Psychometric Validation of Continuous Rating Scales from Complex Data.

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting wiskunde
te verdedigen door:

Annouschka LAENEN

Promotor : Prof. dr. Geert Molenberghs

Dankwoord

Toen ik voor het eerst in CenStat arriveerde, had ik nooit durven dromen dat ik er negen jaar later mijn doctoraatswerk zou mogen verdedigen. Het feit dat ik hier vandaag toch sta, heb ik aan de steun van verschillende mensen te danken.

Geert, jij hebt daarin een fundamentele rol gespeeld. Vanaf het eerste moment in CenStat kreeg ik de kans om mezelf te ontwikkelen door de twee masteropleidingen te volgen. Wanneer ik dacht dat ik de boeken bijna kon neerleggen, kwam het idee van een doctoraat. Een van de vele adviezen die je me gaf, was om een link te leggen tussen mijn werk over betrouwbaarheid en het werk van Ariel over 'surrogate markers'. Hoewel we aanvankelijk beide nogal sceptisch waren, moet ik toegeven dat dit inderdaad een 'geslaagd huwelijk' is gebleken :-)

Een dubbel woord van dank is voor jou, Ariel. Door met jou te kunnen samenwerken is het werk voor mij een groot plezier geweest. Zeer concrete problemen eindigden wel eens in lange filosofische discussies. En precies daardoor heb ik vele inzichten verworven en de boeiende wereld van statistiek en weteschappelijk onderzoek beter leren kennen.

Ook mijn ouders verdienen een 'dankuwel'. Zij hebben me mijn hele leven aangemoedigd om dingen te doen waarvan ik dacht ze niet te kunnen. Dank je mama en papa, zonder jullie vertrouwen had ik het niet gekund.

Naast Geert en Ariel wil ik ook mijn collega's en co-auteurs Tony, Helena en Craig bedanken voor een fijne en constructieve samenwerking. Ook dank aan al mijn collega's voor een aangename werksfeer, en speciaal ook aan Saskia die altijd voor me klaar staat. Bedankt!

Annouschka Laenen
Diepenbeek, 12 December 2008

Contents

List of Abbreviations

AIC	Akaike's Information Criterion
AGFI	Adjusted Goodness-of-Fit Index
BPRS	Brief Psychiatric Rating Scale
CGI	Clinician's Global Impression
CAIC	Consistent Akaike's Information Criterion
CP	Coverage Probability
CTT	Classical Test Theory
ECT	Electroconvulsive Therapy
FDA	Food and Drug Administration
HAMD	Hamilton Depression Rating Scale
HAMA	Hamilton Anxiety Rating Scale
ICC	Intraclass Correlation Coefficient
IRT	Item Response Theory
LMM	Linear Mixed Model
MADRS	Montgomery-Åsberg Depression Rating Scale
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
PANSS	Positive And Negative Syndrome Scale
REML	Restricted Maximum Likelihood
REMLE	Restricted Maximum Likelihood Estimator
RMR	Root Mean Square Residual
SBC	Schwarz's Bayesian Criterion

List of Tables

List of Figures

Publications

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

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

Laenen, A., Alonso, A., Molenberghs, G., and Vangeneugden, T. (2008). Reliability of a longitudinal sequence of scale ratings. *Psychometrika.* DOI: 10.1007/S11336-008-9079-7

Laenen, A., Alonso, A., Molenberghs, G., Mallinckrodt, C. H., and Vangeneugden, T. (2008). Using longitudinal data from a clinical trial in depression to assess the reliability of its outcome scales. *Journal of Psychiatric Research.* DOI: 10.1016/j.jpsychires.2008.09.010

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

Laenen, A., Alonso, A., Molenberghs, G., Vangeneugden, T., and Mallinckrodt, C. H. (2008). Impact of ignoring serial correlation and memory effect on reliability estimates. (*Submitted for publication*).

Alonso, A., Laenen, A., Molenberghs, G., Geys, H., and Vangeneugden, T. (2008). Reliability of Single Administered Tests: A Unified Approach. (*Submitted for publication.*)

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.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., and Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, **61**, 295–304.

Molenberghs, G., Laenen, A., and Vangeneugden T. (2007). Estimating reliability and generalizability from hierarchical biomedical data. *Journal of Biopharmaceutical Statistics*, **17**, 595–627.

Vangeneugden, T., Molenberghs, G., Laenen, A., Alonso, A., and Geys, H. (2007). Generalizability in non-Gaussian longitudinal clinical trial data based on generalized linear mixed models. *Journal of Biopharmaceutical Statistics*, **18**, 691–712.

Vangeneugden, T., Molenberghs, G., Laenen, A., Geys, H., Beunckens, C., and Sotto, C. (2008). Marginal correlation in longitudinal binary data based on generalized linear mixed models. (*Submitted for publication*).

Chapter 1

Introduction

With the development of psychopharmaca and their evaluation in clinical trials came also the development of rating scales for mental health conditions. Rating scales consist of a list of questions or statements that the doctor or patient answers and each response is given a score. Finally, all the scores are added up to obtain a total score. Before being used, however, such scales need to be checked on their reliability and validity. On the evaluation of these two properties, a whole tradition of psychometric research exists. Nevertheless, the classical psychometric methods are not sufficiently flexible to be applied in clinical trials with complex designs.

The main objective of this thesis is to extend classical psychometric methods for the evaluation of reliability to more general settings, that may include complex data structures such as longitudinal and multivariate measurements.

One of the distinctive characteristics of reliability is that it is a population-dependent concept. Indeed, an instrument that gives reliable measurements for one group of individuals might less do so when it is applied to a different group. Therefore, it is advisable to investigate the reliability of an instrument, not only in the developmental phase, but also each time it is used in a different population. However, since reliability research implies additional investment of time and resources, it is often omitted. In the present work, we look for methods to evaluate the reliability of outcome scales using clinical trial data. Such an approach will bring indubitable advantages. For instance, it will allow to study reliability in more realistic settings, i.e, the settings in which the scales are frequently applied in scientific research and

clinical practice. It will also allow to check the reliability of the scale every time it is used in clinical investigation, increasing our understanding of its performance across different populations. Furthermore, clinical trials are known for their stringent procedures in order to assure the quality of the data and they frequently involved large sample sizes. Therefore, using clinical studies for reliability research will guarantee that accurate results can be obtained.

1.1 Mental Health Measurement

In the Middle Ages mental illness was seen as alienation from God. Confession and penance were essential to the cure. In addition, treatments using purgatives, bloodletting, and such practices as trepanation were used. In the growing cities and towns, facilities were developed where the poor, outcast, and mentally ill could be confined and maintained. However, hospitals specifically for the care of the mentally ill were rare. The 18th century, Enlightenment influences resulted in a more optimistic outlook for the treatment of insanity. Nevertheless, treatments were available only to a select few.

In 1899, Sigmund Freud published The Interpretation of Dreams, and psychoanalysis became one of the most influential treatment methods in the twentieth century, but again, only a few individuals could afford it (Merkel 26.6.2008). By the mid-1940s, treatment of the mentally ill took a new turn, with the advent of electroconvulsive therapy (ECT), insulin shock therapy, and the use of frontal lobotomy. In modern times, insulin shock therapy and lobotomies are viewed as being almost as barbaric as earlier "treatments", though in their own context they were seen as the first options which produced any noticeable effect on their patients. ECT is still used in the West, but it is seen as a last resort for treatment of mood disorders, and is administered much more safely than in the past (NCLS 26.6.2008). By the mid-1950s, the first psychiatric drugs became available for the treatment of mental illness which revolutionized psychiatric care and provided new ways for many of the severely mentally ill to return to normal life in society. Newly developed antidepressants were used to treat cases of depression, and the introduction of muscle relaxants allowed ECT to be used in a modified form for the treatment of severe depression and a few other disorders. Nowadays a combination of drug treatment and psychotherapy is a common approach to mental illness. Psychotherapy has evolved to a discipline with several systems, such as psychodynamic, existential, cognitive, behavioral and systemic therapy.

Since World War II, clinical trials have evolved into a standard procedure in the evaluation of new drugs. Its features include the use of a control group of patients that do not receive the experimental treatment, the random allocation of patients to the experimental or control group, and the use of blind or masked assessment so that neither the researchers nor the patients know which subjects are in either group at the time the study is conducted. A difficulty in clinical trials for psychopharmacological drugs, however, is the outcome measurement. Even though there is agreement on the existence of a biological basis for several mental diseases, laboratory tests to measure the condition of a patient do not exist. Therefore, outcome measurement is based on rating scales that assess the presence and severity of symptoms.

The use of rating scales in clinical research in psychiatry developed increasingly in the late 1950s with the introduction of antipsychotics and antidepressants. To evaluate the effectiveness of these new drugs when compared to placebo in randomised clinical trials, it became important to use instruments with a sufficiently high degree of reliability and validity. The rating scales most widely used in the 1960s and 1970s were the Brief Psychiatric Rating Scale (BPRS) which was mainly used to evaluate the effectiveness of antipsychotics, and the Hamilton Depression and Anxiety Scales (HAMD and HAMA) which were used to evaluate the effectiveness of the antidepressants and the antianxiety drugs (Bech and Jha 26.6.2008).

Besides being used as the primary outcome measurement in psychopharmacological trials, rating scales are extensively used in the measurement of quality of life in general pharmacological research. In the context of cancer research, the FDA has stated that efficacy with respect to overall survival and/or improvements in quality of life might provide the basis for drug approval (O'Shaughnessy *et al* 1991).

1.2 Psychometrics

Psychometrics is the scientific discipline concerned with the theory and technique of psychological measurement. Much of the early theoretical and applied work in psychometrics was undertaken in an attempt to measure intelligence. Pioneers in this field were Charles Spearman and L.L. Thurstone. In their research on intelligence, both psychometricians made important contributions to the theory and application of factor analysis.

More recently, psychometric theory has been applied in the measurement of personality, attitudes and beliefs, academic achievement, and in health-related areas.

Measurement of these unobservable phenomena is difficult, and much of the research in this domain has been developed in an attempt to properly define and quantify such intangible traits.

Psychometric theory involves several distinct areas of study. First, psychometricians have created a large body of theory used in the development of mental tests and analysis of data collected from these tests. This work can be roughly divided into classical test theory and item response theory. Second, psychometricians have developed methods for working with large matrices of correlations and covariances. Techniques in this general tradition include factor analysis, multidimensional scaling, data clustering, and more recently, structural equation modelling and path analysis.

It is within the context of mental test development, studied in classical test theory (CTT), that the concepts of reliability and validity play a key role. Reliability refers to the extent in which a measurement is free of measurement error. A reliable scale is therefore an instrument that measures consistently. A valid scale, on the other hand, measures what it is supposed to measure in the context in which it is applied. An instrument may be consistent without necessarily being valid. A broken ruler, for example, may always under-measure a quantity by the same amount each time (consistently), but the resulting quantity is still wrong, that is, invalid.

In the literature, different concepts are captured by the general term reliability. A first concept is *reproducibility*, that indicates the degree in which a repetition of a measurement, under the same conditions, gives the same results as the first measurement. When the repeated measurement is performed by a second rater, we talk about *inter-rater reliability*. When the repeated measurement is taken at a later time point, it is called *test-retest reliability*. The Pearson correlation coefficient and the intra-class correlation coefficient (ICC) are the most commonly used statistics in this area. A second concept that is frequently mentioned under the reliability label is *internal consistency*, indicating to which extent the different items of an instrument measure the same underlying construct. Internal consistency may be assessed by correlating performance on two halves of a test (split-half reliability). A commonly used measure is Cronbach's α, which is equivalent to the mean of all possible split-half coefficients.

Also validity can be studied in different ways. We can differentiate *content validity*, *construct validity*, and *criterion validity*. Content validity can be defined as the extent to which the instrument assesses all the relevant or important content or domains. The term *face validity* is used to indicate whether the instrument appears to be assessing the desired qualities at face. This form of validity consists of a judgement by experts

in the field. To study construct validity the investigator examines whether a measure is related to other variables as required by theory. The most commonly used methods to explore construct validity are the analysis of extreme groups (for example, an instrument is applied to cases and non-cases), convergent and discriminant validity testing (correlation with other measures of this construct and no correlation with dissimilar or unrelated constructs) and the multitrait-multimethod matrix (Campbell and Fiske 1959). Criterion validity can be assessed by correlating measures with a criterion instrument known to be valid. When the criterion measure is collected at the same time as the measure being validated the term *concurrent validity* is used; when the criterion is collected later one refers to *predictive validity*. The most commonly used method to assess criterion validity is by calculation of the Pearson correlation coefficient.

A different but nowadays widely used approach in mental test development is found in item response theory (IRT). Item response theory is typically used in educational assessment to measure abilities in domains such as reading, writing, and mathematics. Item response models are latent trait models in which the probability of correct responses are modelled as function of examinee's ability and the item characteristics, such as their difficulty. Psychometricians apply IRT in order to achieve tasks such as developing and refining exams, maintaining banks of items for exams, and equating for the difficulties of successive versions of exams, for example, to allow comparisons between results over time. In spite of having been developed mainly for educational assessment, IRT can be also applied in many other areas.

1.3 Structure of the Thesis

In this thesis the focus lies on the development of psychometric techniques to be used in clinical trials where repeated measurements are taken. Because methods will be illustrated on real case studies, we start by introducing two clinical studies in Chapter 2. The following three chapters provide a general theoretical background. In Chapter 3, a summary on the classical approach to reliability is provided, whereas Chapter 4 discusses important alternative approaches. In Chapter 5 we introduce the general modelling framework used in the thesis.

In Chapter 6 begins our search for methods to extend the classical psychometrical techniques to more general settings such as the ones provided by clinical trials. In this chapter we extend the intraclass correlation coefficient, a commonly used method

to calculate reliability, to more complex situations. In Chapter 7 we take one step back by reexamining the definition of reliability and by formulating a set of basic properties that any measure for reliability should fulfill. In the same chapter we introduce a new measure for reliability, the R_T coefficient. The application of this measure in a real case study is extensively illustrated in Chapter 8. Chapter 9 places the R_T coefficient in a broader framework and Chapter 10 introduces another measure for reliability, R_Λ, that is mathematically very close to the first measure, but bears a whole different interpretation. The common basis of both new measures is dealt with in Chapter 11, as well as their link with existing reliability measures. In Chapter 12 we argue that these new measures and the modelling framework on which they are based can deal with some previously unresolved problems in reliability research. In Chapter 13 we further illustrate that the previously introduced methodology forms a complementing set of measures that are easy to obtain as well as easy to interpret.

Where all the methodology introduced so far is developed in a longitudinal framework, in Chapter 14 we show that the proposed measures can be directly applied in a multivariate cross-sectional setting.

Finally, in Chapter 15 we summarize the most important conclusions and we further reflect on some questions in the area of psychometric validation of scales that are still open for future research.

Chapter 2

Motivating Case Studies

In this chapter we introduce two clinical studies on the evaluation of psychopharmacological drugs, the first one for the treatment of schizophrenia and the second one for the treatment of a major depressive disorder. In later chapters, newly developed methods will be applied to study the reliability of the outcome scales that were used for the evaluation of the patients in these trials.

2.1 Schizophrenia

Schizophrenia is a chronic, severe, and disabling mental illness. It has long been described as a complex and heterogeneous condition with variable symptoms. Two distinct syndromes in schizophrenia are often discerned; the positive syndrome is composed of florid symptoms, such as delusions and hallucinations. The negative syndrome is characterized by deficits in cognitive, affective and social functions including blunting of affect, poverty of speech and passive withdrawal. Other groups of symptoms are cognitive symptoms (disorganized thoughts, difficulty concentrating, memory problems,...) and affective symptoms (mainly depression).

Schizophrenia affects about 1 percent of people all over the world. Onset of the disorder typically occurs in late adolescence or early adulthood, with males tending to show symptoms earlier than females. The cause of schizophrenia is unknown and schizophrenia cannot be cured, but it can be treated. There are various theories to explain the development of this disorder. Genetic factors may play a role, but also

psychological and social factors may have an influence.

The clinical study contains five double-blind randomized clinical trials, comparing the effects of risperidone to conventional antipsychotic angents for the treatment of schizophrenia. Since the label in most countries recommends that risperidone is most effective in schizophrenia at doses ranging from 4 to 6 mg/day, we include in our analyses only patients who received either these doses of risperidone or an active control like haloperidol, levomepromazine, perphenazine, or zuclopenthixol. Depending on the trial, treatment was administered for a duration of 4–8 weeks. For example, in the international trials by Peuskens *et al* (1995), Chouinard, Jones, and Remington (1993), and Hoyberg *et al* (1993) patients received treatments for 8 weeks; in the study by Blin, Azorin, and Bouhours (1996) patients received treatments for 4 weeks, while in the study by Huttunen *et al* (1995) patients were treated over a period of 6 weeks. The sample sizes were 453, 176, 74, 49, and 71, respectively. Measurements were taken at baseline and, depending on the trial, after 1, 2, 3, 4, 6, and 8 weeks.

Three different rating scales were used as outcome measures to assess the patient's condition: the Positive and Negative Syndrome Scale (PANSS), the Brief Psychiatric Rating Scale (BPRS), and the Clinical Global Impression (CGI).

The Brief Psychiatric Rating Scale (BPRS) developed by Overall and Gorham (1962) is one of the most widely used scales in psychiatric research. The original 16-item instrument was expanded to 18 items in 1966. Each item represents a symptom that is scored from 1 (not present) to 7 (extremely severe). The assessment is based on interview with the patient and on observations of the patient's behavior over the previous 2 to 3 days or on reports of the patient's behavior from family members or carers.

The PANSS is composed of the entire BPRS supplemented by 12 items from the Psychopathology Rating Schedule (Sing and Kay 1975), resulting in a 30-item instrument (Kay, Fiszbein, and Opler 1987). The scale was developed to provide a balanced representation of positive and negative symptoms. The positive-and negative-symptom item groups are often reported separately. Each item is scored on the same seven-point severity scale as used in the BPRS.

The CGI (Guy 1976) is a global assessment tool and is designed to assess global severity of illness and change in the clinical condition over time. The scale is used widely in psychopharmacology trials, and it is not specific for schizophrenia. There are three subscales; severity of illness, global improvement, and efficacy index. The

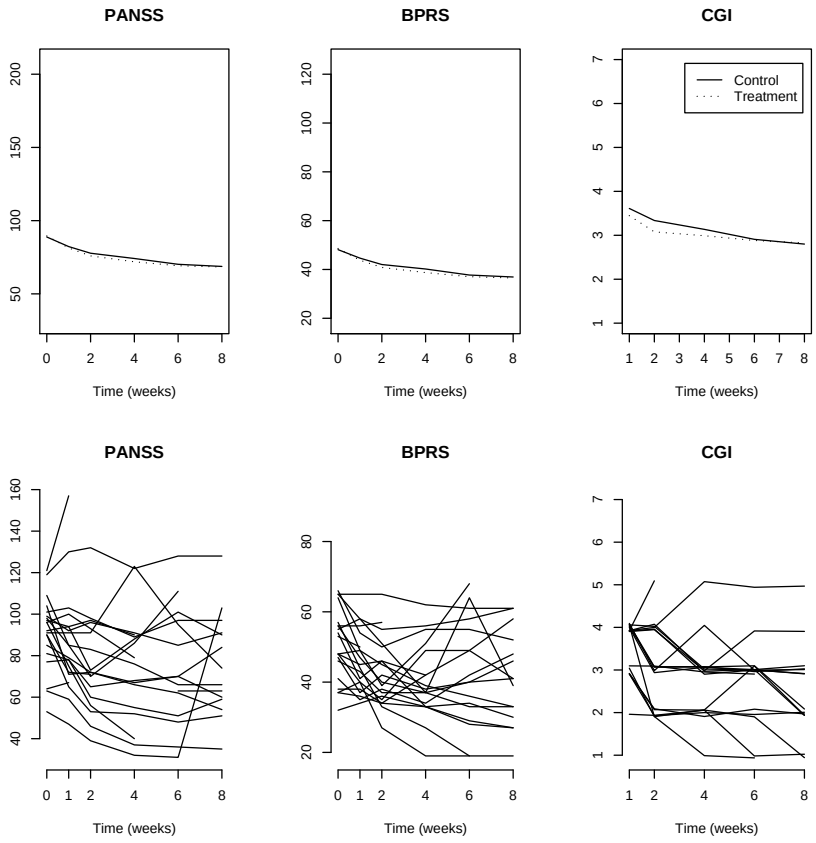

Figure 2.1: *Schizophrenia data. Mean profiles (top) and individual profiles for 20 randomly selected patients (bottom).*

global improvement scale measures the change versus baseline measurement on a scale with scores ranging from 1 (very much improved) to 7 (very much worsened).

The study data are illustrated in Figure 2.1, based on the first trial (Peuskens *et al* 1995). The three graphs on top show the mean profiles over time for the three different rating scales. It can be seen that on average the total scale scores decrease over time, indicating that patients improve. The graphs below plot the individual profiles of 20 randomly selected patients, illustrating that different patients evolve in fairly different ways. It can also be seen that not for all patients all six measurements were taken.

2.2 Major Depressive Disorder

Major depression is a serious medical illness. Unlike normal emotional experiences of sadness, loss, or passing mood states, major depression is persistent and can significantly interfere with an individual's thoughts, behavior, mood, activity, and physical health. There is not one specific way that people look and behave when they have major depression. However, most people will either have depressed mood or a general loss of interest in activities they once enjoyed, or a combination of both. In addition they will have other physical and mental symptoms that may include fatigue, difficulty with concentration and memory, feelings of hopelessness and helplessness, headaches, body aches, and thoughts of suicide.

Among all medical illnesses, major depression is the leading cause of disability in the western world. In adults, major depressive disorder affects twice as many women as men. Within an entire lifetime, major depression will affect 10 to 25 percent of women and 5 to 12 percent of men. At any one point in time, between 5 and 9 percent of women and between 2 and 3 percent of men are likely to be clinically depressed.

There is no single cause of major depression. Psychological, biological, and environmental factors may all contribute to its development. Scientists have also found evidence of a genetic predisposition to major depression. Whatever the specific causes of depression, scientific research has firmly established that major depression is a biological, medical illness. Although major depression can be a devastating illness, it is highly treatable. Between 80 and 90 percent of those diagnosed with major depression can be effectively treated and return to their usual daily activities and feelings.

The case study data come from two identical randomized double-blind clinical trials to investigate the efficacy of duloxetine in the treatment of major depressive disorder (study 5 and study 6 in Mallinckrodt *et al* 2003). The primary efficacy measure was the total score on the Hamilton Depression Rating scale (HAMD). Secondary measures were the total scores on the Hamilton Anxiety Rating Scale (HAMA) and the Montgomery-Åsberg Depression Rating Scale (MADRS). The first trial contained a total of 354 patients of which 90 were assigned to the placebo group, 91 received Duloxetine (40 mg/d), 84 received Duloxetine (80 mg/d) and 89 received Paroxetine (20 mg/d). The sample size of the second trial was 353 with 89, 86, 91, and 87 patients in the respective treatment groups. Measurements were taken at baseline and after 1, 2, 4, 6, 8 and 10 weeks.

The HAMD scale was developed in the late 1950s to assess the effectiveness of the

first generations of antidepressants. The scale quickly became the standard measure of depression severity for clinical trials of antidepressants and is until now the most commonly used measure for depression. The original rating form included 21 items, although Hamilton (1960) indicated that only 17 items should contribute to the total scale score because 1 of the last 4 items represented depressive type rather than depression severity, and 3 other items did not occur with sufficient frequency. Nine of the 17 items are rated from 0 to 4, whereas 8 items are rated 0 to 2. Several other versions have been developed later on, but the most commonly used version is the 17-item scale.

Concurrently, Hamilton (1959) developed one of the first rating scales to quantify the severity of anxiety symptomatology: HAMA. The scale consists of 14 items, each defined by a series of symptoms. Each item is scored on a scale of 0 (not present) to 4.

Several conceptual and psychometric problems with the HAMD scale have been described in the literature. The MADRS was designed to address the limitations of the HAMD scale, and was supposed to measure contemporary definitions of depression and to be more sensitive to change (Montgomery and Åsberg 1979). The scale is a 10-item checklist.

Figure 2.2 illustrates the depression data based on the first trial. The mean profiles on top illustrate that on average patients improve over time, in all the treatment groups. The individual profiles of 20 randomly selected patients illustrate that not only between different patients large differences appear, but also the profiles of individual patients can be rather unstable.

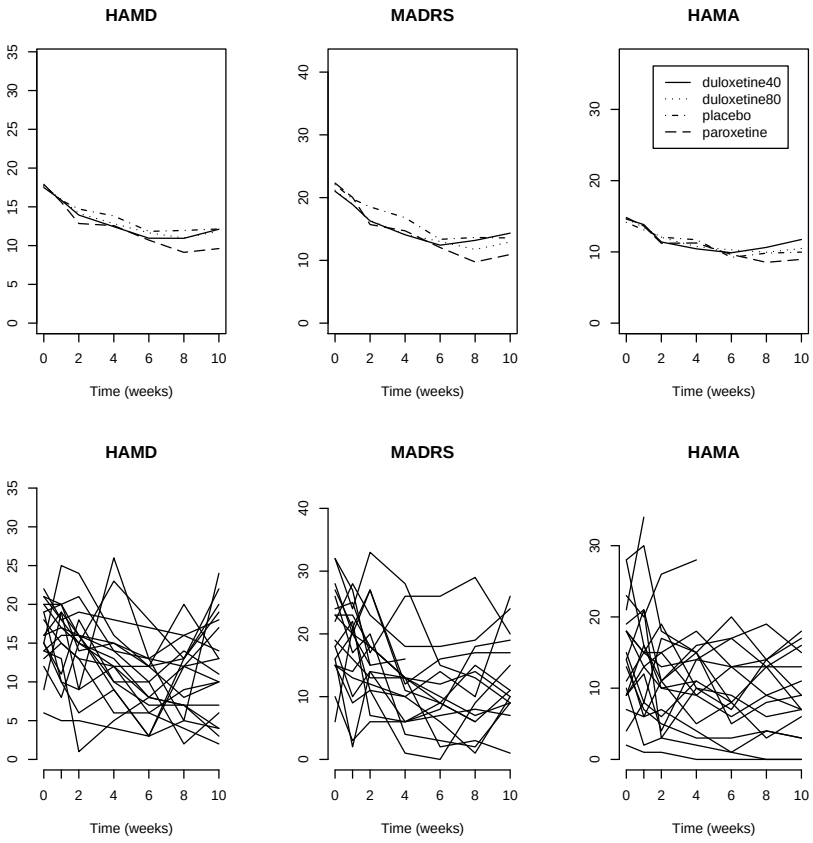

Figure 2.2: *Depression data. Mean profiles (top) and individual profiles for 20 randomly selected patients (bottom).*

Chapter 3

Classical Approach to Reliability

In Chapter 1 we have seen that research on the measurement of unobservable human phenomena dates back more than one hundred years. Psychometric methods have been developed to assess the quality of such measurements. In this chapter we summarize some of the most important contributions to the study of reliability. We will also stress the importance of having reliable measurements.

3.1 Early Psychometric Literature

The first attempt to quantify the reliability of measurements was carried out by Charles Spearman at the beginning of the 20th century. The concept of correlation was already known, but Spearman was the first to consider various hidden underlying causes affecting the true correlation. He proposed a formula to correct for attenuation when finding the true relationship between two variables (Spearman 1904).

In 1910, Spearman introduced the term *reliability coefficient* as "the correlation between one half and the other half of several measures of the same thing". The Spearman (1910) and Brown (1910) computational formula for estimation of reliability is based on splitting the test into two halves, a and b, typically by selecting the odd

and the even numbered items, to obtain the split-half coefficient of reliability

$$\rho = \frac{2r_{ab}}{1 + r_{ab}},$$

where r_{ab} equals the Pearson correlation between the obtained scores on the two halves. The constant 2 in the numerator of the formula is associated with the fact that the original test was halved. As the result was an estimate of the reliability of a test twice as long as each half, the formula became known as a prophecy formula. In a more general form, the Spearman-Brown formula can be written as

$$\rho = \frac{k\rho_{XX'}}{1 + (k-1)\rho_{XX'}},$$

where ρ is the reliability of a composite of k parallel tests, and $\rho_{XX'}$ is the correlation between two parallel tests which is assumed to be constant for all pairs of tests. When all the items are interpreted as separate tests, ρ is the reliability of a test consisting of k items, and the $\rho_{XX'}$ is the common reliability of the items. This formula clearly shows that the reliability of a test depends on the true underlying correlation across items as well as on the number of items. In fact, it illustrates that the reliability of a test is an increasing function of its number of items.

Apart from the assumption of equal correlations between all pairs of tests, the Spearman-Brown formula for composite tests further assumes that these tests all measure the same dimension equally (e.g., an assortment of math problems of equal difficulty), and that all test variances are equal. These strong assumptions underlying the Spearman-Brown formula have led to many criticisms, but despite this, the formula was used in the psychological, educational, and sociological research for decades. The method was simple and no real alternative appeared to exist. For a long time in a pre-computers era, simplicity was an important issue, because the calculations were done mostly by hand.

In 1937, Kuder and Richardson introduced a collection of new reliability measures, of which one became especially popular: the *Kuder-Richardson formula 20* or, KR-20 for dichotomous items

$$\rho = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} p_j q_j}{\sigma_u^2}\right),$$

with k the number of items, p_j is the proportion of correct responses on item j, $q_j = 1 - p_j$, and σ_u^2 the variance of the total score. In this approach, items are

compared with each other, rather than comparing one half of the items with the other half. It can be shown mathematically that the Kuder-Richardson reliability coefficient is actually the mean of all split-half coefficients resulting from different splittings of a test. Alternatively, KR-21 assumes that all the items are equally difficult and its expression is very similar to the one of the KR-20 but substituting p_i by the average proportion of correct responses. For both measures it is assumed that all items measure the same thing. Thanks to ease of calculation and uniqueness of estimate, compared to split-half methods, the KR-20 became a classic tool for the evaluation of reliability.

Alternative forms of the formula were suggested, but the most famous variation, known as *alpha*, was presented by Cronbach (1951). Cronbach's alpha extends the previous measure to non-dichotomous items, and equals KR-20 in the dichotomic case

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k}\sigma_{x_j}^2}{\sigma_u^2}\right).$$

Here $\sigma_{x_j}^2$ denotes the variance of the jth item. As previous proposals, this coefficient was introduced as a measure for reliability. However, only under quite restrictive conditions the coefficient α is an estimate of the reliability of a composite measurement. Whenever these assumptions are not met, alpha only provides a lower bound to the reliability (Novick and Lewis 1967). Nevertheless, many textbooks in psychology and education, as well as practical investigations of test reliability, continued to regard coefficient alpha as an estimator of reliability. In the years to follow, test theorists began to think of coefficient alpha, as well as of its predecessors, as mere measures of internal consistency or item homogeneity. High values then indicate that the items in the test form a relatively homogeneous set.

3.2 The Classical Test Theory

Much of the literature on reliability originates in the classical test theory (CTT). Important contributions in this field came from many scholars, such as Spearman, Yule (1912), Guttman (1945, 1953), and Gulliksen (1950). However, most referred to in this context is the work of Novick (1966) and Lord and Novick (1968).

In CTT, the outcome of a test for subject i, $i = 1, \ldots, n$, is modelled as

$$X_i = \tau_i + \varepsilon_i, \tag{3.1}$$

where X_i represents the observed score, τ_i is the true score and ε_i the corresponding measurement error. One rarely thinks of τ_i as an actual true score, but it is often defined as the expected value of X_i if the subject were remeasured an infinite number of times. Further, it is assumed that the measurement errors are mutually uncorrelated as well as uncorrelated with the true scores. Under these assumptions

$$\mathrm{Var}(X_i) = \mathrm{Var}(\tau_i) + \mathrm{Var}(\varepsilon_i).$$

The reliability of a measurement instrument is then defined as the ratio of the true score variance to the observed score variance

$$R = \frac{\mathrm{Var}(\tau_i)}{\mathrm{Var}(X_i)} = \frac{\mathrm{Var}(\tau_i)}{\mathrm{Var}(\tau_i) + \mathrm{Var}(\varepsilon_i)}. \tag{3.2}$$

It also equals the squared correlation between the observed and the true scores

$$\mathrm{Corr}(X_i, \tau_i)^2 = \frac{[\mathrm{Cov}(X_i, \tau_i)]^2}{\mathrm{Var}(X_i)\mathrm{Var}(\tau_i)} = \frac{\mathrm{Var}(\tau_i)}{\mathrm{Var}(X_i)} = R \tag{3.3}$$

as well as the correlation of two *parallel tests*, i.e., two tests with equal true scores and equal error variances

$$\mathrm{Corr}(X_{1i}, X_{2i}) = \frac{\mathrm{Cov}(X_{1i}, X_{2i})}{\sqrt{\mathrm{Var}(X_{1i})}\sqrt{\mathrm{Var}(X_{2i})}} = \frac{\mathrm{Var}(\tau_i)}{\mathrm{Var}(\tau_i) + \mathrm{Var}(\varepsilon_i)} = R. \tag{3.4}$$

From the theoretical definition of reliability (3.2), and taking into account that variances cannot be negative, the upper and lower limit of the reliability coefficient can easily be derived as $0 \leq R \leq 1$, and $R = 0$ if the test contains nothing but measurement error. On the other hand, if no measurement error is present, the observed-score variance equals the true score variance and the measurement instrument is perfectly reliable (assuming that there is true-score variation). In this scenario the reliability coefficient reaches its upper bound, i.e., $R = 1$. It is important to point out that the observed-score variance is population dependent, as is the reliability coefficient. Indeed, the variability of the true scores is a population-specific parameter and therefore, every time a scale is used in a new population its reliability should be reassessed.

Notice also that in the previous modelling framework, the true scores τ_i are unobservable latent quantities what makes the direct estimation of their variability impossible. As a consequence, the direct estimation of the reliability coefficient (3.2) is

also problematic. One way to circumvent this problem is to estimate reliability by correlating the test with a parallel test, as expressed in (3.4). Such parallel tests might be formed by different versions of a test containing different items but measuring the same underlying construct, for example two halves of a test, or by repeating the same test more than once. However, in both situations it might be practically unfeasible to obtain the required conditions of equal true scores and equal error variances.

To overcome the stringent assumptions of the parallel model, relaxations to the model have been proposed. Two tests are said to be *tau-equivalent* if the true scores are equal but the error variances differ. When the true scores only differ by a constant, the tests are said to be *essentially tau-equivalent*. It is under the assumptions of the essentially tau-equivalent model that Cronbach's alpha equals the reliability of a composite measurement (Novick and Lewis 1967). A major limitation of essential tau-equivalence is that it requires equal covariances between test parts, which will rarely be encountered in practice. A further relaxation is allowed in the case of *congeneric tests*: the true scores of the tests can now be linearly related so that true score variances, error variances and population means can differ.

The analysis of congeneric measures, as developed by Jöreskog (1971), can serve as an alternative to coefficient α if items are suspected to have different true variances. Congeneric measures have pairwise perfectly correlated true scores, but may have different true variances. For congeneric tests, the true scores can be written in terms of a latent variable τ. This implies that for test j we have

$$\tau_j = \mu_j + \beta_j \tau.$$

Further, the observed score is expressed as the sum of the true score and the error

$$X_j = \mu_j + \beta_j \tau + \varepsilon_j.$$

Without loss of generality, $\mathrm{Var}(\tau)$ can be set to 1. The observed variance for the jth test can then be written as

$$\mathrm{Var}(X_j) = \beta_j^2 + \sigma_j^2$$

Hence, the reliability of the jth test is equal to the square of the slope divided by the total variance

$$\rho_{jj} = \frac{\beta_j^2}{\mathrm{Var}(X_j)} = \frac{\beta_j^2}{\beta_j^2 + \sigma_j^2}. \tag{3.5}$$

In some situations one may want to combine some of the tests into a linear composite. If $\boldsymbol{a}' = (a_1, a_2, \ldots, a_p)$ is a vector of relative weights then we can define the new test

as $Y = \boldsymbol{a}'\boldsymbol{X} = \boldsymbol{a}'\boldsymbol{\mu} + (\boldsymbol{a}'\boldsymbol{\beta})\tau + \boldsymbol{a}'\boldsymbol{\varepsilon}$, and the reliability of Y is

$$\rho = \frac{(\boldsymbol{a}'\boldsymbol{\beta})^2}{(\boldsymbol{a}'\boldsymbol{\beta})^2 + \boldsymbol{a}'\boldsymbol{\Theta}\boldsymbol{a}},$$

where $\boldsymbol{\Theta}$ a diagonal matrix with error variances. Maximum likelihood estimation is proposed by Jöreskog for parameter estimation. Note that if $\boldsymbol{a}$ equals a vector of ones the composite is formed by the simple sum of test scores. The weights can further be selected in such a way that the reliability of this composite measurement is maximized.

With more than three tests, the assumption that tests are congeneric can be evaluated. It is possible that the congeneric model does not fit, for example when the one-factor model is not valid, i.e., when all items or subtests are not measuring exactly the same underlying construct. Then a structural model with more than one dimension can be fit. Werts, Rock, Linn and Joreskög (1978) tried to enhance the procedure and find a way to estimate the reliability of a factorially more complex composite. For a composite test composed of p tests they used the measurement model

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\tau}$ is a vector of order p of the true scores, $\boldsymbol{B}$ is a $p \times p$ identity matrix and $\boldsymbol{\varepsilon}$ is a p-dimensional vector of the measurement errors. Further, they let $\text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$ (diagonal) and $\text{Cov}(\boldsymbol{\tau}) = \boldsymbol{\Gamma}$.

The true score vector $\boldsymbol{\tau}$ is assumed to have an underlying factor model with k common factors, so that $\boldsymbol{\tau} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\eta}$, where $\boldsymbol{\xi}$ is a vector of order k of common factors, $\boldsymbol{\eta}$ is a vector of order p of the unique factors and $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of factor loadings. It is assumed that $E(\boldsymbol{\eta}) = \boldsymbol{0}$, $E(\boldsymbol{\xi}\boldsymbol{\eta}') = \boldsymbol{0}$, $\text{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Psi}$ (diagonal), and $\text{Cov}(\boldsymbol{\xi}) = \boldsymbol{\Phi}$. The covariance matrix of the p observed variables is given by

$$\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{\Gamma}\boldsymbol{B}' + \boldsymbol{\Theta} = \boldsymbol{B}(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})\boldsymbol{B}' + \boldsymbol{\Theta},$$

the identity matrix $\boldsymbol{B}$ is there only for reasons of compatibility with earlier introduced models by the same authors. The reliability of a composite test $Y = \mathbf{a}'\boldsymbol{X}$, with $\boldsymbol{a}$ a vector of weights, is then

$$\rho = \frac{\boldsymbol{a}'\boldsymbol{\Gamma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'\boldsymbol{a} + \boldsymbol{a}'\boldsymbol{\Psi}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}.$$

In this proposal, the effect of the unique variances is included in the true variation. Tarkkonen and Vehkalahti (2005) argue that this choice can cause problems in identification and they suggest to include the unique variance in the error variance.

3.3 Estimating Reliability in Practice

The estimation of the reliability of an instrument will always rely on a repetition of measurement results. This repetition can be achieved either by considering different subtests or items on a single measurement occasion, or by repeating the whole measurement. In Section 3.1, several simple methods have been described that can be applied in the single-occasion case. For example, the KR-20 and KR-21 formulas resulted from attempts to determine reliability from a single administration of a test. This quest held considerable appeal for many people, because the test-retest and parallel-forms methods were time-consuming and expensive.

However, further research in this area has clearly shown that these methods quantify the reliability of an instrument only when a restrictive set of assumptions is satisfied. For instance, the Spearman-Brown formula requires parallel measurements, whereas Cronbach's α and KR-20 need essentially tau-equivalence. Whenever these assumptions are not met, these methods provide merely a lower bound to reliability. Additionally, the lower bound will be very poor, except for tests that are relatively homogeneous or long (Novick and Lewis 1967). The value of the Cronbach's α coefficient can even get negative if the sum of all item covariances is negative. Better lower bounds than coefficient alpha have been found, such as λ_4, proposed by Guttman (1945), and the greatest lower bound (glb) to reliability, by Jackson and Agunwamba (1977).

The methods described in Section 3.1 are nowadays mainly considered as estimates for the internal consistency of an instrument, which indicates the homogeneity of the items, or, an indication of how much they measure the same underlying construct.

A second approach is based on the idea of repeating the measurements using the same test. An entire measurement can be repeated by either asking two different raters to evaluate the same group of subjects, or by administering the same test to the same subjects at two points in time. The former is referred to as *inter-rater reliability*, the latter as *test-retest reliability*. The repeated measurements are then assumed to be parallel (equal true scores and equal error variances). Following (3.4), reliability can then be derived as the intraclass correlation (ICC). ICC may be conceptualized as the ratio of between-groups variance to total variance and can be obtained from a one-way analysis of variance with subjects as factor (Bartko 1966, Fleiss 1986). Note that sometimes the product-moment correlation instead of ICC is used for estimating reliability. ICC is preferred over Pearson's correlation only when sample size is small

(< 15) or when there are more than two tests (Shrout and Fleiss 1979).

Also for estimating the inter-rater reliability, it might be difficult to fulfill the required assumptions. One rater might, for example, tend to give higher scores than his colleague. In that case extensions can be made to a two-way analysis of variance. Rater can then be considered as a random effect so that a two-way random effects model is obtained. If rater is considered as fixed, a two-way mixed effects model can be applied (Dunn 1989).

Test-retest reliability has its specific difficulties as well. Generally it is found that the shorter the time interval the higher the estimate of reliability. For a simple replication study with two measurements, it is often advised to take the time interval between two measurement occasions sufficiently short so that it is safe to assume that the underlying process is unlikely to have changed. Off course, the appropriate length of the interval depends on the stability of the trait that is being measured. Nevertheless, if both measurements are taken sufficiently close in time, it is also quite likely that the rater will recall the previous ratings and the assessments will be influenced by them. Usually the rater will give similar ratings in each of the replications, making them appear more consistent than they in fact are (Dunn 1989, Streiner and Norman 1995). A second and related problem has to do with the assumption that the errors of measurement are uncorrelated, while correlated error terms are very common among repeated measurements (Bohrnstedt 1983).

A third problem is related to the assumption of equal true scores. Whenever measuring living organisms, it is clear that the characteristics being measured might change from one replication to another. In this case, stability of the trait or characteristic being measured will be confounded with test reliability. If one wishes to disentangle the effects of lack of stability from the effects of poor instrument reliability, then more data are needed. Wiley and Wiley (1970) formulated a model in which a subject's true score at a particular moment is linearly related to the true score of another moment of observation. At least three measurement occasions are necessary to estimate the reliabilities at the different moments of observation. Interestingly, this approach also shows that the reliability of a test may change with time. Unfortunately, the approach does not allow for correlated measurement error across time.

3.4 Consequences of Low Reliability

In scientific research, population parameters are mostly estimated based on a random sample. Statistical models include the sampling variation to correct for uncertainty induced by the sampling process. A second source of uncertainty can be found, however, in the measurement process, through the occurrence of random measurement error. In many cases this source of uncertainty is ignored by the assumption that subjects are measured without error.

Clearly, some variables are more likely to be measured without or with negligible error, like many biological parameters, compared to others that need, for example, subjective judgement of a clinician. In spite of a widely accepted biological basis for many psychiatric disorders, no laboratory tests exist that can be used in treatment efficacy evaluation in this area. As a consequence, psychopharmacological studies generally rely on the use of rating scales for outcome measurement. Evaluation of the impact of measurement error is therefore crucial in such studies.

Fleiss (1986) stated that "the most elegant design of a clinical study will not overcome the damage by unreliable or imprecise measurement." In clinical trials, one typically wants to differentiate between treatments. However, if the reliability of the outcome measurements is low, the ability to differentiate between the different subjects in various treatment groups decreases. Fleiss (1986) and Lachin (2004) discuss a number of consequences of low reliability in clinical studies. To start with, measurement error may result in biased sample selection in clinical studies when patients are selected with a minimum level of a certain measurement. Second, the correlation between two variables, X and Y, is affected by the reliability with which they are measured, as expressed by

$$\text{Corr}(\tau_X, \tau_Y) = \frac{\text{Corr}(X, Y)}{\sqrt{R_X R_Y}},$$

with τ_X and τ_Y referring to the true scores of the two variables X and Y, and R_X and R_Y to the respective reliability coefficients. This expression is known as the attenuation formula (Lord and Novick 1968). Knowing that reliability coefficients lie between 0 and 1, the formula shows that the observed correlation might be seriously decreased compared to the true correlation if one or both variables are measured with considerable error. Further continuing to regression models, the reliability of a covariate affects its estimated effect. There is a direct relationship between the reliability of the measurement and the power of a study. Indeed, the lower the reliability of the

measurements, the larger the sample size that is needed to achieve certain power. One can easily show that for a paired t-test, the required sample size becomes $n = n^*/R$ where n^* denotes the required sample size for a measurement without error. When sample size calculations ignore information on the reliability of the measurements, the power of the study might be lower than expected. Using unreliable measurement scales can therefore conceal treatment effects and might lead to the rejection of promising treatments.

Additionally, Lord and Novick (1968) link the concept of reliability to the one of validity by proving that

$$\text{Corr}(X, Y) \leq \text{Corr}(\tau_X, X) = \sqrt{R_X}.$$

This means that the validity of a measure X in relation to a second measure Y cannot exceed the square root of its reliability. The reliability of a measurement thus defines an upper bound for the validity of this measurement.

In the present section we have mainly focussed on the impact of reliability on scientific studies. However, besides its central role in scientific research, measurement is also part of every day life and important decisions might be taken based on the results of it. Think, for instance, of medical screening tests, or psychological tests as part of a selection process for a job. Obviously, also in these situations the effect of measurement error needs to be restricted to a minimum.

Chapter 4

Alternative Approaches to Reliability

In this chapter, we briefly discuss two important methods in the field of psychometrics: generalizability theory (G-theory) and item response theory (IRT). Both approaches play a prominent role in research and applications and offer alternative visions on how reliability can be defined and estimated.

Technically, classical, generalizability, and item response theory are not directly comparable against each other because they have different foci. In IRT the interest lies in the unobserved theoretical latent trait and the primary goal is to estimate a subject's score on this trait. In classical and in G-theory the interest lies in the observed score from the test and one aims at evaluating the quality of this score by estimating reliability coefficients and standard errors. Also, whereas the fundamental unit of analysis for IRT is the item, the unit of analysis for both classical and G-theory is the overall score.

In spite of the fact that G-theory was initiated more recently than IRT, the latter is often referred to as the "modern test theory"; perhaps due to its many recent expansions and its unique ability to work with modern computerized adaptive tests. In the next section we will briefly describe the main ideas and concepts behind IRT.

