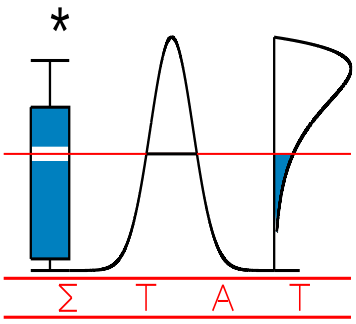
T E C H N I C A L

R E P O R T

0656

ACCOUNTING FOR THE RATER'S MEMORY EFFECT IN RELIABILITY ESTIMATION

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



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

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

Accounting for the rater's memory effect in reliability estimation

Annouschka Laenen, Ariel Alonso, and Geert Molenberghs

Center for Statistics, Hasselt University, Agoralaan, B-3590 Diepenbeek, Belgium

Email: annouschka.laenen@uhasselt.be

Tony Vangeneugden

Tibotec, Johnson & Johnson, 2800 Mechelen, Belgium

Abstract

A difficult point in the design of a test-retest reliability study has always been the length of the time interval between both measurements. To ensure that patients do not evolve on the trait being measured during the time interval, a short period is often recommended. However, if the period is too short, raters might recall their previous answer and tend to repeat it, the infamous memory effect, resulting in an overestimation of the reliability.

It has been shown by Vangeneugden et al. (2004) that a modeling approach to reliability can cope with changes of subjects over time. In this paper, we establish that further adding a so-called serial correlation component to a linear mixed model is a convenient way of correcting for a rater's memory effect. *Key Words:* Brief Psychiatric Rating Scale; Inter-rater Agreement; Linear Mixed Model; Positive and Negative Syndrome Scale; Test-retest Reliability.

1 Introduction

In classical test theory (CTT), the reliability of a measurement is defined as the ratio of the true score variability over the total variability (Lord and Novick 1968). Under certain assumptions, reliability equals the correlation between two measurements on the same subject. Essentially, these assumptions state that for both measurements: (i) the true scores are equal; (ii) the error variances are equal; (iii) the measurement errors are independent. In this framework, the test-retest reliability of a measurement can then be

estimated by rating a group of subjects at two occasions, separated by a time interval and calculating Pearson's correlation coefficient, based on these two sets of measurements.

It is fair to say that test-retest reliability has always been controversial. Indeed, 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 be safe to assume that the underlying process is unlikely to have changed. However, if both measurements are taken sufficiently close in time, it is also quite likely that the rater will recall his previous ratings and his assessments will be influenced by them. Usually he will give similar ratings in each of the replications. The results would overstate the raters' performance. This effect of memory is not limited to the case where raters make subjective decisions; it can also occur in a second attempt on a cognitive ability test, or when filling in a questionnaire on political attitudes (Dunn 1989, Streiner and Norman 1995).

Vangeneugden et al. (2004) and Laenen et al. (2006b) proposed a modeling approach to the estimation of reliability that solves the problem emanating from a probable change in the subject's condition over time. These authors showed that, using linear mixed effects models, it is possible to incorporate this change in the mean's fixed-effects structure, while the variance-covariance part of the model provides us with the components necessary to calculate reliability. Valid reliability estimates can therefore be obtained in more general scenarios when more than two replications per subject are available and the underlying true scores change over time.

In this paper, we will illustrate that, additionally, the problems stemming from a potential rater's memory effect can be tackled by using linear mixed models. The memory effect is accounted for by adding a serial correlation component to the model. Meaningful reliability estimates can then be obtained from repeated measurements that are contaminated with a memory effect.

In Section 2, we introduce a modeling framework suitable for studying reliability in

the presence of memory effect. In Section 3, the impact of such a memory effect of the rater on the reliability estimation is explored by means of a simulation study. Section 4 illustrates the methodology by means of a case study.

2 A Modeling Approach

Linear mixed models (LMM) allow extension of the CCT modeling framework to more general settings, applicable whenever the stringent assumptions described in Section 1 become implausible. Assuming a balanced study design, with the same number of measurements taken at a common set of time points for all study subjects, this model can be written as

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, \tag{1}$$

where $\boldsymbol{Y}_i$ is the p-dimensional vector of responses Y_{ij} for subject $i = 1, \ldots, n$ and occasions $j = 1, \ldots, p$. X_i and Z_i are fixed $p_i \times q$ and $p_i \times r$ dimensional matrices, respectively, of known covariates, $\boldsymbol{\beta}$ is the q-dimensional vector of fixed effects, $\boldsymbol{b_i} \sim N(\mathbf{0}, D)$ is the r-dimensional vector containing the random effects, and $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma)$ is a p-dimensional vector of errors. Additionally, D is a general $r \times r$ covariance matrix and Σ is a $p \times p$ variance-covariance matrix. Finally, the vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are assumed independent.

