A measure for the reliability of a rating scale based on longitudinal clinical trial data

LAENEN, Annouschka; ALONSO ABAD, Ariel & MOLENBERGHS, Geert (2007) A measure for the reliability of a rating scale based on longitudinal clinical trial data. In: PSYCHOMETRIKA, 72(3). p. 443-448.

DOI: 10.1007/s11336-007-9002-7
Handle: http://hdl.handle.net/1942/7816

A MEASURE RELIABILITY OF A RATING SCALE BASED ON LONGITUDINAL CLINICAL TRIAL DATA.

Annouschka Laenen

hasselt university, belgium.

Ariel Alonso

hasselt university, belgium.

Geert Molenberghs

hasselt university, belgium.

Correspondence should be sent to

E-Mail: annouschka.laenen@uhasselt.be
Phone: +32 11 268292
Fax: +32 11 268299

Acknowledgements: 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."

A MEASURE FOR THE RELIABILITY OF A RATING SCALE BASED ON LONGITUDINAL CLINICAL TRIAL DATA.

Abstract

A new measure for reliability of a rating scale is introduced, based on the classical definition of reliability, as the ratio of the true score variance and the total variance. Clinical trial data can be employed to estimate the reliability of the scale in use, whenever repeated measurements are taken. The reliability is estimated from the covariance parameters obtained from a linear mixed model. The method provides a single number to express the reliability of the scale, but allows for the study of the reliability's time evolution. The method is illustrated using a case study in schizophrenia.

Key words: Reliability, Linear mixed model, Longitudinal data, Psychiatry, Rating scale.

1. Introduction

Many measurements in medical practice and research are based on observations made by clinicians using rating scales. The subjective nature of these scales inquires investigation of their psychometric qualities, such as the validity and the reliability.

Classically, test-retest reliability is estimated as the correlation between two consecutive measures assuming a steady state condition in the subjects. However, rating scales are often used longitudinally in clinical practice, where this assumption is unlikely. Vangeneugden et al. (2004) have shown that linear mixed models allow to model the change in the subjects' condition within the fixed effects structure, estimating it simultaneously with the covariance parameters needed for the calculation of the intraclass correlation (ICC).

The ICC is the ratio of the between-subject variability and the total variability, and is easy to obtain from a linear mixed model with only a random intercept. Vangeneugden et al. (2004) extended the calculation of the ICC to more complicated models where a random slope for time and a component of serial correlation are allowed. Depending on the complexity of the model the reliability is then estimated as a single number, a correlation depending on the time lag between two measurements, or a correlation matrix for any pair of measurements. Even though this approach allows us to investigate reliability in a general setting, using a correlation matrix to quantify it can lead to interpretational difficulties. The problem is further aggravated when we want to compare two or more

scales based on their reliabilities. It is not straightforward how the corresponding matrices should be compared in order to determine which is the most reliable instrument.

In this paper, the basic definition of reliability as "the ratio between the true score variance to the observed score variance" is the starting point for a new definition of reliability, delivering a single yet meaningful measure, independently of the model used to fit the data, facilitating its interpretation and applicability. Section 2 presents the underlying model, and Section 3 introduces this new measure. In Section 4 we apply the methodology to clinical trial data on schizophrenia.

2. The Linear Mixed Model

Linear mixed models allow repeated measurements to be described entirely in terms of means, variances, and covariances (Laird and Ware, 1982; Verbeke and Molenberghs, 2000). One can distinguish between three components of variability. Part of the covariance structure arises from subject-specific random effects explaining the heterogeneity between individuals. Another component of variability is the serial correlation, accounting for the fact that measurements taken closer in time tend to be stronger correlated. The third component is the measurement error. The model can be written as

