitting covariance structure for the data was based on model building guidelines laid out in Verbeke and Molenberghs (2000). Regarding the random effects we considered models with: (a) constant true scores over time, i.e., a subject-specific intercept; (b) true scores varying linearly over time, i.e., subject-specific intercept and slope; and (c) true scores varying as a quadratic function over time, i.e., subject-specific intercept, linear slope and quadratic slope. Additionally, for the measurement-error terms, the correlation structures considered are: (a) autoregressive; (b) exponential; (c) serial Gaussian; (d) power; and (e) banded unstructured. The latter structure, in contrast to the other four, only allows correlation between errors of measurements taken at adjacent occasions and assumes zero correlations for other pairs of measurements. Structure (e) further assumes heterogeneity of the error variances, whereas the structures (a)–(d) were fitted with homogeneous as well as heterogeneous error variances. This distinction can also be found in the two remaining error variance-covariance structures without error correlation: (f) features an unstructured main diagonal, while (g) is a so-called ‘simple’ or ‘variance-components’ structure, both with the off-diagonal elements equal to zero. For details and examples on the covariance structures we refer to Verbeke and Molenberghs (2000).

Akaike’s Information criterion was used for model selection. The parameter estimates were calculated using restricted maximum likelihood (REML). Table 4 shows the best fitting models for the two scales, for both trials separately. Note

that, within each trial, the same model was selected for the two different scales HAMD and MADRS. In trial 1 a linear random effects model was selected whereas in trial 2 a quadratic model resulted in the best fit. Further, in all four cases, a heterogeneous auto-regressive correlation structured was selected. Graphical exploration (not shown) indicated that the models capture the most important data features reasonably well.

Table 4 further presents the reliability estimates for each of the four cases. SAS macro's for the calculation of R_T, R_Λ and the corresponding confidence intervals can be obtained from the first author. We observe that, within each of the trials, both scales HAMD and MADRS perform very similarly. In the first trial, the point estimates of R_T and R_Λ are slightly higher for HAMD, while in the second trial we observe the opposite. In both trials, the confidence intervals around the reliability estimates for the two rating scales largely overlap. Hence, we do not find evidence of MADRS being a more reliable scale than HAMD, or vice versa. Similar results were found by Maier et al. (1988) for inter-rater reliabilities. They compared the HAMD and MADRS based on three different studies, but did not find differences in reliabilities in any of them.

Further, note that the reliability estimates for the two scales are clearly higher in the second trial than in the first one. Reliability is known to be a population-dependent concept, and will generally be estimated higher in more heterogeneous groups. However, it is highly unlikely that this can explain the observed difference between the two trials since both studies were developed from one protocol and they were identical in every way. Other factors might have

had an influence as well, such as training, experience, and quality of the raters. Also on this matter, equality of the two trials was aimed for. At a single start up meeting, all sites in both studies were present to be trained on the protocol and to qualify raters. Investigative sites were randomly selected to be part of either trial, but there is no guarantee that this random assignment truly equalized quality of sites and raters.

Even though it is difficult to identify the reasons for the differences in reliability between the two trials, it is very interesting to relate this finding to the clinical outcomes of the studies. Both studies tested 3 treatments with what are now proven to be effective doses of anti-depressants. Trial 1, however, had worse separation from placebo than trial 2 (Mallinckrodt et al., 2003). The finding that the reliability of the measurements was also lower in the first trial might explain why the clinical effects were stronger in the second trial. This finding illustrates that measurement error or low reliability can have an effect on the results found in clinical studies, as emphasized by Fleiss (1987) and Lachin (2004).

The average reliabilities per time point (R_T) that were found for HAMD and MADRS for the two trials are lower than the reliabilities generally mentioned in the literature (Bagby et al., 2004). Also Zimmerman, Posternak and Chelminski (2005) report that, in spite of other psychometric flaws of HAMD, the inter-rater and test-retest reliabilities are mostly good. The fact that the obtained R_T values are lower than their counterparts reported in the literature can have several reasons. As indicated before, reliability is a population-dependent concept and tends to be lower in more homogeneous populations. The studies, on which the

present estimates are based, only included patients suffering from a major depressive disorder, likely reducing variability between the patients. It is not always clear on which populations the reliability estimates in the literature are based.

Note also that, in our case study, a serial correlation term was present for all scales in both trials. Our simulations showed that ignoring this type of correlation can lead to a serious overestimation of the reliability parameters, what can be a plausible explanation for the relatively higher values reported in the literature.

