Measuring the Reproducibility of and Comparability between Physiological and Psychological responses in Exercise Testing

Shahram Zare

A Dissertation Submitted to the

University of Glasgow

for the degree of

Doctor of Philosophy

Department of Statistics

November 1997

Abstract

The main topics covered in this thesis involve the estimation of the Reproducibility of variables and the Comparability between two variables, with specific application to Exercise Testing in Coronary Care and in Sports Science.

Reproducibility refers to the consistency of scores obtained from a given trait or characteristic of a given individual. Its importance in health and medical measurements is well recognized and scientific progress can hardly be achieved in the absence of reproducible data. *Comparability* between two variables arises when, as in Exercise Testing, two different and distinct aspects of an underlying characteristic of an individual are measured at a succession of points during an exercise test. For example, one aspect may be a direct measure of breathlessness such as Ventilation while the other is an indirect quantitative impression of the subject's perception of his/her breathlessness. Any relationship between these two is likely to vary not only from subject to subject but even within repeat exercise tests on the same subject. The aim of this part of the thesis is to quantify and estimate an overall measure of the relationship between the two aspects for the 'typical exercise test' of the 'typical individual' i.e. the Comparability of the two aspects/variables.

The essential practical rationale for this is that, in Exercise Testing, the physiological measurement of breathlessness using a Douglas Bag is not a realistic option for a Coronary Care patient whereas

his/her subjective assessment of breathlessness and/or fatigue can be easily reported by the patient through the test without additional stress. The key question therefore is whether such a subjective assessment is, in general, an adequate 'mirror' of the underlying physiological profile.

Chapter 1 gives a brief background to Exercise Testing and its importance as well as a literature review of relevant topics including reproducibility, comparability, components of variance and the estimation of common correlation; the latter two are essential building blocks for the estimation of Comparability.

Chapter 2 deals with the estimation of measurement reproducibility of data from mixed effects models involving two variance components. Two approaches, one based on sums of squares and the other on Profile Likelihood are used for the separate cases of balanced and unbalanced data. This is carried out in two distinct contexts, one for simple replication and the other assuming an order effect to the replications. Applicability of the approaches to Exercise Testing data shows that while point estimates from both approaches are often identical, interval estimates from the Profile Likelihood approach tend to be narrower.

Chapter 3 involves a simulation study to investigate and assess the performances of the two approaches. Data are simulated from a variety of underlying configurations and the performances then compared according to three statistical criteria. The results of this study again favour the Profile Likelihood approach.

The estimation of Comparability between two variables is the other aspect of the thesis put forward in chapter 4 where, first of all, the estimation of a common correlation coefficient from a population of correlation coefficients is considered. Five different methods for point and interval estimation of a common correlation coefficient

are introduced. An illustrative example using data from an Exercise Testing procedure is used to compare the performances of the methods.

Further investigation on the performances of the five methods was carried out by means of a simulation study across a variety of underlying configurations. The overall results suggest the 'Fisher method' as the best method of point and interval estimate of common correlation.

The Comparability between two variables is then modelled, in chapter 5, by developing structures for 'pooling' correlation coefficients across individuals and replicate visits. Illustrative examples from Exercise Testing are used to investigate the applicability of the models on real data. A comprehensive simulation study across a variety of configurations was then carried out and the performances of the models assessed. The results show that the Multiplicative Fisher model in the One-Stage modelling and the Components of Variance model in the Two-Stage modelling are the best approaches to estimating 'Comparability'.

Finally, chapter 6 outlines the conclusions from the previous chapters and suggests some ideas for further work.

To my Parents,
my wife, Ziba,
and my children, Armin and Azin

Acknowledgements

I would first like to express my appreciation and deep respect and gratitude to my supervisor *Mr. Thomas C. Aitchison* for suggesting the topic of this research, advice, encouragement, support and more importantly his patience throughout this research.

I would also like to thank the head of the Statistics department *Mr. Peter Breeze* and former head of department *Professor Ian Ford* and the staff members in the department, particularly *Miss Mary Nisbet* and *Mrs. Myra Smith* for their kindness.

I feel also grateful to my fellow research students for providing me a very comfortable working environment.

I also owe a debt of gratitude to the Iranian Ministry of Health and Medical Education, Bandar Abbas University of Medical Sciences and the Iranian Embassy in London for their financial support.

Finally, my heartfelt acknowledgements are due to my parents, my wife, *Ziba*, and our children for their understanding, patience and support.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Introduction

The *Reproducibility* of any form of measurement is an important issue in any scientific research. Cardiologists and sports scientists, for example, use physiological as well as psychological measures in *Exercise Testing* to evaluate the functional performance and capacity of the cardiovascular system. Reproducibility studies are essential in such contexts to assess the validity and usefulness of these measurements.

Further, it may be expected that any differences in physiological stimuli would result in changes in psychological responses to such stimuli. Therefore, it is important to assess the *Comparability* between psychological and physiological measurements of, allegedly, the same underlying aspect of a subject.

In this chapter, a brief description of different aspects of Exercise Testing is given. Measurement reproducibility is then defined and standard approaches to estimating reproducibility are discussed.

The basic ideas involved in the comparability of variables through the use of correlation coefficients are outlined and finally, a brief layout of the thesis is presented.

1.2 Exercise Testing

1.2.1 Importance of Exercise Testing in Medical and Physical Sciences

Physical activities are among the most common physiological stressors. *'Exercise Testing'* provides a distinctive and practical means of assessing the body's capacity for physical activity. It can define the functional capacities of a symptomatic patient as well as the limits of an athlete's performance (Skinner, 1987). Usually two categories of person are employed in exercise testing, athletes and those with a health problem (e.g. Coronary Heart Disease) unable to exercise but for whom assessment, perhaps, of their cardiovascular system is required, over a treatment period.

Coronary heart disease continues to be the most frequent cause of death in economically developed countries. Participation in an *'Exercise Testing'* program allows the clinical status and exercise capacity of such patients to be assessed.

Traditionally, cardiologists have used a maximum exercise tolerance test for:

- disease prediction, prognosis and severity,
- evaluation of surgical and medical treatments,

- assessment of functional capacity or maximum oxygen consumption,
- exercise prescription (Naughton, et al, 1973)

The reason for carrying out maximum exercise tolerance tests is the hope that coronary heart failure patients can attain a representative 'maximum' effort which reflects the true limits of the cardiovascular system. From another point of view, although health, fun or fitness may result from exercise, the primary goal of most athletes is to improve their performance. Exercise scientists usually determine the specific characteristic of the activity in which an athlete is going to compete and decide on the details of the exercise program in order to develop the physiological potential of an athlete.

1.2.2 Clinical Application of Exercise Testing

An *Exercise Test* can be used to evaluate patients that currently have chest pain or sensations or cardiac abnormalities, patients with a history of myocardial infarction (MI) or chest pain, or patients with other findings tending to suggest of Coronary Heart Disease (CHD). In addition, the successful performance of a test after an acute myocardial infarction can be reassuring and is the first step in rehabilitation (Sivarajan, et al, 1977). An exercise test, which reflects the normal life of coronary heart failure patients and can evaluate symptoms, may be of value in long term monitoring of coronary heart failure patients. Its results may also be used to establish the risk of morbidity and mortality to compare potential treatment responses or to categorise the severity of heart failure.

On the other hand, healthy individuals can undergo exercise testing to identify the individuals at high risk for CHD or even to evaluate the safety of participation in an exercise program or the performance of other activities (Skinner, 1987)

1.2.3 Variables Measured in Exercise Testing

Patients with coronary heart failure often exhibit symptoms of distress during exercise. Quantification of these during the exercise may be of value in the evaluation of both performance of the patient and the efficacy of the prognosis. During the performance of an exercise test, a variety of physiological as well as psychological symptoms are often considered.