4.1 Item Response Theory

The central feature of item response theory is that it relates item responses to characteristics of individual persons (latent traits) and characteristics of the assessment (item parameters). The latent trait is the human capacity or attribute measured by the test. Since most of the research has dealt with variables such as scholastic, reading, and mathematical abilities, the generic term "ability" is often used in IRT to refer to such latent traits. The most important item parameter is the item location which, in the case of attainment testing, is referred to as the item difficulty. Often also the discrimination of the item is estimated, that is, the degree to which the item discriminates between persons in different regions on the latent continuum. For items such as multiple choice, a third parameter can be introduced to account for the effect of guessing on the probability of a correct response. For example, in the three parameter logistic (3PL) model (Lord 1980), the probability of a correct response to item i is given by

$$p_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-Da_i(\theta - b_i)}}$$

where θ is a person-specific random effect describing the latent trait the test tries to assess (ability), a_i denotes the item discrimination parameter, b_i denotes the item difficulty parameter, and c_i is the guessing parameter. Further, D denotes a constant with value 1.701 which rescales the logistic function to closely approximate the cumulative normal ogive. The model was originally developed using the normal ogive but the logistic model with the rescaling provides virtually the same results while greatly simplifying computations involved with its application. Simplifications with respect to the 3PL model are the two parameter logistic (2PL) model (Birnbaum 1968) in which $c_i = 0$ and thus no correction for guessing is made, and the one parameter logistic (1PL) model where $c_i = 0$ and $a_i = a$, i.e., all items are assumed to have equal discrimination capacity. The 1PL model is sometimes also referred to as the Rasch model. Even though the development of the Rasch model (Rasch 1960) was independent of the 1PL model, they both have similar features and are mathematically equivalent. Extensions of these models have been made, among others for polytomous responses. Actually, most extant item response models are special cases of generalized linear or nonlinear mixed models (GLMM and NLMM), which form two general and flexible model families for repeated categorical data (De Boeck and Wilson 2004).

Many IRT models are based on the assumption that the items are measuring a single continuous latent variable. The unidimensionality of a scale can be evaluated

by performing a factor analysis to explore the factor structure underlying the observed covariation among item responses. However, extensions to multidimensional IRT models have been made. A second assumption of IRT models is local independence: when the abilities influencing the test performance are held constant, subjects' responses to any pair of items are assumed to be statistically independent. When this property holds items can be linked together on a common metric allowing the creation of questionnaires that may use a different set of items depending on the target audience of responders. In other words, two responders that administered two different assessments can have scores comparable on a similar metric. IRT is therefore the foundation of computerized adaptive testing. For parameter estimation, marginal maximum likelihood (Bock and Aitkin 1981) or conditional maximum likelihood (Andersen 1972) are frequently used, however, a Bayesian approach is available as well (Swaminathan and Gifford 1986).

Compared to CTT, IRT has a different view on the concept of reliability. It is assumed that precision is not uniform across the entire range of item scores. Scores at the edges of the test's range, for example, generally have more error associated with them than scores closer to the middle of the range. Item response theory uses the concept of item and test information to replace reliability. Information is also a function of the model parameters. For example, according to Fisher's information theory, the item information supplied in the case of the Rasch model for dichotomous response data is simply the probability of a correct response multiplied by the probability of an incorrect response (q_i)

$$I(\theta) = p_i(\theta)q_i(\theta).$$

The standard error of estimation is then the reciprocal of the test information, at a given trait level

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

Thus more information implies less error of measurement. For other models, such as the two and three parameter models, the discrimination parameter plays an important role in the function. The item information function for the two parameter model is

$$I(\theta) = a_i^2 p_i(\theta)q_i(\theta).$$

In general, item information functions tend to look bell-shaped. Highly discriminating items have tall, narrow information functions; they contribute greatly but over a narrow range. Less discriminating items provide less information but over a wider

range. Because of local independence, item information functions are additive. Thus, the test information function is simply the sum of information functions of the items on the test (Lord 1980).

Item response theory has recently been the center of much attention among psychometricians and test specialists. One possible reason for such attention, is the fact that IRT has gained considerable visibility for its use in many prominent, large-scaled testing programs, such as the National Assessment of Educational Progress (NAEP), the Scholastic Aptitude Test (SAT), and the Graduate Record Examination (GRE) in the United States. Additionally, it has also been applied in large international assessment programs such as the Third International Math and Science Survey (TIMSS) and the Programme of International Student Assessment (PISA). Another possible reason for the attention given to IRT is the technical challenges presented by its many new statistical developments and developments in computer and other technologies (Suen and Lei 2007). However, in spite of the recent attention given to IRT, the classical test theory and G-theory remain important tools in many commercial, psychological and academic tests; particularly in small-scaled testing programs. In the following section, we will introduce some of the main ideas underlying G-theory.

4.2 Generalizability Theory

Generalizability Theory was originally introduced by Cronbach and colleagues (1963, 1972) in response to the limitations of the true-score model of classical test theory. While this model may be reasonable for carefully equated parallel forms of tests, it is overly restrictive and often unrealistic in situations where, for instance, raters differ in the central tendency and variance, observations depend on the context in which they occur, and constructs are obviously heterogeneous. In classical test theory an observation is assumed to be a combination of an individual's true score and a random measurement error. The sources of variation in the measurements are not explicitly modelled. In G-theory one sets out to systematically investigate the sources of variation of measurements. Each of the multiple forms of reliability that were mentioned in the previous chapter identifies and quantifies only one source of error variance at a time.

G-theory is a theory regarding the adequacy with which one can generalize from a sample of observations to a universe of observations from which the sample was drawn. This theory acknowledges that the reliability of an observation depends on the

universe about which the investigator wants to draw inferences. Because a particular measure may conceivably be generalized to many different universes, a measure may vary in how reliably it permits inferences about these universes and, therefore, be associated with different reliability coefficients. "Facets" is the term used in G-theory for the variables or factors that might contribute to the variability in the observations. Examples of facets are time, alternate test forms, raters, etc. The term "conditions" is used to refer to the levels of the factors. Facets may be considered fixed or random. If fixed, the specified conditions are the only conditions of interest; one generalizes only to them. If random, one generalizes to a population which has been sampled. In that case the levels of the facet included in the generalizability study must be representative of the population (universe).

Let us start by considering a test X composed by a set of n_β different items $\boldsymbol{I} = \{i_1, i_2, \ldots, i_{n_\beta}\}$. Further, let us assume that the following measurement model holds

$$\begin{aligned} & X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ & \alpha_i \sim N(0, \sigma^2_\alpha) \\ & \beta_j \sim N(0, \sigma^2_\beta) \\ & \varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon) \\ & \alpha_i, \beta_j, \varepsilon_{ij} \quad \text{are independent for all} \quad (i, j), \end{aligned} \tag{4.1}$$

where X_{ij} denotes the observed score for subject i at item j, μ denotes a general mean, α_i denotes the subject's effect, β_j denotes the item's effect, and ε_{ij} is the measurement error. Note that model (4.1) could be enlarged by adding an interaction term $(\alpha\beta)_{ij}$. However, given the fact that every subject answers every item only one time, this interaction term will be essentially unidentifiable with respect to the error term ε_{ij}.

In G-theory, the items that conform the test are considered a sample from a generic population of items. Even though thinking of a population of items can seem at first odd, this assumption emanates naturally from a model that contemplates a random effect for the item. Basically, we should think about this population as the one defined by the set of all items of potential interest for the domain under investigation.

In practice, decisions usually will be based on multiple observations rather than on a single observation. G-theory typically uses the mean score metric (e.g., mean of multiple item scores) rather than the total score metric (e.g., sum of multiple item scores). We will denote the mean of the observed scores for subject i over a sample

of n_β items $\boldsymbol{I}$ by $\bar{X}_{i\cdot}$, i.e.,

$$\bar{X}_{i\cdot} = \frac{1}{n_\beta}\sum_{j=1}^{n_\beta} X_{ij}.$$

G-theory recognizes that the decision maker might want to make two types of decisions based on a behavioral measurement: relative or norm-referenced and absolute or domain-referenced. A *relative decision* concerns the relative ordering of the individuals (e.g., norm-referenced interpretations of test scores). In this scenario the score of a subject i ($\bar{X}_{i\cdot}$) is only used to evaluate his/her relative performance with respect to other individuals. Therefore, we are not interested in the absolute value of the score $\bar{X}_{i\cdot}$ but only in how it ranks with respect to other scores $\bar{X}_{i'\cdot}$. If all the individuals answer the same set of items $\boldsymbol{I}$ then the complexity of the items, captured by the random effects β_j, becomes irrelevant. In this case we can ignore the effect of the items and we can condition on $\boldsymbol{I}$. If we calculate the average of the X_{ij} over $\boldsymbol{I}$ then from (4.1) we get

$$Y_i = \mu + \alpha_i + \tilde{\varepsilon}_i, \tag{4.2}$$

where $Y_i = \bar{X}_{i\cdot}$ and $\tilde{\varepsilon}_i = \bar{\beta}_\cdot + \bar{\varepsilon}_{i\cdot}$, with $\bar{\beta}_\cdot = \dfrac{1}{n_\beta}\displaystyle\sum_{j=1}^{n_\beta}\beta_j$ and $\bar{\varepsilon}_{i\cdot} = \dfrac{1}{n_\beta}\displaystyle\sum_{j=1}^{n_\beta}\varepsilon_{ij}$.

Notice that if subjects i and i' are evaluated using the same set of items $\boldsymbol{I}$ then

$$Y_i - Y_{i'} = (\alpha_i - \alpha_{i'}) + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{i'\cdot}),$$

and this expression clearly shows that the effect of the items is irrelevant for relative decisions. We can then condition on $\boldsymbol{I}$ which leads to

$$\mathrm{Var}(Y_i|\boldsymbol{I}) = \mathrm{Var}(\alpha_i|\boldsymbol{I}) + \mathrm{Var}(\tilde{\varepsilon}_i|\boldsymbol{I}).$$

Owing to the independence between α_i and β_j, $\mathrm{Var}(\alpha_i|\boldsymbol{I}) = \sigma^2_\alpha$. Further,

$$\mathrm{Var}(\tilde{\varepsilon}_i|\boldsymbol{I}) = \mathrm{Var}(\bar{\beta}_\cdot|\boldsymbol{I}) + \mathrm{Var}(\bar{\varepsilon}_{i\cdot}|\boldsymbol{I}).$$

Obviously, conditional on $\boldsymbol{I}$, $\bar{\beta}_\cdot$ is a constant, hence $\mathrm{Var}(\bar{\beta}_\cdot|\boldsymbol{I}) = 0$. On the other hand,

$$\mathrm{Var}(\bar{\varepsilon}_{i\cdot}|\boldsymbol{I}) = \frac{1}{n_\beta^2} n_\beta \sigma^2_\varepsilon = \frac{\sigma^2_\varepsilon}{n_\beta}.$$

This implies that the variance of the observed scores is

$$\mathrm{Var}(Y_i|\boldsymbol{I}) = \sigma^2_\alpha + \frac{\sigma^2_\varepsilon}{n_\beta}.$$

Using model (4.2), we can now define the generalizability coefficient for relative decisions, which is analogous to the reliability coefficient in classical test theory, i.e.,

$$\mathrm{E}\rho^2 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{n_\beta}}. \tag{4.3}$$

Note that, like with the Spearman-Brown prophesy formula, (4.3) is an increasing function of the number of items. Therefore, the larger the number of items we condition on, the more reliable the instrument will be for relative discrimination.

On the other hand, an *absolute decision* focuses on the absolute level of an individuals's performance independently of the performance of other subjects. Hence, in this scenario we are mainly interested in the absolute interpretation of $\bar{X}_{i\cdot}$, regardless the values of other scores $\bar{X}_{i'\cdot}$. Obviously, in this setting the complexity of the items plays a prominent role and, as a consequence, we can not condition on $\boldsymbol{I}$. Like before, we will base our calculations on model (4.2). The main difference is that in the present setting $\bar{\beta}_\cdot$ will not have variance zero because we are not conditioning on $\boldsymbol{I}$. We then have

$$\mathrm{Var}(Y_i) = \mathrm{Var}(\alpha_i) + \mathrm{Var}(\tilde{\varepsilon}_i),$$

where, similar as before, $\mathrm{Var}(\alpha_i) = \sigma_\alpha^2$ and

$$\begin{aligned}
\mathrm{Var}(\tilde{\varepsilon}_i) &= \mathrm{Var}(\bar{\beta}_\cdot) + \mathrm{Var}(\bar{\varepsilon}_{i\cdot}) \\
&= \frac{1}{n_\beta^2}\sum_{j=1}^{n_\beta}\mathrm{Var}(\beta_j) + \frac{1}{n_\beta^2}\sum_{j=1}^{n_\beta}\mathrm{Var}(\varepsilon_{ij}) \\
&= \frac{1}{n_\beta^2} n_\beta \sigma_\beta^2 + \frac{1}{n_\beta^2} n_\beta \sigma_\varepsilon^2 = \frac{\sigma_\beta^2}{n_\beta} + \frac{\sigma_\varepsilon^2}{n_\beta}.
\end{aligned}$$

The variance of the observed scores takes then the form

$$\mathrm{Var}(Y_i) = \sigma_\alpha^2 + \frac{\sigma_\beta^2}{n_\beta} + \frac{\sigma_\varepsilon^2}{n_\beta},$$

which leads to the absolute reliability coefficient, also called the index of dependability

$$\Phi = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\beta^2}{n_\beta} + \frac{\sigma_\varepsilon^2}{n_\beta}}. \tag{4.4}$$

All the previous calculations can be extended in a straightforward manner to a setting where more facets are included in model (4.1) (Brennan 2001, Webb, Shavelson and Haertel 2007).

In G-theory two different types of studies are identified. In a generalizability (G) study, a sample is used to estimate the variance components related to different facets of measurement error. A decision (D) study uses the information provided by the G-study to design the best possible application of the measurement for a particular purpose. The number of items used in a D-study does not necessarily need to equal the number of items used in the G-study. Therefore the previous expressions are also valid if a different number of items n'_β is used.

The flexibility of the modelling framework used in G-theory has allowed its application to different types of data structure. For instance, it has been used to evaluate reliability in a longitudinal framework. Nevertheless, longitudinal data present some of the most difficult challenges for evaluating reliability, an issue that will be further discussed in the next chapter.

Chapter 5

Setting the Modelling Framework

Frequently in clinical practice and clinical trials patients are measured repeatedly over time. For instance, in psychiatry, this type of longitudinal evaluations constitutes a very powerful tool to obtain precise diagnostics as well as to evaluate the efficacy of new treatments or therapeutic procedures. However, longitudinal studies also bring some methodological challenges, especially from a statistical modelling perspective. Indeed, in such studies, patients usually exhibit a systematic change or evolution over time in addition to an individualized evolution that is characterized by correlated subject-specific effects. Moreover, serial correlation and heterogenous variance components are frequently present as well (Verbeke and Molenberghs 2000). A longitudinal modelling framework should be able to address the special characteristics of this type of data in order to avoid estimation bias. Chapter 3 has clearly outlined the restrictions of the classical test theory with regards to complex data structures. But also the more flexible generalizability theory has some limitations when repeated measurements need to be analyzed (Shavelson, Webb and Rowley 1989). Essentially, the G-theory modelling framework can be applied to a longitudinal setting only if strong and unrealistic assumptions are made. In what follows, we will address in some detail these assumptions and illustrate their restrictive nature.

No true-score change over time: One of the main problems we face when applying

G-theory to a longitudinal framework is the need to assume that the true scores are stable across time. In most of the longitudinal studies, such an assumption will be unrealistic. Indeed, most longitudinal studies are designed to model developmental changes or evolutions over time, not stability. It is also very implausible that patients in a clinical trial or in medical practice will not exhibit a systematic change over time as a result of the treatment they received or any other intervention. Ignoring this time evolution in the model will result in biased estimates for the variance components (Diggle, Liang, and Zeger 1994, Verbeke and Molenberghs 2000) and this will lead to biased estimates of the G-coefficients. Typically, the systematic variability not taken into account in the mean structure of the model will be "absorbed" into the variability of the measurement error and the actual reliability will be underestimated.

Uncorrelated error structure: Correlated error structures occur frequently in longitudinal studies. Usually, observations close in time exhibit a stronger association than observations that are further apart. Ignoring this type of correlation will induce bias in the variance-component estimates and, as a consequence, in the generalizability coefficients. This has been described by some authors. For example, Smith and Luecht (1992) investigated the effect of ignoring correlated errors in a longitudinal framework. Their results show that not taking into account this correlation will lead to an overestimation of the variance of the subject-specific parameter and an overestimation of the generalizability coefficient as a result. In their simulations, Smith and Luecht (1992) considered a stationary correlated error structure, i.e., the error terms were correlated but they had equal variance over time. Bost (1995) studied this issue further by examining the effect of both stationary and non-stationary autoregressive error covariance matrices. His results showed that, in the presence of non-stationary autoregressive 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-component 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 induces bias. Unfortunately, the classic modelling paradigm used in G-theory does not take into account this type of associations and assumes equal variance over time for the error terms.

Uncorrelated random effects: Another assumption underlying the G-coefficients is the independence of the random effects used in the model. This assumption can also be unrealistic in many longitudinal studies. Let us consider, for example, a

study in which patients have both a random intercept, characterizing their status at the beginning of the study, and a random slope, describing their personalized time evolution. Typically, the time evolution of a patient is related to his/her initial status or condition. Ignoring this association will once again induce bias in the estimation of the variance components and G-coefficients.

Missing data problem: Missing data are an omnipresent problem in clinical research. Frequently, in longitudinal studies, some patients miss one or more of the measurements originally planned or even drop out from the study altogether after a number of visits, thus creating a missing data problem. We will address this point, but first set out with some preliminary reflections on the estimation method. In its most classical formulation, G-theory estimates the variance components by calculating the mean square for each effect from an analysis of variance model and then equating each source to its expectation (Cronbach *et al.* 1972, Shavelson and Webb 1991, Brennan 2001). This expected mean square (EMS) estimation method for the variance components in G-theory has many severe limitations specifically in a longitudinal setting. For instance, the EMS method frequently produces negative estimates for the variance components which has led to ad hoc rules, such as setting negative variance estimates equal to zero (Cronbach *et al.* 1972). An even more serious limitation for its application in a longitudinal framework is that the EMS is only applicable to balanced designs without missing data (Marcoulides 1987, Searle, Casella and McCulloch 1992). Further, the EMS estimation procedure assumes that the error terms are independent across time. In situations where the balance of the study has been broken due to missing values, it has often been recommended to randomly discard some observations in order to recover balance (Shavelson and Webb 1991). This approach will not only imply an important loss of information but it also fully ignores the missing data generating mechanism. Indeed, such a procedure will only be valid under a missing completely at random mechanism (MCAR), a very strong and unrealistic assumption (Little and Rubin 2002, Verbeke and Molenberghs 2000, Molenberghs and Kenward 2007). In general, like all frequentist methods, EMS will be biased when data are incomplete, unless the strong and hence unrealistic MCAR assumption holds. It is fair to point out that, if all other assumptions are met, the classical G-theory analysis of variance model with random effects could still be applied in an incomplete data setting. However, we should then abandon the EMS procedure and use instead a likelihood or Bayesian approach.

Many proposals have appeared over the last decades to solve some of these mod-

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

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

In the present work, we will outline our proposals for quantifying reliability within a linear mixed models framework. This modelling paradigm will allow us to incorporate many of the previously discussed features, such as varying true scores, correlated error terms, including different types of serial correlation, heteroscedastic error components, and correlated random effects, in a very natural way (Laird and Waire 1982, Verbeke and Molenberghs 2000). Accounting for all of these complexities within the same modelling paradigm is of the utmost importance to guarantee unbiased results when estimating reliability. For instance, we can incorporate the systematic variability of the true scores into the fixed-effects structure of the model in a very flex-

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

$$\begin{aligned}
&\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i}, \qquad (5.1)\\
&\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{D}),\\
&\boldsymbol{\varepsilon}_{(1)i} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{Ri}),\\
&\boldsymbol{\varepsilon}_{(2)i} \sim N(\mathbf{0}, \mathbf{T}\boldsymbol{H}_i\mathbf{T}),\\
&\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n, \boldsymbol{\varepsilon}_{(1)1}, \ldots, \boldsymbol{\varepsilon}_{(1)n}, \boldsymbol{\varepsilon}_{(2)1}, \ldots, \boldsymbol{\varepsilon}_{(2)n} \text{ independent},
\end{aligned}$$

where $\boldsymbol{Y}_i$ is the p_i-dimensional vector of responses for subject i, with n denoting the number of subjects, p_i denoting the number of observations per subject, and $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ denoting fixed $(p_i \times q)$ and $(p_i \times r)$ dimensional matrices of known covariates. Further, $\boldsymbol{\beta}$ denotes the q-dimensional vector containing the fixed effects, $\boldsymbol{b}_i$ is the r-dimensional vector containing the random effects, and $\boldsymbol{\varepsilon}_{(2)i}$ is an p_i-dimensional vector of components of serial correlation. The error $\boldsymbol{\varepsilon}_{(1)i}$ is an p_i-dimensional vector of residual components. Moreover, $\boldsymbol{D}$ is a general symmetric $(r \times r)$ covariance matrix, $\boldsymbol{\Sigma}_{Ri}$ is an $(p_i \times p_i)$ covariance matrix, $\boldsymbol{H}_i$ is an $(p_i \times p_i)$ correlation matrix, and $\mathbf{T} = \text{diag}(\tau_j)$. In many cases, however, $\tau_j = \tau$ for all j, so that $\mathbf{T}\boldsymbol{H}_i\mathbf{T} = \tau^2\boldsymbol{H}_i$. The matrices $\boldsymbol{H}_i$ and $\boldsymbol{\Sigma}_{Ri}$ depend on i only through their dimension p_i, i.e., the set of unknown parameters will not depend upon i. For a more complete and detailed account about linear mixed models we remit the reader to, for example, Diggle, Liang and Zeger (1994) and Verbeke and Molenberghs (2000).

Model (5.1) implies the marginal model $\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i)$, where

$$\boldsymbol{V}_i = \boldsymbol{\Sigma}_{D_i} + \boldsymbol{\Sigma}_i \qquad (5.2)$$

with $\boldsymbol{\Sigma}_{D_i} = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'$ and $\boldsymbol{\Sigma}_i = \mathbf{T}\boldsymbol{H}_i\mathbf{T} + \boldsymbol{\Sigma}_{Ri}$. Note that the total variability is decomposed into a component stemming from the subject-specific random effects and a residual variability component. The remaining variability is the sum of a serial correlation part and an error part, but we will generically refer to it as the error variability.

Based on this modelling framework the next chapter proposes an extension of the concept of reliability to a longitudinal setting, using a generalization of the intraclass correlation coefficient.

Chapter 6

Generalizing the Intraclass Correlation Coefficient

One frequently used method to estimate reliability is to set up a simple replication study, and to calculate the intraclass correlation coefficient (ICC) based on an analysis of variance model with subject as factor. This procedure is based on the equivalence between reliability and correlation that emanates from classical test theory and that is expressed in (3.4). This method is valid under the assumptions of parallel measurements. Specifically, these assumptions include that (1) the errors are mutually uncorrelated, (2) the errors are uncorrelated with the true scores, (3) the error variances are homogeneous, and (4) the true scores are stable over the two test occasions.

In the present chapter, we investigate how the ICC can be generalized as a measure for reliability in a longitudinal scenario, based on more complex models where these stringent assumptions do not hold. We will propose several models of increasing complexity and we will extend (3.4) to these more general settings. The methodology will be illustrated by applying it to the data of the schizophrenia study, described in Section 2.1, to derive the reliability of the PANSS. Using this rating scale, the patient's global condition was assessed at several occasions. Obviously, the assumptions described above cannot be fulfilled in such a clinical setting. We will study how we can cope for that by using a more flexible model for this specific data structure.

6.1 Model 1

First, we assume a linear mixed model with a random intercept. In that case, the repeated PANSS scores for subject i satisfy model (5.1) with $\boldsymbol{X}_i$ the design matrix for the fixed effects which includes an intercept term, time, treatment and the interaction between time and treatment. Time is modelled as a factor with seven levels such that we obtain a saturated cell means model for time and treatment. $\boldsymbol{Z}_i$ is a p_i-dimensional vector of ones, $\boldsymbol{b}_i \sim N(0, \sigma_b^2)$ and $\boldsymbol{\varepsilon}_{(1)i} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, $\boldsymbol{\varepsilon}_{(2)i} = 0$.

At the level of an individual measurement we can write the model as

$$Y_{ij} = \mu_{ij} + b_i + \varepsilon_{ij}, \tag{6.1}$$

where Y_{ij} is the observed score at time point j for subject i; μ_{ij} groups the fixed-effects structure, b_i is the random intercept and ε_{ij} is the measurement error.

For model (6.1) we assume that (1) the errors are mutually uncorrelated, (2) the errors are uncorrelated with the true scores, (3) the error variances are homogeneous and (4) the individual-specific component is stable over different time points.

Applying (3.4) to the random-intercept model, we obtain for measurements at time points s and t

$$\begin{aligned} R &= \mathrm{Corr}(Y_{is}, Y_{it}) \\ &= \frac{\mathrm{Cov}(\mu_{is} + b_i + \varepsilon_{is}, \mu_{it} + b_i + \varepsilon_{it})}{\sqrt{\mathrm{Var}(b_i + \varepsilon_{is})}\sqrt{\mathrm{Var}(b_i + \varepsilon_{it})}} \\ &= \frac{\mathrm{Cov}(b_i, b_i)}{\sigma_b^2 + \sigma^2} \\ &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}. \end{aligned} \tag{6.2}$$

Reliability is thus calculated based on the estimated variance components of the model. Like in the simple setting considered in CTT, (6.2) expresses the ratio of the variance explained by the subjects (true scores) to the total observed variance. For data containing two measurements per subject, this measure is equivalent to the test-retest reliability of the instrument. In general, for any series of repeated measurements, if model (6.1) holds then (6.2) gives a global measure of reliability. For the PANSS data, the estimated variance parameters, based on restricted maximum likelihood (REML), equal $\hat{\sigma}_b^2 = 311.00$ and $\hat{\sigma}^2 = 125.14$, which leads to a global

reliability measure of $\hat{R} = 0.713$ (s.e. 0.012). The standard error is calculated using the delta method.

Note that the assumption of parallel measurements is not met. The mean PANSS score decreases from 92.4 at baseline to 68.8 at the last measurement. Even though classical reliability studies usually require the assumption of parallel measurements, the present approach, due to the flexibility of the linear mixed model, obviates the need for this, since the mean and variability structures can be clearly separated. In particular, the linear mixed model will account for systematic time and treatment effects by including them into the fixed effects component of the model. Although the steady state is not taken care of by design as it would be in the classical test-retest design, the steady state is provided through modelling at the analysis stage. A conceptually useful way to think about this is through the two-stage approach as the mixed effects model has been introduced historically, by Laird and Ware (1982). If we derive the individual residuals for a linear regression model including the fixed-effects parameters as in (6.1) and subsequently apply a random intercept model on the obtained residuals without a fixed effect component ($\mu_{jk} = 1$), the same estimates for σ_b^2 and σ^2 would be obtained. Furthermore, as stated in Chapter 5, there are additional advantages of using the linear mixed model: this model can be applied when (1) not all subjects have the same number of measurements (due to missingness or irregularly spaced measurement times), and (2) more complicated variance-covariance structures within subjects exist. To study these advantages further, we will consider more elaborate models in subsequent sections.

6.2 Model 2

The use of a random intercept in the assessment of reliability dates back to Bartko (1966) and has been described by Dunn (1989). Model 1 builds upon this work. In addition, we will introduce serial correlation and then generalize the calculation of reliability to this situation. Explicitly, the second model combines a random intercept with serial correlation. This component takes into account that the correlation between pairs of measurements depends on the distance, or time lag, between these measurements. The assumption we made in Section 6.1 that the errors are independent is then violated, or $\text{Cov}(\varepsilon_{is}, \varepsilon_{it}) \neq 0$.

Typical choices for such serial correlation structures for unequally spaced data are based on exponentially or Gaussian decaying processes. In order to choose the covari-

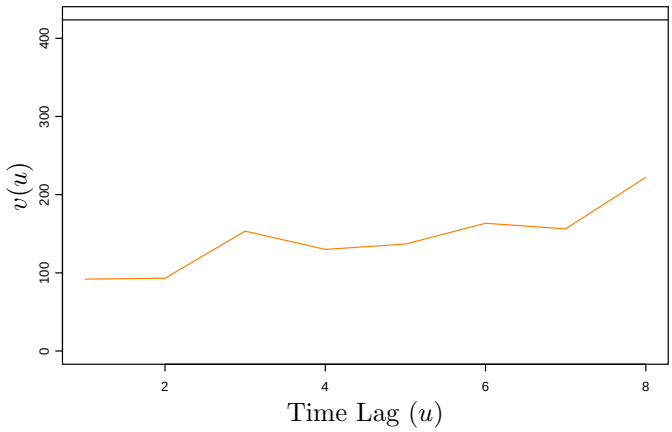

Figure 6.1: *Schizophrenia study. Empirical variogram of the PANSS data.*

ance structure that best fits the data, an empirical variogram was created (Diggle, Liang and Zeger 1994, Verbeke and Molenberghs 2000), shown in Figure 6.1. The value of the variogram at time lag zero is an indication for the relative importance of the measurement error, the discrepancy between the variogram at the largest time lag, and the process variance (represented as a level straight line at the top of the plot) is an indication for the importance of the random intercept. The strength of the serial correlation process is indicated by the amount of increase between zero and the maximum time lag, while the shape of the curve is indicative of the shape of the process of serial decay.

Figure 6.1 suggests that the largest component of variability is attributable to a random intercept. However, there is a hint that a serial component may be present as well. We opt for the Gaussian serial process. Model (5.1) still applies, with $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, $\boldsymbol{b}_i$, and $\boldsymbol{\varepsilon}_{(1)i}$ as in Section 6.1, however now $\boldsymbol{\varepsilon}_{(2)i} \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{H}_i)$. Then $\boldsymbol{\Sigma}_i$, the variance-covariance matrix grouping the measurement error and serial components, equals $\sigma^2 \boldsymbol{I}_i + \tau^2 \boldsymbol{H}_i$ with the following diagonal and off-diagonal elements

$$\begin{aligned} \Sigma_{iss} &= \tau^2 + \sigma^2, \\ \Sigma_{ist} &= \tau^2 \exp\left(\frac{-u_{st}^2}{\rho^2}\right), \quad s \neq t, \end{aligned}$$

where σ^2 denotes the measurement error variance and the remaining part is the serial

variance component with u_{st} denoting time lag between measurements Y_{is} and Y_{it} for subject i, and ρ indicates the strength of the serial correlation.

At the level of the individual measurement, the model can then be written as

$$Y_{ij} = \mu_{ij} + b_i + w_{ij} + \varepsilon_{ij}, \tag{6.3}$$

where w_{ij} is the serial correlation component with $w_{ij} \sim N(0, \tau^2)$. For time points s and t, it then follows that

$$\text{Var}(Y_{is}) = \sigma_b^2 + \tau^2 + \sigma^2 = \text{Var}(Y_{it}) \tag{6.4}$$

and

$$\text{Cov}(Y_{is}, Y_{it}) = \sigma_b^2 + \tau^2 \exp\left(\frac{-u_{st}^2}{\rho^2}\right).$$

In this model we no longer assume the errors to be mutually uncorrelated, instead we correct for dependence in the model. We do assume that (1) the errors are uncorrelated with the true scores, (2) the error variances are homogeneous ($\Sigma_{ss} = \tau^2 + \sigma^2$ for all s), and (3) the individual-specific component is stable over different time points.

Note that, in case assumption (2) is too stringent for the data at hand, compound symmetry can be relaxed further. Instead of a Gaussian serial process a more general structure can be chosen, in such a way that the variances on the main diagonal of the variance-covariance matrix are allowed to vary. In such a case assumption (2) could be restricted to stating that the *residual* error variances are homogeneous ($Var(\varepsilon_{(1)i}) = \sigma^2 \boldsymbol{I}$).

Extending the expression for the ICC (3.4) to the present model, the reliability can then be calculated as a function of time lag u_{st} between two measurements at time points s and t

$$\begin{aligned} R(u_{st}) &= \text{Corr}(Y_{is}, Y_{it}) \\ &= \frac{\text{Cov}(Y_{is}, Y_{it})}{\sqrt{\text{Var}(Y_{is})}\sqrt{\text{Var}(Y_{it})}} \\ &= \frac{\sigma_b^2 + \tau^2 \exp\left(\frac{-u_{st}^2}{\rho^2}\right)}{\sigma_b^2 + \tau^2 + \sigma^2}. \end{aligned} \tag{6.5}$$

The estimated covariance parameters of this model, applied to the PANSS data, are $\hat{\sigma}_b^2 = 103.21$, $\hat{\tau}^2 = 274.97$, $\widehat{\rho} = 6.38$, and $\hat{\sigma}^2 = 65.21$. After correction for the fixed time and treatment effects, the covariance parameter estimates show a considerable

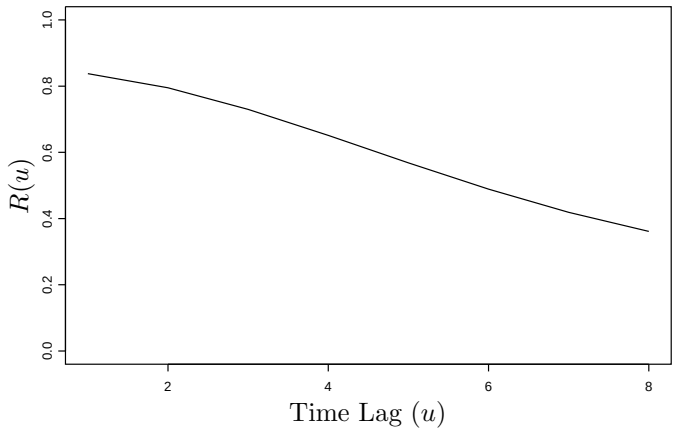

Figure 6.2: *Schizophrenia study. Reliability as a function of the time-lag u between any two measurements.*

remaining serial component in the PANSS data. As can be seen from equation (6.5), a strong serial effect will lead to a fast decreasing correlation for increasing time lags. Figure 6.2 shows that the correlation is 0.80 or higher for measurements no further apart than two weeks but declines rapidly thereafter. If we carry forward the equivalence between reliability and correlation implied by CTT, then the previous graph will also indicate a fast decreasing of reliability as a function of the time lag. Even though this decreasing tendency can be easily elucidated from the correlation perspective, it is a bit more difficult to intuitively grasp its meaning when these correlations are interpreted as reliability coefficients.

6.3 Model 3

After adding serial correlation in model 2 to the random intercept model (model 1), we now add a random slope for time. Model (5.1) therefore still holds, with $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, $\boldsymbol{\varepsilon}_{(1)i}$, and $\boldsymbol{\varepsilon}_{(2)i}$ as in Section 6.2, but $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$, with

$$\boldsymbol{D} = \begin{pmatrix} \sigma_{b0}^2 & \sigma_{b0b1} \\ \sigma_{b0b1} & \sigma_{b1}^2 \end{pmatrix},$$

where σ_{b0}^2 is the variance of the random intercepts b_{i0}, σ_{b1}^2 is the variance of the random slopes b_{i1} and σ_{b0b1} is the covariance between intercepts and slopes.

The model can now be written as follows

$$Y_{ij} = \mu_{ij} + b_{i0} + b_{i1}t_j + w_{ij} + \varepsilon_{ij}, \tag{6.6}$$

where t_j refers to time point j. For time points s and t, we then have

$$\begin{aligned} \text{Var}(Y_{is}) &= \boldsymbol{z}_s \boldsymbol{D} \boldsymbol{z}_s' + \tau^2 + \sigma^2, \\ \text{Var}(Y_{it}) &= \boldsymbol{z}_t \boldsymbol{D} \boldsymbol{z}_t' + \tau^2 + \sigma^2, \\ \text{Cov}(Y_{is}, Y_{it}) &= \boldsymbol{z}_s \boldsymbol{D} \boldsymbol{z}_t' + \tau^2 \exp(-u_{st}^2/\rho^2), \end{aligned}$$

where $\boldsymbol{z}_s$ is the design row in $\boldsymbol{Z}$ corresponding to time s. Note that considering random slopes in addition to random intercepts extends beyond compound symmetry in the sense that the overall variance becomes a non-constant function of time.

The assumptions of this model are that (1) the errors are uncorrelated with the true scores and (2) the residual error variances are homogeneous ($Var(\varepsilon_{(1)i}) = \sigma^2 \boldsymbol{I}$).

The test-retest reliability for observations at time point s and time point t and time lag u_{st} between them, can be derived as the following extension of (3.4)

$$R(u_{st}) = \text{Corr}(Y_{is}, Y_{it}) = \frac{\boldsymbol{z}_s \boldsymbol{D} \boldsymbol{z}_t' + \tau^2 \exp(\frac{-u_{st}^2}{\rho^2})}{\sqrt{\boldsymbol{z}_s \boldsymbol{D} \boldsymbol{z}_s' + \tau^2 + \sigma^2}\sqrt{\boldsymbol{z}_t \boldsymbol{D} \boldsymbol{z}_t' + \tau^2 + \sigma^2}}. \tag{6.7}$$

Equation (6.7) can be used to calculate the different reliabilities for any specific time point and for any given time lag. However, fitting model 6.6 to the data leads to a non-positive definite Hessian matrix. For this reason the results will not be presented here, instead we will investigate a simpler model.

6.4 Model 4

Only the random intercept and the random slope are retained in (6.6). For this model it is assumed that (1) the errors are mutually uncorrelated, (2) the errors are uncorrelated with the true scores, and (3) the residual error variances are homogeneous. Note that the third assumption does not imply compound symmetry; the present model contains random slopes for time that allows for non-constant variance in function of time. The model can be written as

$$Y_{ij} = \mu_{ij} + b_{i0} + b_{i1}t_j + \varepsilon_{ij}. \tag{6.8}$$

Table 6.1: *Schizophrenia study. Estimated test-retest reliabilities of PANSS using random intercept + random slope model.*

	Time point								
Time point	0	1	2	3	4	5	6	7	8
0	0.80	0.79	0.76	0.72	0.68	0.62	0.57	0.52	0.47
1	.	0.79	0.79	0.76	0.73	0.69	0.65	0.61	0.57
2	.	.	0.80	0.79	0.78	0.75	0.72	0.69	0.66
3	.	.	.	0.81	0.81	0.80	0.78	0.75	0.73
4	.	.	.	.	0.82	0.82	0.82	0.80	0.79
5	.	.	.	.	.	0.84	0.84	0.84	0.83
6	.	.	.	.	.	.	0.86	0.86	0.86
7	.	.	.	.	.	.	.	0.87	0.88
8	.	.	.	.	.	.	.	.	0.89

Subsequently, the reliability of measurements observed on time s and time t is

$$R(s,t) = \frac{\boldsymbol{z}_s \boldsymbol{D} \boldsymbol{z}_t'}{\sqrt{\boldsymbol{z}_s \boldsymbol{D} \boldsymbol{z}_s' + \sigma^2}\sqrt{\boldsymbol{z}_t \boldsymbol{D} \boldsymbol{z}_t' + \sigma^2}}. \tag{6.9}$$

The estimated covariance parameters for the PANSS data are $\hat{\sigma}^2_{b0} = 315.21$, $\hat{\sigma}_{b0b1} = -8.01$, $\hat{\sigma}^2_{b1} = 7.07$, $\hat{\sigma}^2 = 79.63$. Table 6.1 displays the reliability coefficients estimated from the random intercept and slope model; only the upper diagonal is shown for this symmetric generalized *test-retest reliability matrix*. Not surprisingly we observe again that the reliability is decreasing with increasing time lag. Another result that occurs is a slight increase in the reliability measures as time goes by, but for a fixed time lag. A possible interpretation for this phenomenon is a learning effect of the raters.

Table 6.2 summarizes the parameter estimates and the log likelihood of the different models described in this and previous sections.

6.5 Conclusion

In this chapter we have attempted a generalization of the concept of reliability based on two pillar elements: 1) the linear mixed model and 2) the equivalence between reliability and correlation found in classical test theory and captured by expression

Table 6.2: *Schizophrenia study. Estimated variance components for various models.*

Effect		RI	RI+SC	RI+RS
Random effects:				
Var. rand. int.	σ^2_{b1}	311.00	103.21	315.21
Cov. (rand. int., rand. slope)	σ_{b0b1}			-8.01
Var. rand. slope	σ^2_{b1}			7.07
Residual variance:				
Serial process variance	τ^2		274.97	
Serial process corr. par.	ρ		6.38	
Measurement error var.	σ^2	125.14	65.21	79.63
-2 log likelihood		33870.7	33232.4	33331.4

RI = Random Intercept, RS = Random Slope, SC = Serial Correlation

(3.4). In general, this is a very appealing approach. Indeed, as stated in Chapter 3, the concept of correlation was at the core of the first reliability ideas developed by Charles Spearman at the beginning of the 20th century. Additionally, correlation is a well defined and understood probabilistic concept that has been successfully applied in many different areas. The equivalence between correlation and reliability, obtained in CTT, is another important element that suggested the extension proposed in the present chapter. Finally, pairwise correlations between the different observations of an individual profile are easily obtainable from the fitted LMM.

Nevertheless, our previous analyses have shown that such an extension can lead to conclusions with an unclear intuitive interpretation, specially for models with a complicated correlation structure as the one used in Section 6.2. Furthermore, the main output of this approach is an entire $p \times p$ dimensional matrix of correlations what can hinder the interpretation of the results.

In the next chapter we will attempt the extension from a totally different perspective. Essentially, we will not use as starting point the equivalence between correlation and reliability observed in CTT but we will focus on the main intuitive properties one would expect a meaningful measure of reliability should satisfy.

Chapter 7

Reliability: An Axiomatic Approach

In this chapter, we will propose an *axiomatic* definition of reliability. The idea is to extend the concept through its fundamental properties rather than mimicking any specific functional expression or relationship. This approach has been successfully applied in many different areas, especially in mathematics, statistics, and probability. We will try to exemplify the general idea using two very well-known examples. When extending the classical concept of distance from the plane or the three-dimensional space to more general and complex mathematical structures, mathematicians used a very similar procedure. Omitting technical details, they essentially defined a distance as any function d satisfying the following three properties: (i) a distance should be positive, $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$; (ii) a distance should be symmetric, $d(x, y) = d(y, x)$; and (iii) a distance should satisfy the triangle inequality, $d(x, z) \leq d(x, y) + d(y, z)$. A second important example is the classical definition of probability density function used in probability and statistics. Here again, the concept is defined by a set of properties. Basically, a probability density function is a function f that satisfies: (i) $f(x) \geq 0$ for all x and (ii) $\int f(x)dx = 1$.

In the classical setting the reliability of a test has been defined in three equivalent ways as (1) the squared correlation between observed scores and true scores, (2) the proportion of the total observed score variance that is due to variance in the true

scores, and (3) one minus the proportion of total variance that is due to error variance (Lord and Novick 1968). Based on these formulations, some properties automatically follow. From (1) we obtain that a measure of reliability lies between 0 and 1. From (2) we conclude that if the true score variance is equal the total variance the reliability equals 1. Further, from (3) it becomes clear that in case the error variance is equal to the total variance, the reliability equals 0. These three basic properties are going to be the cornerstone of our approach.

7.1 An Axiomatic Definition

Along the lines discussed above, we propose the following *axiomatic* definition of reliability. Following the notation introduced in Chapter 5 we will state that R is a measure of reliability if it satisfies

i. $0 \leq R \leq 1$,

ii. $R = 0$ if and only if there is only measurement error: $\boldsymbol{V}_i = \boldsymbol{\Sigma}_i$,

iii. $R = 1$ if and only if there is no measurement error: $\boldsymbol{\Sigma}_i = \mathbf{0}$,

iv. When model (3.1) holds, the classical expression for reliability (3.2) is recovered.

The first property defines a range for the values of the measurement. Note that most of the previous reliability measures are also confined to the $[0, 1]$ interval with some important exceptions like the Cronbach α. Properties (ii)–(iii) establish that R should reach its extreme values, zero and one, when only measurement error or no measurement error, respectively, is present in the observations. Finally, (iv) states that the new measures should allow recovery of the appealing, classical definition of reliability when the necessary assumptions are met. Once these defining properties are given, the most imperative task is to find and study measures of reliability that satisfy them. The next section introduces one of such measures.

7.2 A Measure for Reliability R_T

We will now introduce a new measure of reliability that fulfills the properties (i) – (iv) presented in the previous section. Following the notation described in Chapter 5

the so-called R_T coefficient is given by

$$R_T = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{tr}(\boldsymbol{V}_i) - \text{tr}(\boldsymbol{\Sigma}_i)}{\text{tr}(\boldsymbol{V}_i)},$$

where $\text{tr}(\boldsymbol{A})$ denotes the trace of the matrix $\boldsymbol{A}$. Even though it is not explicitly required by the defining properties (i)–(iv), the previous expression closely resembles the formula of reliability used in CTT. Indeed, in this expression $\text{tr}(\boldsymbol{V}_i)$ accounts for the total variability in the observations for patient i, whereas $\text{tr}(\boldsymbol{\Sigma}_i)$ accounts for the measurement error variability. Therefore

$$\frac{\text{tr}(\boldsymbol{V}_i) - \text{tr}(\boldsymbol{\Sigma}_i)}{\text{tr}(\boldsymbol{V}_i)},$$

is the proportion of all the variability in the observations of subject i that is not due to measurement error. This becomes clearer if the expression for R_T is rewritten as

$$\begin{aligned} R_T &= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\text{tr}(\boldsymbol{\Sigma}_i)}{\text{tr}(\boldsymbol{V}_i)}\right) \\ &= 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\text{tr}(\boldsymbol{\Sigma}_i)}{\text{tr}(\boldsymbol{V}_i)}. \end{aligned} \tag{7.1}$$

Notice that the R_T coefficient can be seen as the average of all patients' contributions. In case of a balanced study design where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ and $\boldsymbol{V}_i = \boldsymbol{V}$ for all i, the following simplification follows

$$R_T = 1 - \frac{\text{tr}(\boldsymbol{\Sigma})}{\text{tr}(\boldsymbol{V})}. \tag{7.2}$$

In the following developments, and without loss of generality, the assumptions that lead to (7.2) are going to be considered to simplify the notation. It is important to point out that these assumptions are frequently encountered in clinical trials, precisely the scenario we are working in.