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon}_{(1)\boldsymbol{i}} + \boldsymbol{\varepsilon}_{(2)\boldsymbol{i}} \tag{1}$$

$$\boldsymbol{b_i} \sim N(\mathbf{0}, D), \quad \boldsymbol{\varepsilon}_{(1)\boldsymbol{i}} \sim N(\mathbf{0}, \Sigma_{Ri}), \quad \boldsymbol{\varepsilon}_{(2)\boldsymbol{i}} \sim N(\mathbf{0}, \tau^2 H_i),$$

$$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_{(1)1}, \ldots, \boldsymbol{\varepsilon}_{(1)N}, \boldsymbol{\varepsilon}_{(2)1}, \ldots, \boldsymbol{\varepsilon}_{(2)N} \text{ independent,}$$

where $\boldsymbol{Y}_i$ is the p_i dimensional vector of responses for subject i, with N subjects and p_i observations per subject. X_i and Z_i are fixed $(p_i \times q)$ and $(p_i \times r)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is the q-dimensional vector of fixed effects, $\boldsymbol{b_i}$ is the r-dimensional vector containing the random effects, $\boldsymbol{\varepsilon}_{(2)\boldsymbol{i}}$ is a p_i-dimensional vector of components of serial correlation, and $\boldsymbol{\varepsilon}_{(1)\boldsymbol{i}}$ is a p_i-dimensional vector of residual errors. Additionally, D is a general $(r \times r)$ covariance matrix, H_i is a $(p_i \times p_i)$ correlation matrix, τ^2 is a variance parameter, and Σ_{Ri} is an $(p_i \times p_i)$ covariance matrix. Furthermore, H_i and Σ_{Ri} depend on i only through their dimension p_i, i.e., the parameters will not depend upon i.

Model (1) implies the marginal model $\boldsymbol{Y_i} \sim N(X_i\boldsymbol{\beta}, V_i)$ where $V_i = \Sigma_{Di} + \Sigma_i$ with $\Sigma_{Di} = Z_i D Z_i'$ and $\Sigma_i = \tau^2 H_i + \Sigma_{Ri}$. The total variability is decomposed into two parts: the first one (Σ_{Di}) accounts for the variability of the subject-specific parameters or true scores, whereas the second one (Σ_i) includes all the remaining sources of variability.

3. Generalizing Reliability

Extending the concept of reliability we want to find a balance between a general and flexible definition and, at the same time, keeping the intuition and appealing interpretation behind the concept used in the classical test theory (CTT) (Lord and Novick, 1968). Therefore, based on the different formulations of reliability within the CTT we propose the following

set of defining properties that any measure of reliability R should satisfy: (i) $0 \leq R \leq 1$, (ii) $R = 0$ if and only if there is only measurement error: $V_i = \Sigma_i$, (iii) $R = 1$ if and only if there is no measurement error: $\Sigma_i = 0$, and (iv) in a cross-sectional setting R equals the true score variance to observed variance ratio.

3.1. R_T: A Measure for Reliability

We now summarize the variability of the repeated measurements on the scale by the trace of its variance-covariance matrix. Similarly, we summarize the error variability by the trace of the variance-covariance matrix associated with the error vectors $\boldsymbol{\varepsilon}_{(1)\boldsymbol{i}}$ and $\boldsymbol{\varepsilon}_{(2)\boldsymbol{i}}$. We then propose to quantify the reliability by the measure R_T:

$$R_T = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{tr}(V_i) - \text{tr}(\Sigma_i)}{\text{tr}(V_i)}. \tag{2}$$

Note the connection of (2) to the equation $D[D+(Z_i'\Sigma_i^{-1}Z_i)^{-1}]^{-1}$, proposed by Bock (1966, 1983) as the "multivariate analogue of reliability". In (2), $\text{tr}(V_i)$ accounts for the total variability in the observations for subject i, whereas $\text{tr}(\Sigma_i)$ accounts for the measurement error variability in this subjects' observations. Therefore, $\frac{\text{tr}(V_i) - \text{tr}(\Sigma_i)}{\text{tr}(V_i)}$ is the proportion of all the variability in the observations of subject i that is not due to measurement error. R_T can then be interpreted as the average of all subjects' contributions. It is easy to show that R_T satisfies properties (i) – (iv). Also, when model (1) reduces to a random intercept

model, R_T reduces to the classical case, i.e. the true score variance to observed variance ratio. Additionally, in the balanced setting, and assuming that $\Sigma_i = \Sigma$ and $\Sigma_{Di} = \Sigma_D$, R_T takes the simpler form:

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

3.2. Estimating R_T

A maximum likelihood estimate $\hat{R}_T$ can be obtained by replacing V_i (or V) and Σ_i (or Σ) in (2) (or (3)) by the matrices' maximum likelihood estimates, $\hat{V}_i$ (or $\hat{V}$) and $\hat{\Sigma}_i$ (or $\hat{\Sigma}$), respectively. A confidence interval for R_T can then be obtained by applying the delta method. According to this method we have: $\hat{R}_T \sim N\left(R_T, \Delta\Sigma_P\Delta'\right)$, where Σ_P is the variance-covariance matrix of the variance-covariance parameter estimates and $\Delta' = \left(\frac{\partial R_T}{\partial D}, \frac{\partial R_T}{\partial \tau^2}, \frac{\partial R_T}{\partial \Sigma_R}\right)$. A detailed derivation of the different elements of Δ can be obtained from the authors' website (www.censtat.uhasselt.be/staff). To avoid confidence limits that exceed the $[0, 1]$ range, a logit transformation is applied to R_T so that $l = \log\left(\frac{R_T}{1 - R_T}\right)$. A $(1 - \alpha)\%$ confidence interval for R_T takes the form $[L_1, L_2]$ with $L_i = \frac{e^{l_i}}{1 + e^{l_i}}$, and $l_i = l + (-1)^i \frac{z_{1-\alpha/2}}{R_T(1 - R_T)} \sqrt{\Delta\Sigma_P\Delta'}, \quad i = 1, 2.$

4. A Case Study

The case study is a clinical trial with 453 patients, comparing risperidone to conventional antipsychotics for the treatment of chronic schizophrenia. Patients were evaluated at baseline and after 1, 2, 4, 6, and 8 weeks, by means of three rating scales. The Positive and Negative Syndrome Scale (PANSS) is a 30-item scale, of which the Brief Psychiatric Rating Scale (BPRS) is essentially a shorter version, with 18 items, and the Clinical Global Impression (CGI) is a 7-grade scale to indicate the patients' change compared to baseline. Reliability estimates were derived for the three scales.

Since interest primarily lies in the covariance structure, a complex fixed effects structure was adopted (Diggle, Liang & Zeger, 1994), containing time (categorically), treatment, and treatment by time interaction. For the random-effects structure, we considered (a) a random intercept, (b) a random intercept and time, and (c) a random intercept, time, and time squared. Even though model (1) is fully general, and allows to decompose the error variability into a serial correlation and a residual variance component, in many practical applications this will lead to identifiability issues. For the case study, the variance-covariance matrix (here: $\Sigma_i = \Sigma = \tau^2 H + \Sigma_R$) was modeled by (a) only a serial correlation component $\tau^2 H$ (spatial gaussian, exponential, and power), (b) only a residual component Σ_R (diagonal with either homogeneous variances or heterogeneous variances depending on the time point), or (c) both (with Σ_R constrained to be equal to $\sigma^2 I$). The model with the

TABLE 1.
Reliability estimates for three scales.

		95 % confidence interval	
Parameter	$\hat{R}_T$	lower limit	upper limit
PANSS	0.846	0.825	0.865
BPRS	0.821	0.797	0.842
CGI	0.737	0.700	0.771

lowest AIC was selected. Restricted maximum likelihood was used for parameter estimation (Verbeke and Molenberghs 2000). For all three scales, the final model has the general form:

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

where Y_{ij} denotes the outcome for subject i at time point 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 1 presents the reliability estimates for the three scales and a 95% confidence interval. A SAS macro for these calculations is available on the authors' website (www.censtat.uhasselt.be/staff).

PANSS, the most extended scale, has the highest reliability. Remarkably, BPRS, which is 12 items shorter, has a reliability of a similar magnitude. Historically, PANSS was

conceived as a completion of BPRS, but these results illustrate that this additional complexity does not bring much gain in reliability. Analogous results were found by Alonso et al. (2002) when studying criterion validity. Similar values were obtained for trial-level validity and individual validity for PANSS and BPRS. R_T for the one-item CGI is, not surprisingly, somewhat below the former results.