The most common physiological variables measured at various fixed time points in an exercise test are:

- VO_2 max: the amount of oxygen extracted from inspired air during a progressive Exercise Test. VO_2 max virtually defines the pumping capacity of the heart. Therefore it is of major importance in the evaluation of severity of heart disease and is an excellent indication of fitness level.
- Heart Rate:
- Ventilation: the volume of respiratory gas exchange during the exercise test.

The self-assessed psychological variables measured at various fixed time points in an exercise test and measured on a subjective scale of 0 to 100 are:

- Breathlessness, which is defined as breathless, out of breath, air hunger, unable to breath enough, and

and

- General Fatigue, which is described as overall tiredness or overall fatigue.

1.2.4 Exercise Testing Protocols

Exercise testing may be carried out by means of different devices such as treadmill, bicycle ergometer, step and arm ergo-meter.

Among the most commonly used devices for exercise testing is a motor-driven treadmill (Figure 1.1). It provides flexibility because the speed of the belt and the slope of the treadmill can be varied either independently or simultaneously.

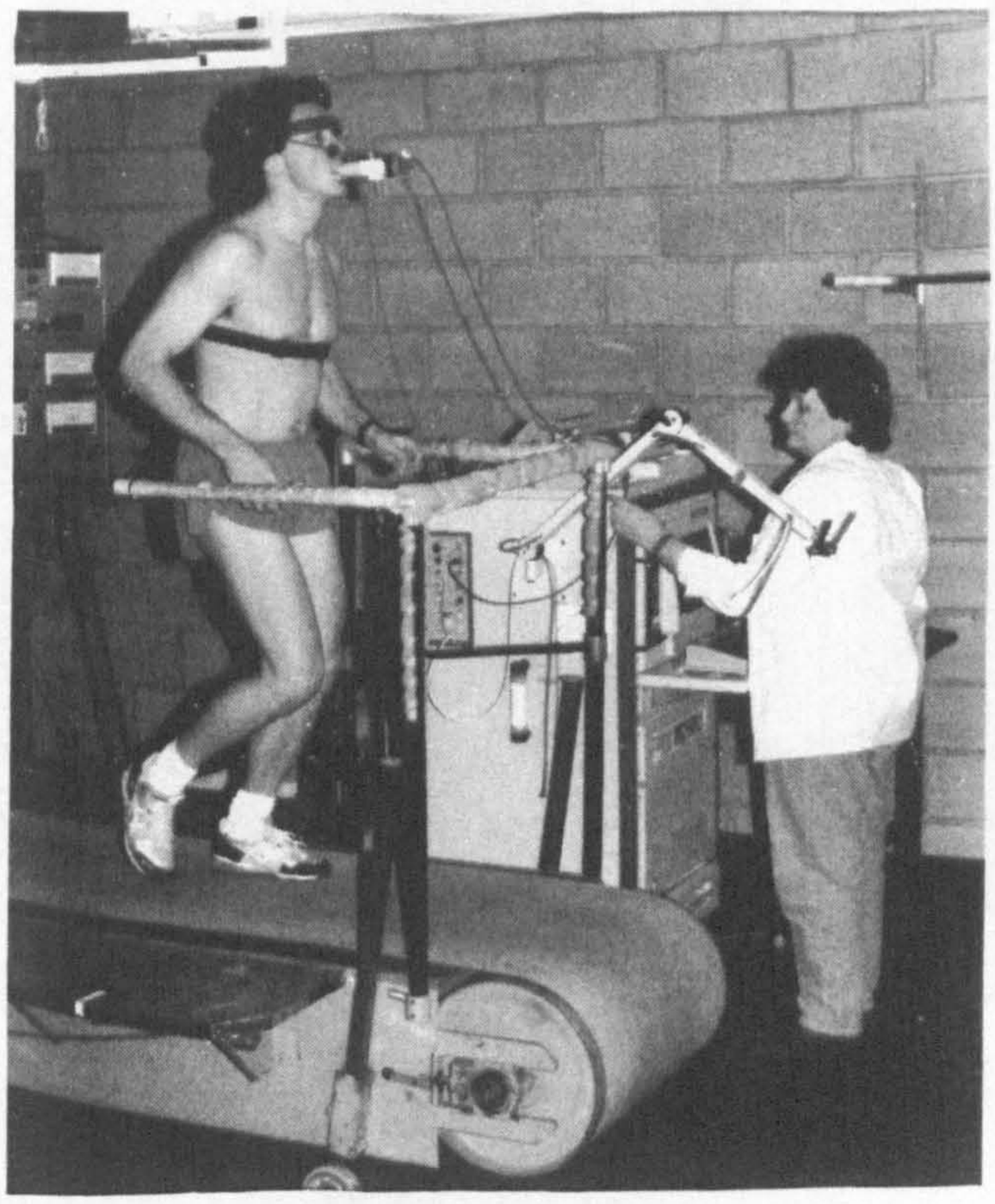

Figure 1.1: Treadmill

Before starting the maximal treadmill tests, subjects are familiarised with the treadmill and the use of the subjective scales. The psychological scales are displayed on a colour computer monitor in front of the subject while he or she exercises on the treadmill. At different stages of the test (e.g. at the end of specific time intervals) the subjects are asked to quantify their perception of breathlessness and general fatigue.

During the exercise and at the specific time points, expired air was collected and, using an automated gas analysis system, analysed in order to measure the relevant respiratory variables. An Electrocardiogram was used to monitor cardiac performance and in particular heart rate.

1.3 Measurement Reproducibility

1.3.1 Definition and Its Importance

Medicine, manufacturing industry and research in many sciences all require an ability to measure a characteristic under study. As discussed by Shrout and Fleiss (1979), all measurements, particularly those made by humans, are usually subject to error. These errors can sometimes seriously affect statistical analysis and hence the interpretation of the results of any study. It is therefore not surprising that a great deal of time and effort is directed toward the development of new or 'improved' and hence more reliable measurement and techniques.

Measurement Reproducibility is the consistency with which a variable assesses a given trait or characteristic of a given subject. In other words, it is formally defined as the proportion of the total variation that is not attributable to random error or 'natural' variability. However, even if a measurement is precise, reproducibility in its use is required to provide a scientific evaluation.

The importance of reproducibility of data in medical and health measurements is well recognized, and scientific progress can hardly be achieved in the absence of reliable assessment of the variables under consideration in any such problem. Failure to reproduce a series of measurements usually implies that the assessments are affected by some sources of variation other than that of the subject attribute under study.

In clinical treatments, for instance, assessment of risk factors using unreliable measurements can produce either overestimation or underestimation of the strength of association of different factors

(Goldberg, 1975). However, large sample sizes are no protection against the systematic biases that sometimes hide strong associations or create associations when there are none and may result in mistaken treatment of patients (Shrout et al, 1987).

In theory, the assessment of measurement reproducibility requires the use of independent measurement procedures. However, in practice, completely independent measurements are rarely possible since replicate values are often affected by the previous measurements when assessed by human observers.

The design of a reproducibility study clearly depends on the context in which the study is being undertaken, what sort of measurement instruments are being used or compared, what properties or characteristics are being measured and finally what sources of variations need to be estimated. On the other hand, it is not enough to choose an instrument and a measurement technique without monitoring and evaluating its performance during its routine use. In exercise testing, for example, physiological or psychological variables may be frequently measured to evaluate the performance of the patient. The measurements need to be monitored and evaluated in an equally rigorous way.

There appear to be three main roles for reproducibility studies, these are:

1. as an aid to instrument development (including training of interviewers, raters or examiners),
2. as an aid to the choice of measurement instrument or choice of the condition in which the measurements are to be made,

and

3. as a way of monitoring individual performance .