If data from K clinical trials are available, then it is possible to show that

$$\begin{aligned} R_T &= \sum_{k=1}^{K}\frac{n_k}{n}\left(\frac{\text{tr}(\boldsymbol{V}_k) - \text{tr}(\boldsymbol{\Sigma}_k)}{\text{tr}(\boldsymbol{V}_k)}\right), \\ R_T &= \sum_{k=1}^{K}\frac{n_k}{n}R_{Tk}, \end{aligned}$$

where n_k denotes the sample size of the k^{th} trial and R_{Tk} is the corresponding value of R_T in that trial. Basically, this meta-analytic version of the R_T coefficient is just

a weighted sum of the different trial contributions, and the weights are proportional to the size of the trial. This expression will allow us to study reliability in a meta-analytic context by combining the information collected in different studies. It also shows that, without loss of generality, one can concentrate on the study of the single trial setting.

At the beginning of this section it has been mentioned that the R_T coefficient satisfies the four defining properties for a measure of reliability. A formal proof for this statement can be found in Appendix A.1. In the next section we construct a point estimate and an asymptotic confidence interval for R_T.

7.3 Estimating R_T

If $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{\Sigma}}$ denote the maximum likelihood estimator for $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ respectively, then the maximum likelihood estimator (MLE) for R_T can be obtained as

$$\hat{R}_T = 1 - \frac{\text{tr}(\hat{\boldsymbol{\Sigma}})}{\text{tr}(\hat{\boldsymbol{V}})}.$$

Further, under general regularity conditions, the delta method implies that asymptotically

$$\hat{R}_T \sim N\left(R_T, \boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}'\right),$$

where $\boldsymbol{\Sigma}_P$ is the variance-covariance matrix of the variance covariance parameter estimates and $\boldsymbol{\Delta}' = \dfrac{\partial R_T}{\partial \boldsymbol{\psi}}$ with $\boldsymbol{\psi}$ a vector containing all parameters in $\boldsymbol{D}$, $\mathbf{T}$, and $\boldsymbol{\Sigma}_R$. A $(1-\alpha)\%$ confidence interval for R_T can then be given by

$$\left[\hat{R}_T \pm z_{1-\frac{\alpha}{2}}\sqrt{\boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}'}\right].$$

However, the upper and lower limits of this asymptotic interval can lie, in some circumstances, outside the $[0, 1]$ range. To avoid this issue the following logit transformation was used,

$$l(R_T) = \log\left(\frac{R_T}{1-R_T}\right)$$

and this implies

$$\begin{aligned}\frac{\partial l(R_T)}{\partial \boldsymbol{\psi}} &= \left[\frac{1-R_T}{R_T}\right]\frac{\frac{\partial R_T}{\partial \boldsymbol{\psi}}(1-R_T(\boldsymbol{\psi}))+\frac{\partial R_T}{\partial \boldsymbol{\psi}}R_T}{(1-R_T)^2}\\ &= \frac{\frac{\partial R_T}{\partial \boldsymbol{\psi}}-\frac{\partial R_T}{\partial \boldsymbol{\psi}}R_T+\frac{\partial R_T}{\partial \boldsymbol{\psi}}R_T}{R_T(1-R_T)}\\ &= \frac{1}{R_T(1-R_T)}\frac{\partial R_T}{\partial \boldsymbol{\psi}}.\end{aligned}$$

A $(1-\alpha)\%$ confidence interval for $l(R_T)$ is then given by

$$\left[l(\hat{R}_T) \pm z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{\partial l(R_T)}{\partial \boldsymbol{\psi}}\boldsymbol{\Sigma}_P\frac{\partial l(R_T)'}{\partial \boldsymbol{\psi}}\right)}\right]$$

or

$$\left[l(\hat{R}_T) \pm \frac{z_{1-\frac{\alpha}{2}}}{R_T(1-R_T)}\sqrt{\boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}'}\right].$$

A restricted $(1-\alpha)\%$ confidence interval for R_T can then be obtained by transforming back the previous interval, leading to

$$\left[\frac{e^{l_1}}{1+e^{l_1}}, \frac{e^{l_2}}{1+e^{l_2}}\right],$$

with l_1 the lower limit and l_2 the upper limit for $l(R_T)$. More details on the derivation of the different elements of $\boldsymbol{\Delta}$ can be found in Appendix B.1.

7.4 R_T and the Number of Measurements

We have defined R_T as a measure of reliability in a longitudinal setting. The most distinctive characteristic of a longitudinal design is the repeated evaluation over time of all the patients included in the study. It is therefore appealing to investigate the relationship between this new measure and the number of repeated measurements per patient.

Let us start by calculating R_T for a random intercept model. Then, model (5.1) holds with $\boldsymbol{Z}$ a p-dimensional vector of ones, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \sigma_b^2)$ and $\boldsymbol{\varepsilon}_{(1)i} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, $\boldsymbol{\varepsilon}_{(2)i} = 0$. It follows that $\boldsymbol{V} = \sigma_b^2\boldsymbol{J} + \sigma^2\boldsymbol{I}$ and $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. In this setting R_T takes the form

$$R_T = 1 - \frac{\text{tr}(\boldsymbol{\Sigma})}{\text{tr}(\boldsymbol{V})} = 1 - \frac{p\sigma^2}{p(\sigma_b^2+\sigma^2)} = \frac{\sigma_b^2}{\sigma_b^2+\sigma^2}. \tag{7.3}$$

Note that, under the random intercept model, R_T also coincides with the classical definition of reliability, exactly as in the cross-sectional setting. From the above expression, it can be seen that for this simple model R_T does not depend on the number of time points p. This result can help us to understand the intuitive meaning of the R_T coefficient. Indeed, under the random intercept model the response Y_{ij} of subject i at time point j is given by the equation (6.1). This implies that at each time point the reliability of Y_{ij} equals $R(t_j) = \dfrac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$, precisely the value found in (7.3). This hints on interpreting the R_T coefficient as some kind of *average* reliability over the different time points.

Let us now move forward to study the effect of increasing the number of measurements under a fully general model, where multiple random effects are considered as well as a general error covariance structure. If model (5.1) holds and every patient was evaluated at p different times points, then we can write R_T as follows

$$R_{Tp} = 1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_p)}{\mathrm{tr}(\boldsymbol{\Sigma}_{Dp}) + \mathrm{tr}(\boldsymbol{\Sigma}_p)} = 1 - \frac{\dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_p)}{\mathrm{tr}(\boldsymbol{\Sigma}_{Dp})}}{1 + \dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_p)}{\mathrm{tr}(\boldsymbol{\Sigma}_{Dp})}} = 1 - \frac{x_p}{1 + x_p},$$

with $x_p = \dfrac{\mathrm{tr}(\boldsymbol{\Sigma}_p)}{\mathrm{tr}(\boldsymbol{\Sigma}_{Dp})}$. If we define $f(x_p) = \dfrac{x_p}{1 + x_p}$, then the derivative of this function equals $f'(x_p) = \dfrac{1}{(1 + x_p)^2} \geq 0$, and this implies that $f(x_p)$ is an increasing function of x_p. Hence, $R_{Tp} = 1 - f(x_p)$ decreases when x_p increases.

When a new time point $p + 1$ is added, then x_{p+1} takes the form

$$x_{p+1} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_p) + \sigma_{p+1}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_{Dp}) + \boldsymbol{z}_{p+1}\boldsymbol{D}\boldsymbol{z}_{p+1}'}.$$

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

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_p)}{\mathrm{tr}(\boldsymbol{\Sigma}_{Dp})} > \frac{\sigma_{p+1}^2}{\boldsymbol{z}_{p+1}\boldsymbol{D}\boldsymbol{z}_{p+1}'}.$$

Essentially, this implies that the expanded sequence of observations will have a higher reliability if and only if the ratio of error variance to true variance of the new observation is smaller than the ratio of error variance to true variance of the previous p measurements. Therefore, the R_T coefficient can either increase or decrease when a new observation is added, depending on the "quality" of the new measurement. Clearly, the previous findings confirm the intuitive interpretation of R_T as the *average* reliability over an entire sequence of measurements.

7.5 A Simulation Study

We further investigate the performance of the point estimator and the asymptotic confidence interval for R_T under various conditions via simulations. We considered 36 different simulation settings. In a first stage, the data were generated based on the following linear mixed model with random intercept

$$Y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 Z_i + b_i + \varepsilon_{ij}, \tag{7.4}$$

where Y_{ij} refers to an observation for subject i at time t_j, and Z_i is the treatment indicator variable. Further, $b_i \sim N(0, \sigma_b^2)$, $\varepsilon_{ij} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$, $\sigma_b^2 = 300$, $\sigma^2 = 30$, 300, or 3000 and the sample size was set to $n = 50$, 150 or 300. These choices for σ_b^2 and σ^2 allow us to study the performance of R_T when the error variance is 9%, 50%, and 90% of the total variance, respectively. These settings correspond to high, medium, and low reliability. In a second stage, data were generated based on a linear mixed model with random intercept and random slope for time

$$Y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 Z_i + b_{1i} + b_{2i} t_j + \varepsilon_{ij}, \tag{7.5}$$

where $(b_{1i}, b_{2i})' \sim N(\mathbf{0}, \boldsymbol{D})$, and $\varepsilon_{ij} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ and

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

The same choices for σ^2 and n are made as before. The norm $||\boldsymbol{D}||$ was used as an indication of the "size" of the random-effects variance and based on this, the values of the error variance account again for 9%, 50%, and 90% of the total variance.

The mean parameters were fixed at $\beta_0 = 85$, $\beta_1 = 2.5$, and $\beta_2 = 3$. These values are based on the results obtained when the previous models were fitted using the schizophrenia case study data. We considered $p = 3$, 6, and 9 time points of measurement and 500 data sets were simulated in each setting.

Table 7.1 presents the true values, estimated values, and the coverage probabilities for a 95% confidence interval for R_T, where the random intercept model has been used as a data generating mechanism. Table 7.2 presents the results for the data coming from a model with random intercept and slope.

Let us first look at the closeness of the point estimates to the real value of R_T. In general, the point estimates are always very close to the true values, even for small sample sizes. Only when the measurement error accounts for 90% of all the variability

Table 7.1: *Simulation study on R_T: random intercept model (7.4). Effect of sample size (n), number of repeated measurements (p), and error percentage (%) on the estimate for R_T ($\hat{R}_T$), and the coverage probabilities (CP) for a 95% confidence interval.*

			$n = 50$		$n = 150$		$n = 300$	
%	p	R_T	$\hat{R}_T$	CP	$\hat{R}_T$	CP	$\hat{R}_T$	CP
9	3	0.909	0.908	95.0	0.909	96.8	0.909	98.2
9	6	0.909	0.907	96.6	0.909	97.4	0.909	97.8
9	9	0.909	0.907	96.8	0.909	97.8	0.909	97.6
50	3	0.500	0.510	91.2	0.508	93.6	0.506	96.4
50	6	0.500	0.502	95.2	0.504	95.2	0.501	95.0
50	9	0.500	0.499	94.6	0.502	96.0	0.501	95.2
90	3	0.091	0.129	86.5	0.101	93.1	0.098	94.2
90	6	0.091	0.101	93.8	0.096	96.0	0.094	97.2
90	9	0.091	0.096	96.4	0.094	96.6	0.093	97.6

in the data and only three repeated measurement per patient were taken, lager samples sizes are required to achieve a good estimate. The bias in this problematic setting seems to be larger for the more complicated model, i.e., the model including a random intercept and a slope. Note further the values for R_T, based on a random intercept model with homogeneous error variances, do not depend on the number of time points, as previously shown.

Let us now look at the coverage probabilities of the confidence intervals. The tables show that coverage probabilities below 95% appear almost exclusively in the settings where only three repeated measurements per subject were taken and the measurement error accounted at least for 50% of the total variability. As one would expect the situation worsens when the error variability increases to 90%. Further, smaller sample sizes, larger error variability, and more complex data seem to increase the chance of a low coverage probability. However, it is important to indicate that in all the settings, the point estimator and asymptotic confidence interval perform reasonably well when a sample size of 150 patients and 6 repeated measurements where considered.

Finally, we draw the attention to the fact that both, the point estimators and the

Table 7.2: *Simulation study on R_T: random intercept and slope model (7.5). Effect of sample size (n), number of repeated measurements (p), and error percentage (%) on the estimate for R_T ($\hat{R}_T$), and the coverage probabilities (CP) for a 95% confidence interval.*

			$n = 50$		$n = 150$		$n = 300$	
%	p	R_T	$\hat{R}_T$	CP	$\hat{R}_T$	CP	$\hat{R}_T$	CP
9	3	0.917	0.917	92.5	0.917	93.2	0.916	95.2
9	6	0.940	0.940	97.2	0.940	98.8	0.940	99.4
9	9	0.961	0.960	99.2	0.960	99.8	0.961	100
50	3	0.523	0.574	83.6	0.556	83.4	0.536	90.8
50	6	0.612	0.616	94.5	0.614	94.8	0.612	96.0
50	9	0.711	0.710	96.0	0.710	96.0	0.709	95.0
90	3	0.099	0.248	61.8	0.179	73.6	0.145	82.0
90	6	0.136	0.173	88.3	0.152	95.8	0.144	96.2
90	9	0.197	0.213	95.5	0.202	96.6	0.199	95.1

confidence intervals, are based on the asymptotic properties of maximum likelihood, or restricted maximum likelihood, estimators of the variance components. The results of the simulations illustrate that these asymptotic results work pretty well, even with small sample sizes, and the bias is negligible in almost all settings considered.

7.6 Conclusion

In the present chapter we introduced an axiomatic definition of reliability. The general idea is to capture the fundamental characteristics of the concept in a reduced and simple set of properties. If successful, these type of definitions can usually bring a lot of flexibility while keeping the intuitive interpretation of the concept one tries to extend. The definition also brings a degree of consistency by requiring that all measures of reliability, no matter how different they could be, should satisfy a minimum set of properties.

One obvious issue that such a definition rises, is the evaluation of the suitability of the chosen properties. However, in any axiomatic approach it is logically impossible to

prove theoretically that the selected set is the most appropriate one. Nevertheless, the R_T coefficient that emanates from this definition seemed to give sensible results when applied in simulations, reinforcing our confidence on the plausibility of the proposed definition.

One of the main advantages of the R_T coefficient is that it allows to summarize the reliability of an entire sequence of observations in a single yet meaningful measure. This, however, does not preclude the possibility of calculating the measure at each time point in order to construct a reliability function over time. In the next chapter we will further study the performance of the R_T coefficient, this time by applying it to the case study in schizophrenia.

Chapter 8

Estimating Reliability of Three Rating Scales for Schizophrenia

In this chapter the methodology introduced in Chapter 7 is applied to evaluate the reliability of the rating scales used in the schizophrenia case study described in Chapter 2. Data from the clinical trial by Peuskens *et al* (1995) were used to estimate the reliability of the three outcome scales; the Positive and Negative Syndrome Scale (PANSS), the Brief Psychiatric Rating Scale (BPRS), and the Clinical Global Impression (CGI). All three scales are regularly used for measuring the severity of schizophrenia. The comparison of reliability estimates, based on a single population, is therefore an interesting excercise. More so given the relatively large differences in size and therefore assessment time between the three scales.

The methodology is entirely model-based, model building is thus a crucial step towards reliability estimation. To this effect, model building guidelines, as laid out in, for example, Verbeke and Molenberghs (2000, Ch. 9) ought to be followed. In the following sections we will give an outline of the model building exercise for each of the scales. In a final section we will summarize the results of the reliability estimations.

The clinical trial contains 453 patients with chronic schizophrenia, randomly assigned to treatment with risperidone or a conventional antipsychotic drug. Patients

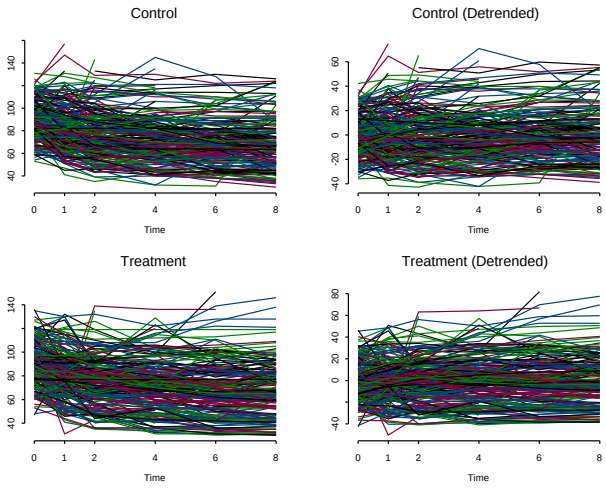

Figure 8.1: *PANSS. Individual profiles per treatment group.*

were evaluated at baseline and after 1, 2, 4, 6, and 8 weeks.

8.1 Model Building for PANSS

The Positive and Negative Syndrome Scale (PANSS) contains 30 items to be scored in 7 grades. As a consequence, the total score of the scale ranges between 30 and 210, with higher scores indicating worse conditions.

8.1.1 Exploratory Data Analysis

The individual profiles are displayed in the left panel of Figure 8.1. The graph suggests a subject-specific nonlinear downward trend over time in both treatment groups. Additionally, the figures also indicate differences between the subjects at the beginning of the study. From the individual profiles it can also be learned that some patients dropped out before the end of the study.

In addition to the average evolution, the covariance structure is also important to build up an appropriate longitudinal model. Notice that properly modelling the covariance structure is especially relevant in this application, given the crucial role of the variance components in the estimation of reliability coefficients. The right panel

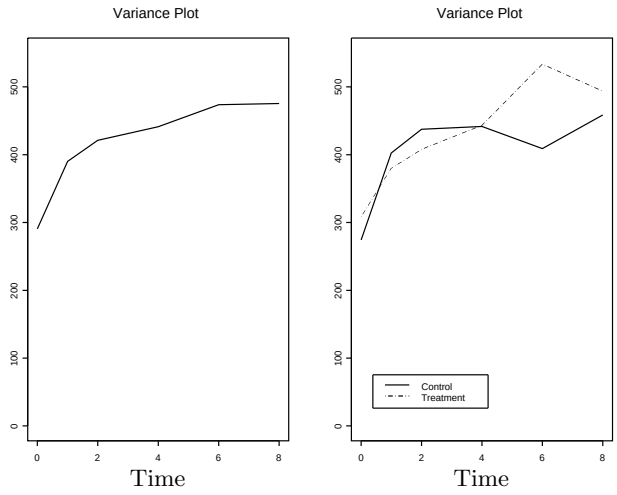

Figure 8.2: *PANSS. Variance of detrended observations for all patients (left), and per treatment group (right).*

of Figure 8.1 shows the detrended profiles and the corresponding variance function is plotted in Figure 8.2. On the left, the variance plot is given for all data and on the right it is plotted separately per treatment group. The overall variance function increases over time, suggesting a model with a random time effect. Interestingly, the right graph in Figure 8.2 shows a different variance profile for each treatment group. The difference results mainly from a higher variability in the treatment group compared to the control group at week 6. This is an unexpected peculiar feature that deserves some attention. As a first step, we decided to explore whether missingness might lie at the basis of this finding.

Table 8.1 reveals some interesting patterns, for instance, it shows that there is more missingness in the control group than in the treatment group and this difference is largest, precisely, at week 6. Additionally, Figure 8.3 illustrates another prominent issue. Indeed, the bottom left graph in Figure 8.3 clearly shows that the patients dropping out at week 6 in the control group are those with the worse average profile. The boxplots for the control group confirm the higher PANSS scores for patients dropping out at week 6, and further show a larger variability in the group of patients dropping out compared to the patients that stayed in the study. In the treatment group, the difference between patients dropping out at week 6 and patients staying

Table 8.1: *PANSS. Number (percentage) of missing values per time point of measurement and for both treatment groups.*

	Baseline	Week 1	Week 2	Week 4	Week 6	Week 8
Control	0	6 (2.7)	21 (9.3)	33 (14.6)	52 (23.0)	61 (27.0)
Treatment	0	1 (0.4)	8 (3.5)	23 (10.1)	31 (13.7)	48 (21.2)

in the study is much less pronounced. The fact that a relatively large number of control-group patients drops out at week 6 could explain why the observed variability decreases at week 6 for the control group. Basically, we observed that at this week a large proportion of the patient with a bad evolution abandoned the study in the control group. This group of patient is fairly variable and their departure redounded in a more homogeneous subsample of patients in the control group at week 6. This pattern is not observed in the treatment group where much less patients dropped out from the study at that point. We therefore believe that the characteristics of the missing data process can explain the peculiarity of the variance function displayed in Figure 8.2.

The previous discussion clearly shows the importance of missing data. Missing data are an almost unavoidable problem in longitudinal studies. In the next section, linear mixed models will be adopted, as proposed in Chapter 5. However, because fitting linear mixed models has a likelihood basis, the ensuing inferences are valid for both balanced as well as unbalanced data. Also, when the data are incompletely observed, the methodology remains statistically valid if the missing data mechanism is missing at random (Rubin 1976), in the sense that missingness is allowed to depend on observed data but, given these, not further on unobserved data. All analyses discussed in this chapter are performed under this assumption. The finding that dropping out is more likely for patients with worse evolutions, as shown in Figure 8.3, gives additional support to this assumption.

Further, we explored the intra-subject correlation by displaying individual scatter plots of standardized residuals, as shown in Figure 8.4. The graph seems to show that a slowly decaying correlation over time is present in the data.

Finally, we evaluated if subject-specific profiles could be described by a linear regression model. As a first exploratory tool, we calculated the subject-specific co-

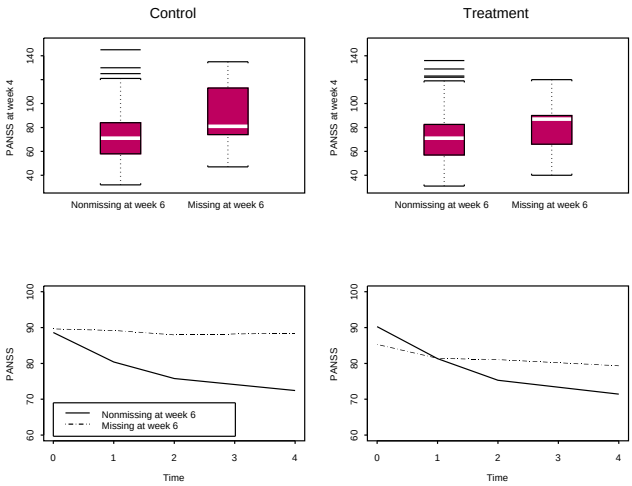

Figure 8.3: *PANSS. Top: boxplots at week 4 for patients without and with missing value at week 6. Bottom: mean profiles for patients without and with missing value at week 6.*

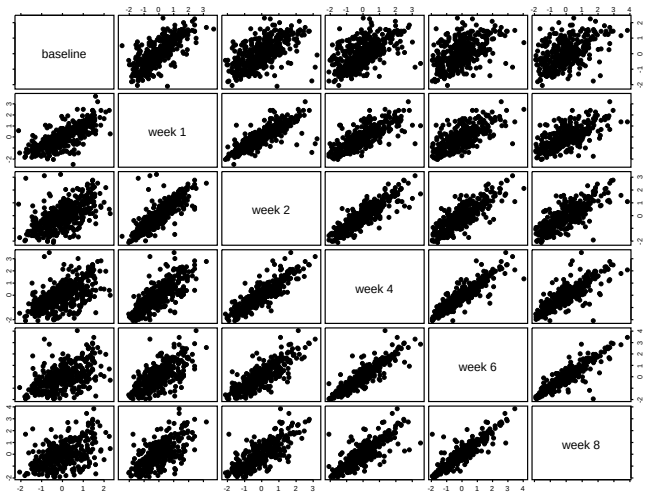

Figure 8.4: *PANSS. Scatter plot matrix of detrended observations.*

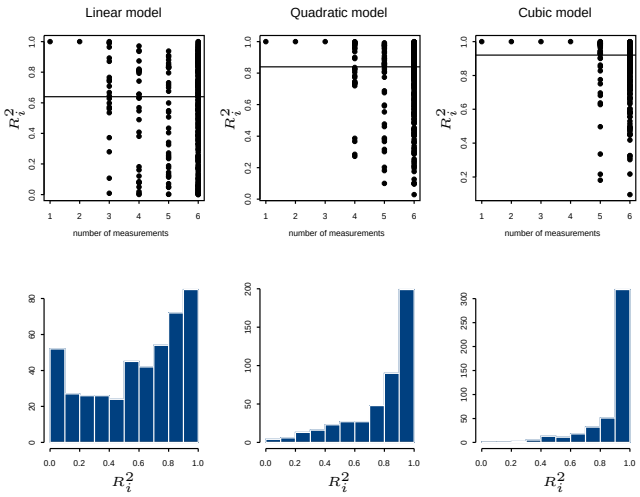

Figure 8.5: *PANSS. Subject-specific coefficients R_i^2 of multiple determination and the overall coefficient R_{meta}^2 of multiple determination for first-stage models which assume linear (left), quadratic (middle) and cubic (right) subject-specific profiles.*

efficient of multiple determination R_i^2 as well as the overall coefficient of multiple determination R_{meta}^2, for three different linear regression models (Verbeke and Molenberghs 2000); with linear, quadratic, and cubic time effect. We obtain, respectively, $R_{meta}^2 = 0.6380$, $R_{meta}^2 = 0.8430$, and $R_{meta}^2 = 0.9185$. These values strongly suggest that a model with quadratic time effect fits the data better than a model with linear subject-specific time trend. Furthermore, a model with a cubic trend still seems to improve the fit to a certain degree. These results are illustrated in Figure 8.5, where a scatter plot of R_i^2 values against the number of time points on top and a histogram of the R_i^2 values in the bottom are shown. From the left-hand histogram it can be observed that for a large number of subjects a linear trend does not fit well, represented by low R_i^2 values. A clearly smaller amount of subjects have low R_i^2 values in the middle and right-hand histograms. However, it is important to point out that missingness might partially distort this picture. Indeed, when there are only two measurements for a subject, a linear time effect leads to a perfect fit, which is captured by $R_i^2 = 1$ for this individual. The same happens with a quadratic time effect in case of three measurements and with a cubic effect in case of 4 measurements, which is clearly visible in the three scatter plots in Figure 8.5. Table 8.1 however shows that

Table 8.2: *PANSS. Model building results.*

	Random effects	Residual covariance	−2Res. LogL.	AIC
1	Cubic	simple	18994.3	19016.3
2	Quadratic, by treat.	simple	19032.0	19058.0
3	Quadratic	simple	19054.3	19068.3
4	Quadratic, by treat.	banded main diagonal	18962.0	18998.0
5	Quadratic	banded main diagonal	18987.1	19011.1

6% of the patients have no more than 3 measurements whereas 12% has no more than 4.

As a second exploratory tool, we used an F test to compare the different first-stage models (Verbeke and Molenberghs 2000). Comparing the first model with intercept and time to the second model which assumes quadratic subject-specific evolutions yields $F_{meta} = 3.3851$ on 426 and 1105 degrees of freedom, which is significant on the 5% level ($p < 0.0001$). This confirms that the second model fits the data better than the first one. Further we compared the second model to the third one resulting in $F_{meta} = 1.6143, 403$ and 702 degrees of freedom ($p < 0.0001$). This informal test thus suggests a cubic random-effects structure.

8.1.2 Model Fitting

We opted for a saturated mean structure with one parameter for each treatment by time combination. This choice was motivated by the fact that interest primarily lies in the estimation of the covariance structure. Eventually, such a general structure for the fixed effects should help to guarantee unbiased estimates for the parameters of the variance components, which are the building blocks of the reliability coefficients (Diggle, Liang and Zeger 1994).

The exploratory data analysis suggested a model with a random time effect, more precisely a quadratic or even cubic effect of time. Furthermore, we have observed that the variance plot takes different shapes for the two treatments, what might indicate a different random-effects structures for both groups. A model building exercise was carried out to investigate which of the random-effects structures, suggested by the

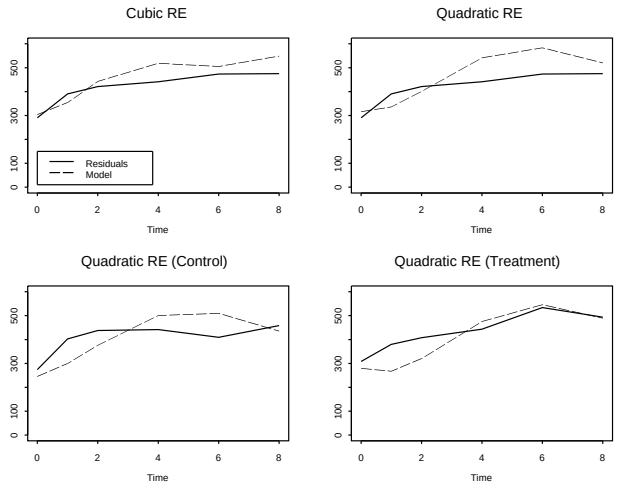

Figure 8.6: *PANSS. Variance plot for the detrended observations together with variance functions for model 1 (top left), model 3 (top right), and model 2 (bottom; control left, treatment right).*

exploratory analysis, describes the data best.

We fitted a random-effects structure with a linear, quadratic, and cubic time effect. Both a general and a treatment-specific random-effects structure were explored. The errors were assumed to follow a *simple* structure, with equal variances over time and zero covariances. The time variable (originally in weeks from 0 to 8) was centralized to stabilize the computations. Restricted maximum likelihood was used for parameter estimation (Verbeke and Molenberghs 2000) and the Akaike Information Criterion (AIC) was used to select the best model. The best results were obtained with the models 1–3 in Table 8.2. Figure 8.6 shows the variance of the detrended observations (as in Figure 8.2) together with the estimated variance function for the three models: top left for model 1 (cubic random-effects model), top right for model 3 (quadratic random-effects model), bottom left for the control arm of model 2 (quadratic random effects, per treatment group), and bottom right for the treatment arm of model 2. The two graphs in the bottom suggest that including separate random effects structures for the two treatment groups does not drastically improve the fit of the variance structure. This finding adds to the hypothesis that missing data lie at the basis of the difference in the variance plots for both treatment groups, observed in Figure 8.2.

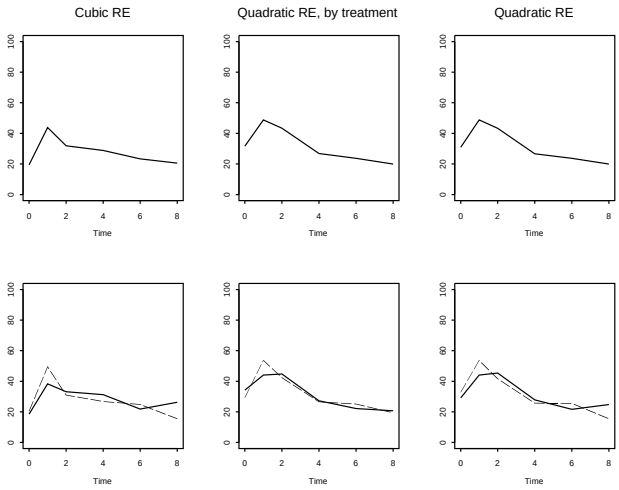

Figure 8.7: *PANSS. Variance plots for models 1, 2 and 3. For all patients (top), and per treatment group (bottom).*

Apart from the random-effects structure, the covariance structure of the errors also needs to be correctly specified. It is then important to know whether the error variances are homogeneous and if a serial correlation is present. To get an idea about the homogeneity of the error variances we further investigated the residual variance of the models 1–3. Figure 8.7 shows the variance of the residuals for these models over time; for all data (top), and per treatment group (below). The graphs show obviously that there remains some heterogeneity in the error variances for all three random-effects structures.

Finally, we explore the correlation structure among the residuals. Figure 8.8 shows the scatter plot matrix of the standardized residuals for model 1. Looking at this figure, no remaining correlation seems to be present between the residuals of pairs of measurements. This indicates that a serial correlation component in the residual covariance structure would not be needed. Scatter plot matrices for models 2 and 3 (not shown) had a similar form. In Figure 8.4 we clearly observed a strong correlation between measurements coming from the same subject. The absence of such a correlation in Figure 8.8 indicates that the within-subject correlation is entirely captured by the random-effects structure in the model.

Based on the above findings, we opted for a residual variance-covariance ma-

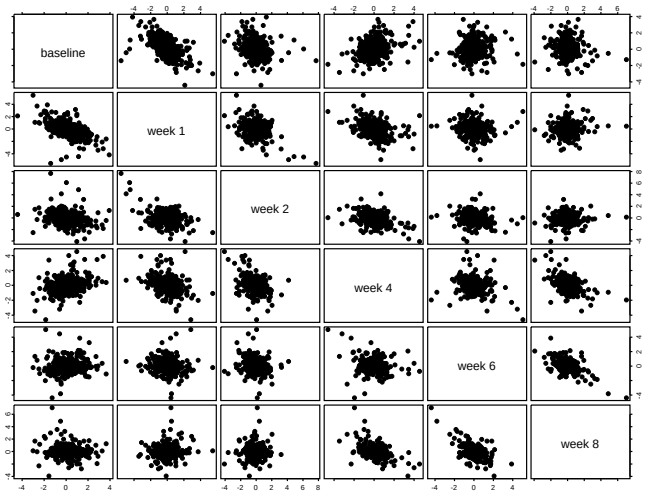

Figure 8.8: *PANSS. Scatter plot matrix of residuals for model 1.*

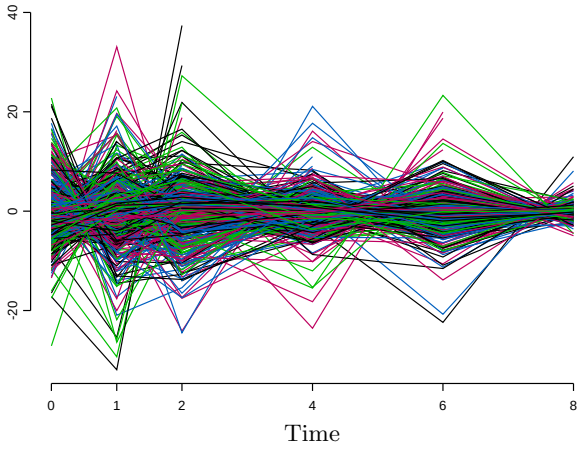

Figure 8.9: *PANSS. Individual residual profiles for model 5.*

trix with different elements on the main diagonal and zeros elsewhere (banded main diagonal). Fitting this covariance structure for the residuals together with a cubic random-effects structure lead to convergence problems. In combination with a

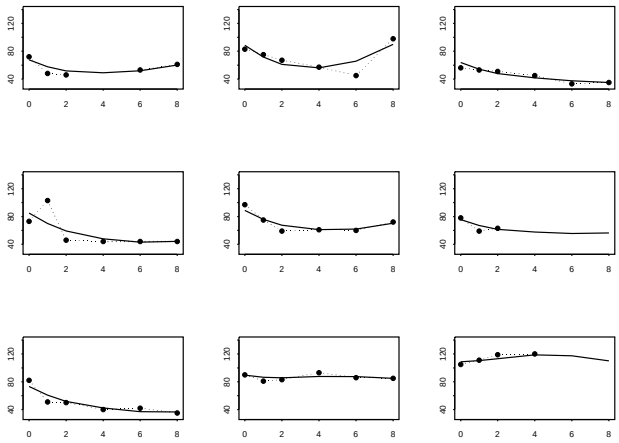

Figure 8.10: *PANSS. Individual observed (dots) and fitted (solid line) profiles for 9 randomly selected patients, based on model 5.*

quadratic random-effects structure (by treatment and general) it leads to the result presented in the lower half of Table 8.2. Following the AIC, model 4 thus emerged as the best fitting model. Several arguments, however, lead us to select model 5 as the final model to be used for reliability estimation. First, model 5 is much more parsimonious and the difference in AIC is relatively small. Second, from a clinical point of view, model 5 is more meaningful than model 4. Essentially, random effects capture subject-specific characteristics not explained by the covariates included in the model. Since the patients in the study were randomly allocated to either treatment group, there is no scientific reason to believe that differences in these characteristics may exist between both treatment groups. Third, it is not unlikely that missingness in the control group lies at the basis of the difference in variance profiles for the two treatment groups, as argumented in Section 8.1.1. This hypothesis was further supported in Figure 8.6, showing that separate random effects structures do not lead to an obviously better fit of the variance profiles. We therefore conclude by selecting model 5 as the final model, and we further present two additional graphs in support of this model.

Figure 8.9 shows the individual residual profiles for the final model. Essentially, no systematic pattern can be detected in this plot what hints on the appropriateness

Table 8.3: *BPRS. Model building results.*

	Random effects	Residual covariance	−2Res. LogL.	AIC
1	Cubic	simple	16428.5	16450.5
2	Quadratic, by treat.	simple	16470.2	16496.2
3	Quadratic	simple	16488.4	16502.4

of the chosen model. Figure 8.10 plots the individual observed (dots) and fitted (solid line) profiles for nine randomly selected patients. Also these graphs show a good fit for the individual profiles.

8.2 Model Building for BPRS

Since PANSS contains all 18 items of BPRS complemented with 12 additional items, it is not surprising that both scales are strongly correlated and exhibit very similar behavior. A model building exercise as performed for PANSS in Section 8.1 therefore delivered very similar results. For that reason we will restrict this section to the presentation of the best fitting models and the selection of the final one. Note that also here the time variable was centralized to stabilize the computations.

Like in Section 8.1.2, a saturated model was selected for the means structure. Table 8.3 presents the three best random-effects models, assuming a simple variance-covariance structure for the residuals. Among these models the first one with cubic random effects leads to the best fit. However, a very large condition number reveals an ill-conditioned $\boldsymbol{D}$ matrix for this model. The second model, with treatment-specific random-effects structure, has a slightly lower AIC value than the third model, with one general structure for the random effects.

As for PANSS, the graphical exploration suggests a residual covariance structure with heterogeneous diagonal elements and off-diagonal elements equal to zero (banded main diagonal). Models with quadratic random effects (by treatment and general) and this residual covariance structure lead, however, to large condition numbers for the residual covariance matrix indicating that it was not positive definite. Basically, the estimated variance for the last time point was close to zero.

Following the AIC, model 2 is then the best model. For the same reasons as

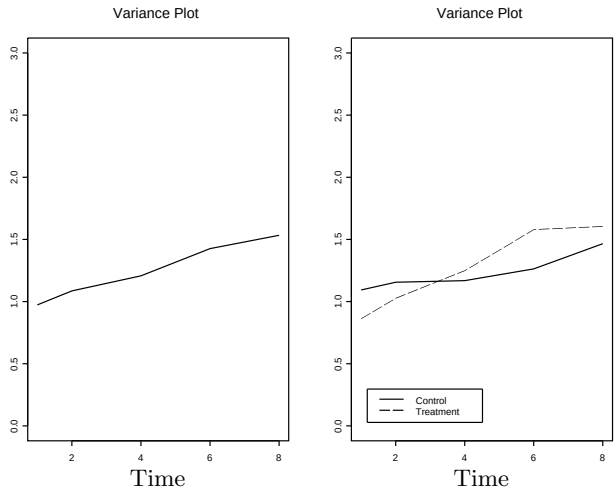

Figure 8.11: *CGI. Variance plots for all patients (left), and per treatment group (right).*

mentioned in Section 8.1.2, however, we select model 3 as the final model.

8.3 Model Building for CGI

The CGI scale is a one-item instrument indicating the change of a patient with respect to his/her baseline condition at each follow-up time. The scale has 7 grades with the following interpretation: 1) 'very much improved', 2) 'much improved', 3) 'minimally improved', 4) 'unchanged', 5) 'minimally worse', 6) 'much worse', 7) 'very much worse'.

It is clear that the scale outcome is essentially ordinal. Whether such data should be analyzed by linear models has been a topic of heated debate between statisticians and measurement theorists for the latest semi-century (Gaito 1980, Townsend and Ashby 1984, Abelson and Tukey 1963). Model (5.1) assumes that the observed scores are of a continuous nature, i.e., it is assumed they are measured on an interval or ratio scale. Such a *strong* type of measurements is very rare in psychology and psychiatry. Therefore, some will argue that statistical procedures based on the assumption of continuous responses would be inadequate in this setting. Statisticians have generally rejected the proscription of statistical methods based only on this type of measurement

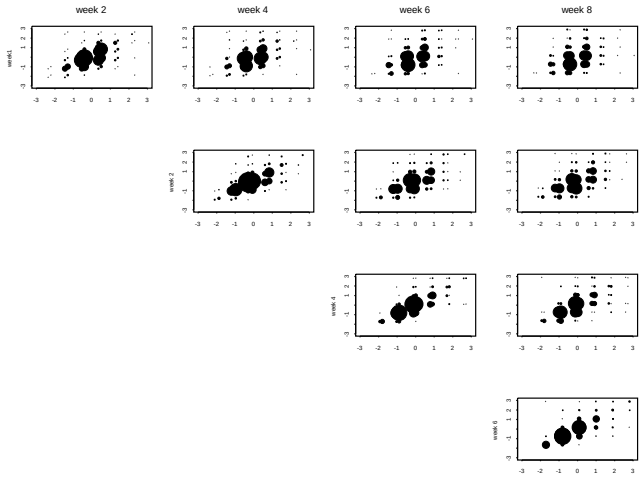

Figure 8.12: *CGI. Scatter plot matrix of detrended observations.*

considerations. We adopt this point of view in the present work and follow the viewpoints of Tukey (1961, 1962), stating that statistical procedures should not be seen as sanctification and rubber stamping for approval, but merely as nevertheless valuable approximations rooted in reality. He argued that science in general and statistics in particular rely upon the test of experience as the ultimate standard of validity. We, therefore, feel encouraged by the many applications of linear models to analyze CGI data that have given very useful and meaningful practical results in full agreement with the specific knowledge of the field.

8.3.1 Exploratory Data Analysis

The variance function of the detrended observations is plotted in Figure 8.11, for all data (left), and per treatment group (right). An increase in variance is observed in both treatment groups, suggesting again a model with a random time effect. Two other exploratory tools, the overall coefficient of multiple determination R^2_{meta} and an F test to compare the different first-stage models (Verbeke and Molenberghs 2000), point in the same direction. They suggest a quadratic or even cubic random-effects structure. Note also that Figure 8.11 indicates distinct variance functions for the two treatment groups. As for the PANSS data, there is clearly less variability in the

Table 8.4: *CGI. Four models for the random-effects structure assuming a saturated mean model and a simple error variance structure.*

	Random-effects structure	−2Res. LogL.	AIC
1	Cubic	5001.7	5023.7
2	Quadratic	5011.0	5025.0
3	Linear	5050.0	5058.8
4	Intercept	5229.9	5233.9

control group than in the treatment group at week 6. The fact that this observation is repeated for all three scales is in agreement with the hypothesis that it might be caused by a disproportionate drop-out of a specific subgroup of patients. In such a case, one should not correct for this difference in the model. As a result, models with different random-effects structures for both treatment groups will no longer be considered.

To explore the correlation structure, Figure 8.12 displays individual scatter plots of standardized residuals obtained from pairs of measurement occasions. The size of the dots in the graphs indicate the number of observations it represents. One can clearly observe the within-subject correlations for pairs of measurements, as well as a decrease of these correlations as the time between two measurements goes up.

8.3.2 Model Fitting

Once more a saturated mean structure was considered. Further, we successively fitted models with random intercept, linear, quadratic, and cubic time effect. The time variable (originally in weeks from 1 to 8) was centralized to stabilize the computations. For the residuals we assumed a simple variance-covariance structure. The results are shown in Table 8.4, indicating an almost equal fit for models with cubic and quadratic random-effects structure, both however obviously better than a model with linear subject-specific trend.

We further investigated the residual variance plots for model 1 and model 2 (Figure 8.13). The graphs show very low remaining variance, after the random effects have been added to the model. Furthermore, the variance seems to be homogeneous over

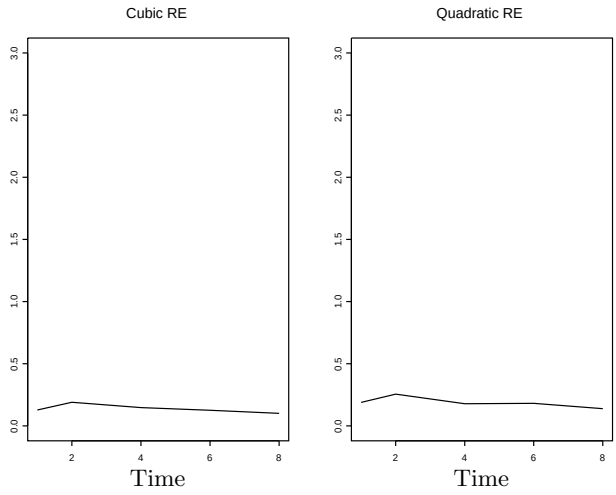

Figure 8.13: *CGI. Variance plots for model 1 (left) and model 2 (right).*

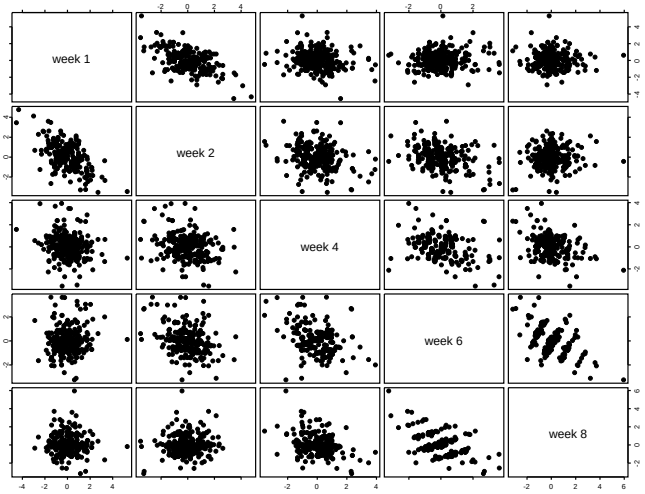

Figure 8.14: *CGI. Scatter plot matrix of residuals for model 2.*

time.

Finally, we explored the correlation structure among the residuals. Figure 8.14 shows the scatter plot matrix of the residuals for model 2. From this figure we learn

Table 8.5: *CGI. Five best fitting models.*

	Random effects	Residual covariance	−2Res. LogL.	AIC
5	Linear	unstructured banded	4992.1	5016.1
6	Quadratic	banded main diagonal	4995.7	5017.7
1	Cubic	simple	5001.7	5023.7
7	Linear	Toeplitz 2 bands	5013.9	5023.9
2	Quadratic	simple	5011.0	5025.0

that there remains a correlation between the residuals of measurements that are one time point apart (e.g., between the residuals of week 6 and 8). Between measurements further apart, no residual correlation is suggested by the graph. Similar results were obtained when this graph was constructed for model 1. Based on the scatterplot, two correlation structures can be suggested: 'Toeplitz with two bands' and 'banded unstructured'. They have the following forms, respectively

$$\begin{pmatrix} \sigma^2 & \sigma_1 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & \sigma_1 & \sigma^2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_3^2 \end{pmatrix}.$$

Additionally, we also considered models with a full serial correlation structure, following gaussian, exponential, and power patterns. In this final model building step we included models with linear, quadratic, and cubic subject-specific random slope for time. The fit statistics for the 5 best models are shown in Table 8.5.