Finally, the results illustrate that a scale with poor performance, in general or in a certain population, can still be used to obtain reliable information if the measurement is repeated over time. Of course, the lower the average reliability per time point, the more repeated measurements will be necessary to achieve a sufficiently high level of global reliably. In the first trial of the case study, seven repeated measurements were needed to obtain a cumulative reliability R_Λ of around 0.80. In the second trial, 4 and 3 measurements, respectively, sufficed to reach the same target for HAMD and MADRS.

Discussion

Longitudinal studies are becoming a standard tool in psychiatric clinical practice as well as in research. Therefore, it is important to evaluate the reliability of rating scales within a longitudinal framework. Laenen et al. (2007, 2009) introduced two measures of reliability, the so-called R_T and R_Λ, that allow quantification of reliability in such a longitudinal scenario. However, measures of

reliability are model-based quantities and their scope and applicability will never venture beyond these of the model they are based on. One of the characteristic issues of longitudinal studies is the presence of intra-subject correlation. This correlation can emanate from many different sources, including the presence of a rater's memory effects. In the present paper, we have studied the impact of ignoring this intra-subject correlation on R_T and R_Λ.

Our conclusions fully coincide with the results found by Smith and Luecht (1992) and Bost (1995) in their study about the effect of ignoring a stationary correlated error structure on the estimation of the G-coefficients. This misspecification can seriously affect both, the point estimates of the reliability parameters and the inferential procedures related to R_T and R_Λ. However, the more general modeling framework on which they are based allows us to adjust for the presence of such a correlation structure. Clearly, our results together with the findings of Smith and Luecht (1992) and Bost (1995) suggest the use of linear mixed models and R_T and R_Λ as a very appropriate choice for the evaluation of reliability in a longitudinal scenario. At the same time, our simulations have illustrated the importance of a very careful model building exercise.

In general, the mean structure can be treated as a nuisance within reliability research. However, an inappropriate mean structure can result in biased estimates for the variance components (Diggle et al., 2002; Verbeke & Molenberghs, 2000) and, as a consequence, it can introduce bias in the reliability coefficients as well. It is therefore advisable to consider sufficiently versatile

mean structures in this setting. This can be done using linear mixed models, for instance, by modeling the systematic evolution over time with fractional polynomials (Royston & Altman, 1994) or via non-parametric approaches such as, for example, smoothing splines (Verbyla, Cullis, Kenward, & Welham, 1999). Another possibility is to model the time evolution in a fully unstructured way, including a parameter for each time-by-group combination, exactly as in our case study.

Further, the covariance structure needs to be modeled carefuly. Here again, linear mixed models offer a lot of flexibility, allowing for correlated error terms, including different types of serial correlation (Gaussian, first-order auto-regressive, exponential, m-dependent structures, to name but a few), and heteroscedastic error components.

Finally, we put a strong focus on the problem of memory effect. In presence of such an effect, the condition of the subject at consecutive and/or close measurement times will appear more similar than they actually are. This effect is one typical source of serial correlation, providing the opportunity to accommodate it into the model by using the serial correlation structure.

It is useful to recall that the terms ‘memory effect’ and ‘serial correlation’ are not fully interchangeable. In fact, a memory effect is but one of the possible causes leading to serial correlation. Our simulations have shown that, regardless of the actual source of serial correlation, it will distort the reliability estimates and should therefore always be taken into account. Therefore, the results of this paper are, broadly, applicable to serial correlation. The reason we chose to

emphasize memory effect is because it has permeated reliability research for the longest time. Many attempts to solving this problem were circumscribed to finding an optimal length for the interval between two consecutive observations. The issue of finding this optimal length has been largely based on knowledge specific to the area of application and is only applicable when solely two repeated measurements per subject are taken. In the present work, we approached the problem from a statistical modeling perspective by considering more general hierarchical models that can account for both the time evolution of the patients, as well as a potential memory effect.

References

Bagby, R.M., Ryder, A.G., Schuller, D.R., & Marshall M.B. (2004). The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, *161*, 2163–2177.

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

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

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 & Sons.

DeShon, R.P., Ployhart, E., & Sacco, J.M. (1998). The estimation of reliability in longitudinal models. *International Journal of Behavioural Development*, 22, 493–515.

Diggle, P.J., Heagerty, P.J., Liang, K.-Y., & Zeger, S.L. (2002). *Analysis of longitudinal data (2nd ed.)*. Oxford: Clarendon Press.

Dunn, G. (1989). *Design and analysis of reliability studies: The statistical evaluation of measurement errors*. New York: Oxford University Press.