1.3.2 Approaches to Measurement Reproducibility Estimation

Reproducible measurements and reliable methodology are accepted as necessary to scientific research and hence, an appropriate use of statistically sound techniques for assessing reproducibility is crucially important.

A 'Reproducibility' experiment would typically involve a series of replicate measurements on a group of randomly chosen subjects from a target population.

For the situations where the measurements are on a *quantitative scale*, the technique which is mostly used in the literature is based on 'Components of Variance'.

'Components of Variance' models may be traced back to the work of Fisher who introduced the term 'variance' and 'analysis of variance' in the literature and implicitly employed variance component models (Khuri and Sahai, 1985).

In the Components of Variance models, typically, some facets of the measurement process may be regarded as having fixed levels, whilst others may be regarded as having been selected at random from a population.

1.3.2.1 Balanced Data

A 'balanced data' set is one in which there are the same number of replicates per subject. In this subsection different approaches to Components of Variance estimation are considered.

A: *Analysis of Variance Approach*

The traditional analysis of variance methods of estimating Components of Variance involve equating the sums of squares in the analysis of variance to their expected values, thereby formulating a set of equations with unknown parameters. The most desirable feature of these estimators is the ease with which their values can be estimated. The usefulness of the analysis of variance approach is impaired by the (frequent) occurrence of negative estimates of variance components, which naturally are nonnegative quantities.

B: *Likelihood Approach*

An alternative approach to variance components estimation is that of 'maximum likelihood'. This approach is also based on assuming normality of the data and maximizing the likelihood function over the parameter space. Maximum likelihood approach has received little attention in the literature which may be mainly attributed to the complexity of computations associated with the solution of the likelihood equations. Hartley and Rao (1967) proposed a maximum Likelihood method for estimating Components of Variance.

Miller (1980) showed that any balanced mixed model in which all effects, fixed or random, are nested has explicit maximum likelihood with and without the variance components constraints. However, explicit maximum likelihood estimators of variance components cannot always be obtained in balanced models (Khuri and Sahai, 1985). Furthermore, it has been shown that the balanced two-way crossed classification random models, with or without interaction, does not have explicit maximum likelihood estimators.

1.3.2.2 Unbalanced Data

'Unbalanced data' refer to a ' reproducibility experiment' in which not all subjects have the same number of observations, e.g. due to drop-outs from the study. This subsection is concerned with the different approaches to components of variance estimation in the case of unbalanced data.

A: *Analysis of Variance Approach*

Fisher in 1925 extended analysis of variance to unbalanced data but did not propose the estimation of variance components. The analysis of variance method of variance components estimation for unbalanced data was later made clear by Tippett (1931). Winsor and Clark in 1940 proposed a so-called analogue of the analysis of variance method for unbalanced data. The essence of their work was a pair of quadratic expressions, similar to the analysis of variance sums of squares for balanced data, which were equated to their expected values and the resulting equations solved for the unknown variance components (Khuri & Sahai, 1985).

However, Components of Variance estimation in the case of unbalanced data is much more complicated and no known exact procedures exist. The main difficulty stems from the fact that in unbalanced data, the partitioning of the total sum of squares can be done in a variety of ways, hence, there is no unique way to write the analysis of variance table as in the case with balanced data. Furthermore, the sums of squares in an analysis of variance table for an unbalanced case, with the exception of the residual sum of squares, do not, in general, have "known" distributional properties and are not even independently distributed.

Probably the basic paper dealing with the variance component es-

timation from unbalanced data is that by Henderson (1953). It established three different sets of quadratic form that could be used for components of variance estimation. All the three sets are closely related to the sums of squares of analysis of variance calculations for unbalanced data and may produce negative estimates. These three methods are known as Henderson's method I, II and III.

In brief, *Method I* uses quadratic forms that are similar to sums of squares of balanced data. Estimates of variance components are then obtained by equating the quadratic forms to their expected values and solving for the equations for the unknown variance components. The method, however, is not suitable for mixed models, in which case it yields biased estimates. In addition, under the usual normality assumptions, the distribution of the estimators, except for the variance error which is usually proportional to a χ^2, cannot be specified in closed form.

Method II is an adaptation of Method I that takes into account the fixed effects in the model. The data are adjusted by using some estimates of the fixed effects based on the observed data and then the method is applied to estimate the variance components from the adjusted data. The method, however, is not applicable when the mixed model contains interactions between fixed and random effects. Furthermore, as in Method I, no closed form expressions are available for sampling variances of estimators.

Method III is the method which is applicable for both random and mixed models, even if interactions exist between fixed and random effects. This method is based on 'borrowing' sums of squares from the analysis of fixed effects models. The procedure uses a sufficient number of reductions in sums of squares due to fitting various submodels of it. These reductions are chosen so that their expected values are free from any mixed effects. Estimates of the variance

components are then obtained by solving the equations which result from equating these reductions to their expected values. Method III produces unbiased estimators, but may require extensive computations. On the other hand, sampling variabilities by this methods can be calculated numerically, but specific closed form expressions are not available.

A complete description of Henderson's methods with its merits and demerits are given in Searle et al, (1992).

However, several authors noted that the analysis of variance methods of estimation of components of variance, including Henderson's methods, are deficient in the sense that they can produce negative estimates. They proposed a number of alternative approaches which include, replacing the negative estimated values by zero, using alternative methods of estimation or changing the design of the data (Smith et al, (1984), Khuri et al, (1985), Searle (1987) and (1992)).

B: *Likelihood Approach*

Likelihood procedures for the estimation of the variance components and the fixed effects in a general mixed model were considered by Hartley and Rao (1967). The solution of the likelihood equations is usually obtainable via iteration, which in some cases can be computationally cumbersome because of the need to invert a variance-covariance matrix of large order. Hemmerle et al, (1973) discussed the Newton-Ramphson algorithm and other iteration methods which have been proposed for computation of maximum likelihood estimates.

Searle (1987) and (1992) provide an up-to-date survey of modern estimation methods for different models. Other methods of point and interval estimation of components of variance, including minimum

norm quadratic unbiased estimation, minimum variance quadratic unbiased and restricted or residual maximum likelihood are discussed in Searle et al.(1992). Readers are referred to Khuri and Sahai (1985) for a review of this and a comprehensive review of further developments in the area of variance components.

1.4 Comparability of Variables

The problem of combining information from independent studies permeates almost all fields of science (Olkin, 1995). In many areas of scientific researches, it may be useful to assess the relationship between continuous variables and correlation coefficients have been used extensively as an index of the linear relationship between variables. In Exercise Testing, for example, physiological and psychological variables will be measured at a number of fixed time points through the test, and this may be repeated in different occasions. Interest focuses on the link between these two variables.

Tests and inferences for the correlation coefficient are frequently based on the assumption of approximate normal distribution for Fisher's z-transformation (Kraemer, 1975). Kraemer introduced a test and confidence interval procedure for a single sample correlation coefficient, and showed that there is a little difference in results whether one use the z-transformation or the procedure described in her paper. She used the likelihood ratio test to test the homogeneity of correlation coefficients from k independent bivariate normal data, and showed that the χ^2 approximation to the likelihood ratio statistic is reasonable even for relatively small sample sizes.

Tests of homogeneity of independent correlation coefficients based on the simple forms of the normal and t approximation to the dis-

tribution of the correlation coefficients in terms of robustness, size, and power were under more investigation in Kraemer (1979). She demonstrated that neither procedure was totally robust under bivariate non-normality. Furthermore, it showed that the procedure based on the t-distribution appeared somewhat biased but more powerful. However, neither the difference in bias or power were of magnitude to make a difference in practical application.

Kowalski (1972), with the aid of simulation and density estimation techniques, concluded that the distribution of estimates of correlation coefficients may be quite sensitive to non-normality and that normal correlation analyses should be limited to situations in which the data is (at least very nearly) normal. However, it showed that the distribution of sample correlation coefficient need not agree well with normal theory when the population correlation coefficient is zero.