Contrary to what Figure 8.13 suggested, the two best fitting models, 5 and 6, contain heterogeneous error variances. The table further indicates almost equal fit for these two models. Without any practical or clinical arguments in favor of either model, we follow the AIC to select model 5 as the final model, which contains a linear random-effects structure and a 'banded unstructured' variance-covariance structure for the residuals.

Figure 8.15 shows the individual residual profiles for model 5 and Figure 8.16 plots the individual observed (dots) and fitted (solid line) profiles for nine randomly selected patients for this model. These graphs indicate a good fit of the final model.

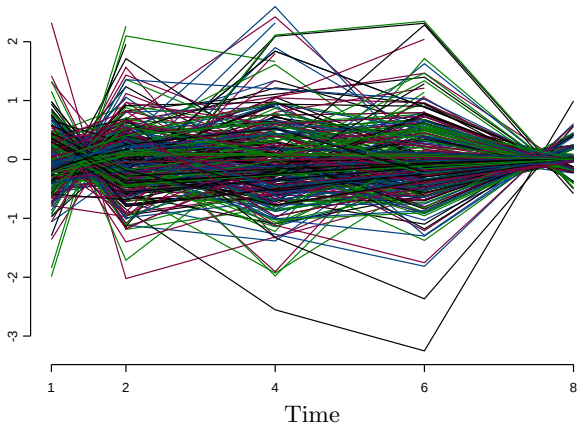

Figure 8.15: *CGI. Individual residual profiles for model 5.*

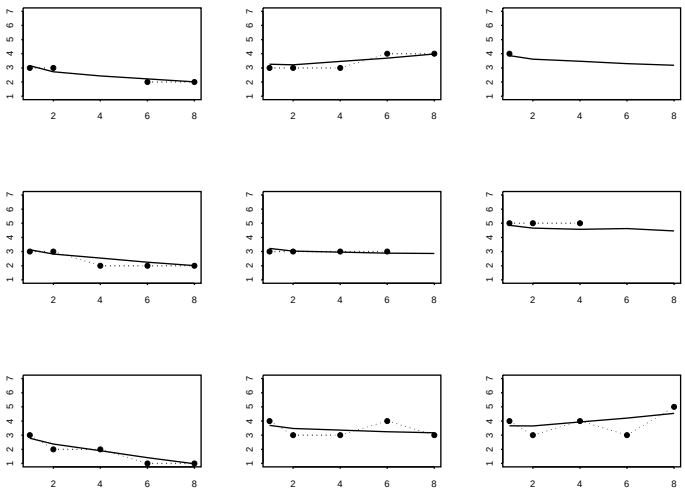

Figure 8.16: *CGI. Individual observed (dots) and fitted (solid line) profiles for 9 randomly selected patients, based on model 5.*

Table 8.6: *Schizophrenia study. Reliability estimates and 95% confidence interval for PANSS, BPRS, and CGI.*

Scale	$\hat{R}_T$	95 % confidence interval	
		lower limit	upper limit
PANSS	0.890	0.871	0.907
BPRS	0.856	0.839	0.871
CGI	0.733	0.622	0.822

8.4 Reliability Estimation

Once the best fitting models for the observed data are selected, reliability estimates can be obtained from the resulting covariance parameter estimates, following the methodology elaborated in Chapter 7. Table 8.6 presents the reliability estimates for PANSS, BPRS and CGI, together with 95% confidence intervals.

Clearly both, PANSS and BPRS, have very high reliabilities, characterized by estimates of R_T that largely exceed 80%. These results are in agreement with findings of Kay, Fiszbein, and Opler (1987) and Bell *et al* (1992) that reported test-retest reliabilities of similar magnitudes. They also agree with the empirical evidence found in clinical practice. Indeed, the ample empirical experience with these scales in common clinical practice have clearly validated them as very useful instruments for the evaluation of psychiatric patients.

As expected, we observe that PANSS has higher reliability than BPRS. More remarkably, however, is that the difference is very small. 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. In that setting, similar values were obtained for trial-level validity and individual validity for PANSS and BPRS. This may suggest that in some practical situations the use of a simpler scale like BPRS could be more advisable. Nevertheless, we should point out that the choice between different instruments usually is not based only on statistical considerations and clinical aspects must be taken into account as well.

The reliability estimate for the CGI scale was based on model 5 in Table 8.5.

However, as stated in Section 8.3.2, the fit of model 6 in this table was almost equally good. This illustrates that sometimes it can be difficult to select a single 'best' model. In such a case, it is advisable to estimate reliability based on all models that give similar fit and compare the results. When the resulting reliability estimates are similar, conclusions are straightforward. If the estimates coming from different models exhibit large differences, results ought to be interpreted with care. The reliability estimate and confidence interval for the CGI based on model 6 in Table 8.5 is given by $R_T = 0.789$ (0.743; 0.829). This value is a bit higher than the one based on model 5, but the interpretation of the results essentially does not change much. The CGI scale is obviously less reliable than PANSS and BPRS, a result that is not surprising given the simplicity of the scale. The fact that we find values above 70% indicates, however, that also this scale has an acceptable reliability when applied in a population of chronic schizophrenic patients.

As pointed out in Section 7.6, we can also compute the R_T coefficient at each time point. Such an analysis can help us to evaluate the evolution of reliability over time and is an important complement to the overall estimate. Such values are plotted in Figure 8.17, and are calculated as

$$R_{Tj} = \frac{z_j \boldsymbol{D} z_j'}{z_j \boldsymbol{D} z_j' + \sigma_j^2}$$

for time point t_j and express the reliability at each of the measurement occasions separately.

It can be observed that although BPRS performs a bit better than PANSS at the beginning of the study, it is outperformed by this scale at later observations. Noticeably, PANSS exhibits a substantial increase of its reliability over time. In the same way, BPRS finds its reliability growing over time as well, but the grow is much less pronounced. We speculate that this increasing reliability over time could be the result of a learning effect of the rater. Such a learning effect could also explain the relative performance of both scales at the beginning of the study. Indeed, BPRS is used more frequently than PANSS in clinical practice and, therefore, it is better known by clinicians. It is also a simpler scale and the combination of these two factors might well explain why it leads to more reliable results than PANSS at the beginning of the study. This effect is reversed once the rater gets more experience in the use of PANSS somewhere after the second measurement. Apart from a gain in experience when using the scale, enhanced familiarity with a patient during follow-up could also lie at the basis of the increasing reliability over time. It is important to point out that

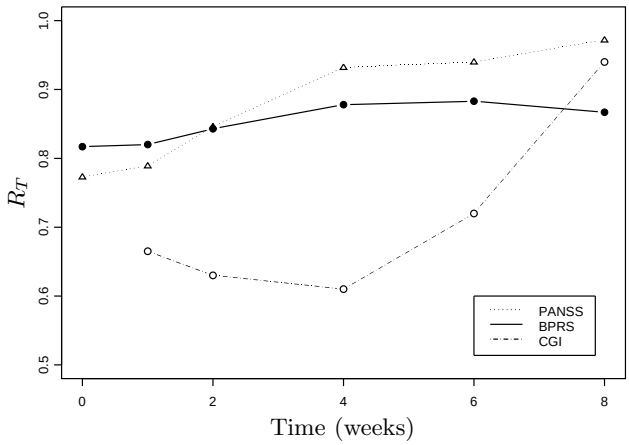

Figure 8.17: *Schizophrenia study.* R_T *over time for PANSS, BPRS and CGI.*

these are only plausible interpretations of the patterns we observed in the data but, of course, they are just speculative, a posteriori explanations and should be taken very carefully.

Also for CGI we observe higher reliabilities at later time points. Given the simplicity of the scale, a learning effect might not be a very likely explanation, however, also here it may play that a larger observation period leads to more information about the patient, what in its turn could result in a better judgement. Recall that CGI scores the change of a patient's condition compared to the baseline measurement. Intuitively, one would indeed expect that this gets easier as the patients change more, thus leading to less measurement error.

The results for PANSS, as presented in Table 8.6 are based on model 5 in Table 8.2, including one general $\boldsymbol{D}$-matrix for all patients. Model 4, with separate $\boldsymbol{D}$-matrices for the two treatment groups, however, fitted the data slightly better as indicated by the AIC. Note that different random-effects structures can imply different reliability estimates for the two groups. However, Table 8.7 shows that the estimated reliabilities for the two groups are identical, and extremely similar to the ones found for model 5. The same results were obtained for BPRS, when considering model 2 in Table 8.3.

Table 8.7: *PANSS. Reliability estimates and 95% confidence interval for PANSS, separate for the two treatment groups.*

	$\hat{R}_T$	95 % confidence interval	
		lower limit	upper limit
Control	0.892	0.869	0.912
Treatment	0.892	0.870	0.911

8.5 Conclusion

In the present chapter we analyzed a real case study to illustrate how the methodology introduced in Chapter 7 can be used for estimating the reliability of rating scales using clinical trial data. A crucial step in this methodology is the selection of the model that most closely describes the true data generating mechanism. For that reason an appropriate model building exercise is of utmost importance. Once more, one should remark that all reliability measures are, essentially, model based quantities. Therefore, their scope and applicability will never surpass the scope and applicability of the model they are based on.

We estimated the reliability of three outcome scales used to measure the severity of schizophrenia: PANSS, BPRS, and CGI. High reliabilities were found for all three the scales, a finding that is in full agreement with the literature and clinical practice. Interestingly, despite the fact that PANSS is almost twice the size of BPRS, their reliabilities are strikingly similar. Further, given the simplicity of CGI, it is not a surprise that its overall performance is a bit lower compared to the other two scales. However, after an observation period of eight weeks, the CGI proved to be very useful for distinguishing between patients that have improved and the ones that have not.

Chapter 9

A Family of Measures for Reliability

The definition introduced in Chapter 7 does not lead to a unique measure of reliability. In fact, in the present chapter we will illustrate that the R_T coefficient can be framed into a more general family of measures for reliability. Further, we will study some special member of this family in detail.

9.1 The Omega Family

Alonso *et al* (2004) introduced a family of parameters to evaluate the criterion validity of psychiatric symptom scales based on canonical correlations. In the evaluation of criterion validity a new scale is compared to a criterion scale with known performance. If this is applied in a longitudinal setting, canonical correlations are a useful tool to quantify the amount of information shared by both instruments. In the context of reliability, we study the reproducibility of a single scale, which implies that canonical correlations are no longer applicable. Nevertheless, we will show that the role played by canonical correlations in the validity research, is in the reliability context assumed by the generalized eigenvalues associated with specific variance-covariance matrices. Let us start by introducing the following theorem.

Theorem 1 *Given the function* $q(\lambda) = |\boldsymbol{\Sigma} - \lambda \boldsymbol{V}|$*, if model (5.1) holds then: (i) all roots of* $q(\lambda) = 0$*, the so-called generalized eigenvalues, are real, and (ii) if* λ_j *is a root of* $q(\lambda) = 0$ *then* $0 \leq \lambda_j \leq 1$*.*

A detailed proof of Theorem 1 can be found in Appendix C.1. Based on this theorem we can now define the following family

$$\Omega = \left\{ \theta : \theta = \sum_{j=1}^{p} w_j \rho_j^2 \quad \text{with} \quad w_j > 0 \quad \sum_{j=1}^{p} w_j = 1 \right\}. \tag{9.1}$$

The elements w_j are weights assigned to the parameters ρ_j^2 where $\rho_j^2 = 1 - \lambda_j$ and the λ_j's are the roots of the equation $q(\lambda) = 0$. It is useful to note that the λ_j's could be equivalently defined as the eigenvalues of the matrix $\boldsymbol{H} = \boldsymbol{V}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{V}^{-1/2}$, where $\boldsymbol{V}^{1/2}$ denotes the symmetric square root of $\boldsymbol{V}$. The matrix $\boldsymbol{H}$ is symmetric and, therefore, it can be written as $\boldsymbol{H} = \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P}'$, where $\boldsymbol{P}$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \text{diag}\{\lambda_j\}$. Using the previous results it immediately follows that

$$\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2} \boldsymbol{H} \boldsymbol{V}^{1/2} = \boldsymbol{V}^{1/2} \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P}' \boldsymbol{V}^{1/2} = \boldsymbol{Q}' \boldsymbol{\Lambda} \boldsymbol{Q}, \tag{9.2}$$

$$\boldsymbol{V} = \boldsymbol{Q}' \boldsymbol{Q}, \tag{9.3}$$

where $\boldsymbol{Q} = \boldsymbol{P}' \boldsymbol{V}^{1/2}$. Using Theorem 1 it is easy to prove that all the members of the Ω family satisfy properties (i)–(iv), introduced in Section 7.1. This proof is provided in Appendix A.2.

This family is structurally similar to the family introduced by Alonso *et al* (2004) in the validity context. The main difference is that here the ρ_j^2 are not the canonical correlations associated with the new and criterion scales, but rather a function of the generalized eigenvalues associated with the total and error variance-covariance matrices.

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

specific, important members of the Ω family in more detail and we will try to shed light on their specific meaning and interpretation.

9.2 R_T as Member of the Ω Family

A first special member of the Ω family is the measure R_T, introduced in Chapter 7. Indeed, plugging (9.2) and (9.3) into expression (7.2) for R_T in a balanced setting, we obtain

$$R_T = 1 - \frac{\text{tr}(\boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q})}{\text{tr}(\boldsymbol{Q}'\boldsymbol{Q})} = 1 - \frac{\text{tr}(\boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{\Lambda})}{\text{tr}(\boldsymbol{Q}\boldsymbol{Q}')}.$$

If we call $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{P}'\boldsymbol{V}\boldsymbol{P}$, we have

$$R_T = 1 - \frac{\text{tr}(\boldsymbol{S}\boldsymbol{\Lambda})}{\text{tr}(\boldsymbol{S})} = 1 - \sum_{j=1}^{p} w_j \lambda_j,$$

with $w_j = \dfrac{s_{jj}}{\text{tr}(\boldsymbol{S})}$ and s_{jj} the jth element in the diagonal of $\boldsymbol{S}$. Note that $\boldsymbol{V}$ is positive definite and, as a consequence, $s_{jj} > 0$ for all j. Further, we also have

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

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

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

9.3 Other Members of the Ω Family

The uncountable nature of the Ω family implies that the choice of some special members to be scrutinized further must be based on pragmatic considerations. Retaining R_T is evident. Another intuitive choice is to set all weights equal to $w_j = 1/p$. We then have that

$$\begin{aligned} R_p &= \sum_{j=1}^{p} \frac{1}{p} \rho_j^2 = \sum_{j=1}^{p} \frac{1}{p} (1 - \lambda_j) \\ &= 1 - \frac{1}{p} \sum_{j=1}^{p} \lambda_j = 1 - \frac{1}{p} \text{tr}(\boldsymbol{\Sigma}\boldsymbol{V}^{-1}). \end{aligned}$$

The above expression applies for a single-trial setting and a balanced study design, assuming that $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ and $\boldsymbol{V}_i = \boldsymbol{V}$. In general, we can write R_p as

$$R_p = 1 - \sum_{i=1}^{n} \frac{1}{np} \mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1}), \tag{9.4}$$

where i denotes the subject and n denotes the total number of subjects.

It would also be appealing to consider the elements of Ω corresponding to the largest and smallest generalized eigenvalues, let us say, $\widetilde{\theta}_{\max} = \rho^2_{(1)}$ and $\widetilde{\theta}_{\min} = \rho^2_{(p)}$, where $\rho^2_{(j)} = 1 - \lambda_{(j)}$ and $\lambda_{(j)}$ is the jth largest generalized eigenvalue. However, the restrictions placed on the weights ($w_j > 0$) imply that $\widetilde{\theta}_{\max}$ and $\widetilde{\theta}_{\min}$ are not members of the Ω family. Actually, $\widetilde{\theta}_{\max}$ and $\widetilde{\theta}_{\min}$ can be interpreted as an upper and lower bound of Ω in the sense that for any given scale, and independently of the element of Ω that one may use in the analysis, the reliability of the instrument will always lie in the interval $(\widetilde{\theta}_{\min}, \widetilde{\theta}_{\max})$. Nevertheless, we can approximate $\widetilde{\theta}_{\max}$ and $\widetilde{\theta}_{\min}$ using elements of the family by defining

$$\theta_{\max} = \sum_{j=1}^{p} w_j \rho^2_{(j)} \quad \text{with} \quad w_1 \gg w_j \quad \text{for } j \neq 1, \tag{9.5}$$

$$\theta_{\min} = \sum_{j=1}^{p} w_j \rho^2_{(j)} \quad \text{with} \quad w_p \gg w_j \quad \text{for } j \neq p. \tag{9.6}$$

Note that, if the weights w_j are carefully chosen, we can be rather confident that in any practical situation if θ denotes any arbitrary element of Ω then $\theta_{\min} \leq \theta \leq \theta_{\max}$. In the next section we will study in some more detail the special measures previously proposed, via simulations.

9.4 A Simulation Study

In this section, we investigate the characteristics of the "special members" of the Ω family, introduced in sections 9.2 and 9.3, by mean of simulations. To study the behavior of the different measures under various conditions, data were generated with various amounts of measurement error, different numbers of repeated measurements per patient, and different sample sizes. For a detailed description of the simulation study we entirely refer to Section 7.5, on the simulation study for R_T. Essentially, the same data were used for the investigation of these members of Ω.

Moreover, the parameters $\theta_{\min}$ and $\theta_{\max}$ were specified in the following way

Table 9.1: *Simulation study on $\theta_{\max}$: random intercept model (7.4). Effect of sample size (n), number of repeated measurements (p), and error percentage (%) on the estimate for $\theta_{\max}$ ($\hat{\theta}_{\max}$), and the coverage probabilities (CP) for a 95% confidence interval.*

			$n = 50$		$n = 150$		$n = 300$	
%	p	$\theta_{\max}$	$\hat{\theta}_{\max}$	CP	$\hat{\theta}_{\max}$	CP	$\hat{\theta}_{\max}$	CP
9	3	0.967	0.966	99.0	0.967	98.4	0.967	99.6
9	6	0.983	0.982	100	0.983	100	0.983	100
9	9	0.988	0.988	99.6	0.988	100	0.988	100
50	3	0.749	0.752	95.6	0.754	97.0	0.753	98.6
50	6	0.856	0.854	97.2	0.857	97.6	0.857	97.8
50	9	0.899	0.897	97.2	0.899	98.8	0.899	97.6
90	3	0.231	0.290	87.9	0.245	93.9	0.242	94.6
90	6	0.375	0.383	93.8	0.383	95.8	0.378	95.8
90	9	0.473	0.468	96.4	0.476	95.4	0.477	96.0

- $\theta_{\min} = \sum_{j=1}^{p} w_j \rho^2_{(j)}$ where $w_j = 0.999$ for $j = p$ and $w_j = \dfrac{0.001}{p-1}$ otherwise, and
- $\theta_{\max} = \sum_{j=1}^{p} w_j \rho^2_{(j)}$ where $w_j = 0.999$ for $j = 1$ and $w_j = \dfrac{0.001}{p-1}$ otherwise.

A confidence interval, based on the delta method, can be derived for all members of the Ω family, assuming the weights are known constants. Details on this can be found in Appendix B.2. Note, however, that this assumption is not fulfilled by the R_T coefficient. Therefore, confidence intervals for R_T were calculated as described in Section 7.3. Using restricted maximum likelihood, we can then obtain the point estimates, the confidence intervals, and the coverage percentage (CP) of the confidence intervals. For each of the measures, two tables are presented: Tables 9.1 and 9.3 contain the results for $\theta_{\max}$ and R_p that arise from the random intercept model given in (7.4), whereas Tables 9.2 and 9.4 show the findings for the data generated by the random intercept and random slope model given in (7.5). All tables display the true values, estimated values, and the coverage probabilities for a 95% confidence interval

Table 9.2: *Simulation study on $\theta_{\max}$: random intercept and slope model (7.5). Effect of sample size (n), number of repeated measurements (p), and error percentage (%) on the estimate for $\theta_{\max}$ ($\hat{\theta}_{\max}$), and the coverage probabilities (CP) for a 95% confidence interval.*

			$n = 50$		$n = 150$		$n = 300$	
%	p	$\theta_{\max}$	$\hat{\theta}_{\max}$	CP	$\hat{\theta}_{\max}$	CP	$\hat{\theta}_{\max}$	CP
9	3	0.969	0.969	99.0	0.969	99.2	0.969	100
9	6	0.988	0.988	100	0.988	100	0.988	100
9	9	0.994	0.994	100	0.994	100	0.994	100
50	3	0.759	0.783	95.5	0.778	94.6	0.766	97.8
50	6	0.896	0.895	97.5	0.896	97.8	0.895	97.4
50	9	0.952	0.951	98.6	0.952	99.6	0.952	99.6
90	3	0.240	0.423	70.4	0.331	80.8	0.289	86.5
90	6	0.464	0.493	92.5	0.477	96.7	0.468	97.2
90	9	0.670	0.670	96.6	0.670	98.2	0.670	95.7

for the respective measure. To see the results for R_T, we refer to Tables 7.1 and 7.2 in Chapter 7. The parameter $\theta_{\min}$ took values very close to 0 in all settings and this measure is therefore not further considered.

The results of this simulation study clearly show that accurate point estimates for all parameters can be obtained with a relative small sample size of 50 patients. A larger sample size, as expected, produces narrower confidence intervals. Furthermore, the coverage probabilities for all the asymptotic confidence intervals are generally around the pre-specified 95% level. Only when a large amount of measurement error is present and a limited number of patients is available, the asymptotic confidence intervals fail to reach the pre-specified level of confidence. However, the problem is solved when the sample size increases.

Considering the values of the point estimates, Tables 7.1 and 7.2 in Chapter 7 show that the R_T coefficient produces results in line with intuition. We obtain values close to 1 when the error variance is small compared to the total variance, we settle for values in the neighborhood of 0.50 in case the error variance is half of the total variance, and values are close to 0 when error variances are large.

Table 9.3: *Simulation study on R_p: random intercept model (7.4). Effect of sample size (n), number of repeated measurements (p), and error percentage (%) on the estimate for R_p ($\hat{R}_p$), and the coverage probabilities (CP) for a 95% confidence interval.*

			$n = 50$		$n = 150$		$n = 300$	
%	p	R_p	$\hat{R}_p$	CP	$\hat{R}_p$	CP	$\hat{R}_p$	CP
9	3	0.323	0.322	99.2	0.323	100	0.323	100
9	6	0.164	0.164	100	0.164	100	0.164	100
9	9	0.110	0.110	100	0.110	100	0.110	100
50	3	0.250	0.251	94.6	0.251	98.2	0.251	99.4
50	6	0.143	0.143	98.6	0.143	99.4	0.143	99.8
50	9	0.100	0.100	100	0.100	100	0.100	100
90	3	0.077	0.097	86.8	0.082	95.6	0.081	96.6
90	6	0.063	0.064	92.8	0.064	96.0	0.063	97.6
90	9	0.053	0.052	95.8	0.053	97.0	0.053	98.4

Interestingly, $\theta_{\max}$ takes higher values in all settings. This is to be expected when we consider the definition of the measure. From (9.5) it can be seen that it is based on the maximum of the elements ρ_j^2, and can therefore be interpreted as the maximum obtainable reliability measure of the Ω family. No other member of this family will provide higher values. In Tables 9.1 and 9.2 it can further be seen that the $\theta_{\max}$ values increase with an increasing number of time points. This happens, in contrast to R_T, also in the random intercept model. To gain intuition about this behavior, let us recall that $\theta_{\max} \approx \rho_{(1)}^2 = 1 - \lambda_{(1)}$ and consider the random intercept model, where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ and $\boldsymbol{V} = \sigma_b^2 \boldsymbol{J} + \sigma^2 \boldsymbol{I}$. It can be shown that in this scenario

$$\theta_{\max} \approx \rho_{(1)}^2 = \frac{p\sigma_b^2}{p\sigma_b^2 + \sigma^2}. \tag{9.7}$$

From (9.7), it can be seen that this measure increases with the number of time points.

Turning to the third measure, R_p, we observe again a totally different pattern. This measure generally gives low values. Even when the error variance is small compared to the total variance, R_p reaches values far below 1. Studying R_p under the random intercept model, it can easily be shown that, if $\sigma^2 \neq 0$

Table 9.4: *Simulation study on R_p: random intercept and slope model (7.5). Effect of sample size (n), number of repeated measurements (p), and error percentage (%) on the estimate for R_p ($\hat{R}_p$), and the coverage probabilities (CP) for a 95% confidence interval.*

			$n = 50$		$n = 150$		$n = 300$	
%	p	R_p	$\hat{R}_p$	CP	$\hat{R}_p$	CP	$\hat{R}_p$	CP
9	3	0.509	0.506	97.1	0.508	97.2	0.509	98.2
9	6	0.313	0.312	99.0	0.313	99.8	0.313	99.8
9	9	0.216	0.215	99.8	0.216	99.8	0.216	100
50	3	0.291	0.339	91.6	0.319	90.9	0.303	96.0
50	6	0.224	0.222	95.7	0.222	98.2	0.225	96.2
50	9	0.177	0.175	99.6	0.177	99.6	0.177	99.6
90	3	0.084	0.199	72.4	0.151	76.6	0.124	82.8
90	6	0.090	0.110	91.1	0.099	96.7	0.095	95.4
90	9	0.091	0.098	94.6	0.093	98.7	0.092	99.1

$$R_p = \frac{\sigma_b^2}{p\sigma_b^2 + \sigma^2}. \tag{9.8}$$

Note that, unlike $\theta_{\max}$, R_p is a decreasing function of the number of time points. The expression further shows that, even when the error variance is very small, the measure R_p can never exceed $1/p$. Additionally, R_p is not a continuous function of σ^2 for $\sigma^2 = 0$. Indeed,

$$\lim_{\sigma^2 \to 0} R_p = \frac{1}{p} \neq 1 = R_p(\sigma^2 = 0).$$

The previous discussion clearly illustrates that these measures convey different type of information. The R_T coefficient seems to be closer to the classical idea of reliability. Indeed, it seems to express best the ratio between the true score variability and the error variability. On the other hand, $\theta_{\max}$ exhibits a totally different behavior, it is an increasing function of the number of measurements and it always leads to very high quantification of reliability. This last characteristic can be logically derived from the fact that $\theta_{\max}$, by definition, is the maximum attainable reliability. Finally, the R_p

coefficient appears to behave in a very counter-intuitive manner. The entire meaning of these new proposals and their interpretation will be further studied and clarified in the subsequent chapters.

9.5 Conclusion

In this chapter, we have defined an entire family of which all members satisfy the four defining properties. The family is built based on the generalized eigenvalues related to the error and total variance-covariance matrices. Different weights assigned to these eigenvalues lead to different members of the family.

The uncountable nature of the Ω family naturally rises the question of finding an "optimal" element. Indeed, having an infinitude of parameters to evaluate reliability, faces us with the problem of choosing the most appropriate element to be used in a given application. We believe, however, that posting this optimality problem is inappropriate in the present context. Basically, we argue that such a general optimal element does not exist.

To illustrate this point let us recall the example previously introduced over the mathematical definition of distance. The set of properties used to define a mathematical distance does not lead to a unique quantification in any given space. For instance, the Euclidian distance, the Mahalanobis distance or the distance based on the uniform norm all satisfy the defining properties. We believe that such a diversity of measures is one of the strengths of this type of axiomatic definitions. However, the question about the "optimal" distance function is somewhat sterile if it is set in a general way, because its answer will essentially depend on the specific application.

A second important example, previously mentioned, is the classical definition of probability density function used in probability and statistics. Here again, the concept is defined by a set of properties. Once more, the set of defining properties is satisfied by an uncountable class of functions. The normal distribution, the chi-squared, the beta, and the Cauchy distributions are merely some examples. As before, this diversity gives the possibility of approaching several practical and theoretical problems in a very flexible way. Note that some densities can even have very counterintuitive properties. For instance, the Cauchy density does not have a finite mean or variance but it nevertheless plays an important role in some applications in physics. Here again, it is impossible to define a general optimal density and choosing one over another one will depend on the specific application.

Our simulations have clearly illustrated that different elements of the Ω family may have different interpretations and they seem to capture different aspects of the reliability of the observations. Therefore, in the next chapters we will argue that none of them can be considered optimal in a general sense and, like in the the previous examples, their utility will depend on the specific problem we are working on.

In a similar fashion, a family of parameters has been introduced to evaluate the criterion validity of psychiatric symptom scales (Alonso *et al* 2004). It is appealing to see that the two most important psychometric characteristics of a scale can be investigated using similar methodologies.

The previous arguments will be further sustained by the developments in the next chapter where a new measure of reliability will be introduced and studied. This new measure is strongly related to the ones presented in this chapter and will help us to get a better insight about some elements of the Ω family.

Chapter 10

Reliability of a Sequence of Ratings

In Chapter 7 we proposed an axiomatic definition of reliability and introduced a measure that satisfied it, the so-called R_T coefficient. Notably, and even though it was not required by the definition, the R_T coefficient mimics the general functional form of the classical definition of reliability. Indeed, the trace of a variance-covariance matrix is usually regarded, in multivariate analysis, as a plausible generalization of the univariate concept of variance. From this perspective, it is easy to see that the functional form of the R_T coefficient is very similar to the one used in CTT. In the present chapter, we will summarize the variability in this multivariate setting, using another plausible generalization of the concept of variance: the determinant of the variance-covariance matrix. Remarkably enough, such a change leads to a completely new measure of reliability, with different mathematical properties and interpretation.

10.1 An Alternative Measure for Reliability: R_Λ

As stated above, in multivariate analysis the generalized variance of a random vector can be defined using either the trace or the determinant of the corresponding variance-covariance matrix. Replacing the trace in the definition of the R_T coefficient by the

determinant leads to the following expression for reliability

$$R_\Lambda = \frac{1}{n}\sum_{i=1}^{n}\frac{|\boldsymbol{V}_i| - |\boldsymbol{\Sigma}_i|}{|\boldsymbol{V}_i|},$$

or, equivalently,

$$R_\Lambda = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{|\boldsymbol{\Sigma}_i|}{|\boldsymbol{V}_i|} = 1 - \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}|.$$

Without loss of generality and in a similar fashion as in the previous chapters, in what follows we will focus on the single-trial setting with balanced design, assuming that $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ and $\boldsymbol{V}_i = \boldsymbol{V}$. In this scenario, the R_Λ coefficient can be written as

$$R_\Lambda = 1 - |\boldsymbol{\Sigma}\boldsymbol{V}^{-1}|. \tag{10.1}$$

Note that R_Λ is closely related to the Wilks' Lambda statistic (Johnson and Wichern 1998), well-known in multivariate analysis.

Interestingly, this apparently small replacement introduces fundamental changes. For instance, the R_Λ coefficient does not fully satisfy the definition proposed in Chapter 7. In fact, this new measure satisfies properties (i), (ii), and (iv), but not (iii). Notwithstanding, it satisfies a milder version of (iii) that states: (iii') $R_\Lambda = 1$ if and only if $|\boldsymbol{\Sigma}| = 0$. Only if $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, (iii') is equal to (iii). In general, property (iii') contains (iii) in that, if θ satisfies (iii), then it will also satisfy (iii') and therefore, the latter provides a more flexible defining set of properties for reliability. Basically, (iii') implies certain degeneracy in the distribution of the error terms which, at the same time, implies a deterministic relationship between a linear combination of the observed scores and one of the true scores. Therefore, we argue that properties (i), (ii), (iii'), and (iv) lead to a more general definition of reliability and in what follows we will adopt them as such. The proof that R_Λ satisfies the properties (i), (ii), (iii'), and (iv) is given in Appendix A.3. The reason why we still consider R_Λ a useful tool in the study of reliability will become clear in the following sections.

Details on the estimation of R_Λ and the calculation of an asymptotic confidence interval for this measure can be found in Appendix B.3.

10.2 R_Λ: The Reliability of an Entire Sequence

To acquire a better insight into this new measure, as well as to better understand its relationship with the R_T coefficient, we will study its behavior in an important

special case: the random intercept model. Let us then start by assuming that model (5.1) holds with $b_i \sim N(0, \sigma_b^2)$ and $\boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i} = \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$. In Section 7.4 we have shown that, in this setting, $R_T = \dfrac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$, and does not depend on the number of time points p.

We will now also derive the expression for R_Λ under this model, where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ and $\boldsymbol{V} = \sigma^2 \boldsymbol{I} + \sigma_b^2 \boldsymbol{J}$. From Section 10.1 we know that

$$R_\Lambda = 1 - |\boldsymbol{\Sigma} \boldsymbol{V}^{-1}|.$$

Additionally, applying the identity $(a\boldsymbol{I}_p + b\boldsymbol{J}_p)^{-1} = \dfrac{1}{a}\left(\boldsymbol{I}_p - \dfrac{b}{a+pb}\boldsymbol{J}_p\right)$ for $a \neq 0$ and $a \neq -pb$, we obtain that $\boldsymbol{V}^{-1} = \dfrac{1}{\sigma^2}\left(\boldsymbol{I} - \dfrac{\sigma_b^2}{\sigma^2 + p\sigma_b^2}\boldsymbol{J}\right)$. Using the previous result we get

$$\begin{aligned} \boldsymbol{\Sigma} \boldsymbol{V}^{-1} &= \sigma^2 \boldsymbol{I} \frac{1}{\sigma^2}\left(\boldsymbol{I} - \frac{\sigma_b^2}{\sigma^2 + p\sigma_b^2}\boldsymbol{J}\right) \\ &= \boldsymbol{I} - \frac{\sigma_b^2}{\sigma^2 + p\sigma_b^2}\boldsymbol{J}. \end{aligned}$$

Using now that $|a\boldsymbol{I}_p + b\boldsymbol{J}_p| = a^{-1}(a + pb)$ we have

$$|\boldsymbol{\Sigma} \boldsymbol{V}^{-1}| = \left|\boldsymbol{I} - \frac{\sigma_b^2}{\sigma^2 + p\sigma_b^2}\boldsymbol{J}\right| = 1 - \frac{p\sigma_b^2}{\sigma^2 + p\sigma_b^2},$$

so that

$$\begin{aligned} R_\Lambda &= 1 - |\boldsymbol{\Sigma} \boldsymbol{V}^{-1}| = \frac{p\sigma_b^2}{p\sigma_b^2 + \sigma^2}, \\ R_\Lambda &= \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma^2}{p}}. \end{aligned} \tag{10.2}$$

This expression is very interesting from a theoretical as well as a practical point of view. First, let us note that (10.2) is similar to the Spearman-Brown prophecy prediction formula (Spearman 1910, Brown 1910) and it implies that reliability increases with the number of observations. A second important issue is that R_Λ goes to one as the number of time points goes to infinity. This shows that, unlike R_T, the R_Λ coefficient does not capture the average reliability but rather the reliability of the sequence as a whole. Increasing the number of measurements, we also increase the amount of useful information about the patient, even if it comes contaminated with

measurement error. Actually, (10.2) confirms an old clinical truth: the longer one follows a patient, the more reliable will be the conclusions about that patient that one can make.

The practical implications of this result are also appealing. We can study the number of repeated measurements needed to obtain a certain level of reliability R_Λ, which can be derived as

$$p = \frac{\sigma^2}{\sigma_b^2} \frac{R_\Lambda}{1 - R_\Lambda}.$$

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

Until now, we have used the random-intercept model to gain some insight into the meaning of the measure R_Λ. However, the assumptions on which this model is based will be too restrictive in many real applications. The following theorem extends the previous result to a totally general scenario and confirms our interpretation for this measure.

Theorem 2 *Let us assume that model (5.1) holds for a balanced study design in which p time points have been considered. Further, let us denote by $R_\Lambda(p)$ the corresponding value of the R_Λ coefficient in this setting. If q additional observations are taken, then the new value of R_Λ for the $p+q$ time points sequence satisfies that $R_\Lambda(p+q) \geq R_\Lambda(p)$.*

The theorem proves that increasing our information about the patients can only increase the reliability of our conclusions, a very plausible and appealing result. A proof of an equivalent result is provided in Appendix C.5.

It is important to point out that the usual approach followed to estimate reliability in a longitudinal framework is based on the calculation of the reliability at each time point separately (Tisak and Tisak 1996, Wiley and Wiley 1970, Raykov 2000). This typically leads to a function of reliability that changes over time. Note further that both, the R_T and the R_Λ coefficients, can also be calculated at each time point, leading again to a general function of reliabilities across time. However, they also offer a global measure of reliability that nicely complements their functions over time. We believe this is an important issue because having a global measure of reliability, valid under such a general scenario, can substantially facilitate the interpretation of the results when two or more scales are compared and can expedite the understanding of their

psychometric properties. It is intuitively clear that a single meaningful measure is much easier to analyze, understand and interpret, than several functions of changing reliabilities over time.

We believe that in a longitudinal framework, a measure as R_Λ might be more attractive than the classical reliability functions previously proposed. Indeed, the main objective of longitudinal studies is to get information from the entire profile and not to analyze each time point separately. The R_Λ coefficient quantifies precisely this, i.e., the reliability of the whole profile we have at hand.

10.3 The Relationship Between R_Λ and Ω

We will now investigate the link between the measure R_Λ and the Ω family that was introduced in Chapter 9. We have seen that every member of Ω is a weighted sum of the elements $\rho_j^2 = 1 - \lambda_j$. Actually, R_Λ can also be written as a function of these elements. Let us note first that from (9.2) and (9.3) we have

$$\begin{aligned} |\boldsymbol{\Sigma}| &= |\boldsymbol{Q}'||\boldsymbol{Q}||\boldsymbol{\Lambda}| = |\boldsymbol{Q}|^2|\boldsymbol{\Lambda}|, \\ |\boldsymbol{V}| &= |\boldsymbol{Q}|^2, \end{aligned}$$

and therefore

$$R_\Lambda = 1 - \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{V}|} = 1 - \prod_{j=1}^{p} \lambda_j = 1 - \prod_{j=1}^{p} (1 - \rho_j^2).$$

Let us further look at the relationship between the R_Λ coefficient and the elements $\theta \in \Omega$. If $w_j > 0$ and $\sum w_j = 1$ then

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

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

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

This final inequality shows that $\theta \leq R_\Lambda$ for all $\theta \in \Omega$ and therefore the R_Λ coefficient can be interpreted as an upper bound for the family. This result totally coincides

with our interpretation of R_Λ as a measure of reliability for the entire sequence and our interpretation of the Ω family as summary measures of "average" reliability.

Let us also look at the relationship between R_Λ and some special members of Ω that were considered in Chapter 9. Notice first that under the random intercept model, the expressions for R_Λ (10.2) and $\widetilde{\theta}_{\max}$ (9.7) are identical. We have seen that $\widetilde{\theta}_{\max}$, like R_Λ, is an upper bound of the Ω family. Additionally, and similarly to the R_Λ coefficient, $\widetilde{\theta}_{\max}$ does not satisfy the definition given in Chapter 7 but the more general version introduced in Section 10.1. Indeed, it is easy to show that $\widetilde{\theta}_{\max}$ does not satisfies (iii) but the most general (iii'). In general we have

$$\begin{aligned} R_\Lambda &= 1 - \prod_{j=1}^{p}(1-\rho_j^2) = 1 - (1-\rho_{(1)}^2) \prod_{\rho_j^2 \neq \rho_{(1)}^2} (1-\rho_j^2) \\ &\geq 1 - (1-\rho_{(1)}^2) = \widetilde{\theta}_{\max}, \end{aligned}$$

and therefore, $R_\Lambda \geq \widetilde{\theta}_{\max}$. The previous expression indicates that $\widetilde{\theta}_{\max}$ can be interpreted as an approximation of R_Λ when $\rho_j^2 \approx 0$ with $j \neq (1)$. In spite of the preceding inequality, $\widetilde{\theta}_{\max}$ and R_Λ are frequently close as illustrated in Figure 10.1, that gives the true values for both measures under the various simulation settings, described in Section 7.5.

We will further look at the relationship between R_Λ and R_p, the other special member of Ω. Combining (10.2) with (9.8) we can find the following functional relationship between R_p and R_Λ for the random intercept model

$$R_p = \frac{R_\Lambda}{p}. \tag{10.3}$$

Formula (10.3) helps us to clarify the interpretation of R_p. If R_Λ represents the reliability of an entire sequence with p time points, then R_p quantifies the "average" contribution of each time point to the total reliability of the sequence. Basically, the R_p coefficient can be interpreted as the "efficiency" with which the total reliability, quantified by R_Λ, is obtained.

To further explain this point let us notice that for the random intercept model the R_Λ coefficient increases with the number of time points and, in principle, any value of reliability can be achieved if a sufficiently long sequence is considered. Further, we have that $R_p \leq \frac{1}{p}$ and, therefore, the R_p coefficient is a decreasing function of p. If for certain sequence we have a low value of R_p, this will be an indication of a "poor/inefficient" scale that will require a long follow up to achieve a high value of

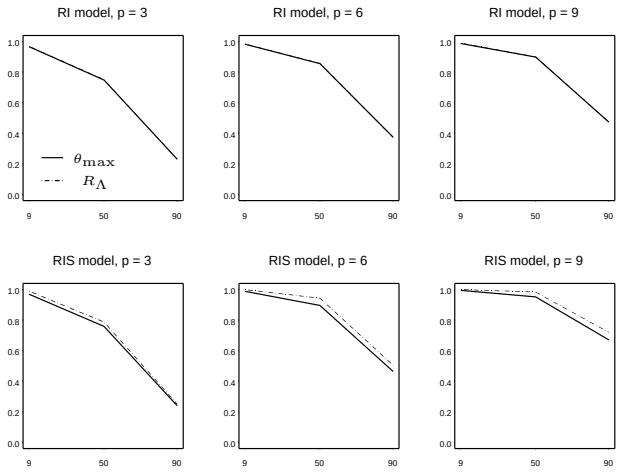

Figure 10.1: *Simulation study.* $\theta_{\max}$ *and* R_Λ *in case of 9, 50, and 90% of error variance, for the random intercept (RI) model (top) and random intercept and slope (RIS) model (bottom); and for 3 (left), 6 (middle), and 9 (right) repeated measurements.*

total reliability. That long follow up is the price we pay to obtain a high reliability with a poor instrument. On the other hand, if R_p is large, this will give evidence for a good instrument from which a few measurements are sufficient to obtain reliable results.

In what follows we will extend the functional relationship between R_p and R_Λ to a more general setting beyond the random intercept model. Let us recall that

$$R_\Lambda = 1 - \prod_{j=1}^{p} \lambda_j \tag{10.4}$$

$$\Rightarrow \qquad \log(1 - R_\Lambda) = \sum_{j=1}^{p} \log \lambda_j. \tag{10.5}$$

If $-1 < x < 1$ then the Maclauring series expansion for $\log(1 - x)$ is

$$-\log(1 - x) = x + \frac{x^2}{2} + \ldots + \frac{x^n}{n} + \ldots \tag{10.6}$$

Using the previous series expansion we get

$$-\log \lambda_j = -\log(1 - (1 - \lambda_j)) = 1 - \lambda_j + \frac{(1 - \lambda_j)^2}{2} + \ldots + \frac{(1 - \lambda_j)^n}{n} + \ldots$$

It has been shown in Appendix C.1 that $0 \leq \lambda_j \leq 1$, this implies that $0 \leq 1 - \lambda_j \leq 1$ and therefore $Q_j = \sum_{k=2}^{\infty} \frac{(1-\lambda_j)^k}{k} \geq 0$ for all j. We can then rewrite $-\log \lambda_j$ as

$$-\log \lambda_j = 1 - \lambda_j + Q_j \Rightarrow \log \lambda_j = \lambda_j - 1 - Q_j.$$

The previous equality leads to

$$\begin{aligned} \log(1 - R_\Lambda) &= \sum_{j=1}^{p} \lambda_j - p - \sum_{j=1}^{p} Q_j \\ &= -p \left(1 - \frac{1}{p} \sum_{j=1}^{p} \lambda_j \right) - \sum_{j=1}^{p} Q_j \\ &= -pR_p - M, \end{aligned}$$

where $M = \sum_{j=1}^{p} Q_j = \sum_{j=1}^{p} \sum_{k=2}^{\infty} \frac{(1-\lambda_j)^k}{k}$. Note that if M can be considered negligible then

$$R_\Lambda \approx 1 - e^{-pR_p}$$

and using again a first order of a series expansion we get

$$R_p \approx \frac{R_\Lambda}{p}. \tag{10.7}$$

This expression generalizes (10.3) for the random intercept model and confirms our previous interpretation for this measure of reliability.

10.4 A Simulation Study

We set up a simulation study to investigate the performance of the point estimator and asymptotic confidence interval for R_Λ under various conditions. For a detailed description of the simulation settings we refer to Section 7.5, on the simulation study for R_T. The same data were used for studying the R_Λ coefficient.

Table 10.1 presents the true values, estimated values, and the coverage probabilities for a 95% confidence interval for R_Λ, where the random intercept model (7.4) has been used as a data generating mechanism. Table 10.2 presents the results when the data were generated using the random intercept and slope model given in (7.5).

We first look at the point estimators. Like before, the estimator seems to work very well in almost all settings. Only when the measurement error accounts for 90% of

Table 10.1: *Simulation study on R_Λ: random intercept model (7.4). Effect of sample size (n), number of repeated measurements (p), and error percentage (%) on the estimate for R_Λ ($\hat{R}_\Lambda$), and the coverage probabilities (CP) for a 95% confidence interval.*

			$n = 50$		$n = 150$		$n = 300$	
%	p	R_Λ	$\hat{R}_\Lambda$	CP	$\hat{R}_\Lambda$	CP	$\hat{R}_\Lambda$	CP
9	3	0.968	0.967	98.0	0.968	97.6	0.968	97.3
9	6	0.984	0.983	99.4	0.984	99.8	0.984	99.8
9	9	0.989	0.989	99.8	0.989	100	0.989	100
50	3	0.750	0.753	95.8	0.754	96.8	0.754	98.6
50	6	0.857	0.855	97.6	0.858	97.8	0.857	98.0
50	9	0.900	0.898	97.4	0.900	98.8	0.900	98.6
90	3	0.231	0.290	87.9	0.245	93.7	0.242	94.6
90	6	0.375	0.383	94.4	0.384	96.6	0.380	97.4
90	9	0.474	0.469	96.2	0.476	95.2	0.477	96.0

all the variability, biased results can be obtained for small sample sizes. The coverage probabilities of the confidence intervals are also good in general. There are only a few instances where it is not in the neighborhood of 95%. This is mainly when the sample size is small in combination with large measurement error.