One of the features making (1) appealing for reliability estimation is its natural capability to simultaneously account for fixed and random effects. The change of a patient's condition over time can then be modeled within the fixed structure, obviating the need to assume a steady-state condition (Vangeneugden et al. 2004, Laenen et al. 2006b). Additionally, LMMs are able to naturally distinguish between different sources of variability (Laird and Ware 1982, Verbeke and Molenberghs 2000), a property very relevant in the light of reliability estimation. For calculating reliability, a distinction needs to be made between variability coming from the true scores of the subjects, a subject-specific or random effects, and the residual variability. Note that LMMs decompose the total variability of the longitudinal observations as $V_i = Z_i D Z_i' + \Sigma_i$ where $Z_i D Z_i'$ accounts for

the variability of the subject-specific parameters or true scores, and Σ_i includes all the remaining sources of variability. In a balanced design, subscripts i can be dropped from the above matrices.

The p diagonal elements of Σ are the variances of the measurement errors at the various time points, whereas the off-diagonal elements are the residual covariances between the measurement errors at any two points in time. We claim that a memory effect can be incorporated into our framework by introducing a serial correlation term into the model. Essentially, a serial correlation structure would imply, as expected in the presence of a memory effect, that observations closer in time are more similar than observations taken further apart. In the simplest scenario, like the one used in classical test theory, all correlations within a subject can be described through the variance of the random effects, the true scores, and the variance of the measurement error. In such a setting, the reliability and the within-subject correlation are equivalent concepts. In more complex situations, like the one considered in the present work, the within-subject correlation cannot be explained fully through the variability of the random effects and uncorrelated measurement error terms. Therefore, correlated error terms should be used and this correlation can be taken into account using a serial correlation structure. Arguably, in this scenario the link between reliability and within-subject correlation breaks down. One commonly used structure is the autoregressive one, where $\Sigma = \tau^2 H$, with $H_{jk} = \rho^{d_{jk}}$. Here, τ^2 is a common error variance for all the time points, H is a correlation matrix with ρ the correlation between two measurements taken one unit of time apart, and d_{jk} the time lag between two measurements taken at time points j and k. A strong memory effect would reflect in a large value of ρ. Including this term into the model will remove the potential effect that it would exert on the estimation of the variance components, were not taken into account.

As stated before, in complex settings, the link between reliability and within-subject correlation breaks down and an extension of the classical concept is needed. Laenen et al. (2006ab) proposed such an extension based on a minimum set of defining properties. Further, these authors introduced two families of parameters designed to evaluate relia-

bility, with building block the cross-eigenvalues associated with the matrices Σ and V. Considering some mathematical arguments and simulation studies they found two "optimal" members, one from each family, which one may want to use in practical situations, the so-called R_T and R_Λ. These two measures are defined as $R_T = 1 - \text{tr}(\Sigma)/\text{tr}(V)$ and $R_\Lambda = 1 - |\Sigma V^{-1}|$. Note that both proposals quantify the proportion of the total variability not owing to measurement error, exactly as in the classical definition of reliability. Actually, when applied in the classical setting, they both reduce to the classical definition. The main difference between both measures is their different approach for summarizing the variability in a variance-covariance matrix. For R_T, this is done by means of the trace of the matrix, and for R_Λ the determinant is used. Laenen et al. (2006b) have shown that these measures for reliability lead to quite distinct interpretations. The R_T should be interpreted as the average reliability over measurements at different time points, whereas the R_Λ can be seen as the overall reliability over the entire sequence of measurements. Owing to this characteristic, this measure increases with the number of repeated measurements. In the following section we will study, via simulations, the impact of a memory effect on the R_T and the R_Λ when this effect either is or is not included into the model.

3 A Simulation Study

To study the consequences of a memory effect on the estimation of the reliability, we set up a simulation study where the data generating mechanism was chosen to resemble a realistic longitudinal study, with scores within subjects changing over time. First, data were generated based on the following random-intercept model:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 Z_i + b_i + \varepsilon_{ij}, \tag{2}$$

with $b_i \sim N(0, \sigma_b^2)$, $\varepsilon \sim N(0, \tau^2 H)$, t_{ij} the time at which measurement j for subject i is taken, and Z_i the treatment allocation for subject i. The value of σ_b^2 was fixed at 300 and τ^2 equaled 100, corresponding to a situation where the error variability accounts for one quarter of the total variability. In this model, the true score of a subject, which is

the sum of the fixed effects and the random intercept, can change over time due to the presence of a time variable in the fixed effects part of the model. However, this change is the same for all subjects. This model would correspond to an *essentially tau-equivalent model*, where the true scores differ only by a "constant", even though such a constant would transform to a function of time in our case. Second, data were generated based on an extended model, including a random slope for time in addition to the random intercept, thus expanding (2) to

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 Z_i + b_{1i} + b_{2i} t_{ij} + \varepsilon_{ij}, \tag{3}$$

where now $b_i \sim N(0, D)$, $\varepsilon \sim N(0, \tau^2 H)$ and

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

The time variable now also appears in the random part of the model, allowing for subject-specific evolutions over time. This model resembles, but is not exactly equal to, a *congeneric model*. A congeneric model would be (3) without the random intercept term b_{1i}.