Whilst the necessity of combining independent statistical results is present in different areas of research, papers on common correlation coefficient are quite few in number.

Viana (1980) described a so-called Z-additive method for combining independent sample correlations from bivariate normal data and suggested a combined estimate of the correlation parameters based on an approximation to Olkin and Pratt's (1958) unbiased estimator to correlation coefficient. In this procedure, both cases of availability of the sample correlation coefficients or availability of the original paired data were considered. Furthermore, a test of homogeneity of the data, based on a Chi-Squared statistic, was discussed.

Bushman and Wang (1995) used a weighted average of sample correlation coefficients to estimate the population correlation coefficient and suggested Fisher transformation of correlation coefficients as a

remedy to solve the problem with the complicated distribution of correlation coefficients.

1.5 Chapter Layout

The major aims of this thesis are to estimate

i) - measurement reproducibility, and

ii) - a summary measure of relationship between physiological and psychological measurements, across subjects, visits and time points into visits.

The breakdown by chapter is as follows:

Chapter 2 concerns two approaches of point and interval estimation to measurement Reproducibility. It deals with the cases of a simple replication model and a replication model with an order effect. Real examples are given throughout the chapter to illustrate the application of the two approaches.

Chapter 3 investigates the performance of the two approaches of estimating measurement reproducibility through a simulation study. This study covers the performance of the two approaches to balanced as well as unbalanced data under the two situations (i.e. replication and order effect replicates). Three statistical criteria are used to compare the performance of the two approaches.

Chapter 4 contains two parts. The first part introduces five methods of point and interval estimates of Comparability under the assumption that this involves a common correlation. A real example is used to illustrate the applicability of the methods. In the second part, performance of the five approaches of estimating a common correlation is investigated by a simulation study. This study is carried out under a number of underlying configurations.

Chapter 5 deals with modelling Comparability in the more realistic context where the correlation coefficient between two variables under consideration will vary not only between tests on the same subject but also across subjects. The Comparability will be, in a sense, the 'average correlation' for the 'typical test on the typical subject'. Two models for the cases of the a one-stage process (i.e. assuming no intra-subject variability in the correlation) and three models for the case of a two-stage process (i.e. allowing for intra-subject variability as well as inter-subject variability in the correlation) are proposed. Applicability of the models are illustrated by real examples. To examine the performance of the models, a simulation study is carried out. Three statistical criteria are used to compare the performance of the models.

Finally **Chapter 6** summarizes the findings of this thesis and points out some directions for possible further work.

Chapter 2

Estimating Measurement Reproducibility: The Different Approaches

2.1 Introduction

Most measurements in medical sciences involve measurement variabilities/errors from a variety of sources, and judgements based on these measurements are usually plagued by this problem. On the other hand, measurement variabilities/errors can seriously affect statistical analysis and in particular interpretation of the data, so, it is important to assess the amount of such variabilities/errors by means of a reproducibility index.

In Sports Sciences, for instance, physiological variables such as ventilation, heart rate and etc. as well as psychological variables like breathlessness or fatigue may be measured at different time points during an Exercise Test and the test may be repeated at a number of visits. In order to assess the performance of a patient or

an athlete, one may be interested in estimating measurement variabilities to evaluate the reliability of the measurements. Imprecise estimation of variabilities of these data, can lead to serious bias in the reliability estimation and weaken the efficiency of a scientific judgement.

In this chapter, the estimation of measurement reproducibility of data from mixed effects models involving two variance components, is investigated using ANOVA-based methods and Profile Likelihood methods for both balanced (equal number of observations per individual) and unbalanced (unequal number of observations per individual) data sets. This is carried out for two different models, one involving simple replication across one of the variance components while the other assumes an order effect to the replications e.g. due to learning or familiarisation with the Exercise Testing procedure.

A : ANOVA-based Approach

2.2 Balanced Data

2.2.1 Simple Replication Model

2.2.1.1 Basic Model

Suppose for each of a random sample of N individuals from a population of interest there are T replicate observations, the model that will be used is:

$$X_{ij} = \mu + \tau_i + e_{j(i)} \tag{2.1}$$

$$i = 1, 2, \ldots, N \quad j = 1, 2, \ldots, T,$$

where

X_{ij} is the j^{th} observation of the i^{th} individual,
μ is a general mean,
τ_i is the 'effect' of the i^{th} individual,
and $e_{j(i)}$ is replicate variability.

Further

τ_i and $e_{j(i)}$ are assumed to be independent and normally distributed both with mean 0 and variances σ_B^2
and σ_W^2, respectively,
i.e.

$$\tau_i \sim N(0, \sigma_B^2) \text{ and } e_{j(i)} \sim N(0, \sigma_W^2) \quad \forall\, i, j.$$

2.2.1.2 Definition of Measurement Reproducibility

Measurement reproducibility of a variable (e.g. VAS on Fatigue) is in fact the consistency with which the variable assesses a given characteristic under 'identical' conditions. Its importance in medical contexts is well recognized in that reliable medical decisions can hardly be expected in the absence of reliable assessments of the variables on which such decisions are based. The measurement reproducibility of a variable, denoted by ρ, is defined as the ratio of the between individual variance to the total variance (i.e. between plus within individual variance) (Dunn, G. 1989) . For example in the model given by 2.1,

$$\rho = \frac{Var(\tau_i)}{Var(X_{ij})} \tag{2.2}$$

The assumptions on the model in the previous section imply that,

$$Var(\tau_i) = \sigma_B^2,$$

and

$$Var(X_{ij}) = Var(\tau_i) + Var(e_{j(i)}) = \sigma_B^2 + \sigma_W^2$$

Hence the measurement reproducibility is

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \tag{2.3}$$

2.2.1.3 Point Estimation of Measurement Reproducibility

Two 'sums of squares' that are the basis of the analysis of variance of balanced data are

$$SSB = T\sum_{i=1}^{N}(\overline{X}_{i.} - \overline{X}_{..})^2 \tag{2.4}$$

$$SSW = \sum_{i=1}^{N}\sum_{j=1}^{T}(X_{ij} - \overline{X}_{i.})^2 \tag{2.5}$$

which are the between individual sum of squares and within individual sum of squares with $(N-1)$ and $N(T-1)$ degrees of freedom, respectively. Between-subject mean square and within-subject mean square are defined as:

$$MSB = \frac{SSB}{N-1} \quad \text{and} \quad MSW = \frac{SSW}{N(T-1)} \tag{2.6}$$

while the Expected Values of these mean squares are

$$E(MSB) = \frac{T}{N-1}E[\sum_{i=1}^{N}(\overline{X}_{i.} - \overline{X}_{..})^2] = T\sigma_B^2 + \sigma_W^2 \tag{2.7}$$

$$E(MSW) = \frac{1}{TN-N}E[\sum_{i=1}^{N}\sum_{j=1}^{T}(X_{ij} - \overline{X}_{i.})^2] = \sigma_W^2. \tag{2.8}$$

Based on these, unbiased estimators of σ_W^2 and σ_B^2 can be derived from the equations

$$MSW = \widehat{\sigma_W^2} \quad \text{and} \quad MSB = T\widehat{\sigma_B^2} + \widehat{\sigma_W^2} \tag{2.9}$$

giving

$$\widehat{\sigma_B^2} = \frac{MSB - MSW}{T} \quad \text{and} \quad \widehat{\sigma_W^2} = MSW \tag{2.10}$$

Finally, by using the definition of the measurement reproducibility in the previous section, this can be estimated by

$$\hat{\rho} = \frac{\widehat{\sigma_B^2}}{\widehat{\sigma_B^2} + \widehat{\sigma_W^2}} \tag{2.11}$$