Let us now compare the performance of R_Λ and R_T. Table 7.1 clearly illustrated that the values of R_T, based on a random intercept model with homogeneous error variances, do not depend on the number of time points. This has also been shown theoretically in (7.3).

On the other hand, Table 10.1 illustrates that the values for R_Λ, under the same model, increase with the number of time points, a result that has also been derived theoretically in (10.2). For example, when the error variability is 50% of the total variability, we still obtain very high values for R_Λ in case 6 or 9 measurements are taken. This means that, when there is a lot of measurement error, still very reliable information can be obtained when the measurement is repeated a sufficient number of times. Even repeating the measurement three times, the combined information could be considered as reliable ($R_\Lambda = 0.79$). A similar result is found for R_Λ under the random intercept and random slope model (Table 10.2).

Table 10.2: *Simulation study on R_Λ: random intercept and slope model (7.5). Effect of sample size (n), number of repeated measurements (p), and error percentage (%) on the estimate for R_Λ ($\hat{R}_\Lambda$), and the coverage probabilities (CP) for a 95% confidence interval.*

			$n = 50$		$n = 150$		$n = 300$	
%	p	R_Λ	$\hat{R}_\Lambda$	CP	$\hat{R}_\Lambda$	CP	$\hat{R}_\Lambda$	CP
9	3	0.986	0.986	99.6	0.986	99.4	0.986	99.8
9	6	0.999	0.999	100	0.999	100	0.999	100
9	9	1	1	100	1	100	1	100
50	3	0.787	0.831	95.2	0.816	94.3	0.799	98.5
50	6	0.943	0.941	98.0	0.942	98.4	0.943	98.2
50	9	0.983	0.982	99.4	0.983	100	0.983	100
90	3	0.250	0.516	74.9	0.410	79.5	0.347	84.4
90	6	0.504	0.576	90.7	0.538	95.8	0.522	95.4
90	9	0.720	0.740	95.8	0.724	98.2	0.722	96.6

Further, even though Table 7.2 shows an increase of R_T for an increasing number of time points, in Section 7.4 we have seen that this does not need to be the case. To illustrate the difference in this respect between R_T and R_Λ we set up an additional simulation study.

We revisit the results of the simulations following the random intercept model for $p = 3$. We further generated data with one extra time point ($p = 4$), in such a way that the extra measurement satisfies

$$\frac{\text{tr}(\boldsymbol{\Sigma}_p)}{\text{tr}(\boldsymbol{\Sigma}_{Dp})} < \frac{\sigma^2_{p+1}}{\boldsymbol{z}_{p+1}\boldsymbol{D}\boldsymbol{z}'_{p+1}},$$

or equivalently, under the present model

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

thereby expecting R_T to decrease compared to the results displayed in Table 7.1. Precisely, the data were generated based on model (7.4), where $b_i \sim N(0, \sigma_b^2)$, $\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma})$, with $\sigma_b^2 = 300$, and with $\boldsymbol{\Sigma}$ a diagonal matrix with the first three diagonal

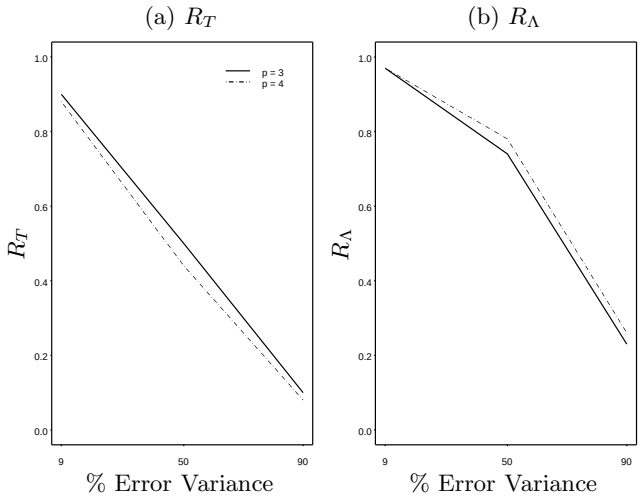

Figure 10.2: *Simulation study. Effect on R_T and R_Λ of adding an additional measurement that has a large error variability compared to three previous measurements.*

elements equal to σ^2 and the fourth diagonal element equal to $2\sigma^2$, with $\sigma^2 = 30$, 300, and 3000. The sample size was set equal to $n = 50$. Figure 10.2 summarizes our findings. We indeed observe that the values for R_T decrease with a larger number of time points. However, the values of R_Λ increase.

This simulation study illustrates once more that R_T and R_Λ should be interpreted in different ways. R_T is an average reliability, taken over a number of measurements. Adding time points with "low" reliability will pull the average down, adding "reliable" measurements will lift the average up. Unlike R_T, R_Λ quantifies the reliability of the whole sequence of measurements. Adding more measurements to the sequence will never decrease our total information about the true scores. Obviously, the magnitude of the increase will depend on the amount of measurement error that contaminates the new observations. Adding measurements with little measurement error will lead to a faster increase of R_Λ than what measurements with a lot of error would do.

10.5 Analysis of the Case Study

In this section we apply the methodology described in this and the previous chapter to the case study data in schizophrenia (Peuskens *et al* 1995). We will estimate the

Table 10.3: *Schizophrenia Study. Reliability estimates and 95% confidence intervals for PANSS, BPRS, and CGI based on three different reliability measures:* R_T, R_p, *and* R_Λ.

	PANSS	BPRS	CGI
R_T	0.890 [0.871; 0.907]	0.856 [0.839; 0.871]	0.733 [0.622; 0.822]
R_p	0.414 [0.381; 0.478]	0.366 [0.347; 0.385]	0.333 [0.270; 0.403]
R_Λ	0.999 [0.996; 1.000]	0.996 [0.995; 0.997]	0.988 [0.726; 1.000]

reliability measures R_Λ and R_p for the three rating scales used in that study, PANSS, BPRS, and CGI. For model building and model selection we refer to Chapter 8. Table 10.3 presents the point estimates and 95% confidence intervals for the R_Λ, R_p, and R_T coefficients.

First, let us look at the results obtained with the R_p coefficient. Earlier, we have argued that this coefficient is an indicator of efficiency, expressing the average contribution of each measurement to the total reliability. Notice that the estimates in Table 10.3 indicate that the approximation given in (10.7) may be imprecise when the model deviates from the random intercept case. However, this does not invalidate the general interpretation derived from this approximation and, therefore, it will be retained in the following discussions. Clearly, the most complex scale, PANSS, exhibits the largest efficiency, followed by BPRS and CGI. This finding is in full agreement with the results obtained earlier when the R_T coefficient was used in Chapter 8 and that are also summarized in Table 10.3.

Turning to the R_Λ coefficient the table shows very large estimates for the three scales, all are close to 1. We have previously interpreted the R_Λ coefficient as the reliability of the entire sequence of measurements, increasing each time an extra measurement is taken. The high values observed here are thus the result of two elements: first, the high average reliability, expressed by high estimates of R_T and second, the fair number of repeated measurements taken in this study. From this finding we can then conclude that all the instruments can provide very reliable information about the patients in the population studied in this clinical trial, and that the impact of measurement error is negligible for all three rating scales in this setting.

We also analyzed the increase of the R_Λ coefficient over the number of measure-

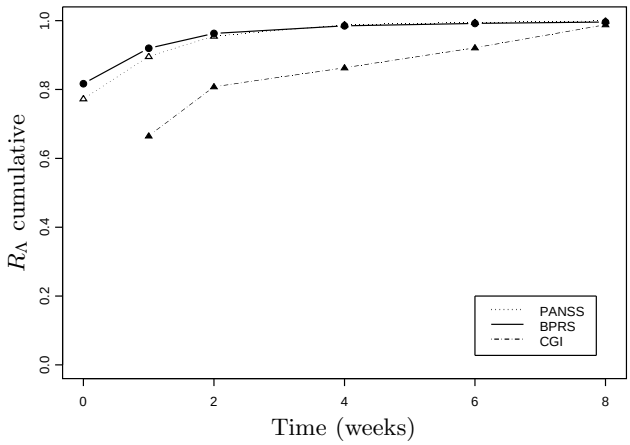

Figure 10.3: *Schizophrenia study.* R_Λ *cumulative over time for the three outcome scales.*

ments, as shown in Figure 10.3. The graph presents the cumulative R_Λ values over time, where the first point indicates the value of R_Λ based on the first measurement alone, the second point indicates the value based on the first and second measurement jointly, and so on. The figure shows clearly that PANSS and BPRS are indistinguishable with respect to their reliabilities. Notice that the same conclusion was drawn in Chapter 8, where we have seen that R_T estimates for both scales are similar. The graph further indicates that taking a second measurement with either of the two scales leads to an increase in reliability of about 10% compared to considering the first measurement alone. Nevertheless, after three measurement a further gain in reliability can hardly be achieved by further increasing the number of measurements.

Based on the same figure, we clearly observe that the CGI scale is less reliable than PANSS and BPRS, a conclusion that was also drawn in Chapter 8 based on the R_T coefficient. Indeed, for the latter two scales it takes only two measurements to arrive at R_Λ values around 0.90, whereas for CGI it takes five measurements to get to the same level. However, it is fair to say that at the end of the study CGI reaches the same level of reliability as the two multi-item scales. We could conclude that when a small number of measurements is taken, PANSS and BPRS are more reliable than CGI, but the difference in reliability fades away if, as in the present case study, a

sufficient number of of repeated measurements is taken.

10.6 Conclusion

In this chapter, we introduced a new parameter to evaluate reliability, the R_Λ coefficient. Mathematically, R_Λ and R_T are very similar, with the only difference being that for the former determinants instead of traces are used to summarize the variability in variance-covariance matrices. Interestingly, we have seen that this leads to a measure that bears a quite distinct interpretation. Unlike R_T, R_Λ quantifies the reliability of the complete sequence of observations. The R_Λ coefficient cannot decrease when the number of measurements goes up, illustrating that even scales with a relatively low average reliability can lead to reliable results if the follow up of the patients is long enough. This is a very important and encouraging result. Indeed, the strong subjective component of many rating scales will frequently produce relatively small or moderate values of reliability when they are administered once. Nevertheless, Theorem 2 shows that such an instrument can still be valuable if it is applied repeatedly over time.

Chapter 11

Connections with Earlier Approaches

In this chapter we will discuss some interesting links between the new approach to reliability, as elaborated in the chapters 7 to 10, and some earlier approaches developed in the framework of the classical test theory and generalizability theory.

11.1 Reliability as a Measure of Association Between True and Observed Scores

Correlation has been at the core of reliability research since the pioneering work of Charles Spearman at the beginning of the 20th century. In CTT it has been shown that reliability equals the squared correlation between the observed and true scores, as expressed in (3.3). Essentially, the reliability of a scale tries to quantify the amount of information that the instrument conveys about the latent, unobserved true scores. Therefore, this equivalence between reliability on one hand and the correlation between true and observed scores on the other hand, is very appealing. In this section we will explore this link further.

As pointed out in Chapter 10, the Ω family and R_Λ are both built based on the same basic elements, namely the $\rho_j^2 = 1 - \lambda_j$ where the λ_j's are the solutions of the equation $q(\lambda) = |\boldsymbol{\Sigma} - \lambda \boldsymbol{V}| = 0$. Nevertheless, from this definition the practical

interpretation of the ρ_j^2 is not totally clear. In this section we approach reliability from an alternative point of view that will help us to clarify the role and interpretation of the elements ρ_j^2.

As previously stated, in the classical test theory, Lord and Novick (1968) have proven that reliability equals the squared correlation between the observed score Y_i and the true score τ_i, i.e.,

$$R = \text{Corr}(Y_i, \tau_i)^2. \tag{11.1}$$

In the previous chapters we extended the classical definition to assess reliability in a more general setting where the steady state condition implied by model 3.1 does not hold and repeated measurements for each subject are available. Nevertheless, given the intuitive appeal of (11.1), it would be natural to explore whether such a connection also holds in the more general scenario considered in chapters 7 to 10. Therefore, in what follows, we will study the relationship between the measures of reliability previously introduced and the squared association between $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$. Let us start by denoting $\boldsymbol{S}_i = \begin{pmatrix} \boldsymbol{Y}_i \\ \boldsymbol{b}_i \end{pmatrix}$. The following theorem will allow us to quantify this association.

Theorem 3 *If model 5.1 holds then* $\boldsymbol{S}_i \sim N(\boldsymbol{\mu}_{0i}, \boldsymbol{\Sigma}_{0i})$ *where*

$$\boldsymbol{\mu}_{0i} = \begin{pmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ \boldsymbol{0} \end{pmatrix} \quad \textit{and} \quad \boldsymbol{\Sigma}_{0i} = \begin{pmatrix} \boldsymbol{V}_i & \boldsymbol{Z}_i\boldsymbol{D} \\ (\boldsymbol{Z}_i\boldsymbol{D})' & \boldsymbol{D} \end{pmatrix}.$$

Proof

Let us first prove that $\boldsymbol{S}_i$ follows a normal distribution. In what follows we will use the following result from Johnson and Wichern (1998, p. 165)

> A random vector $\boldsymbol{X}$ is multivariate normal distributed, if and only if, for any vector $\boldsymbol{a}$ ($\boldsymbol{a} \neq \boldsymbol{0}$), $\boldsymbol{a}'\boldsymbol{X}$ is univariate normal distributed.

We will now consider the general vector $\boldsymbol{a}' = (\boldsymbol{a}_1', \boldsymbol{a}_2')$, where $\boldsymbol{a}_1 \in \mathbb{R}^{p_i}$ and $\boldsymbol{a}_2 \in \mathbb{R}^q$. We want to prove that $\boldsymbol{a}'\boldsymbol{S}_i$ is univariate normal distributed.

$$\begin{aligned} \boldsymbol{a}'\boldsymbol{S}_i &= \boldsymbol{a}_1'\boldsymbol{Y}_i + \boldsymbol{a}_2'\boldsymbol{b}_i = \boldsymbol{a}_1'(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i) + \boldsymbol{a}_2'\boldsymbol{b}_i \\ &= \boldsymbol{a}_1'\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{a}_1'\boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{a}_1'\boldsymbol{\varepsilon}_i + \boldsymbol{a}_2'\boldsymbol{b}_i \\ &= \boldsymbol{a}_1'\boldsymbol{X}_i\boldsymbol{\beta} + (\boldsymbol{a}_1'\boldsymbol{Z}_i + \boldsymbol{a}_2')\boldsymbol{b}_i + \boldsymbol{a}_1'\boldsymbol{\varepsilon}_i. \end{aligned}$$

However

$$\begin{pmatrix} \boldsymbol{b}_i \\ \boldsymbol{\varepsilon}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{D} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_i \end{pmatrix}\right)$$

and, therefore, using again the result from Johnson and Wichern (1998, p. 165) we can conclude that $(\boldsymbol{a}_1'\boldsymbol{Z}_i + \boldsymbol{a}_2')\boldsymbol{b}_i + \boldsymbol{a}_1'\boldsymbol{\varepsilon}_i$ is univariate normal distributed. We have proven that for any $\boldsymbol{a}$ ($\boldsymbol{a} \neq \mathbf{0}$), $\boldsymbol{a}'\boldsymbol{S}_i$ is univariate normal distributed and, as a consequence, $\boldsymbol{S}_i$ is also multivariate normal distributed, with

$$\boldsymbol{\mu}_{0i} = E\left[\begin{pmatrix} \boldsymbol{Y}_i \\ \boldsymbol{b}_i \end{pmatrix}\right] = \begin{pmatrix} E(\boldsymbol{Y}_i) \\ E(\boldsymbol{b}_i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_{0i} = \begin{pmatrix} \mathrm{Cov}(\boldsymbol{Y}_i) & \mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{b}_i) \\ \mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{b}_i)' & \mathrm{Cov}(\boldsymbol{b}_i) \end{pmatrix}.$$

We know that $\mathrm{Cov}(\boldsymbol{Y}_i) = \boldsymbol{V}_i$ and $\mathrm{Cov}(\boldsymbol{b}_i) = \boldsymbol{D}$. Let us now calculate $\mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{b}_i)$.

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{Y}_i, \boldsymbol{b}_i) &= E\{(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{b}_i'\} \\ &= E\{(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i - \boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{b}_i'\} \\ &= E\{\boldsymbol{Z}_i\boldsymbol{b}_i\boldsymbol{b}_i' + \boldsymbol{\varepsilon}_i\boldsymbol{b}_i'\} \\ &= E\{\boldsymbol{Z}_i\boldsymbol{b}_i\boldsymbol{b_i}'\} + E\{\boldsymbol{\varepsilon}_i\boldsymbol{b}_i'\} = \boldsymbol{Z}_i\boldsymbol{D} \end{aligned}$$

The last equality comes from the fact that $E\{\boldsymbol{\varepsilon}_i\boldsymbol{b}_i'\} = \mathbf{0}$. Indeed, $(\boldsymbol{\varepsilon}_i, \boldsymbol{b}_i')$ are independent and $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$, $E(\boldsymbol{b}_i) = \mathbf{0}$ what implies that $E\{\boldsymbol{\varepsilon}_i\boldsymbol{b}_i'\} = \mathbf{0}$. Finally, we get

$\boldsymbol{\Sigma}_{0i} = \begin{pmatrix} \boldsymbol{V}_i & \boldsymbol{Z}_i\boldsymbol{D} \\ (\boldsymbol{Z}_i\boldsymbol{D})' & \boldsymbol{D} \end{pmatrix}$ and therefore $\boldsymbol{S}_i \sim N(\boldsymbol{\mu}_{0i}, \boldsymbol{\Sigma}_{0i})$. □

Theorem 3 states that $\boldsymbol{S}_i$ is multivariate normal distributed. A natural way to quantify the association between $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$ is then to use canonical correlations. From multivariate analysis (Johnson & Wichern 1998) we know that if

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right),$$

then we can quantify the association between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ through the set of canonical correlations which are the eigenvalues of the matrix $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$. If we

now consider the case of a single balanced clinical trial, then the matrix $\boldsymbol{\Sigma}_{0i} = \boldsymbol{\Sigma}_0$ takes the form

$$\boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{V} & \boldsymbol{ZD} \\ (\boldsymbol{ZD})' & \boldsymbol{D} \end{pmatrix},$$

where $\boldsymbol{V} = \boldsymbol{ZDZ}' + \boldsymbol{\Sigma}$. The canonical correlations of $\boldsymbol{S}_i$ are then the eigenvalues of the matrix $\boldsymbol{V}^{-1/2}\boldsymbol{ZDD}^{-1}\boldsymbol{DZ}'\boldsymbol{V}^{-1/2} = \boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}_D\boldsymbol{V}^{-1/2} = \boldsymbol{V}^{-1/2}(\boldsymbol{V} - \boldsymbol{\Sigma})\boldsymbol{V}^{-1/2} = \boldsymbol{I} - \boldsymbol{H}$. We have already proven that the eigenvalues of $\boldsymbol{H}$ are the solutions of the equation $q(\lambda) = |\boldsymbol{\Sigma} - \lambda\boldsymbol{V}| = 0$. On the other hand, it is easy to show that if λ is an eigenvalue of the matrix $\boldsymbol{H}$ then $1 - \lambda$ is an eigenvalue of the matrix $\boldsymbol{I} - \boldsymbol{H}$. The implications of these results are very appealing. In fact, if we want to extend the concept of reliability using the expression (11.1), then the reliability of $\boldsymbol{Y}_i$ should be based on the canonical correlations associated with $\boldsymbol{S}_i$. The previous results show that the canonical correlations between $\boldsymbol{Y}_i$ and $\boldsymbol{b}_i$ equal $1 - \lambda_j$ where λ_j are the solutions of the equation $q(\lambda) = |\boldsymbol{\Sigma} - \lambda\boldsymbol{V}| = 0$. Note that in the definition of the Ω family (9.1), the elements $\rho_j^2 = 1 - \lambda_j$ are just these canonical correlations.

It is appealing to see that two equivalent classical definitions of reliability also concur in this extended setting. These results clearly show that any extension of the classical definition of reliability that wants to retain the interpretation derived from (11.1), should necessarily be based on the ρ_j^2. However, a high-dimensional vector of canonical correlations may be difficult to interpret and difficult to use when comparing two scales regarding their reliabilities. Therefore, aiming at an easier interpretation, we have summarized the information about the reliability, contained in the canonical correlation vector, using meaningful functions of its elements.

Furthermore, with this new interpretation, the Ω family is in a stronger agreement with the similar family introduced by Alonso *et al* (2004) to study criterion validity. In the context of criterion validity the ρ_j^2 are canonical correlations between two rating scales, in the context of reliability they are canonical correlations between true and observed scores.

11.2 Relationship Between the New Proposals and the G Coefficients

One of the most important attempts to estimate reliability in a longitudinal framework is based on G-theory and the use of the G coefficients. In this section, we will study

the relationship between R_T, R_Λ, and the G coefficients when the assumption of the G-theory modelling framework are met.

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

$$Y_{ij} = \mu + b_i + \tau_j + \varepsilon_{ij}, \tag{11.2}$$

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

Note that, using vector notation, model (11.2) can be rewritten as

$$\boldsymbol{Y}_i = \mathbf{1}_p \mu + \mathbf{1}_p b_i + \boldsymbol{\tau} + \boldsymbol{\varepsilon}_i, \tag{11.3}$$

where $\boldsymbol{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$ denotes a column vector with all observations originating from subject i, $\mathbf{1}_p = (1, 1, \ldots, 1)'$ denotes a p-dimensional column vector, $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_p)'$ a column vector with the time effects, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{ip})'$ denotes the column vector with all the error terms associated with subject i. This model can be seen as a special case of the linear mixed model we considered. Indeed, model (11.3) is a linear mixed model with only one subject-specific random effect and the error structure decomposed into a time component (which can be seen as a special type of serial correlation where the $\boldsymbol{H}_i$ matrix reduces to the identity), and a component that captures extra residual variability. As we have stated before, it is important to differentiate the variability emanating from the subject-specific random effects and the one coming from other sources. In this case, we have only one subject-specific random effect b_i and, therefore, for this model the variance-covariance matrix associated with the subject-specific random effects is a scalar; $\boldsymbol{D} = \sigma_b^2$. Using matrix notation, we can now write

$$\boldsymbol{V} = \mathrm{Var}(\boldsymbol{Y}_i) = \boldsymbol{J}_p \sigma_b^2 + \boldsymbol{I}_p(\sigma_\tau^2 + \sigma^2), \tag{11.4}$$

where $\boldsymbol{J}_p = \mathbf{1}_p \mathbf{1}_p'$ and $\boldsymbol{I}_p$ is a $p \times p$ identity matrix. Employing the notation introduced in Chapter 5, we have $\boldsymbol{V} = \boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma}_D = \boldsymbol{J}_p \sigma_b^2$ accounting for the variability coming from the subject-specific effect and $\boldsymbol{\Sigma} = \boldsymbol{I}_p(\sigma_\tau^2 + \sigma^2)$ accounting for the remaining variability. It now follows that

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

Interestingly, this expression equals the index of dependability Φ as expressed in (4.4), in case there is only one measurement, i.e., $p' = 1$. The index of dependability is one of the two reliability-like coefficients widely used in G-theory, and is appropriate when scores are given "absolute" interpretations as in domain-referenced or criterion-referenced situations. In the above, p' refers to the number of time points in a D-study. The latter can be seen as a mind experiment to study the reliability that would be obtained under different circumstances. The R_T coefficient thus equals the expected reliability if we would take only one measurement. This interpretation nicely corresponds to the previously given interpretation for R_T as the average reliability over the time points.

To calculate the value of the R_Λ coefficient we will need the following result: $|a\boldsymbol{I}_p + b\boldsymbol{J}_p| = a^{p-1}(a + pb)$ (Searle 1982). We can now write

$$\begin{aligned} R_\Lambda &= 1 - \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{V}|} = 1 - \frac{(\sigma_\tau^2 + \sigma^2)^p}{(\sigma_\tau^2 + \sigma^2)^{p-1}(p\sigma_b^2 + \sigma_\tau^2 + \sigma^2)} \\ &= 1 - \frac{\sigma_\tau^2 + \sigma^2}{p\sigma_b^2 + \sigma_\tau^2 + \sigma^2} = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\sigma_\tau^2}{p} + \frac{\sigma^2}{p}}. \end{aligned}$$

Note that this expression equals the index of dependability Φ, but now for $p' = p$. The R_Λ coefficient thus equals the reliability that is obtained with the number of measurements equal to the number used in the G-study. However, at first sight, this equivalence seems to disagree with the interpretation of R_Λ as the reliability of the entire longitudinal sequence. Indeed, as stated in Chapter 4, G-theory typically uses the mean score metric. The index Φ therefore expresses the reliability of the mean of the observed scores. On the other hand, we have previously stated that the R_Λ coefficient expresses the reliability of the entire sequence of ratings, this means: the reliability of the vector of observed scores. The question is then raised as to how both coefficients can coincide, if a vector is supposed to contain more information than a mean. To solve this issue, let us define $\varepsilon_{ij}^* = \tau_j + \varepsilon_{ij}$, so that model (11.2) can be written as

$$Y_{ij} = \mu + b_i + \varepsilon_{ij}^*, \tag{11.5}$$

where $\varepsilon_{ij}^* \sim N(0, \sigma_\tau^2 + \sigma^2)$ and $Y_{ij}|b_i \sim N(\mu + b_i, \sigma_\tau^2 + \sigma^2)$. In this case, however, $\bar{Y}_{i\cdot}$ is a sufficient statistic for $\mu + b_i$. In other words, $\bar{Y}_{i\cdot}$ contains the same amount of information about $\mu + b_i$ as does the entire vector $\boldsymbol{Y}_i$. Furthermore, when $p \to \infty$ then $\bar{Y}_{i\cdot} \xrightarrow{P} \mu + b_i$ and $R_\Lambda \to 1$. This explains why, in case model (11.2) holds, the reliability of the mean equals the reliability of the entire vector of measurements.

The index of dependability Φ is applicable when scores are given "absolute" interpretations. A very similar proof can be constructed to illustrate that, after conditioning on the time points, the R_T and the R_Λ coefficients equal the generalizability coefficient $\mathrm{E}\rho^2$, as expressed in (4.3), for 1 and p measurements, correspondingly. $\mathrm{E}\rho^2$ is a measure for reliability that is applicable when only relative decisions ought to be taken. This is when we are only interested in the relative position of different individuals with respect to each other, and not in the absolute values they score on the scale. Nevertheless, within the context of mental health or health in general, the absolute interpretation is usually more useful and appealing.

The previous derivations show that, when the modelling assumptions of G-theory hold, the commonly used G-theory coefficients coincide with the measures of reliability previously proposed. This implies that these G-coefficients also satisfy our defining properties and can be framed within the present approach. Given the seminal success of G-theory in many applications, these results increase our confidence in the newly proposed reliability definition and measures.

Nevertheless, as stated in Chapter 5, the assumptions required by the G-theory modelling framework are often too restrictive to be applicable in a longitudinal scenario. If these assumption are violated then severe bias can appear in our estimates. The next chapter retakes this issue and explores the effect of some of these violations on the new proposals.

Chapter 12

Impact of Ignoring Serial Correlation and Memory Effect on Reliability Estimates

Test-retest studies are one of the most commonly used methods to evaluate reliability. In these studies subjects are tested on two different occasions, and the Pearson correlation or the intraclass correlation coefficient is used as a measure for reliability. This method is valid under the assumptions of the classical test theory, i.e., (i) the true scores are equal; (ii) the error variances are equal; and (iii) the measurement errors are independent. Clearly, a test-retest scheme can be seen as the simplest possible longitudinal design.

However, 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 probable that the characteristics being measured will 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. Nevertheless, if both measurements are taken too close in time, it is quite likely that the rater will recall his/her previous rating and the new assessment could be influenced by the previous one. Usually, the rater will give similar ratings in each of the replications (Dunn 1989, Streiner and Norman 1995).

The problem of memory also appears when we want to study reliability in a more general longitudinal setting, i.e., when subjects are measured at more than two occasions using the same rating scale. If a memory effect emerges in such a setting, then it will imply that observations closer in time are more alike than observations further apart. Basically, this is the effect produced by a serial correlation component, a term used to capture exactly this type of effect in the association structure (Verbeke and 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 chapter, we study via simulations the bias produced by such uncontrolled sources of serial correlation, when employing the reliability coefficients proposed in previous chapters. This study complements previous research that has reported the effect of ignoring intra-subject serial correlation on the G-coefficients within a generalizability theory framework.

12.1 Ignoring Intra-subject Serial Correlation

As stated before, an important attempt to extending the concept of reliability to a longitudinal setting was done using generalizability theory. The utility of G-theory to evaluate reliability in longitudinal studies depends on the adequacy of its underlying model (analysis of variance with random effects) to describe the specific data structure encountered. As has been mentioned in Chapter 5, the G-theory modelling framework can be applied to a longitudinal setting only if strong and unrealistic assumptions are made. One such assumption is the presence of an uncorrelated and homoscedastic error structure. In fact, correlated error structures occur frequently in longitudinal studies. Usually, observations close in time exhibit a stronger association than observations with more time separation. Ignoring this correlation will induce bias in the variance-component estimates and, as a result, in the generalizability coefficients. This has been documented in the literature and a detailed description of these works was presented in Chapter 5. Unfortunately, the classic modelling paradigm used in

G-theory is not designed to capture this type of associations and assumes uncorrelated error terms with equal variance over time.

In previous chapters we proposed a different extension of the reliability concept to a longitudinal framework that is based on hierarchical linear models. This type of models allow to incorporate many of the features of longitudinal data, including varying true scores, correlated random effects, heteroscedastic error components, and correlated error terms. Additionally, the LMM framework conveniently offers a large amount of flexibility for modelling serial correlation. For instance, Gaussian or exponential structures could be used 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.

As stated before, 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 correlation 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 irrelevant from a reliability perspective, because what really matters is the fact that a serial correlation component 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.

In the next section, we will study the impact of ignored sources of serial correlation on the R_T and R_Λ coefficients.

12.2 A Simulation Study

The design of the simulation study was a $2\times3\times2$ complete factorial arrangement with: 2 types of subject specific profiles, (1) random intercept, and (2) random intercept and random slope; 3 levels of auto-regressive serial correlation 0.1, 0.5, and 0.8; and two types of analyses (1) ignoring serial correlation and (2) fitting serial correlation.

The random-intercept model can be expressed as

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 Z_i + b_i + \varepsilon_{ij}, \tag{12.1}$$

Table 12.1: *Instability and serial correlation on reliability measures: correlation coefficients. RI refers to random intercept model with serial correlation (12.1), RIS refers to model (12.1) with random intercept, random slope, 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

with $b_i \sim N(0, \sigma_b^2)$, $\boldsymbol{\varepsilon}_i \sim N(0, \tau^2 \boldsymbol{H})$, t_{ij} denoting the time at which measurement j for subject i is taken, and Z_i the treatment allocation for subject i. We fix $\sigma_b^2 = 300$ 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 can be written as

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 Z_i + b_{1i} + b_{2i} t_{ij} + \varepsilon_{ij}, \tag{12.2}$$

where now $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{H})$ and

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

Values for the fixed effects were set to $\beta_0 = 85$, $\beta_1 = -2.5$, and $\beta_2 = 3$. 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.

Before going to the actual analysis we will illustrate the effect of instability (patient evolution over time) and serial correlation on ordinary reliability estimates, calculated as test-retest correlations. Table 12.1 presents Pearson correlations between the outcome at the first measurement (Y_{i0}) and the outcomes at later measurement occasions ($Y_{i2} - Y_{i10}$), for different strengths of serial correlation (ρ). For the random intercept

model (RI) one can easily obtain the reliability of the measurement as the ratio of the true score variability to the total variability

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

Note that in this model the subject-specific evolution over the various measurements is constant and, therefore, it does not influence the correlation. 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 12.1 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, which equals 0.826. For this model, the subject-specific evolution is no longer constant: 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 whenever subjects evolve differently 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 considered two different scenarios for analysis: (i) a correctly specified model that includes a serial correlation component with an auto-regressive structure and (ii) a misspecified model that assumes an uncorrelated structure for the residual part, i.e., $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. Based on these model fits, we calculated the point estimates and confidence intervals for R_T and R_Λ. Tables 12.2 and 12.3 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. An asterisk indicates that the 95% confidence interval around the coverage percentage does not contain the true value of 0.95. The number of simulations ensures that the width of these confidence intervals is smaller than 0.10 in the expected range. We further have a power of over 80% to detect a difference of 0.05 in the coverage probabilities resulting from the two analysis methods.

Let us first focus on the random-intercept setting, i.e., when the data where generated using model (12.1). The first half of table 12.2 illustrates that, when the model used to fit the data does not include a serial correlation component, both $\widehat{R}_T$ and $\widehat{R}_\Lambda$ 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 is also more seriously affected compared to R_T. Unsurprisingly but with important ramifications, the situation worsens considerably for larger values of serial correlation.

Note that 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 modelling framework used in G-theory, linear mixed models allow for the

Table 12.2: *Effect of ignoring intra-subject correlation on reliability measures: random intercept model (12.1). 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

(*) the 95% confidence interval around the CP does not contain 0.95.

absorption of such a correlation structure. The second part of Table 12.2 shows the results obtained when the models fitted to the data included a serial correlation component. As one would expect, neither the R_T nor the R_Λ point estimates are biased in this case. Furthermore, the confidence intervals now enjoy coverage very close to their nominal level.

Interestingly, the true value of R_Λ decreases when the serial correlation increases, 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. 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 12.3 displays the results obtained under the second setting, i.e., when the data where generated from model (12.2). The conclusions in this case are almost identical to the 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

Table 12.3: *Effect of ignoring intra-subject correlation on reliability measures: random intercepts and slope model (12.2). 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	0.826	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

(*) the 95% confidence interval around the CP does not contain 0.95.

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 was increased to 500 patients.

12.3 Conclusion

The conclusions of the simulation study 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 the R_T and R_Λ coefficients. However, the more general modelling 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 more appropriate choice for the evaluation of reliability

in a longitudinal scenario.

We put a strong focus on the problem of memory effect. In case 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 such an effect using the serial correlation structure of a linear mixed model. The reason we chose to emphasize memory effect is because it has permeated reliability research for a long 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 mainly effective when solely two repeated measurements per subject are taken. In the present work, we approached the problem from a statistical modelling perspective by considering more general hierarchical models that can account for both, the time evolution of the patients and a potential memory effect.

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 always be taken into account.

Chapter 13

Reliability of Outcome Scales in a Depression Trial

We will now apply the methodology introduced in previous chapters to the depression case study, presented in Chapter 2. Basically, we will investigate the reliability of the three rating scales used in this study: the Hamilton Depression Rating Scale (HAMD), the Hamilton Anxiety Rating Scale (HAMA), and the Montgomery-Åsberg Depression Rating Scale (MADRS). The case study contains two identical clinical trials to investigate drug effectiveness in major depressive disorder. A general presentation of the data in trial 1 can be seen in the first row of Figure 13.1, where the individual profiles for each scale are displayed. Reliability will be investigated for both trials separately.

Section 13.1 gives a general outline of the model building exercise that was carried out to find the best fitting models for the observed data. In Section 13.2 we present and discuss the results of the reliability estimation for the different scales.

13.1 Model Building

Because interest primarily lies in the covariance structure, a complex mean model is adopted to avoid bias in the estimation of the variance components (Diggle, Liang and Zeger 1994). We considered a mean structure including time categorically, treatment, investigator, and the interaction between treatment and time. Regarding the random

Table 13.1: *Depression study. Selected models for the three scales, HAMD, MADRS, and HAMA, separately for the two trials.*

	Scale	Random effects structure	Structure of $\mathbf{\Sigma}$
Trial 1	HAMD	linear slope	heterogeneous autoregressive
	MADRS	linear slope	heterogeneous autoregressive
	HAMA	linear slope	banded unstructured
Trial 2	HAMD	quadratic slope	heterogeneous autoregressive
	MADRS	quadratic slope	heterogeneous autoregressive
	HAMA	quadratic slope	autoregressive

effects we considered models with: (a) subject-specific intercept; (b) subject-specific intercept and linear slope over time; and (c) subject-specific intercept and quadratic slope. For the covariance matrix of the error terms, $\mathbf{\Sigma}$, we considered five structures that allow correlation, and two structures that do not allow for such a correlation. 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. The latter structure further assumes heterogeneity of the error variances, whereas the structures (a)–(e) 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. Akaike's information criterion (AIC) was applied for selecting the best fitting model and parameter estimation was based on the restricted maximum likelihood method (REML). Table 13.1 summarizes the structure of the final models obtained for the three scales in each trial. Models selected for the first trial's data encompass a linear subject-specific time trend, the models for the second trial all include a quadratic term, indicating that individual subject profiles tend to be curved. All models further include an error variance-covariance structure $\mathbf{\Sigma}$ that allows correlated errors terms. Given the fact that the measurements are not

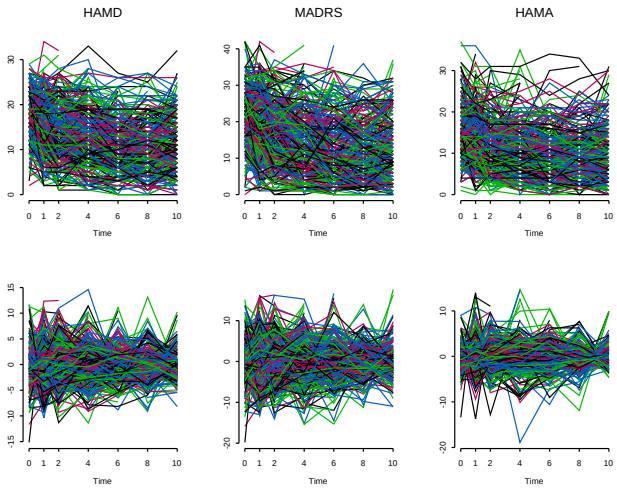

Figure 13.1: *Depression study. Individual patient profiles for three rating scales: observed (top) and residual (bottom) profiles.*

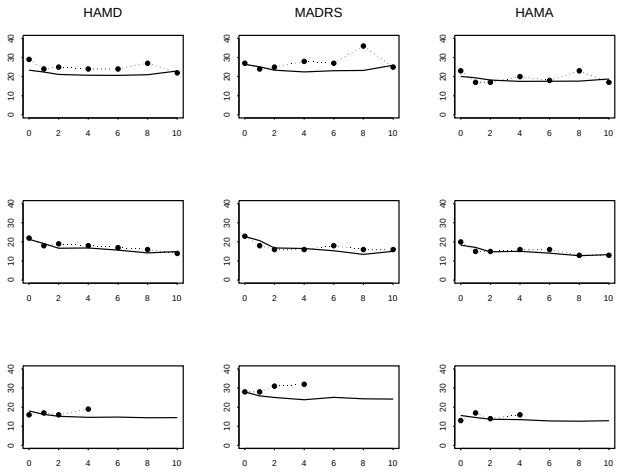

Figure 13.2: *Depression study. Individual observed profiles (dots) and fitted profiles (solid line) for three randomly selected patients.*

Table 13.2: *Depression study. Estimates of R_T and R_Λ with 95 % confidence intervals for the three outcome scales, HAMD, MADRS, and HAMA, separeately for the two trials.*

	Scale	R_T	CI_{R_T}	R_Λ	CI_{R_Λ}
Trial 1	HAMD	0.493	[0.405; 0.581]	0.829	[0.734; 0.895]
	MADRS	0.474	[0.378; 0.571]	0.812	[0.704; 0.886]
	HAMA	0.612	[0.545; 0.676]	0.955	[0.897; 0.980]
Trial 2	HAMD	0.629	[0.513; 0.731]	0.932	[0.872; 0.966]
	MADRS	0.692	[0.603; 0.769]	0.977	[0.957; 0.988]
	HAMA	0.675	[0.601; 0.741]	0.964	[0.930; 0.986]

entirely equally spaced, it is a bit surprising that an autoregressive structure leads to a better model fit than the spacial structures. This indicates most likely that a difference in time lag of one week does not influence the error correlation too much. Further we find that all but one of the selected structures include unequal diagonal elements, indicating heterogeneous error variances for the different time points. The second row of Figure 13.1 shows the residual patient profiles for the three scales, resulting from the best fitting models in trial 1. No systematic pattern seems to emerge from the graphs, indicating that the models capture the most important data features reasonably well. Further, Figure 13.2 plots the predicted and observed values for three randomly chosen patients in trial 1. Here again, a reasonable agreement between the models and the data is observed, reinforcing our confidence in the results of the model building step. Similar results (not shown) were found for trial 2.

Once sufficiently adequate models have been selected, reliability can be estimated using the variance components estimates emanating from these models.

13.2 Reliability Estimation

Reliability estimates are obtained separately for both clinical trials. The general results are presented in Table 13.2, estimates per time point are plotted in Figure 13.3. Let us first compare the HAMD and MADRS depression scales. For the two different trials, the graphs at the top of Figure 13.3 show the R_T values for these scales at

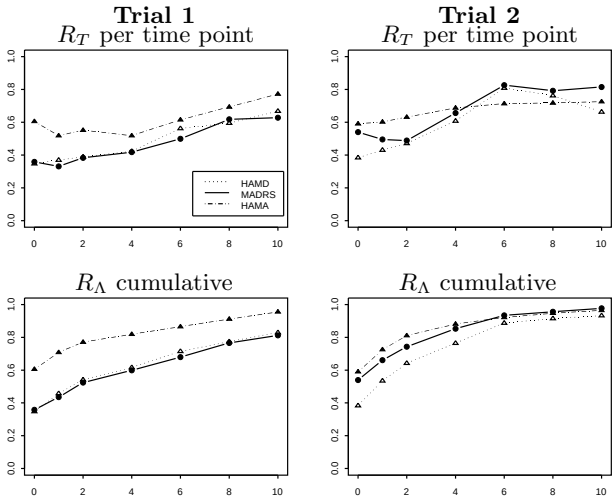

Figure 13.3: *R_T per time point and R_Λ cumulative over time points.*

each time point. These graphs illustrate that both scales perform rather poorly at the beginning of the trials. However, we can see that in both studies, the R_T values increase with time. For trial 1 we observe a gradual increase, whereas in trial 2 the increase is more abrupt. Arguably, such an increase could have been induced by a learning effect of the raters, stemming from gaining experience and/or enhanced familiarity with the patients during follow-up.

To compare the two scales, it is also useful to look at the general R_T values (Table 13.2) that give the average reliability over the different time points. Interestingly, regarding the point estimates in the first trial, HAMD performs slightly better than MADRS, whereas in trial 2 the opposite behavior is observed. Irrespective of these small differences in the point estimates, Table 13.2 reveals that the confidence intervals for R_T of the two scales largely overlap in both trials. Clearly, based on the present data, we encounter no evidence that MADRS is a more reliable scale than HAMD. Taking into account that MADRS was created to address some of the limitations of HAMD, this finding is somehow unexpected. However, similar results were found by Maier *et al* (1988) for inter-rater reliabilities. They compared HAMD and MADRS based on three different studies, but did not find differences in reliabilities in any of them.

It can further be noted 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 take higher values in more heterogenous 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 arms of 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 (1986) 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 were conducted in a patient segment 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. The simulation study in Chapter 12 showed that ignoring this type of correlation can lead to a serious overestimation of the reliability parameters, what may also explain the higher values of reliability reported in the literature.

Let us now turn to the second reliability measure, R_Λ, quantifying the reliability of the *accumulated* observations. As stated in previous chapters, when we measure

a patients once, we obtain a certain amount of information. By measuring a second time, we can only increase the amount of information on the patient even if it comes contaminated by measurement error. This accumulation of valuable information is nicely captured by R_Λ. The lower half of Figure 13.3 shows the cumulative R_Λ values over the different time points. At the first time point, the R_Λ coefficient expresses the reliability of the first measurement, which is equal to the R_T coefficient at the first time point. At the second time point, the R_Λ coefficient captures the reliability of the information contained in the first and the second measurement combined, and so on. The values shown in Table 13.2 present the results for the entire study, capturing the reliability of the whole sequence of observations. The R_Λ coefficient illustrates that, whenever a scale has low reliability, reliable results can still be obtained when the scale is applied repeatedly over time and the repeated outcomes are considered together. Obviously, the lower the reliability of the scale at each time point, the more measurements will be needed to obtain a pre-specified degree of cumulative reliability. Figure 13.3 shows that, in the first trial, a value of 0.80 was reached only at the last measurement. In the second trial 5 and 4 measurements, respectively, were needed to reach the same level of reliability for HAMD and MADRS.

While in the first trial, the cumulative evolutions of R_Λ are very similar for both depression scales, a better performance is observed for MADRS compared to HAMD at the beginning of the second trial. The relatively high reliability for MADRS at the first time point gives this scale a head start. Towards the end of the trial, HAMD has caught up with MADRS, leading to a small difference in the final R_Λ values, as shown in Table 13.2.

To find out whether, in the second trial, the R_Λ's for MADRS and HAMD differ significantly at the beginning of the study, we plot the 95% confidence bands for the cumulative R_Λ values for both scales, as shown in Figure 13.4. The figure shows wide confidence intervals for the earlier time points, while they get narrower towards the end of the study, when more information becomes available. The intervals for the two scales overlap at any of the time points. Hence, we do not find evidence of MADRS being a more reliable scale than HAMD, or vice versa.

Let us finally look at the results for HAMA. This particular scale measures anxiety and should therefore not be compared directly to the two depression scales. Table 13.2 shows somewhat better reliabilities in the second trial, which is in agreement with earlier findings. However, the differences are not too large. The average reliabilities, R_T, are 0.61 and 0.68, respectively, indicating a decent, however not excellent, reli-

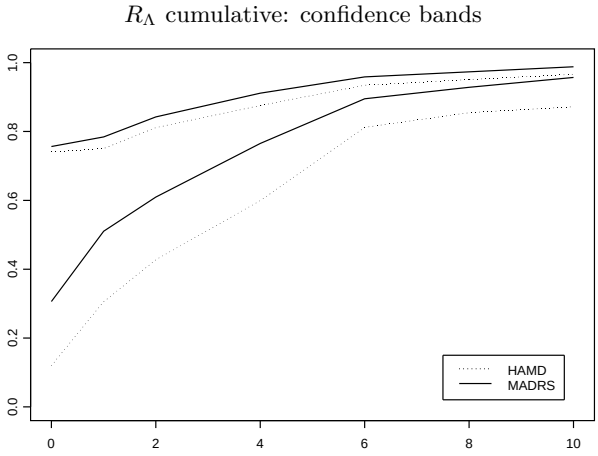

Figure 13.4: *Trial 2. 95% confidence bands around R_Λ cumulative over time points.*

ability. The results for trial 1 clearly illustrate that, even when the R_T values are stable over time or decrease, the total information, as expressed by R_Λ, still increases. When a level of 0.80 is aimed at, four measurements are needed in case of the first trial and three in case of the second trial.