Fleiss, J.L. (1986). *Design and analysis of clinical experiments*. New York: John Wiley & Sons.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurological and Neurosurgical Psychiatry*, *23*, 56–62.

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

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

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

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). Reliability of a longitudinal sequence of scale ratings. *Psychometrika*, DOI: 10.1007/S11336-008-9079-7.

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

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

Maier, W., Philipp, M., Heuser, A., Schlegel, S., Buller, R., & Wetzel, H. (1988). Improving depression severity assessment: Reliability, internal validity and sensitivity to change of three observer depression scales. *Journal of Psychiatric Research*, *22*, 3–12.

Mallinckrodt, C.H., Goldstein, D.J.,Detke, M.J., Lu, Y., Watkin, J.G., & V.Tran, P. (2003). Duloxetine: a new treatment for the emotional and physical symptoms of depression. *Primary Care Companion to the Journal Clinical Psychiatry*, *5*, 19–28.

Montgomery, S.A. & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, *168*, 594–597.

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

Royston, P., & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Applied Statistics*, *43*, 429–468.

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

Streiner, D.L., & Norman, G.R. (1995). *Health measurement scales*. Oxford: Oxford University Press.

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., Breland, H.M., Grandy, L., & 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.

Zimmerman, M., Posternak, A., & Chelminski I. (2005). Is it time to replace the Hamilton depression rating scale as the primary outcome measure in treatment studies of depression? *Journal of Clinical Psychopharmacology*, *25*, 105–110.

Table 1

Instability and serial correlation on reliability measures: correlation coefficients. RI refers to random-intercept model, RIS refers to model with random intercepts and random slopes. ρ is the serial correlation parameter and (Y_{ij}, Y_{ik}) refer to pairs of measurement occasions.

Model	ρ	(Y_{i0}, Y_{i2})	(Y_{i0}, Y_{i4})	(Y_{i0}, Y_{i6})	(Y_{i0}, Y_{i8})	(Y_{i0}, Y_{i10})
RI	0.1	0.770	0.751	0.748	0.748	0.748
RI	0.5	0.871	0.810	0.779	0.764	0.757
RI	0.8	0.948	0.908	0.875	0.850	0.830
RIS	0.1	0.746	0.683	0.617	0.553	0.492
RIS	0.5	0.845	0.734	0.641	0.564	0.498
RIS	0.8	0.921	0.822	0.718	0.624	0.544

Table 2
Effect of ignoring intra-subject correlation on reliability measures: random intercept model. ρ is the correlation coefficient; both reliability measures are considered, with R_T and R_Λ the true values, $\widehat{R}_T$ and $\widehat{R}_\Lambda$ the simulation averages, and CP referring to coverage probability.

Correlation structure	ρ	R_T	$\widehat{R}_T$	CP(R_T)	R_Λ	$\widehat{R}_\Lambda$	CP(R_Λ)
variance components	0.1	0.750	0.757	90.4	0.939	0.949	50.0
variance components	0.5	0.750	0.815	3.2	0.889	0.963	0
variance components	0.8	0.750	0.902	0	0.824	0.982	0
auto-regressive	0.1	0.750	0.748	95.2	0.939	0.938	96.4
auto-regressive	0.5	0.750	0.746	95.2	0.889	0.886	96.0
auto-regressive	0.8	0.750	0.734	95.2	0.824	0.808	96.0

Table 3

Effect of ignoring intra-subject correlation on reliability measures: random intercept model and slope model. ρ is the correlation coefficient; both reliability measures are considered, with R_T and R_Λ the true values, $\widehat{R}_T$ and $\widehat{R}_\Lambda$ the simulation averages, and CP referring to coverage probability.

Correlation structure	ρ	R_T	$\widehat{R}_T$	CP(R_T)	R_Λ	$\widehat{R}_\Lambda$	CP(R_Λ)
variance components	0.1	0.826	0.837	83.2	0.986	0.990	35.2
variance components	0.5	0.826	0.900	0	0.972	0.997	0
variance components	0.8	0.826	0.960	0	0.965	0.999	0
auto-regressive	0.1	0.826	0.825	97.6	0.986	0. 986	96.8
auto-regressive	0.5	0.826	0.821	96.8	0.972	0. 968	97.2
auto-regressive	0.8	0.826	0.812	88.1	0.965	0. 955	91.9

Table 4
Selected models and reliability estimates [95% confidence intervals] for HAMD and MADRS for trial 1 and trial 2.