2.2.1.4 Interval Estimation of Measurement Reproducibility

In the case of balanced data, sums of squares divided by their expected mean squares are, under normality assumptions, distributed independently with χ^2 distributions,

i.e.

$$\frac{SSB}{T\sigma_B^2 + \sigma_W^2} \sim \chi^2_{(N-1)} \tag{2.12}$$

independently of

$$\frac{SSW}{\sigma_W^2} \sim \chi^2_{[N(T-1)]} \tag{2.13}$$

Hence,

$$\frac{\left(\frac{SSB}{T\sigma_B^2+\sigma_W^2}\right)/(N-1)}{\left(\frac{SSW}{\sigma_W^2}\right)/[N(T-1)]} \sim F_{\{N-1,N(T-1)\}} \tag{2.14}$$

i.e.

$$\frac{MSB/(T\sigma_B^2 + \sigma_W^2)}{MSW/(\sigma_W^2)} \sim F_{\{N-1,N(T-1)\}} \tag{2.15}$$

Thus, on defining appropriate lower and upper points of the F-distribution as F_U and F_L, by

$$F_L = F_{\{N-1,NT-N\ ;\alpha/2\}}$$

$$F_U = F_{\{N-1,NT-N\ ;1-\alpha/2\}}$$

one can produce a $100(1-\alpha)\%$ confidence interval using the result

$$Pr\{F_L \leq \frac{\sigma_W^2 \hat{F}}{T\sigma_B^2 + \sigma_W^2} \leq F_U\} = 1 - \alpha \tag{2.16}$$

where $\hat{F} = MSB/MSW$ and hence $(T\widehat{\sigma_B^2} + \widehat{\sigma_W^2})/\widehat{\sigma_W^2}$.
Now, after some algebra, one can have

$$Pr\left(\frac{\frac{\hat{F}}{F_U} - 1}{T} \leq \frac{\rho}{1-\rho} \leq \frac{\frac{\hat{F}}{F_L} - 1}{T}\right) = 1 - \alpha \tag{2.17}$$

where $\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$ as usual. Hence a $100(1-\alpha)\%$ confidence interval for ρ is given by

$$\left(\frac{\frac{\hat{F}}{F_U} - 1}{T + \frac{\hat{F}}{F_U} - 1}, \frac{\frac{\hat{F}}{F_L} - 1}{T + \frac{\hat{F}}{F_L} - 1}\right) \tag{2.18}$$

where T is number of replicates per individual.

2.2.1.5 A Specific Application

To illustrate the above model, data from 12 individuals undergoing a series of exercise efforts were considered, where their Ventilation using a Douglas Bag, was measured at distinct 2-minute intervals during 8 different exercise tests/visits (i.e. N=12 and T=8). For this example data from a specific time point (i.e. 6 minutes into the test) is chosen.

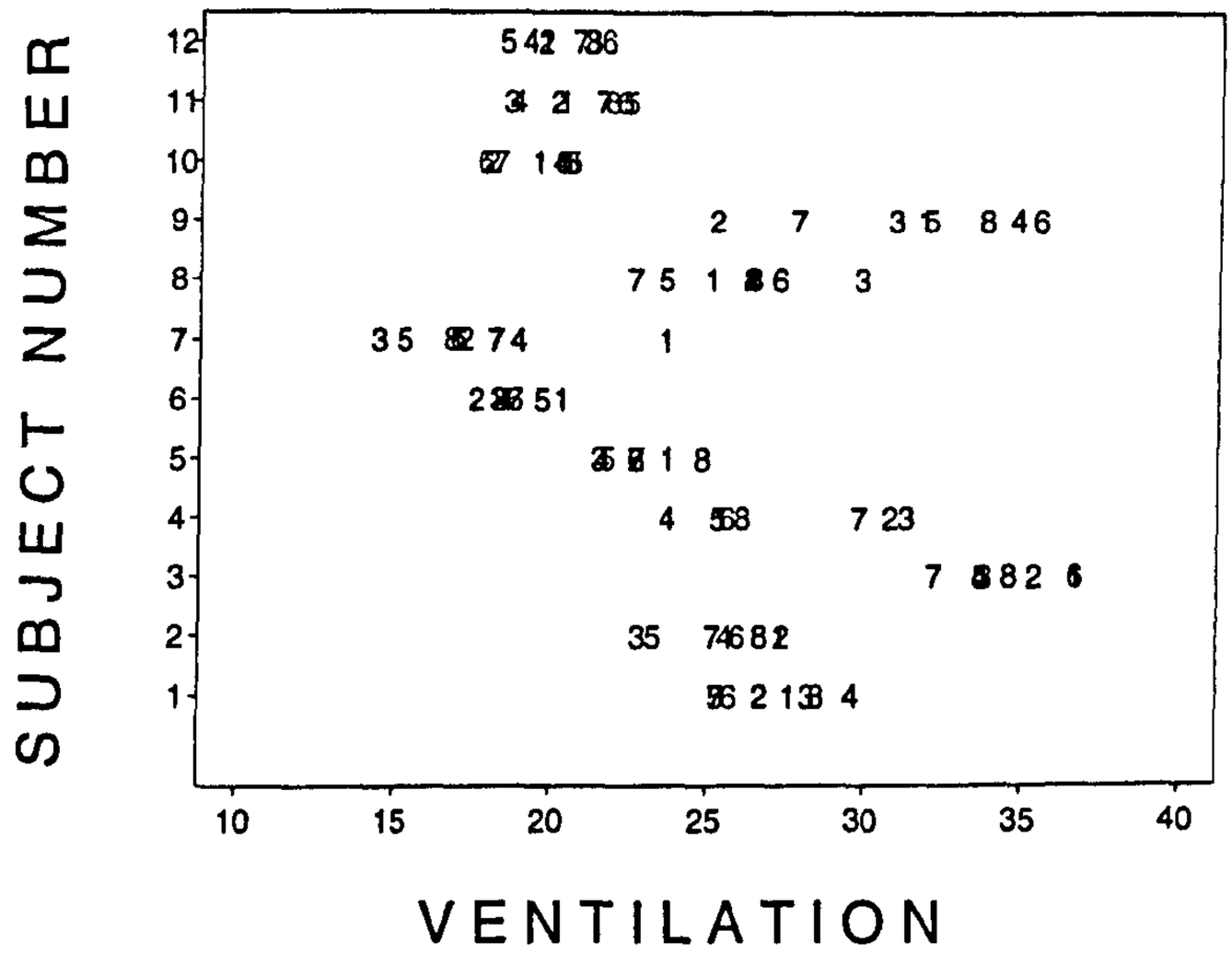

Figure 2.1: Ventilation across each of the 8 visits (labelled in the plot by the order 'number' of the visit) for each of the 12 subjects.

Figure 2.1 shows a scatterplot of these data for each of the 8 visits and for all of the 12 individuals.

Table 2.1 gives point estimates of within and between individuals variabilities as well as point and interval estimates of measurement reproducibility of the data.

$\hat{\sigma}_B$	$\hat{\sigma}_W$	$\hat{\rho}$	95% *C.I. for* ρ
5.03	1.95	0.87	(0.73 , 0.93)

Table 2.1: Point estimates of components of variance and point and interval estimates of measurement reproducibility for the Ventilation data

2.2.2 Replication Model with an Order Effect

2.2.2.1 Model

Since any set of exercise tests is likely to have a familiarisation or learning effect, it is natural to model this in order to ascertain its magnitude as well as to remove its effect from the assessment of measurement reproducibility.

In this case suppose, for each of N individuals through T (ordered) replicates, there is a learning order (or visit) effect giving rise to the following model,

$$X_{ij} = \mu + \tau_i + \beta_j + e_{j(i)}$$
$$i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, T. \tag{2.19}$$

where

X_{ij} is the j^{th} observation of the i^{th} individual,
μ is an unknown general mean,
τ_i is the difference of the i^{th} individual from μ,
β_j is the fixed order effect of the j^{th} replicate
and $e_{j(i)}$ is the measurement error.