13.3 Conclusion

The analysis of this second case study illustrates that the proposed methodology gives meaningful results when applied to real data. The new coefficients not only behave in an stable and coherent way but they also lead to conclusions that are in line with the clinical knowledge and experience.

Even though we have mainly focussed on the evaluation of reliability in a longitudinal framework, in many situations the repeated evaluation of the subjects over time is impossible or impractical. This raises the question about the applicability of these ideas in a cross-sectional setting. The following chapter explores this issue further.

Chapter 14

A Unified Approach to Multi-item Reliability

Hitherto, the study of reliability has mainly followed two parallel lines of research, depending on the structure of the available data, i.e., single administration *versus* multiple administration. As a consequence, and despite the fact that the same concept is targeted in both settings, measures of reliability in these two scenarios are often conceptually different. In this chapter, we aim at bringing some degree of conceptual unity to the evaluation of reliability.

We apply the methodology introduced for a longitudinal framework (Chapters 7 to 10), to estimate reliability in a setting where cross-sectional multivariate measurements are taken. The link with existing literature on reliability in such settings is extensively discussed.

14.1 Single Administration of a Test

Test-retest reliability requires re-measuring which is often time consuming and expensive. This explains the large amount of attention that has gone to the evaluation of reliability based on a single administration of a test. For instance, the Spearman-Brown formula, the Kuder-Richardson formulas, including the well-known KR-20, its slight and famous variation known as Cronbach's α, the five lower bounds introduced

by Guttman, and the measure proposed by Mosier, are some of the proposals to quantify reliability in this context (Spearman 1910, Brown 1910, Kuder and Richardson 1937, Cronbach 1951, Guttman 1945, Mosier 1943). It has been extensively shown, however, that these measures equal reliability only under rather stringent assumptions. Indeed, parallel tests are required for the Spearman-Brown formula and Cronbach's alpha requires essentially tau-equivalent tests (Novick and Lewis 1967). One of the requirements is *unidimensionality* which means that all items of an instrument or composite test measure the same thing. When these assumptions are not met, the previous measures can not be considered a proper quantification of reliability but merely a lower bound for it (Guttman 1945, Novick and Lewis 1967). Therefore, they are nowadays mainly considered as measures for the *internal consistency* of an instrument, which indicates the homogeneity of the items, or, equivalently said, how much they measure a unidimensional underlying construct. In spite of these limitations, the study of these measures has received a lot of attention in the psychometric literature and they are routinely applied in many practical situations (Barchard and Hakstian 1997, Ten Berge and Hofstee 1999, Ten Berge and Soĉan 2004).

To deal with the fact that the items (or parts) of many tests are not unidimensional, Werts *et al* (1978) proposed a procedure for estimating the reliability of instruments derived from a multidimensional scale, based on factor analytic models. A similar approach has been recently proposed by Tarkkonen and Vehkalahti (2005).

We will start by introducing in the next section the measurement model that will be used through the rest of the chapter.

14.2 Measurement Model

We assume that we have a multi-item scale, formed by p items. Further, we assume that for the ith subject the following measurement model holds

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_i, \tag{14.1}$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ denotes the p-dimensional vector of observed scores, $\boldsymbol{\tau}_i = (\tau_{i1}, \tau_{i2}, \ldots, \tau_{ik})'$ is a k-dimensional vector of true scores, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{ip})'$ is a p-dimensional vector of measurement errors, $\boldsymbol{B}$ is a $p \times k$ matrix that describes the functional relationship between the observed and true scores and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ is a vector describing the mean of the observed scores. Additionally, we assume that:

i) $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$ with $\text{Cov}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma}$, ii) $E(\boldsymbol{\tau}_i) = \mathbf{0}$ with $\text{Cov}(\boldsymbol{\tau}_i) = \boldsymbol{D}$, and finally that iii) $\boldsymbol{\tau}_i$ and $\boldsymbol{\varepsilon}_i$ are independent.

Based on the previous assumptions if we define $\boldsymbol{G} = \boldsymbol{BDB}'$, then the variance-covariance matrix of the measured items $\boldsymbol{V} = \text{Cov}(\boldsymbol{X}_i)$ can be written as

$$\boldsymbol{V} = \boldsymbol{G} + \boldsymbol{\Sigma}. \tag{14.2}$$

Model (14.1) comprises many model families. For instance, it contains as a special case the true-score model used in classical test theory. It is also related to factor analysis and the modeling framework used in generalizability theory. However, unlike the measuring model used in CTT and the analysis-of-variance models with random effects typically used in G-theory, the previous model allows a multidimensional vector of correlated random effects for describing the true scores. Note that, stemming from identifiability issues, some restrictions may be needed to estimate the parameters. For instance, if one assumes that $\boldsymbol{D} = \boldsymbol{I}$ and that $\boldsymbol{\Sigma}$ is a diagonal matrix, then model (14.1) reduces to the classical orthogonal factor analytic model. While these connections are appealing and insightful, in what follows, we will work with model (14.1) in its most general form.

Model (14.1) also contains as special cases three models that have played a prominent role in the quantification of reliability of multi-item scales. They all assume a unidimensional true score and can be defined as

1. *Parallel tests*: obtained when μ and τ_i are scalars, $\boldsymbol{B} = \boldsymbol{\beta} = (1, 1, \ldots, 1)' = \mathbf{1}$, and $\text{Cov}(\boldsymbol{\varepsilon}_i) = \sigma^2 \boldsymbol{I}$

2. *Essentially tau-equivalent tests*: obtained when $\boldsymbol{B} = \boldsymbol{\beta} = \mathbf{1}$, τ_i is a scalar and $\text{Cov}(\boldsymbol{\varepsilon}_i) = \text{diag}(\sigma_j^2)$, with $j = 1, \ldots, p$

3. *Congeneric tests*: obtained when $\boldsymbol{B} = \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$, τ_i is a scalar and $\text{Cov}(\boldsymbol{\varepsilon}_i) = \text{diag}(\sigma_j^2)$.

Interestingly, expression (14.2) closely resembles the decomposition of the total variance-covariance matrix in the longitudinal framework, as expressed in (5.2). For the longitudinal setting we introduced an axiomatic definition of reliability based on four properties, aiming at an extension of the concept to more general scenarios. In the present chapter we argue that the same set of defining properties should be valid in a cross-sectional setting, i.e, they should be universally valid for the definition of reliability. It then logically follows that the measures defined in the chapters 7, 9, and

10 could be also applied to estimate reliability in this context. In the next sections we will expand these ideas further.

We will start by analyzing in some detail the parallel, essentially tau-equivalent and congeneric tests. As stated before, these models have played a prominent role in the evaluation of reliability of multi-item scales. Indeed, an important part of the earlier work focused on estimating the reliability of the scale $Y_i = \mathbf{1}'\boldsymbol{X}_i$,—or, more generally $Y_i = \boldsymbol{a}'\boldsymbol{X}_i$ with $\boldsymbol{a} \in \mathbb{R}^p$,—under the conditions defined by models (1)–(3). In the next section, we will apply the R_T and R_Λ coefficients to quantify reliability in these scenarios. It is important to point out that these measures are valid in more general settings than those defined by (1)–(3). However, their performance in these special cases will help to increase our understanding of their properties and interpretation.

14.3 Reliability with Unidimensional True-score Models

Let us first consider the simplest of the three special cases: the parallel test. The assumptions behind parallel tests are very restrictive and unlikely to hold in practice. Under this model, the decomposition of the variance-covariance matrix given in (14.2) takes the form $\boldsymbol{V} = \sigma_\tau^2 \mathbf{1}\mathbf{1}' + \sigma^2 \boldsymbol{I}$. It is then easy to show that

$$R_T = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2}.$$

Note that if we assume that the items of a scale form parallel tests, then each single item satisfies the model used in classical test theory, i.e.,

$$X_{ij} = \mu_j + \tau_i + \varepsilon_{ij},$$

and the reliability of every item equals $\rho_{xx} = \sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$. Earlier, we have described R_T as a measure of average reliability over time points. The previous results show that in the present context this measure retains its interpretation, but now as an average reliability over items.

When applied to this specific setting, R_Λ takes the form

$$R_\Lambda = \frac{p\rho_{xx}}{(p-1)\rho_{xx} + 1}.$$

Remarkably, under these assumptions, the R_Λ coefficient equals the Spearman-Brown formula. Like before, R_Λ quantifies in this context the reliability of an entire vector of observations. However, this vector is now formed by all the item scores instead of a collection of different scores over time. The previous expression for R_Λ clearly indicates that the reliability of the instrument is an increasing function of the number of items. This is an intuitive and appealing result. Obviously, each new item added to the scale will bring certain level of information about the true score τ_i, even if this information is contaminated by measurement error. As a consequence, the expanded scale will always contain more or at least the same amount information about τ_i than the original scale. Intuitively, the reliability of a scale is the amount of information on the true scores that the scale conveys. Therefore, it is reasonable that adding new items to the instrument can only increase the reliability of the conclusions derived from it.

It is important to recall at this point that R_Λ quantifies the reliability of the entire scale, i.e., the multivariate vector $\boldsymbol{X}_i$. However, the Spearman-Brown formula was originally obtained as the reliability of the scale $Y_i = \mathbf{1}'\boldsymbol{X}_i$ under the parallel test assumptions. We thus find that, under these assumptions, the reliability of the entire scale $\boldsymbol{X}_i$ equals the reliability of the simple sum score. Nevertheless, as we will illustrate later, in more general settings Y_i no longer has the same reliability as the entire scale $\boldsymbol{X}_i$ but rather, as expected, the reliability of a summary statistic like Y_i is usually smaller than the one of the entire instrument.

Let us proceed with model (2). Essentially tau-equivalent tests relax the assumptions of parallel tests by allowing item-specific error variances in model (14.1) so that $\boldsymbol{V} = \sigma_\tau^2 \mathbf{1}\mathbf{1}' + \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = \text{diag}(\sigma_j^2)$. Under these assumptions, R_T takes the form

$$R_T = \frac{\sigma_\tau^2}{\sigma_\tau^2 + S},$$

where $S = (\sum_j \sigma_j^2)/p$. Note that R_T is a decreasing function of S and, therefore, if a new item $(p+1)$ is added to a scale, then

$$R_T(p) \leq R_T(p+1) \quad \text{if and only if} \quad \sigma_{p+1}^2 \leq \frac{\sum_j \sigma_j^2}{p}.$$

Essentially, this implies that the expanded instrument will have a higher average reliability if and only if the error variance of the new item is smaller than the average error variance of the other items of the scale. Therefore, the R_T coefficient can either increase or decrease when a new item is added, depending on the "quality" of such

an addition. Clearly, the previous findings confirm the intuitive interpretation of R_T as the *average* item reliability of the scale.

Turning to R_Λ, we first need to compute the determinant of $\boldsymbol{V}$. It is easy to show that if $\boldsymbol{V} = \sigma_\tau^2 \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Sigma}$ then

$$|\boldsymbol{V}| = (1 + \sigma_\tau^2 \boldsymbol{\beta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}) \cdot |\boldsymbol{\Sigma}|. \tag{14.3}$$

For essentially tau-equivalent tests $\boldsymbol{\beta} = \mathbf{1}$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_j^2)$, and the previous expression for the determinant leads to

$$R_\Lambda = \frac{S}{1+S},$$

which is an increasing function of S, with $S = \sum_{j=1}^{p} \sigma_\tau^2/\sigma_j^2$. Obviously, adding a new item to the scale can only increase the value of S and, therefore, R_Λ is always an increasing function of the number of items. Note however that, if the new item comes contaminated with a lot of measurement error then $\sigma_\tau^2/\sigma_{p+1}^2$ will be negligible and R_Λ will remain nearly constant.

Finally, congeneric tests are the most general ones among the three special cases considered so far. In this scenario the variance-covariance matrix takes the more general form $\boldsymbol{V} = \sigma_\tau^2 \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Sigma}$. For this specific set of assumptions,

$$R_T = \frac{\sigma_\tau^2}{\sigma_\tau^2 + S},$$

with $S = \sum_j \sigma_j^2 / \sum_j \beta_j^2$. Like before, adding a new item can increase or decrease the value of R_T depending on the impact of the new item on S.

Moreover, with $|\boldsymbol{V}|$ as in (14.3), it easily follows that $R_\Lambda = \sigma_\tau^2 S/(1 + \sigma_\tau^2 S)$, with $S = \sum_i \beta_i^2/\sigma_i^2$, and like for tau-equivalent tests, R_Λ can only increase its value when a new item is added.

The above reflections are a useful aid in understanding the meaning and the complementarity of the two new measures. Whereas R_T provides us with information on the quality of the items in a scale, regardless of their number, the R_Λ coefficient informs us on the amount of information the total package of items contain on the underlying traits.

However, due to the strong assumptions on which they are based, the modelling frameworks analyzed in this section have limited practical value. In the next section we will apply the new measures in the more general scenario defined by model

(14.1). Notice that the weaker assumptions that this model requires enhance its practical value and, as a consequence, the newly proposed measures will also allow us to approach the reliability problem in more general settings.

14.4 Reliability with Multidimensional True-score Models

In models (1)–(3), τ_i is a scalar, which means that unidimensionality of the instrument is assumed. Werts *et al* (1978) extended the measurement model by assuming a factor model for the true scores, thence allowing multiple dimensions in the measurement instrument. The specific factors in their model are considered as part of the true scores, so that the model contains specific factors as well as an error component. Such a model might, however, lead to identifiability problems. In their data example, Werts *et al* (1978) assume the specific factors to be zero. Tarkkonen and Vehkalahti (2005) suggested considering the specific factors as measurement errors.

In general, these authors are mainly concerned with the evaluation of the reliability of a new scale Y_i formed as a weighted sum of the item scores. When model (14.1) holds and $\boldsymbol{a} \in \mathbb{R}^p$, Y_i can be written as

$$Y_i = \boldsymbol{a}'\boldsymbol{X}_i = \boldsymbol{a}'\boldsymbol{\mu} + \boldsymbol{a}'\boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{a}'\boldsymbol{\varepsilon}_i.$$

If $\sigma_Y^2 = \mathrm{Var}(Y_i)$, then (14.1) implies $\sigma_Y^2 = \boldsymbol{a}'\boldsymbol{G}\boldsymbol{a} + \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}$. Tarkkonen and Vehkalahti (2005) proposed to quantify the reliability of Y_i as

$$\rho(\boldsymbol{a}) = \frac{\boldsymbol{a}'\boldsymbol{G}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = 1 - \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}}. \tag{14.4}$$

Notice that the previous expression matches the classical definition (CTT) of reliability for the measure Y_i. Similarly to Chapter 9, we can define $\boldsymbol{H} = \boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2}$, where $\boldsymbol{V}^{1/2}$ denotes the symmetric square root of $\boldsymbol{V}$. Like before, $\boldsymbol{H}$ is a symmetric matrix and, therefore, it can be written as $\boldsymbol{H} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, where $\boldsymbol{P}$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_j\}$. It is easy to show that in this setting the λ_j's coincide again with the generalized eigenvalues associated with the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$. Finally, from the previous developments directly follows that

$$\boldsymbol{\Sigma} = \boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q}, \tag{14.5}$$

$$\boldsymbol{V} = \boldsymbol{Q}'\boldsymbol{Q}, \tag{14.6}$$

where $\boldsymbol{Q} = \boldsymbol{P}'\boldsymbol{V}^{1/2}$. Using these results one can rewrite $\rho(\boldsymbol{a})$ as

$$\rho(\boldsymbol{a}) = 1 - \frac{\boldsymbol{a}'\boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{a}}. \tag{14.7}$$

This last expression for $\rho(\boldsymbol{a})$ will play an important role in the subsequent developments.

Werts *et al* (1978) proposed a quantification of reliability very similar to (14.4), actually, their proposal equals (14.4) when the specific factors, included in their model, are assumed to be zero. Tarkkonen and Vehkalahti (2005) further proved that, in general, $\rho(\mathbf{1}) \geq \alpha$ and the equality is obtained if and only if $\boldsymbol{G} = \sigma_\tau^2 \mathbf{1}\mathbf{1}'$ with $\sigma_\tau^2 > 0$ and $\boldsymbol{\Sigma}$ diagonal, i.e., exactly the conditions defined by (2).

In what follows, we will apply R_T and R_Λ to evaluate the reliability of the previously defined scale Y_i. Furthermore, we will study the relationship between the Ω family and the family of scales formed by different weight vectors $\boldsymbol{a}$.

14.5 R_T, R_Λ and ρ for a Weighted Score

The reliability coefficient proposed by Werts *et al* (1978) and Tarkkonen and Vehkalahti (2005) quantifies the reliability of a univariate weighted sum $Y_i = \boldsymbol{a}'\boldsymbol{X}_i$. When applying the measures R_T and R_Λ to this weighted sum, and assuming that model (14.1) holds, we find that both measures equal the coefficient ρ: $R_T(\boldsymbol{a}) = R_\Lambda(\boldsymbol{a}) = \rho(\boldsymbol{a})$. Obviously, in this univariate scenario, the average and total reliability coincide and, therefore, R_T and R_Λ are equal.

14.6 The Ω Family

As stated before, a considerable part of the psychometric literature has focussed on studying the reliability of a family of scales, constructed as the weighted sums of the items of a multi-item scale $\boldsymbol{X}_i$, i.e., the family $\Psi^* = \{Y_i = \boldsymbol{a}'\boldsymbol{X}_i : \boldsymbol{a} \in \mathbb{R}^p\}$. Moving from a high-dimensional instrument $\boldsymbol{X}_i$ to a univariate version Y_i can considerably facilitate the practical use of the scale and the clinical interpretation of the results. Actually, in clinical practice, psychiatrists and psychologists frequently work with weighted sums of multivariate scales.

On the other hand, in Chapter 9, we introduced a general family of plausible reliability measures Ω. Different measures can be formed by assigning different weights to the generalized eigenvalues $\lambda_1, \ldots, \lambda_p$ in (9.1). The following two theorems will shed

light on the relationship between the family of scales Ψ^* and the family of measures Ω.

Theorem 4 *If model (14.1) holds and $\theta \in \Omega$ then there exists a vector $\boldsymbol{a} \in \mathbb{R}^p$ so that θ equals the reliability of the weighted scale $Y_i = \boldsymbol{a}'\boldsymbol{X}_i$, or $\theta = \rho(\boldsymbol{a})$.*

Given a member of the Ω family $\theta = 1 - \sum_j w_j\lambda_j$, it is possible to show that there is a vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_p)'$ so that $w_j = \delta_j^2 / \sum_m \delta_m^2$. Theorem 4 then becomes an immediate consequence of (14.7) with $\boldsymbol{a} = \boldsymbol{Q}^{-1}\boldsymbol{\delta}$. A detailed proof of the previous result can be found in Appendix C.2. Basically, Theorem 4 shows that any member of Ω can be interpreted as the reliability of certain member of Ψ^*. Actually, different choices of $\boldsymbol{a}$ can lead to scales with the same reliability and, therefore, each $\theta \in \Omega$ is associated with more than one Y_i in Ψ^*. The reverse relationship can be also of interest, i.e., one would like to know whether the reliability of any scale in Ψ^* is a member of Ω. Theorem 5 focuses on this issue.

Theorem 5 *Let us assume that model (14.1) holds. If $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{a} \neq \boldsymbol{0}$, then there exists a $\theta \in \Omega$ so that θ is the reliability of the weighted scale $Y_i = \boldsymbol{a}'\boldsymbol{X}_i$ $\{\theta = \rho(\boldsymbol{a})\}$, if and only if $\boldsymbol{a} \in C$, where $C = \{\boldsymbol{a} : (\boldsymbol{Q}\boldsymbol{a})_j \neq 0 \quad \forall j\}$.*

A proof of this result can be found in Appendix C.3. It is clear from Theorem 5 that Ω does not contain the reliability of all the scales in Ψ^*. Indeed, the previous result shows that the Ω family is only equivalent to the family $\Psi = \{Y_i = \boldsymbol{a}'\boldsymbol{X}_i : \boldsymbol{a} \in C\}$. Formally, Ψ^* will be equivalent to a more general family Ω^* which can be defined as

$$\Omega^* = \left\{ \theta : \theta = 1 - \sum_{j=1}^{p} w_j\lambda_j, \quad w_j \geq 0 \quad \text{and} \quad \sum_{j=1}^{p} w_j = 1 \right\}.$$

Note, however, that the elements of Ω^* do not necessarily satisfy the properties (i)–(iv) introduced in Chapter 7. In what follows we will argue that the Ω family contains the reliabilities of those scales Y_i that are meaningful; or in other words, that only those scales included in Ψ should be considered in general.

Truthfully, not all vectors $\boldsymbol{a}$ will lead to meaningful scales. To illustrate this, let us denote by $\ell(\boldsymbol{B})$ the vector space generated by the columns of $\boldsymbol{B}$. Further, we will consider $\boldsymbol{a} \in \ell^{\perp}(\boldsymbol{B})$, where $\ell^{\perp}(\boldsymbol{B})$ denotes the vector space orthogonal to $\ell(\boldsymbol{B})$. Assuming that model (14.1) holds, we have

$$Y_i = \boldsymbol{a}'\boldsymbol{X}_i = \boldsymbol{a}'(\boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_i) = \boldsymbol{a}'\boldsymbol{\mu} + \boldsymbol{a}'\boldsymbol{B}\boldsymbol{\tau}_i + \boldsymbol{a}'\boldsymbol{\varepsilon}_i = \boldsymbol{a}'\boldsymbol{\mu} + \boldsymbol{a}'\boldsymbol{\varepsilon}_i.$$

Note that this scale does not contain any information about the true scores $\boldsymbol{\tau}_i$. Obviously, such a scale would not have any practical value. We will show that $Y_i \notin \Psi$, or what is the same, that the reliability of this scale $\rho(\boldsymbol{a})$ is not an element of Ω. Indeed, if model (14.1) holds, then from (14.2) we have

$$\boldsymbol{Q}\boldsymbol{a} = \boldsymbol{P}'\boldsymbol{V}^{-1/2}\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'\boldsymbol{a} + \boldsymbol{P}'\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{a} \Rightarrow \boldsymbol{Q}\boldsymbol{a} = \boldsymbol{P}'\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{a}. \tag{14.8}$$

Further, using $\boldsymbol{H} = \boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, we get

$$\boldsymbol{P}'\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{P}'\boldsymbol{V}^{1/2} \Rightarrow \boldsymbol{P}'\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{a} = \boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{a}. \tag{14.9}$$

Substituting (14.9) into (14.8), and denoting $\boldsymbol{\delta} = \boldsymbol{Q}\boldsymbol{a}$, we finally obtain $\boldsymbol{\delta} = \boldsymbol{\Lambda}\boldsymbol{\delta}$ or, equivalently, $\delta_j = \lambda_j\delta_j$ for all j. Note that $\boldsymbol{V} \neq \boldsymbol{\Sigma}$ and therefore there is at least a k so that $\lambda_k \neq 1$ and this immediately implies that $\delta_k = (\boldsymbol{Q}\boldsymbol{a})_k = 0$ and, as a consequence, $Y_i \notin \Psi$, or equivalently, $\rho(\boldsymbol{a}) \notin \Omega$. The previous discussion illustrates that meaningless scales like this one are not elements of Ψ.

To enhance insight into this issue in more generality, let us consider a vector $\boldsymbol{a} \in \mathbb{R}^p$ and like before let $\boldsymbol{\delta} = \boldsymbol{Q}\boldsymbol{a}$. Note that $\boldsymbol{a} \in C$ if and only if $\delta_j \neq 0$ for all j. We then have

$$Y_i = \boldsymbol{a}'\boldsymbol{X}_i = (\boldsymbol{Q}^{-1}\boldsymbol{\delta})'\boldsymbol{X}_i = \boldsymbol{\delta}'(\boldsymbol{Q}')^{-1}\boldsymbol{X}_i.$$

Further, let us denote $\boldsymbol{X}_i^* = (\boldsymbol{Q}')^{-1}\boldsymbol{X}_i$. Note that $(\boldsymbol{Q}')^{-1}$ defines a bijective map from $\mathbb{R}^p$ to $\mathbb{R}^p$. In fact, $(\boldsymbol{Q}')^{-1}$ is an invertible matrix and therefore $\boldsymbol{X}_i^*$ and $\boldsymbol{X}_i$ contain the same amount of information about the true scores $\boldsymbol{\tau}_i$. Actually, as one would expect in such a case, they have the same value of R_Λ,

$$R_\Lambda(\boldsymbol{X}_i^*) = 1 - \frac{|\boldsymbol{V}^*|}{|\boldsymbol{\Sigma}^*|} = 1 - \frac{|\boldsymbol{Q}|^{-2}|\boldsymbol{V}|}{|\boldsymbol{Q}|^{-2}|\boldsymbol{\Sigma}|} = 1 - \frac{|\boldsymbol{V}|}{|\boldsymbol{\Sigma}|} = R_\Lambda(\boldsymbol{X}_i).$$

We can then rewrite Y_i as $Y_i = \boldsymbol{\delta}'\boldsymbol{X}_i^*$. We have already stated that $\boldsymbol{a} \in C$ if and only if $\delta_j \neq 0$ for all j. Therefore, the scales that are not included in Ψ, or equivalently, the scales whose reliabilities do not belong to Ω, are those for which $\delta_k = 0$ at least for one k. Let us assume, without loss of generality, that $k = 1$. Clearly, if $\delta_1 = 0$, then the scale Y_i will not use any information coming from the first item of $\boldsymbol{X}_i^*$. So, we are essentially removing one of the items of $\boldsymbol{X}_i^*$ when calculating the weighted average. Such a choice can only reduce the amount of information about the true scores $\boldsymbol{\tau}_i$ contained in Y_i and, therefore, working with a $Y_i \notin \Psi$ can only imply a loss of information about the true scores.

As stated at the beginning of this section, moving from a high-dimensional scale, like $\boldsymbol{X}_i$, to a univariate counterpart Y_i brings practical and interpretational advantages. However, these profits come at a price. The following section explores this issue further.

14.7 Weighted Score versus Multivariate Score

Different scales can be constructed on the same set of items when different weights are applied. However, the reliability of such a sum is always smaller than or equal to the reliability of the entire scale, as the following theorem establishes.

Theorem 6 *If model (14.1) holds and $\boldsymbol{a} \in \mathbb{R}^p$, then the reliability of $Y_i = \boldsymbol{a}'\boldsymbol{X}_i$ is always smaller than or equal to the reliability of $\boldsymbol{X}_i$, i.e., $\rho(\boldsymbol{a}) \leq R_\Lambda$. Equality is obtained if and only if $\lambda_{(2)} = \lambda_{(3)} = \cdots = \lambda_{(p)} = 1$ and $\boldsymbol{a} = \boldsymbol{V}^{-1/2}\boldsymbol{u}_{(1)}$, where $(\lambda_j, \boldsymbol{u}_j)$ are the eigenvalues and eigenvectors associated with the matrix $\boldsymbol{H}$ and $\lambda_{(j)}$ denotes the j^{th} largest eigenvalue with $\boldsymbol{u}_{(j)}$ its corresponding eigenvector.*

For a detailed proof we refer the reader to Appendix C.4. This theorem clearly underscores the price for simplicity. However, this not necessarily implies that unidimensional versions of $\boldsymbol{X}_i$, like Y_i, should never be considered in practice. Indeed, even though the previous result unequivocally states that the multivariate scale $\boldsymbol{X}_i$ will always convey more information than its univariate counterparts, it can be very difficult to grasp the clinical meaning of a p-dimensional vector of observations. A balanced trade-off between interpretability and reliability may, in many situations, well suggest sacrificing a bit of the latter to increase the former. A comparison between estimates of R_Λ and $\rho(\boldsymbol{a})$ thus provides important information for making such a trade-off. The theorem establishes R_Λ as an upper bound for the reliability of an entire family of instruments constructed from the original set of items. Notice that if the value of this measure is low, then any instrument derived as a weighted sum of the original items will have an even lower reliability and will be basically useless.

Let us expand upon the conditions for the equality of R_Λ and $\rho(\boldsymbol{a})$. For the special case implied by the parallel tests, we have that $\boldsymbol{V} = \sigma_\tau^2\mathbf{1}\mathbf{1}' + \sigma^2\boldsymbol{I}$. It is not difficult to show that $\boldsymbol{\Sigma}\boldsymbol{V}^{-1} = \boldsymbol{I} - \sigma_\tau^2(\sigma^2 + p\sigma_\tau^2)^{-1}\mathbf{1}\mathbf{1}'$. Moreover, $\boldsymbol{\Sigma}\boldsymbol{V}^{-1}$ has eigenvalues 1 with multiplicity $p-1$ and $\sigma^2/(\sigma^2 + p\sigma_\tau^2)$ with multiplicity 1. It is also possible to show that, in this case, $\boldsymbol{v}_1 = \mathbf{1}$, where $\boldsymbol{v}_1$ denotes the eigenvector associated to the smallest eigenvalue of $\boldsymbol{\Sigma}\boldsymbol{V}^{-1}$. Moreover, it can be proven that if

$(\lambda_j, \boldsymbol{u}_j)$ is an eigenvalue-eigenvector pair associated with $\boldsymbol{H}$, then $(\lambda_j, \boldsymbol{V}^{1/2}\boldsymbol{u}_j)$ is the corresponding eigenvalue-eigenvector pair associated with $\boldsymbol{\Sigma}\boldsymbol{V}^{-1}$. All these results explain our previous findings when analyzing the parallel-test setting. Indeed, in this scenario $\lambda_{(2)} = \lambda_{(3)} = \cdots = \lambda_{(p)} = 1$ and $\boldsymbol{a} = \mathbf{1}$ up to a multiplicative constant, and therefore $Y_i = \mathbf{1}'\boldsymbol{X}_i$ has the same reliability as $\boldsymbol{X}_i$. In other words, this explains why, under parallel-test assumptions, R_Λ equals the Spearman-Brown formula, which was proposed as a measure of reliability for the simple sum of item scores. To finish, we will say a few more words on the correspondence between these two measures.

In the special cases considered in Section 14.3, we showed that R_Λ is an increasing function of the number of items. The following theorem extends this result to the more general scenario implied by model (14.1).

Theorem 7 *Let us assume that model (14.1) holds. Further, denote by $R_\Lambda(p)$ the corresponding value of R_Λ for the p-dimensional scale $\boldsymbol{X}_i$. If q additional items are added to $\boldsymbol{X}_i$, then the value of R_Λ for this new $(p+q)$-dimensional scale satisfies $R_\Lambda(p+q) \geq R_\Lambda(p)$.*

A detailed proof is given in Appendix C.5. Note that, as stated before, adding new items to an existing scale can only bring more information about the true scores. This fact is nicely captured by R_Λ, which, like the Spearman-Brown formula in the parallel setting, is an increasing function of the number of items. Actually, all these findings indicate that R_Λ can be interpreted as a generalization of the Spearman-Brown formula, that is applicable in settings where the original formulation would not be valid.

14.8 Analysis of the Case Study in Schizophrenia

We will now illustrate the methodology presented in previous sections using the schizophrenia case study introduced in Chapter 2. Particularly, the reliability of PANSS will be analyzed based on a cross-sectional measurement. First we will provide more background information on the origin of the scale.

14.8.1 The Positive And Negative Syndrome Scale

Schizophrenia is a complex and heterogeneous disorder with variable symptoms. To improve research clarifying the diversity in the disorder, Kay *et al* (1987) developed a

standardized instrument; the Positive And Negative Syndrome Scale (PANSS). The instrument contains 30 items (symptoms), which are all scored on a 7-grade scale ranging from "absent" to "extreme." As reflected by the name of the scale, schizophrenia is often described in terms of positive and negative symptoms. Positive symptoms include hallucinations and delusions and are typically regarded as manifestations of psychosis. Negative symptoms are so-named because they are considered to be the loss or absence of normal traits or abilities, and include features such as blunted affect, apathy, and social withdrawal. Besides these two dimensions, general psychopathology was included as a third, a priori factor in PANSS (Kay *et al* 1987). However, empirical research evidenced the existence of five factors, which can be described as; negative syndrome, positive syndrome, excitement, depressive symptoms, and cognitive dysfunction (Lindenmayer *et al* 1995). Many other studies have confirmed a five-factor structure for this scale (e.g. Van der Gaag *et al* 2006a).

Even though the five-factor model is confirmed by several studies, differences are often found in the exact allocation of the items to the factors. Such differences might be related to the use of different statistical techniques or model assumptions, but also to differences in the investigated populations. Dolfus and Petit (1995), for example, did not observe a depression dimension in an acute population while it was observed in a chronic population. A plausible explanation is that depressive symptoms cannot be expressed when positive symptoms are very severe. In many of the studies investigating the factor structure of PANSS, models have been developed where each item loads only on one factor. The underlying aim is to divide the scale in separate sub-scales composed of clearly distinguished sets of items. Van der Gaag *et al* (2006b) showed, by means of a cross-validation study, that allowing some items to load on more than one factor leads to a better model fit.

14.8.2 Data Analysis

We investigate the reliability of PANSS, based on clinical trial baseline measurements taken from 520 in-patients with a diagnosis of chronic schizophrenia after a single-blind placebo washout period (Chouinard *et al* 1993, Marder and Meibach 1994).

The first step in the reliability analysis is to find a well fitting model for the data, thence providing us with the variance-covariance parameter estimates, necessary for the estimation of reliability. As expressed in (14.2), variability in the observations comes from two sources, the latent variables (random effects) and the measurement errors. Since both are unobserved, model restrictions will be inevitable to avoid

identifiability problems. A factor-analytic approach is applied to fit and compare different models.

We start with an exploratory factor analysis (EFA), where we assume that $\boldsymbol{D} = \boldsymbol{I}$, $\boldsymbol{\Sigma}$ is a diagonal matrix, and $\boldsymbol{B}$ unstructured. This means that the factors are assumed to be independent, as well as the measurement errors, and each item can load on every factor. Models with one until seven factors are compared.

We further use confirmatory factor analysis (CFA) to fit two models that were proposed in the literature. Restrictions are now mainly laid on the $\boldsymbol{B}$ matrix by allowing the items to load only on pre-defined factor(s). In the first model, each of the items loads on one factor only. The model follows the five sub-scales proposed by Marder, Dabis, and Chouinard (1997) where $\boldsymbol{D}$ is an unstructured correlation matrix, indicating that factors are allowed to correlate and $\boldsymbol{\Sigma}$ is a diagonal matrix. The second model is the one proposed by Van der Gaag *et al* (2006b). In this model, several items can load on more than one factor. Further, some factors are assumed to be correlated and also some pre-specified measurement errors can be correlated.

All models were fitted using maximum likelihood estimation. Table 14.1 presents fit statistics for the various fitted models. Two goodness-of-fit measures are based on the direct comparison of the sample and model-implied variance-covariance matrices. The Adjusted Goodness-of-Fit Index (AGFI) is generally a number between 0 and 1 with a better fit when values are closer to 1 (Mulaik *et al* 1989). The Root Mean Square Residual (RMR) is the mean of the squared residuals, with values closer to 0 indicating a better fit. Further we present three likelihood-based goodness-of-fit measures: Akaike's Information Criterion (AIC), the Consistent Akaike's Information Criterion (CAIC), and the Schwarz's Bayesian Criterion (SBC). The latter two incorporate a penalty term based on sample size and therefore tend to select simpler models than does AIC. The model that yields the smallest value of each of these three criteria is considered best.

Comparing the seven EFA models, we find the smallest CAIC value for a model with five factors. The other four fit statistics, however, point in the direction of a more complex seven-factor model. This can be partly due to the large size of the data set. We observe indeed that the SBC value of the 7-factor model does not show a very substantial improvement compared to the 5-factor model.

When we further take into account the CFA models, the smallest CAIC and SBC values are found for the model proposed by Van der Gaag (2006b) (CFA2), obviously a more parsimonious model than the EFA models, but closer to the observed data than

Table 14.1: *Various fit statistics for the models considered. The models are indicated by 'Exploratory Factor Analysis' (EFA) 1 to 7 or 'Confirmatory Factor Analysis' (CFA) 1 and 2. The fit statistics: AGFI: Adjusted Goodness-of-Fit Index; RMR: Root Mean Square Residual; AIC: Akaike's Information Criterion; CAIC: Consistent Akaike's Information Criterion; SBC: Schwarz's Bayesian Criterion.*

model	AGFI	RMR	AIC	CAIC	SBC
EFA1	0.42	0.30	3598	1478	1883
EFA2	0.60	0.18	1956	-13	363
EFA3	0.68	0.12	1211	-611	-263
EFA4	0.74	0.10	707	-973	-652
EFA5	0.83	0.07	304	-1240	-945
EFA6	0.84	0.06	189	-1224	-954
EFA7	0.87	0.05	74	-1213	-967
CFA1	0.73	0.20	1271	-796	-401
CFA2	0.82	0.14	527	-1420	-1048

the even simpler model by Marder *et al* (1997) (CFA1). On the other hand, according to AGFI, RMR and AIC the 7-factor EFA model still fits the data best. In factor analysis, the interpretability of the model is often an important additional criterion for the selection of a model. Preference is then given to a model that corresponds to the knowledge in the field. Taking this into account, preference could then go to the model proposed by Van der Gaag (2006b) or to the five factor EFA model. Looking at the factor loadings after varimax rotation, the latter corresponds very closely to the five-factor model commonly proposed in the literature (e.g., Lindenmayer *et al* 1995).

Selecting the best model based on factor analysis is very difficult. Indeed, such models are heavily latent and specify a lot about the unobserved, leading to different models that seem to fit the data equally well. Table 14.2 presents the reliability estimates and confidence intervals for the three models yielding the best results. For details on the calculation of the confidence intervals, we refer the reader to Appendix B.4. The table shows similar results for the three models, indicating a certain degree of "robustness" for the reliability estimations. The previous results seem to

Table 14.2: *PANSS. Point estimates and 95% confidence intervals for the three reliability measures:* R_T, R_Λ, *and* $\rho(\mathbf{1})$.

model	R_T	R_Λ	$\rho(\mathbf{1})$
EFA5	0.479 [0.436; 0.522]	1.000 [0.999; 1.000]	0.911 [0.872; 0.939]
EFA7	0.521 [0.481; 0.562]	1.000 [1.000; 1.000]	0.918 [0.878; 0.946]
CFA2	0.446 [0.424; 0.468]	1.000 [1.000; 1.000]	0.888 [0.765; 0.952]

indicate that while finding the 'best' model can be hard, it is sufficient to find a good fitting model in order to estimate reliability.

The measure R_T indicates the average item reliability and lies around 0.50, which is, for a single item, certainly an acceptable level. As stated before, R_Λ represents the information available when all items are considered jointly, i.e., it expresses the reliability of the entire multivariate scale. The fact that individual items already achieve a decent reliability level and that PANSS contains no less than 30 items, explains why we obtain values for R_Λ equal to one. Essentially, such a high value of R_Λ indicates that the scale conveys a lot of information on the latent variables.

In practice, the sum score of the PANSS items is mostly used for clinical evaluation and data analysis. We have already shown that working with the sum of the item scores always leads to a certain amount of information loss. Table 14.2, however, shows that the reliability of the sum score, expressed by $\rho(\mathbf{1})$, is indeed lower than R_Λ, but still has a very high value. The results thus show that summing PANSS items leads to a relatively small loss of information. It is important to point out here that these two reliability measures are valid at two different levels. Indeed, the R_Λ coefficient quantifies the amount of information shared by the vector of observed scores and the vector of true scores, whereas $\rho(\mathbf{1})$ quantifies the information shared by a well-chosen linear combination of the observed scores and a corresponding linear combination of the true scores. At any rate, the high reliability of the sum score obtained for this scale suggests that working with the sum for clinical evaluation and data analysis may be a sensible idea given the substantial simplification that it brings.

Interest may also lie in estimating the patients' scores on PANSS sub-scales. For example, Marder *et al* (1997) investigated drug-effectiveness on the different dimensions of schizophrenia. Reliability estimates for the separate sub-scales can then be obtained by replacing the full matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ by sub-matrices, $\boldsymbol{\Sigma}_S$ and $\boldsymbol{V}_S$, re-

Table 14.3: *PANSS. Point estimates of the three reliability measures for the five selected sub-scales.*

	Positive	Negative	Cognitive	Excitement	Depression
R_T	0.401	0.571	0.436	0.590	0.466
R_Λ	0.949	0.942	0.914	0.902	0.858
$\rho(\mathbf{1})$	0.798	0.894	0.829	0.836	0.754

lated to the variances and covariances between the items in the sub-scale. Table 14.8.2 presents the point estimates of the three reliability measures, for each of the five sub-scales. The estimates are based on the five-factor exploratory factor-analytic model. The results show that all five sub-scales have good reliability. Additionally, the sum score reliabilities are all above 0.75. Interestingly, the negative sub-scale clearly has a higher average (R_T) and sum score [$\rho(\mathbf{1})$] reliability than the positive sub-scale, however the R_Λ's are similar. This owes to the fact that the positive sub-scale has 8 items whereas the negative sub-scale has 7. For the positive sub-scale, about 15% of information is lost due to summing the item scores, for the negative sub-scale only 5% is lost.

PANSS is a widely used and appreciated scale to evaluate the severity of schizophrenia and our previous results clearly confirm the quality of this instrument. Obviously, the methodology described in this chapter can also be applied to less widely known scales. For example, all of these measures can be useful tools in the developmental phase of a rating scale. Indeed, during this process, one could calculate the reliability per item (R_{Tj}) and items with low values could then be reconsidered or discarded. Furthermore, in order to find an optimal length for the scale, R_Λ could be very helpful as well. By calculating R_Λ cumulatively, i.e., recalculating its value for each new scale constructed by adding an item to the previous one, the additional gain in information of a new item could be quantified, on top of items already included. Obviously, once a pre-defined level of reliability has been achieved no other items would need to be added. The combination of both measures would allow selecting the most informative items, limiting at the same time, the length of the scale.

Finally, we would like to remark that our measurement model (14.1) assumes that the observed scores are of a continuous nature, i.e., it is assumed they are measured on an interval or ratio scale, whereas the items of PANSS are strictly only ordinal

measurements. In the same way as argued in Section 8.3 for the CGI scale we follow the predominant view among statisticians and nicely expressed by Tukey (1961, 1962), who states that science in general and statistics in particular rely upon the test of experience as the ultimate standard of validity. We, therefore, feel encouraged by the many successful applications of factor-analytic models to rating scale data, among others to PANSS. Results stemming from such applications have given very useful and meaningful practical results in full agreement with the specific knowledge of the field.

14.9 Conclusion

In this chapter we have shown that the concepts previously introduced for the evaluation of reliability in a longitudinal context can also be meaningfully applied in a cross-sectional scenario. We believe this brings some degree of conceptual unity to the evaluation of reliability. Indeed, to our knowledge, and in spite of being targeting the same concept, research in these two settings, i.e, the cross-sectional and longitudinal, has run on parallel lines. We proposed a unifying approach that is based on an axiomatic definition of reliability. Interestingly, the developments derived from this definition can be applied in both, the single- and multiple-administration scenarios.

As previously shown, an uncountable number of reliability measures emerges from this approach, the so-called Ω family. One special member of this family is R_T, a measure that in a cross-sectional setting can be interpreted as the average item reliability. Additionally, we have shown that the elements of Ω account for the reliability of all "meaningful" scales constructed as the weighted sum of the item scores of the original instrument. We have also shown that working with such a univariate counterpart can substantially simplify the clinical interpretation but it always implies a loss of information, which then further translates in a decreased reliability.

We also studied the R_Λ coefficient which is not an element of the Ω family. This measure, originally defined in a longitudinal scenario, provides an upper limit for the Ω family and expresses the amount of information that is available in a multi-item scale. Like the Spearman-Brown formula, R_Λ is an increasing function of the number of items. However, unlike the former, R_Λ is valid in more general settings than the one defined by the parallel tests. Remarkably, under parallel tests, R_Λ equals the Spearman-Brown measure. Basically, the R_Λ coefficient can be seen as a generalization of the Spearman-Brown formula to more complex modelling scenarios.

Chapter 15

Concluding Remarks and Further Research

15.1 Concluding Remarks

Rating scales are frequently used for the primary outcome measurement in psychopharmacological trials. When using such scales in research or in clinical practice, information on their psychometric properties should be available. These properties are generally investigated when a scale is being developed, however, the reliability of a scale is not a fixed characteristic of the instrument, but is rather population dependent. More heterogeneous populations give rise to more reliable measurements. Furthermore, reliability can also depend on other external factors like, for instance, the skills or the level of training of the raters. It is therefore useful to evaluate the reliability of certain rating scale, whenever this scale is applied. However, many approaches for estimating reliability are based on very restrictive modelling frameworks.

A common feature in present-day psychopharmacological trials is the presence of repeated measurements. The modelling frameworks used in CTT or G-theory will frequently be inappropriate to study reliability in this scenario. In the present work we have tried to extend the concept of reliability to this more general setting.

A psychiatric symptom scale will be useful only if it can discriminate among different patients, essentially those who have a mental illness from those that do not,

or those patients who are in a more advanced stage of a disease from those who are in a more primary stage, or those patients who have made progress from those who have not, or did so to a lesser extent. This discriminating capability will be possible only if the scale's values vary more between subjects than what they vary within the same subject. This relation between the within and between-subject variability is what we try to determine when we study the reliability of the scale. The reliability of a scale is therefore the capacity of the scale to discriminate between different subjects or different groups of subjects.

The appraisal of reliability has certainly been among the most central issues in psychometrics during the past century. Despite the fact that they are all targeting the same concept, measures used to quantify reliability have largely depended on the data structure. This lack of a unifying approach has resulted in a myriad of measures, which sometimes lead to different conclusions and varying interpretations. Hitherto, the two main contexts for the appraisal of reliability, i.e., the cross-sectional and longitudinal scenario (single-administration and multiple-administration) have been studied using different approaches. In the present work we have introduced a general definition of reliability based on a simple set of properties. This definition can be equally applied in the cross-sectional and longitudinal setting.

The definition lead to a whole family of reliability measures, the Ω family. All the members of this family are built upon the same basic elements: the roots of the equation $q(\lambda) = |\boldsymbol{\Sigma} - \lambda \boldsymbol{V}| = 0$; where $\boldsymbol{\Sigma}$ expresses the within-subject variability and $\boldsymbol{V}$ the total variability, and therefore $\boldsymbol{V} - \boldsymbol{\Sigma}$ the between-subject variability.