	Scale	Rand. eff.	Structure of $\boldsymbol{\Sigma}$	R_T	R_Λ
1	HAMD	linear	heterog. auto-regressive	0.493 [0.405; 0.581]	0.829 [0.734; 0.895]
	MADRS	linear	heterog. auto-regressive	0.474 [0.378; 0.571]	0.812 [0.704; 0.886]
2	HAMD	quadratic	heterog. auto-regressive	0.629 [0.513; 0.731]	0.932 [0.872; 0.966]
	MADRS	quadratic	heterog. auto-regressive	0.692 [0.603; 0.769]	0.977 [0.957; 0.988]

Response to the Editor

We thank the editor, once more, for his/her useful comments and remarks and for giving us the opportunity of resubmitting and improving our manuscript. We fully agree with the Editor that the previous version of the manuscript was lacking clarity in some aspects. Most of the problems pointed out by the Editor were direct consequences of a lack of clarity regarding the definition of true scores in our modeling framework. We have now explicitly stated what we mean by true scores in our models and their relationship with the proposed measures of reliability. In general, taking into account all the comments, we have now substantially rewritten many sections of the paper.

Below we provide point-to-point answers to all the comments and questions raised by the editor. We further indicate where we have made corresponding changes in the manuscript.

Finally we would like to thank as well reviewer 1 for the positive comments and for constructive comments in the previous report.

Meaning of various terms, including three coefficients and true score

As the Editor rightly points out a clear definition of what was considered true scores was not explicitly given in the previous version. Obviously, this is a very important issue and needed a more careful consideration. We have now substantially rewritten the Methodology section of the paper and explicitly stated what we mean by true scores in our modeling framework.

Essentially, the true scores are subject-specific effects that determine their individualized responses. Clinical practice clearly shows that some individuals tend to score very similar to the average in the population whereas others score higher or lower than the average. This subject-specific behavior is captured by the so-called individual true scores. On of the complications associated with longitudinal studies is that these true scores may vary over time. Indeed, similar to the previous cross-sectional example, some individuals tend to evolve over

time like the average in the population whereas others have personalized evolutions that differ from the average.

To explain this issue further let us denote by τ_{ij} the true score of subject i at time point j. Note that this true score can change over time. If one focuses on a fixed time point, say t_j, then one is back to the cross-sectional setting and the classical definition of reliability can be applied. However, if one wants to study reliability in a longitudinal way, then we are forced to take into account the time evolution of τ_{ij}. Within the linear mixed model framework one can explicitly model the true scores as linear functions of time by considering $\tau_{ij} = \mathbf{z}_j \mathbf{b}_i$ where $\mathbf{z}_j$ is a row-vector that may depend on time and $\mathbf{b}_i$ is a vector of subject-specific coefficients. Similarly to classical linear regression, this function is linear in the subject-specific parameters $\mathbf{b}_i$ but it does not need to be linear in time. For example, one could consider the following expression to model the true scores as a function of time $\tau_{ij} = b_{i0} + b_{i1} t_j \log(t_j)$. In the previous formula $\mathbf{z}_j = (1 \;\; t_j \log(t_j))$ and $\mathbf{b}_i = \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix}$. One of the biggest advantages of this approach is that the time evolution of the subject-specific true scores τ_{ij} is now entirely characterized by a vector of subject-specific coefficients $\mathbf{b}_i$ that does not vary over time. Basically, the τ_{ij} s and $\mathbf{b}_i$ are equivalent quantities and, therefore, $\mathbf{b}_i$ can be treated as a vector of true scores itself.

We have fully clarified this point in the new version of the manuscript and also clearly stated the relationship between the true scores and the different measures of reliability used in the paper.

Equation 2: elements of Y_i

We followed the editors' advice by making this clearer in the text. In the description of model 1 we now refer to $\mathbf{Y}_i$ as follows:

"$\boldsymbol{Y}_i$ is a p-dimensional vector of repeated measurements on certain trait for subject i"

Use of subscripts

We agree with the editor that clarity can be gained by dropping the subscript i from the beginning. We have adapted the manuscript accordingly.

Time function of reliabilities

This remark of the editor is also related to his/her first comment. In the previous version of the manuscript we had not clarified sufficiently what we understood by true scores and this made unclear the meaning of this function of reliability over time. We believe this issue is now much clearer. Essentially, as previously said, in a longitudinal framework the true scores may vary over time. If one focuses on a single fixed time point then one is back into a cross-sectional scenario and the classical definition of reliability can be applied. Repeating this exercise at each time point leads to a function of reliability over time. This has now been clearly stated in the new version of the manuscript.