Further,

τ_i and $e_{j(i)}$ are independent and distributed randomly as normal distributions both with mean 0 and variances σ_B^2 and σ_W^2, respectively,
i.e.

$$\tau_i \sim N(0, \sigma_B^2) \quad \text{and} \quad e_{j(i)} \sim N(0, \sigma_W^2)$$

also, $$\Sigma_{j=1}^{T} \beta_j = 0.$$

2.2.2.2 Point Estimation of Measurement Reproducibility

The only difference between this case and the simple replicate model (i.e. section 2.2.1.3), is the form of SSW and the resulting degrees of freedom which is

$$SSW = \sum_{i=1}^{N}\sum_{j=1}^{T}(X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..})^2 \qquad (2.20)$$

with (NT-N-T+1) degrees of freedom.

The relevant mean square in this case is

$$MSW = \frac{SSW}{NT - N - T + 1} \qquad (2.21)$$

The ANOVA-based method of estimating components of variations is based on equating observed and expected values of mean squares and solving for the estimators. As before,

$$MSB = T\widehat{\sigma_B^2} + \widehat{\sigma_W^2} \qquad (2.22)$$

and since

$$E(MSW) = \frac{1}{NT - N - T + 1}E[\sum_{i=1}^{N}\sum_{j=1}^{T}(X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..})^2] = \sigma_W^2 \qquad (2.23)$$

so,

$$MSW = \widehat{\sigma_W^2} \qquad (2.24)$$

and then,

$$\widehat{\sigma_B^2} = \frac{MSB - MSW}{T} \quad \text{and} \quad \widehat{\sigma_W^2} = MSW \qquad (2.25)$$

Again, using the definition of measurement reproducibility, this can be estimated by

$$\hat{\rho} = \frac{\widehat{\sigma_B^2}}{\widehat{\sigma_B^2} + \widehat{\sigma_W^2}}$$

2.2.2.3 Interval Estimation of Measurement Reproducibility

Similar to section 2.2.1.4, if one define upper and lower points of a different F-distribution as F_U and F_L,

i.e.

$$F_L = F_{\{N-1, NT-N-T+1\ ;\alpha/2\}}$$

$$F_U = F_{\{N-1, NT-N-T+1\ ;1-\alpha/2\}}$$

a $100(1-\alpha)\%$ confidence interval for the measurement reproducibility is:

$$\left(\frac{\frac{\hat{F}}{F_U} - 1}{T + \frac{\hat{F}}{F_U} - 1} , \frac{\frac{\hat{F}}{F_L} - 1}{T + \frac{\hat{F}}{F_L} - 1} \right) \tag{2.26}$$

where $\hat{F} = MSB/MSW = (T\widehat{\sigma_B^2} + \widehat{\sigma_W^2})/\widehat{\sigma_W^2}$

The difference between this case and interval estimation for the Simple Replicate Model (section 2.2.1.4), is the form of MSW and the second degrees of freedom (NT-N-T+1) of the resulting F-distribution.

2.2.2.4 A Specific Application

To illustrate the model with an order effect, each of the 12 individuals underwent 8 separate exercise tests(visits), where for each individual Breathlessness on a Visual Analogue Scale(VAS) was measured at distinct 2-minute intervals during the test. For this example data from a specific time point (i.e. 12 minutes into the test) with a significant learning effect through visits ($P < 0.05$) is chosen.

Figure 2.2 shows a scatterplot of these data for each of the 8 visits and for all of the 12 individuals.

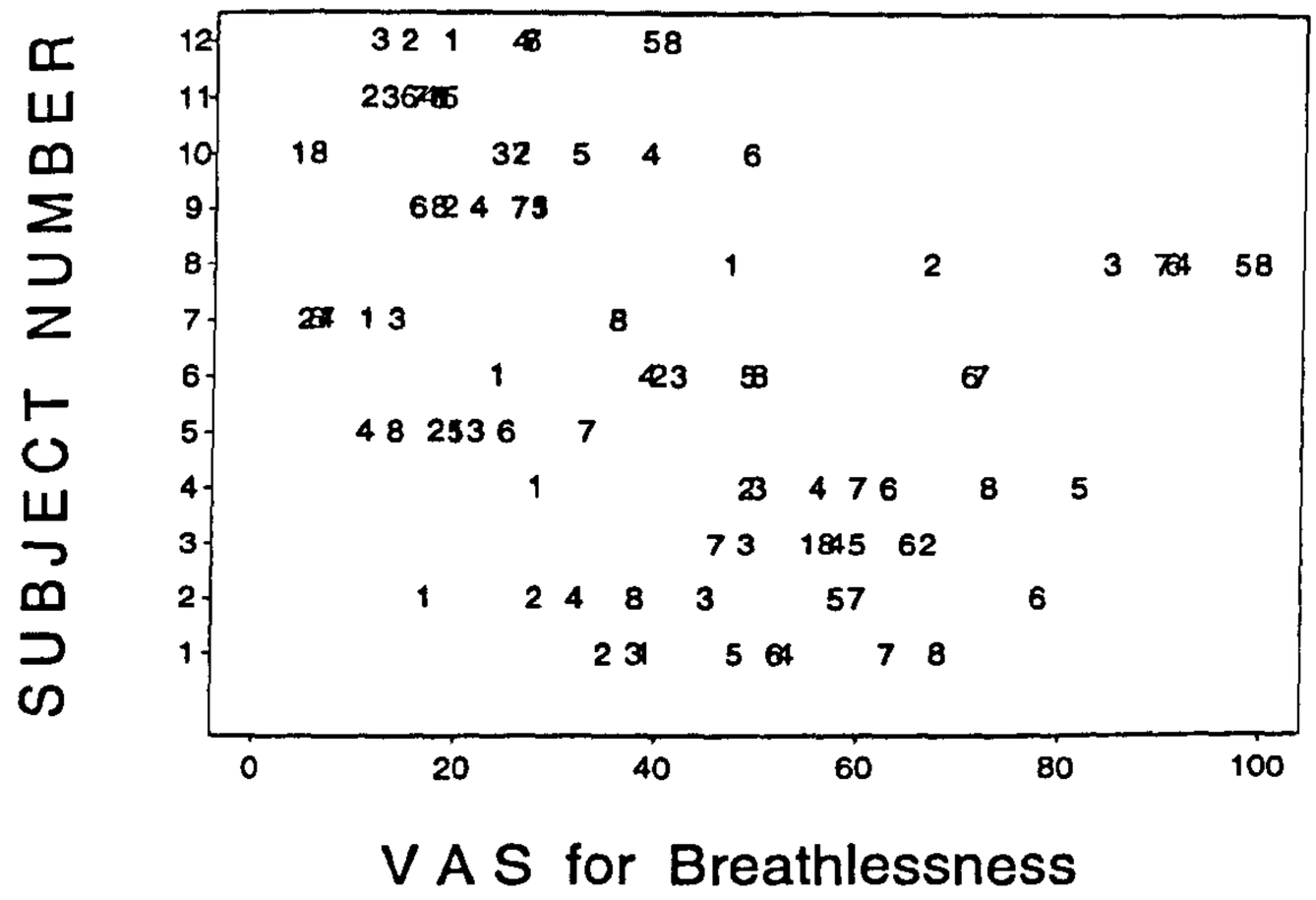

Figure 2.2: VAS for Breathlessness across each of the 8 visits (labelled in the plot by the order 'number' of the visit) for each of the 12 subjects.