In both models (2) and (3), the value of ρ was set to either 0.1, 0.5, or 0.8, corresponding to a small, moderate, and large memory effect, respectively. Values for the fixed effects were set to $\beta_0 = 85$, $\beta_1 = -2.5$, and $\beta_2 = 3$, based on real case study results. Data were generated for six equally spaced time points, at weeks 0, 2, 4, 6, 8, and 10, and sample sizes were equal to 250. This amounts to six settings and 250 data sets for each were generated.

We analyzed the data in two different ways. First, we calculated the Pearson correlation coefficient between the first measurement and the measurements at later occasions. The goal was to study the impact of the change of subjects over time and a memory effect on the classical approach to reliability estimation. Second, we fit the data using two different models: (i) a correctly specified model that includes a serial correlation component with an autoregressive structure and (ii) a misspecified model that includes a variance-components structure for the residual part ($\Sigma = \sigma^2 I$), ignoring the presence of

Table 1: *Instability and memory effect on reliability measures: correlation coefficients. RI refers to random-intercepts model (2), RIS refers to model (2) with random intercepts, random slopes, and serial correlation. ρ is the 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})	$(Y_{i0}, Y_{i,10})$
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

the memory effect. We calculated R_T and R_Λ based on these two models to investigate the impact of the memory effect on both measures for reliability.

Table 1 shows the Pearson correlations between the outcomes of the first measurement (Y_{i0}) and the outcomes at later measurement occasions (Y_{i2}–$Y_{i,10}$). Under the random intercept model (RI, model 2)) the true reliability according to the classical definition, as the ratio of the true score variability to the total variability, can easily be obtained as:

$$R = R_T = R_\Lambda = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} = \frac{300}{300 + 100} = 0.75.$$

In the case of a small memory component, the effect of it already fades away at 4 weeks and the correlation coefficient gives stable and trustworthy results as an estimator of reliability. Since the change of the true score over the various measurements is constant, it does not influence the correlation. The correlation is therefore a valid estimator for reliability in this setting. However, with a memory component of increasing importance, the reliability is strongly overestimated, especially for small time lags, and it takes longer before the effect of memory fades out.

Since the classical definition of reliability does not apply to a model with random intercept and slope (RIS), there is no such thing as the *true* reliability for this model. Therefore, we need to quantify reliability using the R_T and R_Λ concepts, reviewed in Section 2. Table 3 presents the true values of both measures in the simulation study. Different subjects can now change over time in different ways. Table 1 shows that these changes lower the correlations when time lag increases. The effect of memory is also clearly visible in the lower part of Table 1. Obviously, in this scenario, the classical approach to reliability is strongly misleading.

The correlation is a good estimator for reliability if a certain set of stringent conditions is fulfilled. However, when they do not hold, a modeling approach offers a sensible alternative. It has been shown previously that linear mixed models can handle changes of subjects over time, even if evolutions differ across subjects. In what follows, we evaluate the impact of accounting for a memory effect by introducing a serial correlation component into the model. Tables 2 and 3 summarize the results of the modeling approach to reliability, for the RI model (2) and the RIS model (3), respectively. Both tables 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. Both tables show that, when the model does not include a serial correlation component to account for the memory effect, both $\widehat{R}_T$ and $\widehat{R}_\Lambda$ overestimate the real value, exactly as expected. In case of a small memory effect, the estimates are still relatively close to the real value. However, for an increasingly important memory component, the real value is largely overestimated. The estimated confidence interval then almost never contains the true value. Larger sample sizes further confirmed these results: the reliability is then even more strongly overestimated and the coverage probabilities are even lower.

When a serial correlation component is included into the model to account for the memory effect, estimates for R_T and R_Λ are much closer to the real values and coverage probabilities are close to 95%. Only for a very large memory effect ($\rho = 0.8$) are the true values for both measures somewhat underestimated and coverage probabilities below 95%

Table 2: *Memory effect on reliability measures: random intercept model (2). 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
autoregressive	0.1	0.750	0.748	95.2	0.939	0.938	96.4
autoregressive	0.5	0.750	0.746	95.2	0.889	0.886	96.0
autoregressive	0.8	0.750	0.734	95.2	0.824	0.808	96.0

for the RIS model. However, this situation improves for larger sample sizes.