The usefulness of two of these measures, R_T and R_Λ, has been extensively illustrated. In a longitudinal context, the R_T coefficient expresses the average reliability over the different measurement occasions. Having a single measure has the advantage of facilitating interpretation and is very useful whenever two scales should be compared on their reliability. On the other hand, it is possible to obtain R_T values per time point, which can be useful when one is interested in the evolution of reliability over the course of the study. Typically, one observes a slight increase of R_T over time, plausibly due to an increase of the raters' skills and their knowledge about the patients.

The R_Λ coefficient, even though structurally similar to R_T, bears a totally different interpretation. This measure expresses the reliability of the longitudinal sequence as a whole. It captures not the average reliability per time point, but the reliability of the information that is available when considering the repeated measures jointly. As

a consequence, R_Λ will always increase when the number of measurements increases. Relevantly, this implies that we can always obtain a pre-specified level of reliability if the patient is followed long enough. Indeed, even if we only have to our disposal a scale that is permeated by a relatively large amount of measurement error, we can still increase the reliability of our conclusions by repeating the measurement over time.

The previous developments were first considered within a longitudinal framework, and based upon the presence of repeated measurements. However, in psychometric research, much interest has always gone to the study of reliability in the context of cross-sectional, multivariate measurement. We have illustrated that the same measures as proposed in the longitudinal context also apply when studying reliability in a multivariate setting. The R_T coefficient then expresses the average reliability per item whereas the R_Λ coefficient refers to the reliability of the information available in the entire scale.

In practice clinicians frequently work with scales constructed by the (weighted) sum of the item scores. While a loss of information is then unavoidable, interpretability is can be gained. We have seen that any member of the Ω family corresponds to the reliability of different but meaningful weighted sum scores.

Finally, we want to point out that a set of SAS macro's has been written for the calculation of the point estimates and asymptotic confidence intervals for R_T, R_Λ and some elements of the Ω family. Manuals explaining the macros and the interpretation of their results are also available.

15.2 Further Research

The approach to reliability presented in this work is entirely based on the class of linear mixed models which forms a very powerful tool for the analysis of continuous data. Obviously, further extensions for categorical data deserve special attention. In this direction links with IRT are, undoubtedly, an interesting line of research.

The impact of missing data on the performance of the proposed measures is also worth investigating as well as the impact of model misspecifications on the accuracy of their point estimates and the performance of their asymptotic confidence intervals.

Even though the connection between reliability and sample size has been well established in a simple cross-sectional scenario, it has not been studied with more complicated data structures like longitudinal data. Therefore, it would be interesting to explore the relationship between statistical concepts like power and sample size on

one hand and quantifications of reliability like R_T and R_Λ on the other hand.

Clearly, many interesting issues have not been explored in the present work and deserve to be further studied. On the other hand, the evaluation of the new proposals have been limited by time constraints and the availability of real data. Probably, only through the future application and study of the ideas introduced in this work one will be able to fully clarify their potential value as well as their limitations.

References

Abelson, R.P. and Tukey, J.W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics*, **34**, 1347–1369.

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2002). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical Statistics* **12**, 161–179.

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M. (2004). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: Canonical correlation approach, *Biometrics,* **60**, 845–853.

Alonso, A., Laenen, A., Molenberghs, G., Geys, H., and Vangeneugden, T. (2008). Reliability of Single Administered Tests: A Unified Approach. (*Submitted for publication.*)

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, **34**, 42–54.

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

Barchard, K., and Hakstian, A. R. (1997). The effects of sampling model on inference with coefficient alpha. *Educational and Psychological Measurement*, **57**, 893–905.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports,* **19**, 3–11.

Bech, P. and Jha, S.K. (Accessed 26.6.2008). Rating scales in psychiatry. (http://www.cnsforum.com/clinicalresources/ratingscales/ratingpsychiatry)

Bell, M., Milstein, R, Beam-Goulet, J., Lysaker, P., and Cicchetti, D. (1992). The Positive and Negative Syndrome Scale and the Brief Psychiatric Rating Scale: reliability, comparability, and predictive validity. *Journal of Nervous and Mental Disease,* **180**, 723–728.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M.R. (eds) *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Blin, O., Azorin, J. M., and Bouhours, P. (1996). Antipsychotic and anxiolytic properties of risperidone, haloperidol and methotrimeprazine in schizophrenic patients. *Journal of Clinical Psychopharmacology*, **16**, 38–44.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**, 443–459.

Bohrnstedt, G.W. (1983). In: Handbook of Survey Research. New York: Academic Press.

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

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

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

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multi-method matrix. *Psychological Bulletin*, **56**, 85–105.

Chouinard, G., Jones, B., and Remington, G. (1993). A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients. *Journal of Clinical Psychopharmacology*, **13**, 25–40.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* **16**, 297-334.

Cronbach, L. J., Nageswari, R., and Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology* **16**, 137-163.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: John Wiley.

De Boeck P. and Wilson M. (2004). Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Springer-Verlag: New York.

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

Dolfus, S. and Petit, M. (1995). Principal-component analyses of PANSS and SANS-SAPS in schizophrenia: their stability in an acute phase. *European Psychiatry*, **10**, 97–106.

Dunn, G. (1989). *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors.* Oxford University Press: New York.

Fleiss, J. L. (1986). *Design and Analysis of Clinical Experiments.* Wiley: New York.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, **87**, 564–567.

Graybill, F. A. (1983). *Matrices with Applications in Statistics.* Belmont, California: Wadsworth International Group.

Gulliksen, H. (1950). *Theory of Mental Tests.* New York: Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika* **10**, 255-282.

Guttman, L. (1953). Reliability formulas that do not assume experimental independence. *Psychometrika* **18**, 225-239.

Guy, W. (1976). ECDEU Assessment Manual for Psychopharmacology - Revised (DHEW Publ No ADM 76-338). Rockville, MD, U.S. Department of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, NIMH Psychopharmacology Research Branch, Division of Extramural Research Programs, pp 218-222.

Hamilton, M. (1959). The assessment of anxiety states by rating. *British Jounal of Medical Psychology* , **32**, 50–55.

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

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

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

Hoyberg, O. J., Fensbo, C., Remvig, J., Lingjaerde, O. K., Slotei-Nielsen, M., and Salvesen, I. (1993). Risperidone versus perphenazine in the treatment of chronic schizophrenic patients with acute exacerbations. *Acta Psychiatrica Scandinavica*, **88**, 395–402.

Huttunen, M. O., Piepponen, T., Rantanen, H., Larmo, I., Nyholm, R., and Raitasuo, V. (1995). Risperidone versus zuclopenthixol in the treatment of acute schizophrenic episodes: a double-blind parallel-group trial. *Acta Psychiatrica Scandinavica*, **91**, 271–277.

Jackson, P.H., and Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrika*, **42**, 567–578.

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

Johnson, R. A., and Wichern D. W. (1998). *Applied Multivariate Statistical Analysis. Fourth Edition.* Englewood Cliffs, NJ: Prentice-Hall.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika,* 36, 109-133.

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophrenia Bulletin*, **13**, 261–276.

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

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika,* **2**, 151-160.

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

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

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

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

Laenen, A., Alonso, A., Molenberghs, G., and Vangeneugden, T. (2008). Reliability of a longitudinal sequence of scale ratings. *Psychometrika.* DOI: 10.1007/S11336-008-9079-7

Laenen, A., Alonso, A., Molenberghs, G., Vangeneugden, T., and Mallinckrodt, C. H. (2008b). Impact of ignoring serial correlation and memory effect on reliability estimates. *Submitted for publication.*

Laenen, A., Alonso, A., Molenberghs, G., Mallinckrodt, C. H., and Vangeneugden, T. (2008). Using longitudinal data from a clinical trial in depression to assess the reliability of its outcome scales. *Journal of Psychiatric Research.* DOI: 10.1016/j.jpsychires.2008.09.010

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

Lindenmayer, J.P., Bernstein-Hyman, R., Grochowski, S., and Bark, N. (1995). Psychopathology of schizophrenia: initial validation of a 5-factor model. *Psychopathology*, **28**, 22–31.

Little, R.J.A., and Rubin, D.B. (2002). Statistical Analysis with Missing Data (2nd edn). New York: John Wiley & Sons, Inc.

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

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Magnus, J. R., and Neudecker, H. (1994). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley and Sons, Inc.

Maier, W., Philipp, M., Heuser, A., Schlegel, S., Buller, R., and 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., and Tran, P. V. (2003). Duloxetine: a new treatment for the emotional and physical symptoms of depression. *Primary Care Companion to The Journal of Clinical Psychiatry*, **5**, 19–28.

Marcoulides, G. (1987). *An alternative Method for Variance Component Estimation: Applications to Generalizability Theory.* Unpublished doctoral dissertation, University of California, Los Angeles.

Marder, S.R. and Meibach, R.C. (1994). Risperidone in the treatment of schizophrenia *American Journal of Psychiatry*, **151**, 825–835.

Marder, S.R., Dabis, J.M., and Chouinard, G. (1997). The effects of Risperidone on the five dimensions of schizophrenia derived by factor analysis: combined results of the North American trials. *Journal of Clinical Psychiatry*, **58**, 538–546.

Merkel, L. (Accessed 26.6.2008). The history of psychiatry. (http://www.healthsystem.virginia.edu /internet/psych-training/seminars/history-of-psychiatry-8-04.pdf)

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

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

Mosier, C.I. (1943). On the reliability of a weighted composite, *Psychometrika*, **8**, 161–168.

Mulaik, S.A., James, L.R., Van Alstine, J., Bennett, N., Lind, S., and Stillwell, D.C. (1989). Evaluation of goodness of fit indices for structural equation models. *Psychological Bulletin*, **105**, 430–445.

NCLS (Accessed 26.6.2008) Shock therapy makes a comeback: states respond (http://www.ncsl.org/programs/health/shn/2007/sn499c.htm)

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, **3**, 1-18.

Novick, M. R., and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, **32**, 1-13.

O'Shaughnessy, J. A., Wittes, R. E., Burke, G., Friedman, M. A., Johnson, J. R., Niederhuber, J. E., Rothenberg, M. L., Woodcock, J., Chabner, B. A., and Temple, R. (1991). Commentary concerning demonstration of safety and efficacy of investigational anticancer agents in clinical trials. *Journal of Clinical Oncology*, **9**, 2225–32.

Overall, J. E., and Gorham, D. R. (1962). The Brief Psychiatric Scale. *Psychological Reports*, **10**, 799–812.

Peuskens, J. and the Risperidone Study Group (1995). Risperidone in the treatment of patients with chronic schizophrenia: a multi-national multi-centre, double blind, parallel groups study versus haloperidol. *British Journal of Psychiatry*, *166*, 712–726.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: MESA.

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

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

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

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* Wiley: New York.

Searle, S.R., Casella, G., McCulloch, C.E. (1992). *Variance Components.* New York: Wiley.

Shavelson, R. J., Webb, N. M., and Rowley, G. L. (1989). Generalizability theory. *American Psychologist* **44**, 922-932.

Shavelson, R.J., and Webb, N.M. (1991). *A Primer on Generalizability Theory.* Newbury Park, CA: Sage Publications.

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86**, 420-428.

Sing, M. M., and Kay, S. R. (1975). A comparative study of haloperidol and chlorpromazine in terms of clinical effects and therapeutic reversal with benztropine in schizophrenia: theoretical implications for potency differences among neuroleptics. *Psychopharmacologia*, **43**, 103–113.

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

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

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology,* **15**, 72–101.

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

Suen, H. K., and Lei, P. W. (2007). Classical versus generalizability theory of measurement. *Educational Measurement,* **1**, 3-20.

Swaminathan, H. and Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, **50**, 349–364.

Tarkkonen, L. and Vehkalahti, K. (2005). Measurement errors in multivariate measurement scales. *Journal of Multivariate Analysis* **96**, 172–189.

Ten Berge, J.M.F. and Hofstee, W.K.B. (1999). Coefficient alpha and reliabilities of rotated and unrotated components. *Psychometrika*, **64**, 83–90.

Ten Berge, J.M.F. and Soĉan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, **69**, 611–623.

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

Townsend, J.T. and Ashby, F.G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin.* **96**, 394–401.

Tukey, J.W. (1961). Data analysis and behavioral science or learning to beat the quantitative man's burden by shunning badmandments. In *The Collected Works of John W. Tukey, (Vol. III) (1986)*, L.V. Jones (Ed.), Belmont, CA: Wadsworth, Inc., pp. 391–484.

Tukey, J.W. (1962). The future of data analysis. In *The Collected Works of John W. Tukey, (Vol. III) (1986)*, L.V. Jones (Ed.), Belmont, CA: Wadsworth, Inc., pp. 187–389.

Van der Gaag, M., Cuijpers A., Hoffman, T., Remijsen, M., Hijman, R., de Haan, L., van Meijel B., van Harten, P.N., Valmaggia, L., de Hert, M., and Wiersma, D. (2006a). The five-factor model of the Positive and Negative Syndrome Scale I: Confirmatory factor analysis fails to confirm 25 published five-factor solutions. *Schizophrenia Research,* **85**, 273–279.

Van der Gaag, M., Hoffman, T., Remijsen, M., Hijman, R., de Haan, L., van Meijel B., van Harten, P.N., Valmaggia, L., de Hert, M., Cuijpers A., and Wiersma, D. (2006b). The five-factor model of the Positive and Negative Syndrome Scale II: An ten-fold cross-validation of a revised model. *Schizophrenia Research,* **85**, 280–287.

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. Springer: New York.

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

Webb, N.M., Shavelson, R.J. and Haertel E.H. (2007). Reliability coefficients and generalizability theory. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics 26: Psychometrics*. Amsterdam: North Holland.

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

Werts, C. E., Rock, R. D., Linn, R. L., and Jöreskog, K. G. (1978). A general method of estimating the reliability of a composite. *Educational and Psychological Measurement,* **38**, 933-938.

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

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

Yule, G. U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society*, **75**, 579-642.

Zimmerman, M., Posternak, A., and 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.

Appendix A

Four Defining Properties for Reliability Measures

A.1 R_T Satisfies the Four Defining Properties

We will prove the statement that R_T satisfies the properties (i) – (iv) introduced in Section 7.1. Without loss of generality we will provide the proof in the single trial setting with a balanced study design, where $\boldsymbol{V} = \boldsymbol{V}_i$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_i$.

i. $0 \leq R_T \leq 1$
i.1 $R_T \geq 0$
To prove (i.1) it is sufficient to show that $\text{tr}(\boldsymbol{\Sigma}) \leq \text{tr}(\boldsymbol{V})$ so we only have to prove that $\text{tr}(\boldsymbol{\Sigma}_D) \geq 0$. Note that

$$\text{tr}(\boldsymbol{\Sigma}_D) = \text{tr}(\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}') = \sum_{j=1}^{p} \boldsymbol{z}_j \boldsymbol{D} \boldsymbol{z}_j'$$

where p is the number of time points and $\boldsymbol{z}_j$ is the jth row of $\boldsymbol{Z}$. As $\boldsymbol{D}$ is positive definite $\boldsymbol{z}_j \boldsymbol{D} \boldsymbol{z}_j' \geq 0$ for all j and we get (i.1). $\square$
i.2 $R_T \leq 1$ is obvious. $\square$

ii. $R_T = 0$ if and only if $\boldsymbol{V} = \boldsymbol{\Sigma}$
Note that $R_T = 0$ if and only if $\text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{V})$. Additionally, $\text{tr}(\boldsymbol{V}) = \text{tr}(\boldsymbol{\Sigma}_D) + \text{tr}(\boldsymbol{\Sigma})$

and therefore $\text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\boldsymbol{V})$ if and only if $\text{tr}(\boldsymbol{\Sigma}_D) = 0$, or equivalently if $\boldsymbol{z}_j \boldsymbol{D} \boldsymbol{z}_j' = 0$ for all j. Being $\boldsymbol{D}$ positive definite the previous equality can only be obtained in the degenerated case where $\boldsymbol{D} = \boldsymbol{0}$ and as a consequence $\boldsymbol{V} = \boldsymbol{\Sigma}$. □

iii. $R_T = 1$ if and only if $\boldsymbol{\Sigma} = \boldsymbol{0}$ is obvious. □

iv. In the classical setting $R_T = \dfrac{\sigma_b^2}{\sigma_b + \sigma^2}$
In the classical cross-sectional case, model (5.1) reduces to:

$$Y_i = \mu + b_i + \varepsilon_i$$

$$b_i \sim N(0, \sigma_b^2),$$

$$\varepsilon_i \sim N(0, \sigma^2).$$

Now $\boldsymbol{V} = \sigma_b^2 + \sigma^2$ and $\boldsymbol{\Sigma} = \sigma^2$ so that $R_T = 1 - \dfrac{\sigma^2}{\sigma_b^2 + \sigma^2} = \dfrac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$. □

A.2 All Members of Ω Satisfy the Four Defining Properties

We will prove that all the members of Ω (9.1) satisfy the properties (i)–(iv), introduced in Section 7.1.

i. $0 \leq \theta \leq 1$ for all $\theta \in \Omega$

i.1 $\theta \geq 0$

Note first that $\theta = 1 - \sum_j w_j \lambda_j$ and, therefore, $\theta \geq 0$ if and only if $\sum_j w_j \lambda_j \leq 1$. However, from Theorem 1 we have $\sum_j w_j \lambda_j \leq \sum_j w_j = 1$. □

i.2 $\theta \leq 1$ is obvious. □

ii. $\theta = 0$ if and only if $\boldsymbol{V} = \boldsymbol{\Sigma}$

$$\begin{aligned} \theta = 1 - \sum_{j=1}^{p} w_j \lambda_j = 0 \quad &\Leftrightarrow \quad 1 = \sum_{j=1}^{p} w_j \lambda_j \\ &\Leftrightarrow \quad \lambda_j = 1 \quad \text{for all } j \\ &\Leftrightarrow \quad \boldsymbol{\Sigma} = \boldsymbol{V} \quad \square \end{aligned}$$

Note that the last equivalence is a direct consequence of (9.2) and (9.3).

iii. $\theta = 1$ if and only if $\boldsymbol{\Sigma} = \mathbf{0}$

$$\begin{aligned} \theta = 1 - \sum_{j=1}^{p} w_j \lambda_j = 1 \quad &\Leftrightarrow \quad \lambda_j = 0 \quad \text{for all } j \\ &\Leftrightarrow \quad \boldsymbol{\Sigma} = 0 \quad \square \end{aligned}$$

Here again the last equivalence is a direct consequence of (9.2).

iv. In the classical setting $\theta = \dfrac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$ for all $\theta \in \Omega$.

In the classical setting $p = 1$, $\boldsymbol{\Sigma} = \sigma^2$ and $\boldsymbol{V} = \sigma_b^2 + \sigma^2$, so that

$$q(\lambda) = |\sigma^2 - \lambda(\sigma_b^2 + \sigma^2)| = 0$$

and

$$\theta = \sum_{j=1}^{p} w_j \rho_j^2 = \rho_1^2 = 1 - \frac{\sigma^2}{\sigma_b^2 + \sigma^2} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}. \quad \square$$

A.3 R_Λ Satisfies a Modified Set of Properties

We will prove that R_Λ fulfills properties (i), (ii), (iv), introduced in Section 7.1, and property (iii'), as defined in Section 10.1. Let us note first that from (9.2) and (9.3) we have

$$\begin{aligned} \boldsymbol{V} &= \boldsymbol{Q}'\boldsymbol{Q}, \\ \boldsymbol{\Sigma} &= \boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q}, \end{aligned}$$

so that

$$\begin{aligned} |\boldsymbol{\Sigma}| &= |\boldsymbol{Q}'||\boldsymbol{Q}||\boldsymbol{\Lambda}| = |\boldsymbol{Q}|^2|\boldsymbol{\Lambda}|, \\ |\boldsymbol{V}| &= |\boldsymbol{Q}|^2, \end{aligned}$$

and

$$R_\Lambda = 1 - \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{V}|} = 1 - \prod_{j=1}^{p} \lambda_j.$$

We can now prove that R_Λ fulfills the properties (i), (ii), (iii'), and (iv).

i. $0 \leq R_\Lambda \leq 1$

i.1 $R_\Lambda \geq 0$

We have seen already that $0 \leq \lambda_j \leq 1$ so that $R_\Lambda = 1 - \prod_{j=1}^{p} \lambda_j \geq 1 - \prod_{j=1}^{p} 1 = 0$. $\square$

i.2 $R_\Lambda \leq 1$ is obvious. $\square$

ii. $R_\Lambda = 0$ if and only if $\boldsymbol{V} = \boldsymbol{\Sigma}$

$$\begin{aligned} R_\Lambda = 1 - \prod_{j=1}^{p} \lambda_j = 0 \quad &\Leftrightarrow \quad \prod_{j=1}^{p} \lambda_j = 1 \\ &\Leftrightarrow \quad \lambda_j = 1 \quad \text{for all } j \\ &\Leftrightarrow \quad \boldsymbol{\Sigma} = \boldsymbol{V} \quad \square \end{aligned} \tag{A.1}$$

iii'. $R_\Lambda = 1$ if and only if $|\boldsymbol{\Sigma}| = 0$

$$\begin{aligned} R_\Lambda = 1 - \prod_{j=1}^{p} \lambda_j = 1 \quad &\Leftrightarrow \quad \prod_{j=1}^{p} \lambda_j = 0 \\ &\Leftrightarrow \quad \text{there exists } k \text{ so that } \lambda_k = 0 \\ &\Leftrightarrow \quad |\boldsymbol{\Sigma}| = 0. \quad \square \end{aligned}$$

iv. In the classical setting $R_\Lambda = \dfrac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$

In the classical setting $\boldsymbol{\Sigma} = \sigma^2$ and $\boldsymbol{V} = \sigma_b^2 + \sigma^2$ therefore:

$$R_\Lambda = 1 - \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{V}|} = 1 - \frac{\sigma^2}{\sigma_b^2 + \sigma^2} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}. \quad \square$$

Appendix B

Estimation and Asymptotic Confidence Intervals for the Reliability Measures

B.1 Details on the Calculation of an Asymptotic Confidence Interval for R_T

We will provide more details on the derivation of the elements of $\boldsymbol{\Delta}$. Let us first note that in case of a balanced design, and assuming that $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ and $\boldsymbol{V}_i = \boldsymbol{V}$, we have:

$$\begin{aligned} R_T &= \frac{\text{tr}(\boldsymbol{V}) - \text{tr}(\boldsymbol{\Sigma})}{\text{tr}(\boldsymbol{V})} \\ &= \frac{\text{tr}(\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \mathbf{T}\boldsymbol{H}\mathbf{T} + \boldsymbol{\Sigma}_R) - \text{tr}(\mathbf{T}\boldsymbol{H}\mathbf{T} + \boldsymbol{\Sigma}_R)}{\text{tr}(\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \mathbf{T}\boldsymbol{H}\mathbf{T} + \boldsymbol{\Sigma}_R)} \\ &= \frac{\text{tr}(\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}')}{\text{tr}(\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}') + \text{tr}(\mathbf{T}^2) + \text{tr}(\boldsymbol{\Sigma}_R)}. \end{aligned}$$

Note further that $\text{tr}(\mathbf{T}^2) = \sum_{j=1}^{p} \tau_j^2$ and $\text{tr}(\boldsymbol{\Sigma}_R) = \sum_{j=1}^{p} \sigma_{jj'}^2$. We can now derive the different elements of $\boldsymbol{\Delta}$.

In what follows we will calculate $\dfrac{\partial R_T}{\partial z}$, with z a scalar. For z an element of $\boldsymbol{D}$, we

find:

$$\begin{aligned}\frac{\partial R_T}{\partial z} &= \frac{\frac{\partial \mathrm{tr}(\boldsymbol{ZDZ'})}{\partial z}[\mathrm{tr}(\boldsymbol{ZDZ'}) + \mathrm{tr}(\mathbf{T}^2) + \mathrm{tr}(\boldsymbol{\Sigma}_R)] - \frac{\partial \mathrm{tr}(\boldsymbol{ZDZ'})}{\partial z}\mathrm{tr}(\boldsymbol{ZDZ'})}{[\mathrm{tr}(\boldsymbol{ZDZ'}) + \mathrm{tr}(\mathbf{T}^2) + \mathrm{tr}(\boldsymbol{\Sigma}_R)]^2} \\ &= \frac{[\mathrm{tr}(\mathbf{T}^2) + \mathrm{tr}(\boldsymbol{\Sigma}_R)]\frac{\partial \mathrm{tr}(\boldsymbol{ZDZ'})}{\partial z}}{[\mathrm{tr}(\boldsymbol{ZDZ'}) + \mathrm{tr}(\mathbf{T}^2) + \mathrm{tr}(\boldsymbol{\Sigma}_R)]^2},\end{aligned}$$

From Searle (1982) we know that

$$\frac{\partial \mathrm{tr}(\boldsymbol{XA})}{\partial \boldsymbol{X}} = \boldsymbol{A} + \boldsymbol{A'} - \mathrm{diag}(\boldsymbol{A}),$$

so that

$$\frac{\partial \mathrm{tr}(\boldsymbol{ZDZ'})}{\partial z} = \frac{\partial \mathrm{tr}(\boldsymbol{DZ'Z})}{\partial z} = \boldsymbol{Z'Z} + \boldsymbol{Z'Z} - \mathrm{diag}(\boldsymbol{Z'Z}) = 2\boldsymbol{Z'Z} - \mathrm{diag}(\boldsymbol{Z'Z}),$$

and therefore

$$\frac{\partial R_T}{\partial z} = \frac{[\mathrm{tr}(\mathbf{T}^2) + \mathrm{tr}(\boldsymbol{\Sigma}_R)][2\boldsymbol{Z'Z} - \mathrm{diag}(\boldsymbol{Z'Z})]}{[\mathrm{tr}(\boldsymbol{ZDZ'}) + \mathrm{tr}(\mathbf{T}^2) + \mathrm{tr}(\boldsymbol{\Sigma}_R)]^2}.$$

For z an element of $\mathbf{T}^2$, where $z = \tau_j^2$, the following expression is obtained:

$$\frac{\partial R_T}{\partial z} = -\frac{\mathrm{tr}(\boldsymbol{ZDZ'})}{[\mathrm{tr}(\boldsymbol{ZDZ'}) + \mathrm{tr}(\mathbf{T}^2) + \mathrm{tr}(\boldsymbol{\Sigma}_R)]^2}.$$

Finally, for z an element of $\boldsymbol{\Sigma}_R$, we obtain:

$$\frac{\partial R_T}{\partial z} = -\frac{\mathrm{tr}(\boldsymbol{ZDZ'})}{[\mathrm{tr}(\boldsymbol{ZDZ'}) + \mathrm{tr}(\mathbf{T}^2) + \mathrm{tr}(\boldsymbol{\Sigma}_R)]^2}.$$

Note that in practice it frequently occurs that $\tau_j^2 = \tau^2$ for all j, in that case the second and the third formulae simplify to:

$$\begin{aligned}\frac{\partial R_T}{\partial z} &= -\frac{p \quad \mathrm{tr}(\boldsymbol{ZDZ'})}{[\mathrm{tr}(\boldsymbol{ZDZ'}) + p\tau^2 + \mathrm{tr}(\boldsymbol{\Sigma}_R)]^2} \\ \frac{\partial R_T}{\partial z} &= -\frac{\mathrm{tr}(\boldsymbol{ZDZ'})}{[\mathrm{tr}(\boldsymbol{ZDZ'}) + p\tau^2 + \mathrm{tr}(\boldsymbol{\Sigma}_R)]^2}.\end{aligned}$$

B.2 Asymptotic Confidence Interval for the Elements of Ω

To derive a confidence interval for the elements of Ω we will first introduce some results from differential calculus for matrices.

Important results of differential calculus for matrices

Definition 1: Derivative of a matrix with respect to a scalar
Let $\boldsymbol{Y}$ be a $p \times q$ matrix of variables that are functions of z, so that y is a matrix function of z. Then the derivative of $\boldsymbol{Y}$ with respect to z is the $p \times q$ matrix:

$$\frac{\partial \boldsymbol{Y}}{\partial z} = \begin{pmatrix} \frac{\partial y_{11}}{\partial z} & \frac{\partial y_{12}}{\partial z} & \dots & \frac{\partial y_{1q}}{\partial z} \\ \frac{\partial y_{21}}{\partial z} & \frac{\partial y_{22}}{\partial z} & \dots & \frac{\partial y_{2q}}{\partial z} \\ \dots & \dots & \dots & \dots \\ \frac{\partial y_{p1}}{\partial z} & \frac{\partial y_{p2}}{\partial z} & \dots & \frac{\partial y_{pq}}{\partial z} \end{pmatrix}$$

Theorem 8 *Derivative of a product*
Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be $m \times n$ and $n \times r$ matrices of variables which depend on z. The derivative of $\boldsymbol{XY}$ with respect to z is the $m \times r$ matrix:

$$\frac{\partial}{\partial z}(\boldsymbol{XY}) = \frac{\partial \boldsymbol{X}}{\partial z}\boldsymbol{Y} + \boldsymbol{X}\frac{\partial \boldsymbol{Y}}{\partial z}.$$

Theorem 9 *Some important derivatives*

1. $\frac{\partial}{\partial z} \ln |\boldsymbol{Y}| = tr\left(\boldsymbol{Y}^{-1}\frac{\partial \boldsymbol{Y}}{\partial z}\right)$
2. $\frac{\partial}{\partial z} tr(\boldsymbol{XY}) = tr\left(\frac{\partial \boldsymbol{X}}{\partial z}\boldsymbol{Y}\right) + tr\left(\boldsymbol{X}\frac{\partial \boldsymbol{Y}}{\partial z}\right)$
3. $\frac{\partial}{\partial z}\boldsymbol{Y}^{-1} = -\boldsymbol{Y}^{-1}\frac{\partial \boldsymbol{Y}}{\partial z}\boldsymbol{Y}^{-1}$
4. $\frac{\partial}{\partial z}(\boldsymbol{AYB}) = \boldsymbol{A}\frac{\partial \boldsymbol{Y}}{\partial z}\boldsymbol{B}$

Asymptotic Confidence Interval for the Elements of Ω

Let θ be any member of Ω as defined in (9.1). Let $\boldsymbol{\psi}$ be the vector of the covariance parameters of a linear mixed-effects model. We know from ML theory that $\widehat{\boldsymbol{\psi}} \sim N(\boldsymbol{\psi}, \boldsymbol{\Sigma}_P)$ where $\boldsymbol{\Sigma}_P$ is the variance-covariance matrix of $\widehat{\boldsymbol{\psi}}$. Applying now the Delta method to $\widehat{\theta}$ we get: $\widehat{\theta} \sim N(\theta, \boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}')$ where $\boldsymbol{\Delta} = \dfrac{\partial\theta}{\partial\boldsymbol{\psi}}$. A $(1-\alpha)\%$ confidence interval for θ can then be given by

$$\left[\widehat{\theta} \pm z_{1-\frac{\alpha}{2}}\sqrt{\boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}'}\right].$$

To avoid confidence limits that exceed the $[0,1]$ range, a logit transformation is applied, with $l(\theta) = \log\left(\dfrac{\theta(\boldsymbol{\psi})}{1-\theta(\boldsymbol{\psi})}\right)$. A restricted $(1-\alpha)\%$ confidence interval for θ is then given by

$$\left[\frac{e^{l_1}}{1+e^{l_1}}, \frac{e^{l_2}}{1+e^{l_2}}\right],$$

with l_1 the lower limit and l_2 the upper limit of the confidence interval

$$\left[l(\widehat{\theta}) \pm \frac{z_{1-\frac{\alpha}{2}}}{\theta(1-\theta)}\sqrt{\boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}'}\right].$$

In the remainder, we provide more detailed information on the derivation of the different elements of $\boldsymbol{\Delta}$. Let us note that $\theta = 1 - \sum_j w_j\lambda_j$ and

$$|\boldsymbol{\Sigma} - \lambda\boldsymbol{V}| = 0 \quad \Leftrightarrow \quad |\boldsymbol{\Sigma}\boldsymbol{V}^{-1} - \lambda\boldsymbol{I}| = 0,$$

and therefore the λ_j are the eigenvalues of $\boldsymbol{\Sigma}\boldsymbol{V}^{-1}$. This implies that there exists a nonsingular matrix $\boldsymbol{P}$ so that $\boldsymbol{\Lambda} = \boldsymbol{P}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^{-1}\boldsymbol{P}$ with $\boldsymbol{\Lambda} = \text{diag}(\lambda_j)$. On the other hand,

$$\begin{aligned}
\theta &= 1 - \text{tr}(\boldsymbol{W}\boldsymbol{\Lambda}) \quad \text{where} \quad \boldsymbol{W} = \text{diag}(w_j) \\
&= 1 - \text{tr}(\boldsymbol{W}\boldsymbol{P}^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^{-1}\boldsymbol{P}) \\
\theta &= 1 - \text{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{V}^{-1}) \quad \text{where} \quad \boldsymbol{Q} = \boldsymbol{P}\boldsymbol{W}\boldsymbol{P}^{-1}.
\end{aligned}$$

In what follows we will calculate $\dfrac{\partial\theta}{\partial z}$, with z a scalar:

$$\frac{\partial\theta}{\partial z} = -\frac{\partial}{\partial z}\text{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{V}^{-1}).$$

Applying now (2) from Theorem 9 we get:

$$\begin{aligned}\frac{\partial\theta}{\partial z} &= -\frac{\partial}{\partial z}\mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{V}^{-1}) \\ &= -\mathrm{tr}\left(\frac{\partial\boldsymbol{Q}}{\partial z}\boldsymbol{\Sigma}\boldsymbol{V}^{-1}\right) - \mathrm{tr}\left(\boldsymbol{Q}\frac{\partial}{\partial z}(\boldsymbol{\Sigma}\boldsymbol{V}^{-1})\right) \\ \frac{\partial\theta}{\partial z} &= -\mathrm{tr}\left(\boldsymbol{Q}\frac{\partial}{\partial z}(\boldsymbol{\Sigma}\boldsymbol{V}^{-1})\right),\end{aligned}$$

but from the product rule in Theorem 8 we get:

$$\begin{aligned}\frac{\partial}{\partial z}(\boldsymbol{\Sigma}\boldsymbol{V}^{-1}) &= \frac{\partial\boldsymbol{\Sigma}}{\partial z}\boldsymbol{V}^{-1} + \boldsymbol{\Sigma}\frac{\partial\boldsymbol{V}^{-1}}{\partial z} \\ &= \frac{\partial\boldsymbol{\Sigma}}{\partial z}\boldsymbol{V}^{-1} + \boldsymbol{\Sigma}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial z}\boldsymbol{V}^{-1} \\ &= \left(\frac{\partial\boldsymbol{\Sigma}}{\partial z} - \boldsymbol{\Sigma}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial z}\right)\boldsymbol{V}^{-1},\end{aligned}$$

so that

$$\frac{\partial\theta}{\partial z} = \mathrm{tr}\left[\boldsymbol{Q}\left(\boldsymbol{\Sigma}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial z} - \frac{\partial\boldsymbol{\Sigma}}{\partial z}\right)\boldsymbol{V}^{-1}\right],$$

where:

$$\begin{aligned}\frac{\partial\boldsymbol{\Sigma}}{\partial z} &= \frac{\partial\mathbf{T}}{\partial z}\boldsymbol{H}\mathbf{T} + \mathbf{T}\frac{\partial\boldsymbol{H}}{\partial z}\mathbf{T} + \mathbf{T}\boldsymbol{H}\frac{\partial\mathbf{T}}{\partial z} + \frac{\partial\boldsymbol{\Sigma}_R}{\partial z} \\ \frac{\partial\boldsymbol{V}}{\partial z} &= \boldsymbol{Z}\frac{\partial\boldsymbol{D}}{\partial z}\boldsymbol{Z}' + \frac{\partial\boldsymbol{\Sigma}}{\partial z}.\end{aligned}$$

Now, we will give the expression for the derivative of θ with respect to the different parameters, coming from $\boldsymbol{D}$, $\mathbf{T}$, $\boldsymbol{H}$, and $\boldsymbol{\Sigma}_R$. For z an element of $\boldsymbol{D}$, we find:

$$\frac{\partial\theta}{\partial z} = \mathrm{tr}\left[\boldsymbol{V}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{V}^{-1}\left(\boldsymbol{Z}\frac{\partial\boldsymbol{D}}{\partial z}\boldsymbol{Z}'\right)\right].$$

For z an element of $\mathbf{T}$, or $z = \tau_j$, we have:

$$\frac{\partial\theta}{\partial\tau_j} = \mathrm{tr}\left[\boldsymbol{V}^{-1}\boldsymbol{Q}\left(\boldsymbol{\Sigma}\boldsymbol{V}^{-1} - \boldsymbol{I}\right)\frac{\partial\boldsymbol{\Sigma}}{\partial\tau_j}\right]$$

and $\dfrac{\partial\boldsymbol{\Sigma}}{\partial\tau_j} = \boldsymbol{I}_j\boldsymbol{H}\mathbf{T} + \mathbf{T}\boldsymbol{H}\boldsymbol{I}_j$ where $\boldsymbol{I}_j$ is a matrix with zeros everywhere and 1 in the position (j,j). Further, $\dfrac{\partial\theta}{\partial\tau_j^2} = -\dfrac{\partial\theta}{\partial\tau_j}\dfrac{1}{\tau_j^4}$.
For z an element of $\boldsymbol{H}$ we have:

$$\frac{\partial\theta}{\partial z} = \mathrm{tr}\left[\boldsymbol{V}^{-1}\boldsymbol{Q}\left(\boldsymbol{\Sigma}\boldsymbol{V}^{-1} - \boldsymbol{I}\right)\mathbf{T}\frac{\partial\boldsymbol{H}}{\partial z}\mathbf{T}\right],$$

and finally, for z an element of $\boldsymbol{\Sigma}_R$:

$$\frac{\partial \theta}{\partial z} = \operatorname{tr}\left[\boldsymbol{V}^{-1}\boldsymbol{Q}\left(\boldsymbol{\Sigma}\boldsymbol{V}^{-1} - \boldsymbol{I}\right)\frac{\partial \boldsymbol{\Sigma}_R}{\partial z}\right].$$

Note that in practice it frequently occurs that $\tau_j^2 = \tau^2$ for all j, in that case for $z = \tau^2$ we find:

$$\frac{\partial \theta}{\partial \tau^2} = \operatorname{tr}\left[\boldsymbol{V}^{-1}\boldsymbol{Q}\left(\boldsymbol{\Sigma}\boldsymbol{V}^{-1} - \boldsymbol{I}\right)\boldsymbol{H}\right],$$

and for z an element of $\boldsymbol{\Sigma}_R$:

$$\frac{\partial \theta}{\partial z} = \operatorname{tr}\left[\boldsymbol{V}^{-1}\boldsymbol{Q}\left(\boldsymbol{\Sigma}\boldsymbol{V}^{-1} - \boldsymbol{I}\right)\tau^2\frac{\partial \boldsymbol{H}}{\partial z}\right].$$

Finally we will give details on the derivatives of the serial correlations $\dfrac{\partial \boldsymbol{H}}{\partial z}$, for some of the most commonly used serial correlation structures in models for longitudinal data, being the autoregressive, spatial power, spatial exponential, and spatial gaussian structure. These structures have in common that measurements taken closer in time are more strongly correlated than measurements taken further apart, a very common phenomenon in longitudinal measurements.

Note that the autoregressive structure is a special case of the spatial power structure. When the measurements are equally spaced, the spatial power structure reduces to the autoregressive structure. Both structures can be written as $\boldsymbol{H} = (\rho^{d_{st}})$, where ρ is a correlation parameter and d_{st} is the distance between two measurements at times s and t. Then:

$$\frac{\partial \boldsymbol{H}}{\partial \rho} = \frac{1}{\rho}\boldsymbol{A} \odot \boldsymbol{H} \quad \text{where} \quad \boldsymbol{A} \odot \boldsymbol{H} = (c_{st}) = (a_{st}h_{st})$$

sometimes referred to as the Hadamard product, and

$$\boldsymbol{A} = \frac{1}{\ln \rho}\ln(\boldsymbol{H}) \quad \text{with} \quad \ln(\boldsymbol{H}) = (\ln h_{st}).$$

An exponential correlation structure can be written as $\boldsymbol{H} = \left(\exp\left(\dfrac{-d_{st}}{\phi}\right)\right)$, so that:

$$\frac{\partial \boldsymbol{H}}{\partial \phi} = \frac{1}{\phi^2}\boldsymbol{A} \odot \boldsymbol{H} \quad \text{and} \quad \boldsymbol{A} = -\phi \ln(\boldsymbol{H}).$$

Finally, a spatial Gaussian correlation can be written as $\boldsymbol{H} = \left(\exp\left(\dfrac{-d_{st}^2}{\rho^2}\right)\right)$, and thus:

$$\frac{\partial \boldsymbol{H}}{\partial \rho} = \frac{2}{\rho^3}\boldsymbol{A} \odot \boldsymbol{H} \quad \text{and} \quad \boldsymbol{A} = -\rho^2 \ln(\boldsymbol{H}).$$

B.3 Estimation and Asymptotic Confidence Interval for R_Λ

If $\hat{\boldsymbol{D}}$, $\hat{\mathbf{T}}$, $\hat{\boldsymbol{H}}$, $\hat{\boldsymbol{\Sigma}}_R$ denote the MLEs for $\boldsymbol{D}$, $\mathbf{T}$, $\boldsymbol{H}$, and $\boldsymbol{\Sigma}_R$, as defined in (5.1), respectively then the MLE for R_Λ is given by

$$\hat{R}_\Lambda = 1 - |\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{-1}|,$$

where $\hat{\boldsymbol{V}} = \boldsymbol{Z}\hat{\boldsymbol{D}}\boldsymbol{Z}' + \hat{\mathbf{T}}\hat{\boldsymbol{H}}\hat{\mathbf{T}} + \hat{\boldsymbol{\Sigma}}_R$ and $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{T}}\hat{\boldsymbol{H}}\hat{\mathbf{T}} + \hat{\boldsymbol{\Sigma}}_R$.

We will use once more the delta method to obtain a confidence interval for R_Λ. Let $\boldsymbol{\psi}$ be the vector of the covariance parameters of a linear mixed-effects model, with $\widehat{\boldsymbol{\psi}} \sim N(\boldsymbol{\psi}, \boldsymbol{\Sigma}_P)$. Then: $\widehat{R}_\Lambda \sim N(R_\Lambda, \boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}')$ where $\boldsymbol{\Delta} = \dfrac{\partial R_\Lambda}{\partial \boldsymbol{\psi}}$. A $(1-\alpha)\%$ confidence interval for R_Λ can then be given by

$$\left[\widehat{R}_\Lambda \pm z_{1-\frac{\alpha}{2}}\sqrt{\boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}'}\right].$$

As for previous measures, we use a logit transformation to avoid confidence limits that exceed the $[0,1]$ range. with $l(R_\Lambda) = \log\left(\dfrac{R_\Lambda(\boldsymbol{\psi})}{1-R_\Lambda(\boldsymbol{\psi})}\right)$. A restricted $(1-\alpha)\%$ confidence interval for R_Λ is then given by

$$\left[\frac{e^{l_1}}{1+e^{l_1}}, \frac{e^{l_2}}{1+e^{l_2}}\right],$$

with l_1 the lower limit and l_2 the upper limit of the confidence interval

$$\left[l(\widehat{R}_\Lambda) \pm \frac{z_{1-\frac{\alpha}{2}}}{R_\Lambda(1-R_\Lambda)}\sqrt{\boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}'}\right].$$

In what follows we will give more detailed information on the derivation of the different elements of $\boldsymbol{\Delta}$. We will calculate the $\dfrac{\partial R_\Lambda}{\partial z}$ with z a scalar. Let us first note that

$$\begin{aligned} & |\boldsymbol{\Sigma}\boldsymbol{V}^{-1}| = 1 - R_\Lambda \\ \Leftrightarrow \quad & \ln|\boldsymbol{\Sigma}| - \ln|\boldsymbol{V}| = \ln(1-R_\Lambda) = \gamma. \end{aligned}$$

For simplicity we will calculate $\dfrac{\partial \gamma}{\partial z}$. Notice that:

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

and therefore:

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

where

$$\frac{\partial \gamma}{\partial z} = \frac{\partial}{\partial z}\ln|\boldsymbol{\Sigma}| - \frac{\partial}{\partial z}\ln|\boldsymbol{V}|.$$

If we call $\gamma_1 = \ln|\boldsymbol{\Sigma}|$ and $\gamma_2 = \ln|\boldsymbol{V}|$, then:

$$\frac{\partial \gamma}{\partial z} = \frac{\partial \gamma_1}{\partial z} - \frac{\partial \gamma_2}{\partial z}.$$

From (1) in Theorem 9 we have:

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

but $\boldsymbol{\Sigma} = \mathbf{T}\boldsymbol{H}\mathbf{T} + \boldsymbol{\Sigma}_R$ and therefore

$$\frac{\partial \boldsymbol{\Sigma}}{\partial z} = \frac{\partial \mathbf{T}}{\partial z}\boldsymbol{H}\mathbf{T} + \mathbf{T}\frac{\partial \boldsymbol{H}}{\partial z}\mathbf{T} + \mathbf{T}\boldsymbol{H}\frac{\partial \mathbf{T}}{\partial z} + \frac{\partial \boldsymbol{\Sigma}_R}{\partial z}.$$

To calculate the derivative of $\mathbf{T}\boldsymbol{H}\mathbf{T}$ we have applied the product rule of Theorem 8. On the other hand:

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

but $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{\Sigma}$ and therefore:

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

$$\Rightarrow \quad \frac{\partial \gamma_2}{\partial z} = \mathrm{tr}\left[\boldsymbol{V}^{-1}\left(\boldsymbol{Z}\frac{\partial \boldsymbol{D}}{\partial z}\boldsymbol{Z}'\right) + \boldsymbol{V}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial z}\right],$$

and finally

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

where

$$\frac{\partial \boldsymbol{\Sigma}}{\partial z} = \frac{\partial \mathbf{T}}{\partial z}\boldsymbol{H}\mathbf{T} + \mathbf{T}\frac{\partial \boldsymbol{H}}{\partial z}\mathbf{T} + \mathbf{T}\boldsymbol{H}\frac{\partial \mathbf{T}}{\partial z} + \frac{\partial \boldsymbol{\Sigma}_R}{\partial z}.$$

Now, we will give the expression for the derivative of R_Λ with respect to the different parameters, coming from $\boldsymbol{D}$, $\mathbf{T}$, $\boldsymbol{H}$, and $\boldsymbol{\Sigma}_R$. For z an element of $\boldsymbol{D}$, we find:

$$\frac{\partial R_\Lambda}{\partial z} = (1 - R_\Lambda)\mathrm{tr}\left[\boldsymbol{V}^{-1}\left(\boldsymbol{Z}\frac{\partial \boldsymbol{D}}{\partial z}\boldsymbol{Z}'\right)\right].$$

For z an element of $\mathbf{T}$, or $z = \tau_j$, we have:

$$\frac{\partial R_\Lambda}{\partial \tau_j} = (1 - R_\Lambda)\text{tr}\left[(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}^{-1})\frac{\partial \boldsymbol{\Sigma}}{\partial \tau_j}\right]$$

and $\dfrac{\partial \boldsymbol{\Sigma}}{\partial \tau_j} = \boldsymbol{I}_j\boldsymbol{H}\mathbf{T} + \mathbf{T}\boldsymbol{H}\boldsymbol{I}_j$ where $\boldsymbol{I}_j$ is a matrix with zeros everywhere and 1 in the position (j, j). Further, $\dfrac{\partial R_\Lambda}{\partial \tau_j^2} = -\dfrac{\partial R_\Lambda}{\partial \tau_j}\dfrac{1}{\tau_j^4}$.
For z an element of $\boldsymbol{H}$ we have:

$$\frac{\partial R_\Lambda}{\partial z} = (1 - R_\Lambda)\text{tr}\left[(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}^{-1})\mathbf{T}\frac{\partial \boldsymbol{H}}{\partial z}\mathbf{T}\right],$$

and finally, for z an element of $\boldsymbol{\Sigma}_R$:

$$\frac{\partial R_\Lambda}{\partial z} = (1 - R_\Lambda)\text{tr}\left[(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}^{-1})\frac{\partial \boldsymbol{\Sigma}_R}{\partial z}\right].$$

Note that in practice it frequently occurs that $\tau_j^2 = \tau^2$ for all j, in that case the second for $z = \tau^2$:

$$\frac{\partial R_\Lambda}{\partial \tau^2} = (1 - R_\Lambda)\text{tr}\left[(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}^{-1})\boldsymbol{H})\right]$$

and for z an element of $\boldsymbol{\Sigma}_R$:

$$\frac{\partial R_\Lambda}{\partial z} = (1 - R_\Lambda)\text{tr}\left[(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}^{-1})\tau^2\frac{\partial \boldsymbol{H}}{\partial z}\right].$$

B.4 Asymptotic Confidence Intervals for R_T, R_Λ and $\rho(\boldsymbol{a})$ in the Single-Administration Context

Maximum likelihood estimates for R_T, R_Λ and $\rho(\boldsymbol{a})$ can be obtained by filling in the MLE for $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ in, respectively, (7.2), (10.1) and (14.4). Confidence intervals for R_T and R_Λ can be obtained using the delta method, in the same way as explained in sections 7.3 and B.3 in the longitudinal context. One additional element, however, needs to be taken into account. In model (5.1) $\boldsymbol{Z}$ is a fixed design matrix. The corresponding matrix $\boldsymbol{B}$ in model (14.1) is a matrix of estimated factor loadings. This needs to be taken into account in the calculation of $\boldsymbol{\Delta}$. In this section we will address this point for the calculation of a confidence interval for R_T and R_Λ. We will further derive a confidence interval for $\rho(\boldsymbol{a})$.