	$\hat{\sigma}_B$	$\hat{\sigma}_W$	$\hat{\rho}$	95% *C.I. for* ρ
No visit effect	20.70	13.11	0.71	(0.52 , 0.88)
Visit effect	20.93	11.05	0.78	(0.61 , 0.91)

Table 2.2: Point estimates of components of variance and point and interval estimates of measurement reproducibility for the VASB data

Point estimates of within and between individuals variances as well as point and interval estimates of measurement reproducibility of the data for the two situations of without and with considering the learning/visit effect are given in Table 2.2.

For these data, when the learning/visit effect is correctly included in the model, point estimate of measurement reproducibility significantly increases from 0.71 to 0.78 and the interval estimate gets slightly narrower.

2.3 Unbalanced Data

2.3.1 Simple Replication Model

2.3.1.1 Model

Since there may actually be cases where individuals have different numbers of replicates/visits due to circumstances either unconnected with the tests such as holidays, illness etc. or in the case of exercise tests subjects may have stopped the test 'early' due to fatigue, one should consider the situation of unequal numbers of observations for each of N individuals. The appropriate model for this case would be identical to that considered previously, except that

$$X_{ij} = \mu + \tau_i + e_{j(i)} \ ,$$
$$i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, T_i \tag{2.27}$$

In this model T_i is the number of observations observed on individual i. Further let $S = \sum_{i=1}^{N} T_i$ be the total number of observations across all subjects.

2.3.1.2 Point Estimate of Measurement Reproducibility

The 'sums of squares' for unbalanced data are defined as:

$$SSB = \sum_{i=1}^{N} T_i(\overline{X}_{i.} - \overline{X}_{..})^2 = \sum_{i=1}^{N} T_i\overline{X}_{i.}^2 - S\overline{X}_{..}^2 \tag{2.28}$$

$$SSW = \sum_{i=1}^{N}\sum_{j=1}^{T_i}(X_{ij} - \overline{X}_{i.})^2 = \sum_{i=1}^{N}\sum_{j=1}^{T_i} X_{ij}^2 - \sum_{i=1}^{N} T_i\overline{X}_{i.}^2 \tag{2.29}$$

these being the same as in 2.2.1.3 for balanced data, except for having T_i in place of T.

The expected values of within-individuals and between-individuals mean squares are:

$$E(S\overline{X}_{..}^2) = SE\left\{\mu + \frac{\Sigma_{i=1}^{N} T_i\tau_i}{S} + \overline{e}_{..}\right\}^2$$

$$= S\mu^2 + \sigma_B^2 \sum_{i=1}^{N} T_i^2/S + \sigma_W^2$$

$$E(\sum_{i=1}^{N} \overline{X}_{i.}^2) = \sum_{i=1}^{N} T_i E\left\{\mu + \tau_i + \overline{e}_{i.}\right\}^2$$

$$= S\mu^2 + S\sigma_B^2 + N\sigma_W^2$$

and

$$E(\sum_{i=1}^{I}\sum_{j=1}^{n_i} X_{ij}^2) = \sum_{i=1}^{N}\sum_{j=1}^{T_i} E\left\{\mu + \tau_i + e_{ij}\right\}^2$$

$$= S(\mu^2 + \sigma_B^2 + \sigma_W^2).$$

From these results one has

$$E(MSB) = E\left(\frac{SSB}{N-1}\right) = \frac{(S - \Sigma T_i^2/S)\sigma_B^2 + (N-1)\sigma_W^2}{(N-1)}$$

$$E(MSW) = E\left(\frac{SSW}{S-N}\right) = \sigma_W^2$$

and by equating observed mean squares with their expected values, unbiased estimators of σ_B^2 and σ_W^2 can be found as follows

$$\widehat{\sigma_B^2} = \frac{MSB - MSW}{(S - \Sigma_{i=1}^{N} T_i^2/S)/(N-1)} \tag{2.30}$$

$$\widehat{\sigma_W^2} = MSW. \tag{2.31}$$

Hence a 'natural' estimate of ρ would be

$$\hat{\rho} = \frac{\widehat{\sigma_B^2}}{\widehat{\sigma_B^2} + \widehat{\sigma_W^2}} \tag{2.32}$$

2.3.1.3 Interval Estimation of Measurement Reproducibility

In unbalanced data, just as in the balanced case, $\frac{SSW}{\sigma_W^2} \sim \chi^2_{(S-N)}$ and so a confidence interval for σ_W^2 is easily derived in the same manner as in balanced data,
i.e.

$$Pr\left\{\frac{SSW}{\chi^2_{\{(S-N);\ 1-\frac{\alpha}{2}\}}} \leq \sigma_W^2 \leq \frac{SSW}{\chi^2_{\{(S-N);\ \frac{\alpha}{2}\}}}\right\} = 1-\alpha \qquad (2.33)$$

where $\chi^2_{\{(S-N);\ 1-\frac{\alpha}{2}\}}$ and $\chi^2_{\{(S-N);\ \frac{\alpha}{2}\}}$ are upper and lower bounds of the χ^2-distribution. But $\frac{SSB}{E(MSB)}$ does not follow any χ^2-distribution and there is no simple closed distribution or a multiple of it. So, despite independence of SSW and SSB, one could not provide a simple closed form of confidence interval for $\frac{\sigma_B^2}{\sigma_W^2}$ and hence ρ.
An exact confidence interval for ρ was proposed by Wald(1940) as follows:

$$W_i = \frac{\sigma_W^2}{Var(\overline{X}_{i.})} = \frac{\sigma_W^2}{\sigma_B^2 + \sigma_W^2/T_i} = \frac{T_i}{1+T_i\eta} \qquad (2.34)$$

where

$$\eta = \frac{\sigma_B^2}{\sigma_W^2} = \frac{\rho}{1-\rho} \qquad (\rho \neq 1) \qquad (2.35)$$

$$(\text{since} \quad \rho = \tfrac{\eta}{1+\eta})$$

Now define,

$$F^*(\eta) = \frac{h(\eta)}{(N-1)MSW} \qquad (2.36)$$

where,

$$MSW = \frac{\sum_{i=1}^{N}\sum_{j=1}^{T_i}(X_{ij}-\overline{X}_{i.})^2}{S-N}$$

and

$$h(\eta) = \sum_{i=1}^{N} W_i\left(\overline{X}_{i.} - \frac{\sum_{i=1}^{N} W_i\overline{X}_{i.}}{\sum_{i=1}^{N} W_i}\right)^2 \qquad (2.37)$$

Further, it can be shown that $F^*(\eta) \sim F_{\{(N-1),(S-N)\}}$.
Hence

$$F_L \leq \frac{h(\eta)}{(N-1)MSW} \leq F_U, \tag{2.38}$$

where F_L and F_U are lower and upper $\frac{\alpha}{2}$ limits of the F-distribution with (N-1) and (S-N) degrees of freedom, respectively.
From the above equation we have,

$$[(N-1)MSW]F_L \leq h(\eta) \leq [(N-1)MSW]F_U$$

Wald(1940) showed that $h(\eta)$ is a decreasing function of η, so, confidence limits for η, $\hat{\eta}_L$ and $\hat{\eta}_U$, are based on the solution of the following two equations,

$$h(\eta) = [(N-1)MSW]F_U$$

and

$$h(\eta) = [(N-1)MSW]F_L$$

Hence the corresponding induced $100(1-\alpha)\%$ confidence interval for the measurement reproducibility, ρ , is

$$\left(\frac{\hat{\eta}_L}{1+\hat{\eta}_L}\ ,\ \frac{\hat{\eta}_U}{1+\hat{\eta}_U}\right). \tag{2.39}$$

2.3.1.4 An illustrative example

To illustrate the model with unequal number of observations per each individual, data from a set of separate exercise tests on a sample of 12 individuals are considered, where their VO_2, obtained from use of a Douglas Bag, were measured at distinct 2-minute intervals during the tests. For the case where individuals may have different numbers of replicates/visits, data from a specific time point (i.e. 16 minutes into the test when in fact some 'less fit' individuals will have

dropped out) is selected. In this time point there are unequal number of observations per individual and the effect of replicates/visits is not significant($P > 0.05$).
Figure 2.3 shows a scatterplot of these data for each of the 12 individuals across different visits.