Note that the true value of R_Λ decreases when the serial correlation increases. R_Λ has the ability to increase with the number of time points, due to the fact that every new observation brings additional information (Laenen et al. 2006b). However, for an equal number of time points, we have less information when different observations are strongly correlated, explaining lower R_Λ for larger values of ρ.

4 A Case Study in Schizophrenia

Based on data from a clinical trial comprising 453 patients, we estimate the reliabilities of three different rating scales conceived for measuring the severity in schizophrenic patients: the Positive and Negative Syndrome Scale (PANSS), the Brief Psychiatric Rating Scale (BPRS) and the Clinical Global Impression (CGI). The PANSS is a 30-item scale allowing to distinguish between the typical positive and negative symptoms by means of two subscales. The BPRS has 18 items, a subset of PANSS, but the focus of this scale is rather on positive symptoms. The CGI is a general one-item tool that registers the change of

Table 3: *Memory effect on reliability measures: random intercepts and slopes model (3). 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
autoregressive	0.1	0.826	0.825	97.6	0.986	0.986	96.8
autoregressive	0.5	0.826	0.821	96.8	0.972	0.968	97.2
autoregressive	0.8	0.826	0.812	88.1	0.965	0.955	91.9

the patient's condition compared to baseline measurement using seven categories, ranging from 'very much improved' to 'very much worsened'. The trial compared a new treatment against an active control. Study subjects were measured at weeks 0, 1, 2, 4, 6, and 8.

Since interest primarily lies in the covariance structure, an elaborate fixed effects structure was adopted, containing categorical time, treatment, and treatment by time interaction. The selection of the covariance structure was based on the AIC. Restricted maximum likelihood was used for parameter estimation (Verbeke and 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 (PANSS, BPRS, or CGI) 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 $\Sigma = \text{diag}(\sigma_j^2)$. In contrast, for CGI, $\Sigma = \tau^2 H$, with H corresponding to a spatial power serial correlation structure. Table 4 presents

Table 4: *Schizophrenia Study: estimates [95% confidence intervals] for R_T and R_Λ on the Positive and Negative Syndrome Scale (PANSS), the Brief Psychiatric Rating Scale (BPRS), and Clinical Global Impression (CGI).*

	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; 0.995]	0.991 [0.988; 0.993]	0.977 [0.969; 0.983]

the reliability estimates for the three scales and a 95% confidence interval. Interestingly, none of the two multi-item scales give rise to a serial correlation component in the best fitting model. However, the one-item CGI does. A plausible explanation is that such a one-item scale scored by the specialist physician, is more prone to a memory effect than a multi-item scale. Fortunately, by modeling a serial correlation, overestimating the reliability due to such an effect can be avoided. This analysis illustrates that based, on longitudinal scale outcomes, reliability of the scales can be derived in the same way in case of a memory effect as without such an effect.

5 Concluding Remarks

In the simplest scenario, arising, for example, in classical test theory, all correlations within subjects can be described through the variance of the random effects (true scores) and the variance of the measurement error. In such a setting, the reliability and the within-subject correlation are equivalent concepts: the reliability equals the correlation between two measurements on the same subjects. In more complex situations, like the one considered in this paper, the within-subject correlation cannot be explained fully through the variability of the random effects and uncorrelated measurement error terms. Therefore, correlated error terms should be used, or other components should be added to the model, such as, for instance, a serial correlation structure. Arguably, in this scenario,

the link between the reliability and within-subject correlation breaks down.

When a memory effect is present, the condition of the subject at consecutive measurements will appear more similar than they actually are. Incorrectly assuming that the observations are independent will lead to an underestimation of the within-subject variability. However, accounting for the fact that these values are correlated, by including a correlation term, corrects the estimate of this within-subject variability, and allows for estimation of reliability in an unbiased way.

Using data from a clinical trial in schizophrenic patients, we have shown that the ideas developed in this paper can be applied when a memory effect in a rating scale is either present or absent, underscoring the power and flexibility of the method proposed.

While reliability can be summarized by means of more than one measure, each one focusing on different aspects, the measures retain the simplicity of R^2 coefficients, ranging between 0 and 1, so that the intuition coming with the classical approach is retained.

References

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

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

Laenen, A., Alonso, A., Molenberghs, G., and Vangeneugden, T. (2006b). Reliability of a longitudinal sequence of scale ratings. *Submitted for publication.*

Laenen, A., Vangeneugden, T., Geys, H., and Molenberghs, G. (2006c). Generalized reliability estimation using repeated measurements.*British Journal of Mathematical and Statistical Psychology* **59**, 113–131.

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

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

Streiner, D.L. and Norman, G.R. (1995). *Health Measurement Scales.* Oxford: Oxford University Press.

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

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