Note that we can also estimate the reliability at each time point as:

$$R_{Tj} = \frac{\boldsymbol{z}_j D \boldsymbol{z}_j'}{\boldsymbol{z}_j D \boldsymbol{z}_j' + \tau^2 + \sigma_j^2}$$

with $\boldsymbol{z}_j$ the jth row of Z, $j = 1, ..., p$. Figure 1 shows the estimated time point reliabilities for the schizophrenia data. The CGI measures change relative to baseline and therefore does not yield a score at baseline. The graph shows an increasing tendency for the reliability over time for all scales. Note that similar results were obtained for PANSS by Vangeneugden et al. (2004). We speculate that this could be the result of a learning effect of the raters.

5. Discussion

A scale, to be useful in practice, should exhibit small measurement error. Therefore, in the evaluation of the scale, reliability is a concept of the utmost importance. A test-retest reliability study essentially consists of repeating the same measurement. Like Vangeneugden

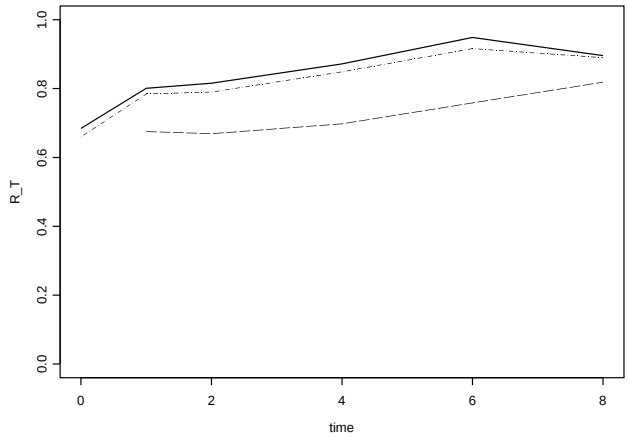

FIGURE 1.
Schizophrenia Study. Reliability per time point for three scales

et al. (2004) we use linear mixed models to account for the different sources of variability but our definition and quantification of reliability differ from their proposal. In the present paper, the approach to reliability starts from the ratio of the true score variance to the total variance. The trace is used to summarize the variability in variance-covariance matrices. Both approaches differ in two main aspects. First, in the methodology proposed by Vangeneugden et al. (2004) the serial correlation is combined with between-subject variation calculating the ICC. In the present approach, only between-subject variability is considered as true-score variability, implying that serial correlation is treated as measurement error or residual variability. Intuitively, reliability can be described as the capacity of a scale to discriminate between subjects. Following this idea we want to quantify how much of the total variability can be explained by differences between the subjects to which

the scale was applied. Since the variability between subjects is fully captured by the random effects, we divide the total variation in: variability coming from the random effects and residual variability. If the former explains a large percentage of the total variability, then the differences observed when using the scale are mainly due to the differences between the subjects evaluated with the instrument. Arguably, such a scale will have a high discriminating capacity and could be considered reliable. A second difference concerns the outcome measure. In Vangeneugden et al. (2004) reliability is expressed as a decreasing function of the time lag between two measurements, or a full correlation matrix when models with complex covariance structures are considered. In our approach, a single measure of reliability is given. Seemingly such a measure can offer interpretational and practical advantages, especially when more than one scale must be compared.

Finally, we would like to underscore the importance of the model building step. In principle, different models could lead to a similar fit, similar AIC values, for the data at hand and, at the same time, to different conclusions regarding reliability. However, our experience with the present method illustrates that such cases are rare. As a general strategy, we recommend that, if different models fit the data equally well, then the R_T should be calculated for all of them as a way of sensitivity analysis. In case discrepant results are observed, other types of considerations, apart from AIC values, could be taken into account to select the most plausible and sensible model. The opinion of the experts in

the field could be of great value in this case. Nevertheless, in such a situation, conclusions should always be taken with care.

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.

Bock, R.D. (1966). Contributions of multivariate experimental designs to educational research. In R.B. Cattell (Ed.). *Handbook of multivariate experimental psychology.* Chicago: Rand-McNally.

Bock, R.D. (1983). The discrete Bayesian. In H. Wainer & S. Messick (Eds.). *Principals of Modern Psychological Measurement.* Hillsdale, NJ: Erlbaum.

Diggle, P.J., Liang, K.-Y., Zeger S.L. (1994). *Analysis of Longitudinal Data.* Oxford Science Publications. Oxford: Clarendon Press.

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.

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* (1), 13–30.

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