Generalizability coefficients

The editor correctly points at one of the essential assumptions of G-theory when applied to a longitudinal setting: the assumption that a subject's true score does not change over time.

We have only compared our coefficients to the G-coefficients in a setting where the data satisfy this and other (e.g. equal error variances at different measurement occasions) assumptions that are posed by G-theory.

In our approach we have simulated this situation by assuming in model 2 that there is only a random intercept and no random slope for time. In such a situation, the true score of a subject would be constant over time.

In case we have longitudinal data that satisfy the assumptions posed by G-theory, it can be shown that the coefficients R_T and R_Λ equal the G-coefficients and have the same interpretation.

We have made this more clear in the present version of the manuscript by giving two examples of G-theory assumptions. On p. 14 we have introduced the following sentence:

*"*Furthermore, when applied in a setting where G-theory assumptions are met, such as non-changing true scores or equal error variances at different time points, the two measures reduce to the index of dependability, and after conditioning on the time points they equal the generalizability coefficient. Both G-coefficients are commonly used to quantify reliability in a longitudinal scenario (Brennan, 2001; Laenen et al., 2009).*"*

Regarding Equation 4

We believe this point has now been clarified in the Methodology section.

Regarding Equation 5

In a similar way we have added a description on what we consider in this model as true scores.

Equation 5: random intercept and random slope coefficients

We thank the editor for having discovered this typo. Random intercept and slope need indeed to be referred to by different parameters. This has now been corrected in the new manuscript.

On page 12, the true value of R_T

The true value of R_T is based on the parameter values that we have chosen for the simulation study. We have clarified this by adding the following sentence:

"Using the parameter values chosen for the simulation study we obtain a value of 0.826."

Changes in the true scores over time

In the new manuscript we have made clearer how we conceptualize the true scores and how they can change over time. To do that we have substantially rewritten the Methodology section.

Figures 1- 3

All figures 1 – 3 were added as graphical evaluations of the model fit. We follow the suggestion of the editor by deleting the graphs in order to gain space. All references to the graphs have been deleted. We now refer in the text to the fact that satisfactory model fits were obtained. We have added the sentence:

"Graphical exploration (not shown) indicated that the models capture the most important data features reasonably well."

Description of final model

In the case study we analyzed two different trials and in each trial two different scales. This results in four models. The final models can be found in Table 4. In this table we summarize the random-effects structure and the structure of the measurement error variance covariance matrix. We have now also extended the description of the final models in the text. We have added the sentence:

"In trial 1 a linear random effects model was selected whereas in trial 2 a quadratic model resulted in the best fit."

Footnotes in Tables 2 and 3

The footnotes are correct; however we agree that they may provoke confusion. We will first explain how they should be interpreted; thereafter we indicate how we have acted to avoid this confusion.

In the simulation study we simulate 250 data sets per setting. For each data set we obtain a point estimate for the measures R_T and R_Λ as well as a 95% confidence interval. This means that we have for each simulation setting 250 confidence intervals. For each confidence interval we analyze whether the true value of the measure (R_T or R_Λ) is within the 95% confidence interval or not. The percentage of times in which this is the case is the coverage probability (CP).

Since we were constructing 95% confidence intervals we would expect this coverage probability to be close to 95% if estimation is working fine.

As a way of testing whether the CP is indeed close to 95% we additionally constructed for each setting a confidence interval around the obtained CP. We decide that the CP is close to 95% if 95% was included in this confidence interval.

As we said, this can indeed bring confusion. On the other hand we also think that the numbers presented in the tables are obvious and can speak for themselves. A formal test is not really needed to draw relevant conclusions. For that reason we have decided to remove the footnotes as well as the stars in the table. We have also removed reference to it in the manuscript.

Confidence bands in Table 4 + software

The confidence intervals for both measures R_T and R_Λ are based on the delta method. The details on their derivation is relatively extensive, and is not retained in the original articles on R_T (Laenen et al., 2007) or R_Λ (Laenen et al., 2009). However, they can be obtained from the authors on request.

SAS macro's are available for the calculation of point estimates as well as confidence intervals for both R_T and R_Λ. With these SAS macro's a small users manual is provided. The macros are available on the following website (www.censtat.uhasselt.be/software/ under Reliability) or can be obtained from the author.

In the case study analysis we have further added the following sentence:

*"*SAS macro's for the calculation of R_T, R_Λ and the corresponding confidence intervals can be obtained from the first author.*"*

APA style

We regret to have missed these details and we have followed the advice of the editor to make the manuscript more conform the APA style.