Obtaining $\boldsymbol{\Delta}$ for R_T

We will derive the elements of $\boldsymbol{\Delta}$ for R_T, with $\boldsymbol{\Delta}' = \dfrac{\partial R_T}{\partial \boldsymbol{\psi}}$ with $\boldsymbol{\psi}$ a vector containing all parameters in $\boldsymbol{B}$, $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$.

$$R_T = \frac{\text{tr}(\boldsymbol{BDB}')}{\text{tr}(\boldsymbol{V})} = \frac{\text{tr}(\boldsymbol{BDB}')}{\text{tr}(\boldsymbol{BDB}') + \text{tr}(\boldsymbol{\Sigma})}.$$

Therefore, in general, with z a scalar:

$$\frac{\partial R_T}{\partial z} = \frac{\frac{\partial \text{tr}(\boldsymbol{BDB}')}{\partial z}\text{tr}(\boldsymbol{V}) - \frac{\partial(\text{tr}(\boldsymbol{BDB}')+\text{tr}(\boldsymbol{\Sigma}))}{\partial z}\text{tr}(\boldsymbol{BDB}')}{\text{tr}(\boldsymbol{V})^2}. \tag{B.1}$$

Let us start by deriving $\dfrac{\partial R_T}{\partial b}$, with b an element of $\boldsymbol{B}$. From (B.1) we obtain:

$$\begin{aligned} \frac{\partial R_T}{\partial b} &= \frac{\text{tr}(\boldsymbol{\Sigma})\frac{\partial \text{tr}(\boldsymbol{BDB}')}{\partial b}}{\text{tr}(\boldsymbol{V})^2} \\ \Rightarrow \qquad \frac{\partial R_T}{\partial b} &= \left(\frac{\text{tr}(\boldsymbol{\Sigma})}{\text{tr}(\boldsymbol{V})^2}\right)\frac{\partial \text{tr}(\boldsymbol{BDB}')}{\partial b} \end{aligned} \tag{B.2}$$

To calculate (B.2), we need to calculate $\dfrac{\partial \text{tr}(\boldsymbol{BDB}')}{\partial b}$. Let us first note that:

$$\text{tr}(\boldsymbol{BDB}') = \text{tr}(\boldsymbol{B}'\boldsymbol{BD}) = \text{tr}(\tilde{\boldsymbol{B}}\boldsymbol{D}) \quad \text{with} \quad \tilde{\boldsymbol{B}} = \boldsymbol{B}'\boldsymbol{B}.$$

Therefore,

$$\begin{aligned} \frac{\partial \text{tr}(\boldsymbol{BDB}')}{\partial b} &= \frac{\partial \text{tr}(\tilde{\boldsymbol{B}}\boldsymbol{D})}{\partial b} \\ &= \text{tr}\left(\frac{\partial \tilde{\boldsymbol{B}}}{\partial b}\boldsymbol{D} + \tilde{\boldsymbol{B}}\frac{\partial \boldsymbol{D}}{\partial b}\right) \\ \Rightarrow \frac{\partial \text{tr}(\boldsymbol{BDB}')}{\partial b} &= \text{tr}\left(\frac{\partial \tilde{\boldsymbol{B}}}{\partial b}\boldsymbol{D}\right) \end{aligned} \tag{B.3}$$

To obtain (B.3) we have applied formula 2 under Theorem 9. Further, we need to

calculate $\frac{\partial \tilde{\boldsymbol{B}}}{\partial b}$. Applying Theorem 8 we obtain:

$$\begin{aligned} & \frac{\partial(\boldsymbol{B}'\boldsymbol{B})}{\partial b} = \frac{\partial \boldsymbol{B}'}{\partial b}\boldsymbol{B} + \boldsymbol{B}'\frac{\partial \boldsymbol{B}}{\partial b} \\ \Rightarrow \quad & \frac{\partial \mathrm{tr}(\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}')}{\partial b} = \mathrm{tr}\left(\frac{\partial \boldsymbol{B}'}{\partial b}\boldsymbol{B}\boldsymbol{D} + \boldsymbol{B}'\frac{\partial \boldsymbol{B}}{\partial b}\boldsymbol{D}\right) \\ \Rightarrow \quad & \frac{\partial \mathrm{tr}(\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}')}{\partial b} = \mathrm{tr}\left(\frac{\partial \boldsymbol{B}'}{\partial b}\boldsymbol{B}\boldsymbol{D}\right) + \mathrm{tr}\left(\boldsymbol{B}'\frac{\partial \boldsymbol{B}}{\partial b}\boldsymbol{D}\right). \end{aligned} \tag{B.4}$$

If we denote

$$\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1q} \\ b_{21} & b_{22} & \dots & b_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ b_{p1} & b_{p2} & \dots & b_{pq} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}' = \begin{pmatrix} b_{11} & b_{21} & \dots & b_{p1} \\ b_{12} & b_{22} & \dots & b_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ b_{1q} & b_{2q} & \dots & b_{pq} \end{pmatrix}$$

then if $b = b_{ij}$, $\frac{\partial \boldsymbol{B}}{\partial b_{ij}}$ is a $p \times q$ matrix with all elements 0 except the element (i, j). Further $\frac{\partial \boldsymbol{B}'}{\partial b_{ij}} = \left(\frac{\partial \boldsymbol{B}}{\partial b_{ij}}\right)'$.

Finally, to summarize:

$$\begin{aligned} & \frac{\partial R_T}{\partial b_{ij}} = \left(\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{V})^2}\right)\mathrm{tr}\left(\frac{\partial \boldsymbol{B}'}{\partial b_{ij}}\boldsymbol{B}\boldsymbol{D} + \boldsymbol{B}'\frac{\partial \boldsymbol{B}}{\partial b_{ij}}\boldsymbol{D}\right) \\ & \frac{\partial \boldsymbol{B}}{\partial b_{ij}} = \{c_{i'j'}\}_{p\times q} \quad \text{with} \quad c_{ij} = \begin{cases} 0 & \text{if} \quad i' \neq i \quad \text{and} \quad j' \neq j \\ 1 & \text{if} \quad i' = i \quad \text{and} \quad j' = j \end{cases} \\ & \frac{\partial \boldsymbol{B}'}{\partial b_{ij}} = \left(\frac{\partial \boldsymbol{B}}{\partial b_{ij}}\right)' \end{aligned} \tag{B.5}$$

Further, from (B.1), if d_{ij} an element of $\boldsymbol{D}$ and σ_{ij} an element of $\boldsymbol{\Sigma}$, we can obtain:

$$\frac{\partial R_T}{\partial d_{ij}} = \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{\mathrm{tr}(\boldsymbol{V})^2}\mathrm{tr}\left(\boldsymbol{B}'\boldsymbol{B}\frac{\partial \boldsymbol{D}}{\partial d_{ij}}\right)$$

and

$$\frac{\partial R_T}{\partial \sigma_{ij}} = \frac{-\mathrm{tr}\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_{ij}}\right)\mathrm{tr}(\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}')}{\mathrm{tr}(\boldsymbol{V})^2}$$

with $\frac{\partial \boldsymbol{D}}{\partial d_{ij}}$ and $\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_{ij}}$ analogous to (B.5).

Obtaining $\boldsymbol{\Delta}$ for R_Λ

In this section we will derive the elements of $\boldsymbol{\Delta}$ for R_Λ. Analogous to Section B.3, we will calculate $\frac{\partial \gamma}{\partial z}$ with $\gamma = \ln(1 - R_\Lambda) = \ln|\boldsymbol{\Sigma}| - \ln|\boldsymbol{V}|$, so that

$$\frac{\partial \gamma}{\partial z} = \frac{\partial \ln|\boldsymbol{\Sigma}|}{\partial z} - \frac{\partial \ln|\boldsymbol{V}|}{\partial z}.$$

If we call $\gamma_1 = \ln|\boldsymbol{\Sigma}|$ and $\gamma_2 = \ln|\boldsymbol{V}|$ then $\frac{\partial \gamma}{\partial z} = \frac{\partial \gamma_1}{\partial z} - \frac{\partial \gamma_2}{\partial z}$ and

$$\begin{aligned} \frac{\partial \gamma_1}{\partial z} &= \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial z}\right) \\ \frac{\partial \gamma_2}{\partial z} &= \operatorname{tr}\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial z}\right). \end{aligned}$$

But $\boldsymbol{V} = \boldsymbol{BDB}' + \boldsymbol{\Sigma}$ and therefore:

$$\begin{aligned} \frac{\partial \boldsymbol{V}}{\partial z} &= \frac{\partial \boldsymbol{BDB}'}{\partial z} + \frac{\partial \boldsymbol{\Sigma}}{\partial z} \\ &= \frac{\partial \boldsymbol{B}}{\partial z}\boldsymbol{DB}' + \boldsymbol{B}\frac{\partial \boldsymbol{D}}{\partial z}\boldsymbol{B}' + \boldsymbol{BD}\frac{\partial \boldsymbol{B}'}{\partial z} + \frac{\partial \boldsymbol{\Sigma}}{\partial z} \\ \Rightarrow \frac{\partial \gamma_2}{\partial z} &= \operatorname{tr}\left[\boldsymbol{V}^{-1}\left(\frac{\partial \boldsymbol{B}}{\partial z}\boldsymbol{DB}' + \boldsymbol{B}\frac{\partial \boldsymbol{D}}{\partial z}\boldsymbol{B}' + \boldsymbol{BD}\frac{\partial \boldsymbol{B}'}{\partial z} + \frac{\partial \boldsymbol{\Sigma}}{\partial z}\right)\right] \end{aligned}$$

Further we can write:

$$\begin{aligned} -\frac{\partial \gamma}{\partial z} &= -\frac{\partial \gamma_1}{\partial z} + \frac{\partial \gamma_2}{\partial z} \\ &= \operatorname{tr}\left[\boldsymbol{V}^{-1}\left(\frac{\partial \boldsymbol{B}}{\partial z}\boldsymbol{DB}' + \boldsymbol{B}\frac{\partial \boldsymbol{D}}{\partial z}\boldsymbol{B}' + \boldsymbol{BD}\frac{\partial \boldsymbol{B}'}{\partial z}\right) + \left(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}^{-1}\right)\frac{\partial \Sigma}{\partial z}\right] \end{aligned}$$

and since

$$\frac{\partial \gamma}{\partial z} = \frac{\partial \ln(1 - R_\Lambda)}{\partial z} = \frac{-1}{1 - R_\Lambda}\frac{\partial R_\Lambda}{\partial z} \Rightarrow -\frac{\partial \gamma}{\partial z} = \frac{1}{1 - R_\Lambda}\frac{\partial R_\Lambda}{\partial z}$$

we obtain the general formula:

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

For the elements d_{ij}, b_{ij} and σ_{ij} of the matrices $\boldsymbol{D}$, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, respectively, we find the following formulas:

$$\begin{aligned}
\frac{\partial R_\Lambda}{\partial d_{ij}} &= (1-R_\Lambda)\mathrm{tr}\left[\boldsymbol{V}^{-1}\left(\boldsymbol{B}\frac{\partial \boldsymbol{D}}{\partial d}\boldsymbol{B}'\right)\right] \\
\frac{\partial R_\Lambda}{\partial b_{ij}} &= (1-R_\Lambda)\mathrm{tr}\left[\boldsymbol{V}^{-1}\left(\frac{\partial \boldsymbol{B}}{\partial b}\boldsymbol{D}\boldsymbol{B}' + \boldsymbol{B}\boldsymbol{D}\frac{\partial \boldsymbol{B}'}{\partial b}\right)\right] \\
\frac{\partial R_\Lambda}{\partial \sigma_{ij}} &= (1-R_\Lambda)\mathrm{tr}\left[\left(\boldsymbol{V}^{-1}-\boldsymbol{\Sigma}^{-1}\right)\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma}\right]
\end{aligned}$$

with $\frac{\partial \boldsymbol{B}'}{\partial b_{ij}} = \left(\frac{\partial \boldsymbol{B}}{\partial b_{ij}}\right)'$ and $\frac{\partial \boldsymbol{B}}{\partial b_{ij}}$ calculated as in (B.5). Further, $\frac{\partial \boldsymbol{D}}{\partial d_{ij}}$ and $\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_{ij}}$ are obtained analogous to (B.5).

A confidence interval for $\rho(\boldsymbol{a})$

Given a cross-sectional measurement of a multi-item scale, the reliability of the sum score of the scale can be derived by

$$R_T(\boldsymbol{a}) = R_\Lambda(\boldsymbol{a}) = \rho(\boldsymbol{a}) = 1 - \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}}.$$

In this section we will obtain a confidence interval for this measure, using the delta method. According to this method we have:

$$\hat{\rho}(\boldsymbol{a}) \sim N(\rho(\boldsymbol{a}), \boldsymbol{\Delta}\boldsymbol{\Sigma}_P\boldsymbol{\Delta}'),$$

where $\boldsymbol{\Sigma}_P$ is the variance covariance matrix of the parameter estimates and $\boldsymbol{\Delta}' = \frac{\partial \rho(\boldsymbol{a})}{\partial \boldsymbol{\psi}}$ with $\boldsymbol{\psi}$ a vector containing all parameters in $\boldsymbol{B}$, $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$. In a similar way as previously, a logit transformation can be applied to avoid that confidence limits exceed the $[0,1]$ range. We will now derive the elements of $\boldsymbol{\Delta}$. For z a scalar that can be any covariance parameter, we can write the general form as follows:

$$\frac{\partial \rho(\boldsymbol{a})}{\partial z} = \frac{\frac{\partial(\boldsymbol{a}'\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'\boldsymbol{a})}{\partial z}(\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}) - \frac{\partial(\boldsymbol{a}'(\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'+\boldsymbol{\Sigma})\boldsymbol{a})}{\partial z}\boldsymbol{a}'\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'\boldsymbol{a}}{[\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}]^2}$$

For the elements d_{ij}, b_{ij} and σ_{ij}, elements of $\boldsymbol{D}$, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ respectively, we obtain the following specific forms:

$$\begin{aligned}
\frac{\rho(\boldsymbol{a})}{\partial b_{ij}} &= \frac{\boldsymbol{a}'\left[\frac{\partial \boldsymbol{B}}{\partial b_{ij}}\boldsymbol{D}\boldsymbol{B}' + \boldsymbol{B}\boldsymbol{D}\frac{\partial \boldsymbol{B}'}{\partial b_{ij}}\right]\boldsymbol{a}\,(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})}{[\boldsymbol{a}'V\boldsymbol{a}]^2} \\
\frac{\rho(\boldsymbol{a})}{\partial d_{ij}} &= \frac{\boldsymbol{a}'\left[\boldsymbol{B}\frac{\partial \boldsymbol{D}}{\partial d_{ij}}\boldsymbol{B}'\right]\boldsymbol{a}\,(\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})}{[\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}]^2} \\
\frac{\rho(\boldsymbol{a})}{\partial \sigma_{ij}} &= \frac{-\boldsymbol{a}'\left[\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_{ij}}\boldsymbol{a}'\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}'\boldsymbol{a}\right]\boldsymbol{a}}{[\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}]^2}
\end{aligned}$$

with $\frac{\partial \boldsymbol{B}'}{\partial b_{ij}} = \left(\frac{\partial \boldsymbol{B}}{\partial b_{ij}}\right)'$ and $\frac{\partial \boldsymbol{B}}{\partial b_{ij}}$ calculated as in (B.5). Further, $\frac{\partial \boldsymbol{D}}{\partial d_{ij}}$ and $\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_{ij}}$ are obtained analogously.

Appendix C

Proofs of Theorems

C.1 Proof of Theorem 1

(i) all roots of $q(\lambda) = 0$, the so-called generalized eigenvalues, are real.

Note first that

$$\begin{aligned} & |\boldsymbol{\Sigma} - \lambda \boldsymbol{V}| = 0 \\ \Leftrightarrow \quad & |\boldsymbol{V}^{-1/2}||\boldsymbol{\Sigma} - \lambda \boldsymbol{V}||\boldsymbol{V}^{-1/2}| = 0 \\ \Leftrightarrow \quad & |\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2} - \lambda \boldsymbol{I}| = 0 \\ \Leftrightarrow \quad & |\boldsymbol{H} - \lambda \boldsymbol{I}| = 0. \end{aligned}$$

The previous equation implies that the generalized eigenvalues associated with the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are just the eigenvalues of the matrix $\boldsymbol{H}$. Finally, one only needs to notice that matrix $\boldsymbol{H}$ is symmetric and, therefore, all its eigenvalues are real.□

(ii) if λ_j is a root of $q(\lambda) = 0$ then $0 \leq \lambda_j \leq 1$.

We will now show that $0 \leq \lambda_j \leq 1$ for all j. Note that $\lambda_j > 0$ is an immediate consequence of (9.2). Indeed, to show that let us assume without loss of generality that $\lambda_1 < 0$, then if $\boldsymbol{e_1} = (1, 0, ..., 0)'$ it follows that $\boldsymbol{e_1'}\boldsymbol{\Lambda}\boldsymbol{e_1} = \lambda_1 < 0$. Further, $\boldsymbol{e_1'}\boldsymbol{\Lambda}\boldsymbol{e_1} = \boldsymbol{e_1'}(\boldsymbol{Q}^{-1})'\boldsymbol{\Sigma}(\boldsymbol{Q}^{-1})\boldsymbol{e_1}$ and if $\boldsymbol{y} = (\boldsymbol{Q})^{-1}\boldsymbol{e_1}$ then $\lambda_1 = \boldsymbol{y}'\boldsymbol{\Sigma}\boldsymbol{y} < 0$. Nevertheless, as a variance-covariance matrix, $\boldsymbol{\Sigma}$ needs to be positive-definite and, therefore, $\boldsymbol{y}\boldsymbol{\Sigma}\boldsymbol{y}' > 0$ for all $\boldsymbol{y}$, thus we have a contradiction and λ_1 cannot be smaller than zero.

Additionally we have

$$\boldsymbol{V} = \boldsymbol{\Sigma}_D + \boldsymbol{\Sigma} \quad \Leftrightarrow \quad (\boldsymbol{Q}')^{-1}\boldsymbol{V}(\boldsymbol{Q})^{-1} = (\boldsymbol{Q}')^{-1}\boldsymbol{\Sigma}_D(\boldsymbol{Q})^{-1} + (\boldsymbol{Q}')^{-1}\boldsymbol{\Sigma}(\boldsymbol{Q})^{-1}.$$

Moreover, from (9.2) and (9.3) we get

$$\boldsymbol{I} - \boldsymbol{\Lambda} = (\boldsymbol{Q}^{-1})'\boldsymbol{\Sigma}_D(\boldsymbol{Q}^{-1}) = \boldsymbol{RDR}' \quad \text{with} \quad \boldsymbol{R} = (\boldsymbol{Q}^{-1})'\boldsymbol{Z}.$$

This implies

$$1 - \lambda_j = \boldsymbol{r_j D r_j}' \geq 0$$

where $\boldsymbol{r_j}$ is the jth row of $\boldsymbol{R}$ and therefore $\lambda_j \leq 1$.□

C.2 Proof of Theorem 4

If $\theta \in \Omega$, then there exists a vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_p)'$ with $w_j > 0$ for all j and $\mathbf{1}'\boldsymbol{w} = 1$ so that $\theta = 1 - \sum_j w_j\lambda_j$. In what follows, we will show that there is at least a vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_p)'$ so that $w_j = \dfrac{\delta_j^2}{\sum_j \delta_j^2}$. Thus, we essentially need to solve the system of equations

$$\left\{\begin{array}{rcl} \delta_1^2 + \delta_2^2 + \cdots + \delta_p^2 &=& \frac{\delta_1^2}{w_1}, \\ \delta_1^2 + \delta_2^2 + \cdots + \delta_p^2 &=& \frac{\delta_2^2}{w_2}, \\ \vdots & & \vdots \\ \delta_1^2 + \delta_2^2 + \cdots + \delta_p^2 &=& \frac{\delta_p^2}{w_p}. \end{array}\right. \tag{C.1}$$

It is easy to see that the previous system of equations does not have a unique solution but rather an infinite number. Indeed, if we let δ_1^2 be a positive number, then

$$\delta_j^2 = \left\{\begin{array}{ll} \delta_1^2 & \text{if} \quad j = 1 \\ \delta_1^2\left(\frac{w_j}{w_1}\right) & \text{if} \quad j = 2, \ldots, p, . \end{array}\right. \tag{C.2}$$

are solutions of (C.1) and therefore

$$\theta = 1 - \sum_{j=1}^{p} \frac{\delta_j^2}{\sum \delta_j^2}\lambda_j = 1 - \frac{\sum_j \delta_j^2\lambda_j}{\sum_j \delta_j^2} = 1 - \frac{\boldsymbol{\delta}'\boldsymbol{\Lambda}\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{\delta}}, \tag{C.3}$$

with $\boldsymbol{\Lambda} = \text{diag}(\lambda_j)$. If we now define $\boldsymbol{a} = \boldsymbol{Q}^{-1}\boldsymbol{\delta}$ with $\boldsymbol{Q}$ like in (14.5)–(14.6), then $\boldsymbol{\delta} = \boldsymbol{Qa}$ and

$$\frac{\boldsymbol{\delta}'\boldsymbol{\Lambda}\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{\delta}} = \frac{\boldsymbol{a}'\boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}},$$

and therefore

$$\theta = 1 - \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = \rho(\boldsymbol{a}),$$

i.e., θ is the reliability of the weighted scale $Y_i = \boldsymbol{a}'\boldsymbol{X}_i$ $\square$

C.3 Proof of Theorem 5

Using (14.5)–(14.6) it is easy to show that

$$\frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{a}} = \frac{\boldsymbol{\delta}'\boldsymbol{\Lambda}\boldsymbol{\delta}}{\boldsymbol{\delta}'\boldsymbol{\delta}},$$

where $\boldsymbol{\delta} = \boldsymbol{Q}\boldsymbol{a}$. This implies

$$\frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = \sum_{j=1}^{p} \left(\frac{\delta_j^2}{\sum_m \delta_m^2} \right) \lambda_j,$$

and therefore

$$\rho(\boldsymbol{a}) = 1 - \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = 1 - \sum_{j=1}^{p} w_j \lambda_j,$$

where $w_j = \dfrac{\delta_j^2}{\sum_m \delta_m^2}$.

Notice that, $\rho(\boldsymbol{a}) \in \Omega$ if and only if $w_j > 0$ for all j and $\sum_j w_j = 1$. The latter is an immediate consequence of the expression of w_j, taking into account that $\boldsymbol{a} \neq \boldsymbol{0}$. Finally, $w_j > 0$ for all j if and only if $\delta_j \neq 0$ for all j or, equivalently, if and only if $\boldsymbol{a} \in C$, where $C = \{\boldsymbol{a} : (\boldsymbol{Q}\boldsymbol{a})_j \neq 0 \quad \forall j\}$ $\square$

C.4 Proof of Theorem 6

Given that that $\rho(\boldsymbol{a}) = 1 - \dfrac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}}$ then

$$\max_{\boldsymbol{a} \neq \boldsymbol{0}} \rho(\boldsymbol{a}) = 1 - \min_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}}.$$

Further,

$$\min_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = \min_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2}\boldsymbol{a}} = \min_{\boldsymbol{z} \neq \boldsymbol{0}} \frac{\boldsymbol{z}'\boldsymbol{V}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{V}^{-1/2}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}} = \min_{\boldsymbol{z} \neq \boldsymbol{0}} \frac{\boldsymbol{z}'\boldsymbol{H}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}},$$

where $\boldsymbol{z} = \boldsymbol{V}^{1/2}\boldsymbol{a}$ and $\boldsymbol{H}$ is like defined in Section 14.4. Note that $\boldsymbol{a} \neq \boldsymbol{0}$ if and only if $\boldsymbol{z} \neq \boldsymbol{0}$. From Johnson and Wichern (2007) we know that

$$\min_{\boldsymbol{z}\neq\boldsymbol{0}} \frac{\boldsymbol{z}'\boldsymbol{H}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}} = \lambda_{(1)},$$

where $\lambda_{(1)}$ is the smallest eigenvalue of $\boldsymbol{H}$ and this minimum is reached for $\boldsymbol{z} = \boldsymbol{u}_{(1)}$, the corresponding eigenvector. Moreover,

$$\lambda_{(1)} \geq \prod_{j=1}^{p} \lambda_j \Rightarrow 1 - \lambda_{(1)} \leq 1 - \prod_{j=1}^{p} \lambda_j = R_\Lambda,$$

but $1 - \lambda_{(1)} = 1 - \min_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}} = \max_{\boldsymbol{a}\neq\boldsymbol{0}} \rho(\boldsymbol{a})$ and therefore,

$$\rho(\boldsymbol{a}) \leq \max_{\boldsymbol{a}\neq\boldsymbol{0}} \rho(\boldsymbol{a}) \leq R_\Lambda.$$

Note that $\rho(\boldsymbol{a}) = \max_{\boldsymbol{a}\neq\boldsymbol{0}} \rho(\boldsymbol{a})$ when $\boldsymbol{z} = \boldsymbol{u}_{(1)}$ or, equivalently, when $\boldsymbol{a} = \boldsymbol{V}^{-1/2}\boldsymbol{u}_{(1)}$. Finally, $\lambda_{(1)} = \prod_{j=1}^{p} \lambda_j$ if and only if $\lambda_{(2)} = \lambda_{(3)} = \cdots = \lambda_{(p)} = 1$. Taking all these elements into account, we conclude that the equality is obtained if and only if $\lambda_{(2)} = \lambda_{(3)} = \cdots = \lambda_{(p)} = 1$ and $\boldsymbol{a} = \boldsymbol{V}^{-1/2}\boldsymbol{u}_{(1)}$. □

C.5 Proof of Theorem 7

Let $\boldsymbol{X}_i^p$ denote a scale with p items and let the $R_\Lambda(p)$ be the reliability of this scale. We will proof that if a new item is added to construct the new scale $\boldsymbol{X}_i^{p+1}$ then

$$R_\Lambda(p) \leq R_\Lambda(p+1).$$

Let us note first that

$$\boldsymbol{V}_p = \boldsymbol{\Sigma}_p + \boldsymbol{B}_p\boldsymbol{D}\boldsymbol{B}_p'$$

where $\boldsymbol{V}_p$ and $\boldsymbol{\Sigma}_p$ are two $p \times p$ matrices, $\boldsymbol{B}_p$ is a $p \times q$ matrix and $\boldsymbol{D}$ is a $q \times q$ matrix, with q the number of true scores. We then have

$$\begin{aligned} & |\boldsymbol{V}_p| = |\boldsymbol{\Sigma}_p||\boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2}| \\ \Rightarrow \quad & \frac{|\boldsymbol{V}_p|}{|\boldsymbol{\Sigma}_p|} = |\boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2}|. \end{aligned}$$

Let us assume now that a new item $p+1$ has been added to the original scale. Then we have

$$\boldsymbol{V}_{p+1} = \boldsymbol{\Sigma}_{p+1} + \boldsymbol{B}_{p+1}\boldsymbol{D}\boldsymbol{B}'_{p+1}$$

where

$$\boldsymbol{B}_{p+1} = \begin{pmatrix} \boldsymbol{B}_p \\ \boldsymbol{b}'_{p+1} \end{pmatrix}, \quad \text{with } \boldsymbol{b}_{p+1} \text{ a } q \times 1 \text{ vector}$$

and

$$\boldsymbol{\Sigma}_{p+1} = \begin{pmatrix} \boldsymbol{\Sigma}_p & \boldsymbol{c} \\ \boldsymbol{c}' & c_0 \end{pmatrix}, \quad \text{with } \boldsymbol{c} \text{ a } p \times 1 \text{ vector.}$$

Note that $\boldsymbol{\Sigma}_{p+1}$ is positive definite and therefore

$$|\boldsymbol{\Sigma}_{p+1}| = |\boldsymbol{\Sigma}_p|(c_0 - \boldsymbol{c}'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{c}) > 0 \quad \Leftrightarrow \quad d = c_0 - \boldsymbol{c}'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{c} > 0.$$

Similarly as before we have

$$\frac{|\boldsymbol{V}_{p+1}|}{|\boldsymbol{\Sigma}_{p+1}|} = |\boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}'_{p+1}\boldsymbol{\Sigma}_{p+1}^{-1}\boldsymbol{B}_{p+1}\boldsymbol{D}^{1/2}|.$$

It is possible to show that

$$\boldsymbol{\Sigma}_{p+1}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_p^{-1} + \dfrac{\boldsymbol{q}\boldsymbol{q}'}{d} & -\dfrac{\boldsymbol{q}}{d} \\ -\dfrac{\boldsymbol{q}'}{d} & \dfrac{1}{d} \end{pmatrix} \tag{C.4}$$

where d is as before and $\boldsymbol{q} = \boldsymbol{\Sigma}_p^{-1}\boldsymbol{c}$. Note further that

$$\begin{aligned} \boldsymbol{B}_{p+1}\boldsymbol{D}^{1/2} &= \begin{pmatrix} \boldsymbol{B}_p \\ \boldsymbol{b}'_{p+1} \end{pmatrix}\boldsymbol{D}^{1/2} = \begin{pmatrix} \boldsymbol{B}_p\boldsymbol{D}^{1/2} \\ \boldsymbol{b}'_{p+1}\boldsymbol{D}^{1/2} \end{pmatrix} \\ \Rightarrow \boldsymbol{\Sigma}_{p+1}^{-1}\boldsymbol{B}_{p+1}\boldsymbol{D}^{1/2} &= \begin{pmatrix} \boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2} + \dfrac{\boldsymbol{q}\boldsymbol{q}'}{d}\boldsymbol{B}_p\boldsymbol{D}^{1/2} - \dfrac{\boldsymbol{q}\boldsymbol{b}'_{p+1}}{d}\boldsymbol{D}^{1/2} \\ -\dfrac{\boldsymbol{q}'}{d}\boldsymbol{B}_p\boldsymbol{D}^{1/2} + \dfrac{\boldsymbol{b}'_{p+1}}{d}\boldsymbol{D}^{1/2} \end{pmatrix} \end{aligned}$$

This implies that

$$\begin{aligned} \boldsymbol{D}^{1/2}\boldsymbol{B}'_{p+1}\boldsymbol{\Sigma}_{p+1}^{-1}\boldsymbol{B}_{p+1}\boldsymbol{D}^{1/2} &= \boldsymbol{D}^{1/2}\boldsymbol{B}'_p\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2} + \frac{1}{d}\boldsymbol{D}^{1/2}\boldsymbol{B}'_p\boldsymbol{q}\boldsymbol{q}'\boldsymbol{B}_p\boldsymbol{D}^{1/2} \\ &\quad -\frac{1}{d}\boldsymbol{D}^{1/2}\boldsymbol{B}'_p\boldsymbol{q}\boldsymbol{b}'_{p+1}\boldsymbol{D}^{1/2} - \frac{1}{d}\boldsymbol{D}^{1/2}\boldsymbol{b}_{p+1}\boldsymbol{q}'\boldsymbol{B}_p\boldsymbol{D}^{1/2} \\ &\quad +\frac{1}{d}\boldsymbol{D}^{1/2}\boldsymbol{b}_{p+1}\boldsymbol{b}'_{p+1}\boldsymbol{D}^{1/2}. \end{aligned}$$

If we define $\boldsymbol{r} = \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{q}$ and $\boldsymbol{s} = \boldsymbol{D}^{1/2}\boldsymbol{b}_{p+1}$ then

$$\begin{aligned}\boldsymbol{D}^{1/2}\boldsymbol{B}_{p+1}'\boldsymbol{\Sigma}_{p+1}^{-1}\boldsymbol{B}_{p+1}\boldsymbol{D}^{1/2} &= \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2} + \frac{1}{d}\boldsymbol{r}\boldsymbol{r}' - \frac{1}{d}\boldsymbol{r}\boldsymbol{s}' - \frac{1}{d}\boldsymbol{s}\boldsymbol{r}' + \frac{1}{d}\boldsymbol{s}\boldsymbol{s}' \\ &= \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2} + \frac{1}{d}(\boldsymbol{r}-\boldsymbol{s})(\boldsymbol{r}-\boldsymbol{s})',\end{aligned}$$

and therefore

$$\boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}_{p+1}'\boldsymbol{\Sigma}_{p+1}^{-1}\boldsymbol{B}_{p+1}\boldsymbol{D}^{1/2} = \boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2} + \frac{1}{d}(\boldsymbol{r}-\boldsymbol{s})(\boldsymbol{r}-\boldsymbol{s})'. \quad \text{(C.5)}$$

Note further that the matrix $\boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2}$ is positive definite and the matrix $\frac{1}{d}(\boldsymbol{r}-\boldsymbol{s})(\boldsymbol{r}-\boldsymbol{s})'$ is semipositive definite. Theorem 22 in Magnus and Neudecker (1994, pag. 21) then implies

$$\begin{aligned}\frac{|\boldsymbol{V}_{p+1}|}{|\boldsymbol{\Sigma}_{p+1}|} &= |\boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}_{p+1}'\boldsymbol{\Sigma}_{p+1}^{-1}\boldsymbol{B}_{p+1}\boldsymbol{D}^{1/2}| \\ &= |\boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2} + \frac{1}{d}(\boldsymbol{r}-\boldsymbol{s})(\boldsymbol{r}-\boldsymbol{s})'| \\ &\geq |\boldsymbol{I} + \boldsymbol{D}^{1/2}\boldsymbol{B}_p'\boldsymbol{\Sigma}_p^{-1}\boldsymbol{B}_p\boldsymbol{D}^{1/2}| = \frac{|\boldsymbol{V}_p|}{|\boldsymbol{\Sigma}_p|}\end{aligned}$$

and finally

$$\frac{|\boldsymbol{V}_{p+1}|}{|\boldsymbol{\Sigma}_{p+1}|} \geq \frac{|\boldsymbol{V}_p|}{|\boldsymbol{\Sigma}_p|} \Rightarrow R_\Lambda(p+1) = 1 - \frac{|\boldsymbol{\Sigma}_{p+1}|}{|\boldsymbol{V}_{p+1}|} \geq 1 - \frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{V}_p|} = R_\Lambda(p)$$

$$\Rightarrow \quad R_\Lambda(p+1) \geq R_\Lambda(p). \square$$

Samenvatting

In de jaren 50 werden de eerste medicijnen ontwikkeld voor de behandeling van psychiatrische stoornissen, zoals antidepressiva en antipschotische medicatie. Rond dezelfde tijd werd farmaceutisch onderzoek meer en meer gebaseerd op gecontroleerde klinische studies. In zulke studies wordt gebruik gemaakt van een controlegroep van patiënten die de experimentele behandeling niet ontvangen, en worden patiënten toevalsgewijs aan de controlegroep dan wel de experimentele groep toegekend. De gezondheidstoestand van beide groepen wordt vervolgens vergeleken, maar net daar schuilt een van de moeilijkheden van het onderzoek naar psychofarmaceutische medicatie. Een nauwkeurige evaluatie van de gezondheidstoestand in het geval van psychiatrische problemen is niet evident. Hoewel er consensus bestaat over een gedeeltelijke biologische oorzaak van verschillende psychiatrische aandoeningen, bestaan er geen laboratoriumtests om de ziekte vast te stellen of de ernst ervan te evalueren. De evaluatie gebeurt daarom op basis van beoordelingsschalen. Zulke schalen bestaan uit een lijst van items of vragen die door een zorgverlener of soms de patiënt zelf beantwoord worden aan de hand van meerkeuze-antwoorden. De scores op elk van de vragen worden vervolgens opgeteld tot een totaalscore die een indicatie geeft van de ernst van de problematiek.

Om een nauwkeurige resultaatsmeting te garanderen moet de beoordelingsschaal aan bepaalde voorwaarden voldoen. Een schaal moet valide zijn, dat betekent dat ze werkelijk meet waarvoor ze bedoeld is. Daarnaast moet een schaal betrouwbaar zijn. Dit houdt in dat de meting met een aanvaardbare precisie gebeurt, of anders gezegd, dat meetfout tot een minimum beperkt wordt.

De evaluatie van de betrouwbaarheid van meetschalen, educatieve en psychologische tests komt uitgebreid aan bod in de klassieke psychometrische literatuur. Nochthans zijn de gangbare methodes vaak niet flexibel genoeg om te worden toegepast in de vaak complexe settings van klinische studies. De doelstelling van deze thesis is

om die klassieke psychometrische methodes uit te breiden naar meer algemene settings, met longitudinale of multivariate metingen.

In de klassieke test-theorie wordt de geobserveerde score van een patiënt beschouwd als een som van de 'ware score' van deze patiënt en een component te wijten aan meetfout. Bij de meting van een groep van patiënten wordt de betrouwbaarheid dan gedefinieerd als de verhouding tussen de variantie die komt van de ware scores van de patiënten en de totale variantie van de observaties (Lord and Novick 1968). Aangezien de totale variantie altijd groter is dan de variantie van de ware scores, is de betrouwbaarheid steeds een getal tussen 0 en 1. Een getal in de buurt van 0 betekent dat de ware-score variantie zeer klein is en de totale variantie bijna volledig wordt verklaard door meetfout. De meting is dan zeer onbetrouwbaar. Een getal in de buurt van 1 betekent dat de totale score variantie bijna volledig verklaard wordt door de ware scores, en dat meetfout een zeer kleine invloed heeft. De meting is dan zeer betrouwbaar. Een schatting van de betrouwbaarheid is maar mogelijk wanneer eenzelfde meting herhaald wordt, door bijvoorbeeld op twee momenten te meten (test-hertest betrouwbaarheid), door twee verschillende beoordelaars te laten meten (inter-rater betrouwbaarheid), of door verschillende maar parallelle instrumenten te gebruiken (interne consistentie).

Bovenstaande benadering is eenvoudig, intuïtief en dus zeer aantrekkelijk. Anderzijds is ze gestoeld op een aantal veronderstellingen die in veel praktische situaties niet kunnen gegarandeerd worden. Een van de veronderstellingen is bijvoorbeeld dat de ware score van een patiënt bij de twee metingen constant is. Het kan nochthans gebeuren dat een patiënt evolueert in de latente variabele die gemeten wordt, denk maar aan depressie. In een longitudinale klinische studie is dat onvermijdelijk. Verder wordt er verondersteld dat de varianties komende van de meetfout constant zijn in de verschillende metingen en dat de meetfouten waargenomen bij eenzelfde patiënt niet met elkaar gecorreleerd zijn. Ook deze twee veronderstellingen zijn zeer weinig plausibel in het geval van herhaalde metingen (Verbeke and Molenberghs 2000).

Om betrouwbaarheid op basis van longitudinale gegevens te kunnen analyseren is het daarvoor essentieel om uit te gaan van een meetmodel dat rekening houdt met de typische kenmerken van dit soort gegevens. Daarom baseren we onze methoden op een linear gemengd model (Laird and Waire 1982, Diggle, Liang and Zeger 1994). Uitgaande van dit meetmodel herdefiniëren we betrouwbaarheid aan de hand van een axiomatische benadering. We stellen dat een maat voor betrouwbaarheid moet beschikken over vier eigenschappen, zijnde (1) de betrouwbaarheid ligt steeds tussen

0 en 1; (2) de betrouwbaarheid is nul enkel in het geval dat de observaties volledig te wijten zijn aan meetfout; (3) de betrouwbaarheid is 1 enkel in het geval er geen meetfout optreedt; en (4) wanneer de veronderstellingen van de klassieke test-theorie correct zijn, moet elke maat voor betrouwbaarheid samenvallen met de maat die in deze klassieke theorie werd voorgesteld. Een meer formele omschrijving vindt men in Hoofdstuk 7.

In hetzelfde hoofdstuk introduceren we de R_T coëfficiënt, een maat voor betrouwbaarheid die voldoet aan de vier bovengenoemde voorwaarden. Verder beargumenteren we dat deze maat in een longitudinaal kader de gemiddelde betrouwbaarheid weergeeft over de verschillende metingen. Op die manier wordt de betrouwbaarheid van de herhaalde metingen in de studie aan de hand van een enkel cijfer samengevat. Zo een beknopte samenvatting kan zeer handig zijn wanneer er een groot aantal metingen werden afgenomen, of wanneer bijvoorbeeld twee meetschalen met elkaar moeten vergeleken worden. Anderzijds laat de methode ook toe om de R_T coëfficiënt afzonderlijk te schatten voor elk van de herhaalde metingen, zodat de evolutie van de betrouwbaarheid over de tijd kan nagegaan worden.

In hoofdstuk 10 introduceren we de R_Λ coëfficiënt; een tweede nuttige maat voor betrouwbaarheid. We beargumenteren dat deze maat een andere, complementaire boodschap geeft. De R_Λ coëfficiënt geeft niet de gemiddelde betrouwbaarheid over de herhaalde metingen weer zoals de R_T coëfficiënt, maar ze geeft de 'totale' betrouwbaarheid weer van de hele reeks van metingen. We kunnen dit als volgt toelichten. Wanneer een beoordelingsschaal eenmaal wordt afgenomen bij een groep van patiënten, levert dat een bepaalde hoeveelheid informatie op. Wanneer de schaal bij diezelfde groep een tweede maal wordt afgenomen, kan dat enkel maar tot meer informatie leiden over de patiënten. Hetzelfde geldt voor een derde meting, enzovoort. Dit intuïtief idee wordt gevat in de R_Λ coëfficiënt. We zien dan ook dat deze maat telkens toeneemt met het aantal herhaalde metingen. Waar de R_T coëfficiënt ons informatie verschaft over de kwaliteit van de schaal, los van de context van de studie, vertelt de R_Λ coëfficiënt ons wat de invloed is van meetfout op het geheel van metingen in een longitudinale studie. Vaak zien we dat een schaal slechts matige betrouwbaarheid vertoont bij een eenmalige meting, maar wanneer de herhaalde metingen samen beschouwd worden, de impact van meetfout minimaal blijkt. Verder kan de R_Λ coëfficiënt ons een hint geven over het aantal herhaalde metingen die met een bepaalde meetschaal en in een bepaalde populatie nodig zijn om de impact van meetfout tot een bepaald niveau terug te dringen.

In hoofdstuk 11 tonen we dat er een sterke link bestaat tussen de door ons voorgestelde benadering en belangrijke eerdere bijdragen uit de psychometrische literatuur. In hoofdstuk 12 zoomen we in op een gekend probleem bij herhaalde metingen, namelijk dat van gecorreleerde meetfouten. We tonen dat onze methode standhoudt waar de klassieke benadering en uitbreidingen ervan tekortschieten.

De methodes worden verder uitgebreid geïllustreerd aan de hand van twee exemplarische studies. De eerste betreft een klinische studie in het domein van schizofrenie waar de betrouwbaarheid van drie verschillende meetschalen wordt geëvalueerd en vergeleken. In het tweede voorbeeld analyseren we de betrouwbaarheid van drie meetschalen om de ernst van een depressie te evalueren.

Het grootste deel van de thesis betreft de analyse van betrouwbaarheid op basis van herhaalde metingen. In Hoofdstuk 14 illustreren we dat de daarvoor geïntroduceerde methodes perfect vertaalbaar zijn naar een multivariate context. In de psychometrische literatuur is veel aandacht gegaan naar de evaluatie van betrouwbaarheid op basis van een eenmalige meting, maar gebruik makende van de afzonderlijke item-scores van de meetschaal. Vertaald naar deze context kan de R_T coëfficiënt geïnterpreteerd worden als de gemiddelde betrouwbaarheid over alle items in de beoordelingsschaal. De R_Λ coëfficiënt geeft anderzijds een indicatie van de hoeveelheid informatie er aanwezig is in de volledige set van items over de onderliggende latente variabelen die men wil meten. Verder onderzoeken we uitgebreid de link tussen onze methodes en bestaande methodes uit de literatuur. We illustreren onze benadering op basis van de 'Positive and Negative Syndrome Scale', een van de beoordelingsschalen voor de evaluatie van schizofrenie.