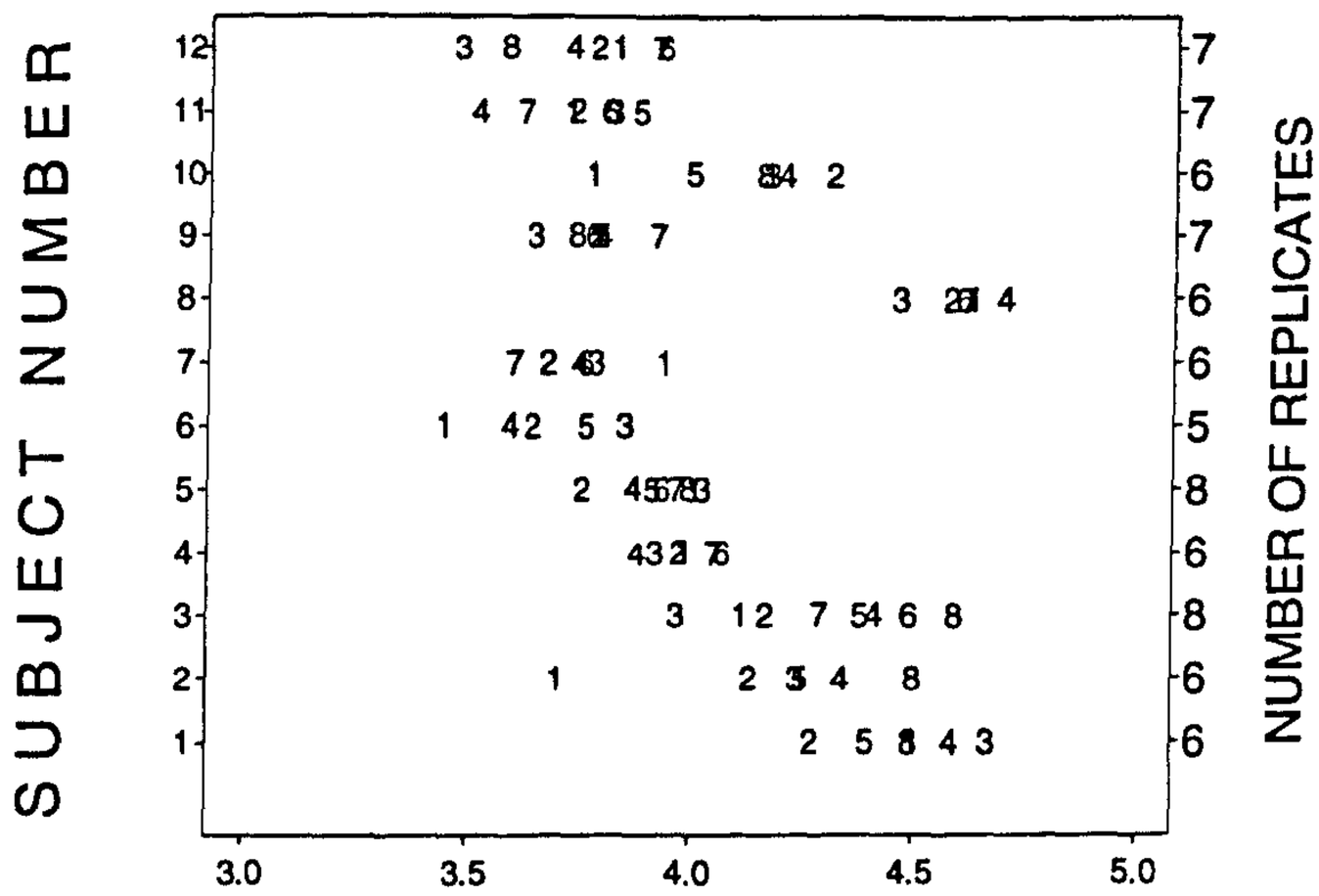

Figure 2.3: VO_2 for each of the 12 individuals and across different visits (labelled in the plot by the order 'number' of the visit) for each individual.

Point estimates of within and between individuals variabilities as well as point and interval estimates of measurement reproducibility of the data are given in Table 2.3 .

$\hat{\sigma}_B$	$\hat{\sigma}_W$	$\hat{\rho}$	95% *C.I. for* ρ
0.35	0.14	0.84	(0.62 , 0.91)

Table 2.3: Point estimates of components of variance and point and interval estimates of measurement reproducibility for the VO2 data

2.3.2 Replication Model with an Order Effect

2.3.2.1 Model

In measurement reproducibility where an order effect of replicates (visits) has to be taken into account, the model is,

$$X_{ij} = \mu + \tau_i + \beta_j + e_{j(i)}, \tag{2.40}$$

$$i = 1, 2, \ldots, N, \; j = 1, 2, \ldots, T_i,$$

In this model T_i is the number of observations on individual i.

2.3.2.2 Point Estimation of Measurement Reproducibility

For the above model an unbiased estimator for σ_W^2 is,

$$\widehat{\sigma_W^2} = \frac{SSW}{S - N - T + 1}$$

while a possible estimator for σ_B^2, based on a similar argument to equating sums of squares in the simple replication model case, is

$$\widehat{\sigma_B^2} = \frac{1}{(S - \sum_{i=1}^{N} T_i^2/S)/(N-1)} \left(\frac{SSB}{N-1} - \frac{SSW}{S - N - T + 1} \right) \tag{2.41}$$

where

T is the maximum number of replications for an individual.

i.e. $$T = \max_{1 \le i \le N} T_i$$

In the above formulas,

$$SSW = \sum_{i=1}^{N} \sum_{j=1}^{T_i} (X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..})^2$$

and SSB is the sum of squares due to τ_i (adjusted for μ).
i.e. if SSW_0 is the residual sum of squares under the model

$$X_{ij} = \mu + \tau_i + e_{j(i)}$$

and SSW_1, the residual sum of squares under the model

$$X_{ij} = \mu + \tau_i + \beta_j + e_{j(i)}$$

SSB in the above formula will be

$$SSB = SSW_0 - SSW_1$$

Again using the definition of measurement reproducibility in section (2.2.1.2), an obvious estimator of ρ is

$$\hat{\rho} = \frac{\hat{\sigma_B^2}}{\hat{\sigma_B^2} + \hat{\sigma_W^2}}. \tag{2.42}$$

2.3.2.3 Interval estimation of measurement reproducibility

To obtain an interval estimate for ρ, use a similar method as that obtained by Wald for the simple replicate model,
i.e. let

$$W_i = \frac{\sigma_W^2}{Var(\overline{X}_{i.})} = \frac{\sigma_W^2}{\sigma_B^2 + \sigma_W^2/T_i} = \frac{T_i}{1 + T_i\eta} \tag{2.43}$$

where

$$\eta = \frac{\sigma_B^2}{\sigma_W^2} = \frac{\rho}{1-\rho} \tag{2.44}$$

Now by defining,

$$F^{**}(\eta) = \frac{h(\eta)}{(N-1)MSW} \tag{2.45}$$

where,

$$MSW = \frac{\sum_{i=1}^{N}\sum_{j=1}^{T_i}(X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..})^2}{S - N - T + 1}$$

with

$$h(\eta) = \sum_{i=1}^{N} W_i \left(\overline{X}_{i.} - \frac{\sum_{i=1}^{N} W_i \overline{X}_{i.}}{\sum_{i=1}^{N} W_i}\right)^2 . \qquad (2.46)$$

Now supposing that $F^{**}(\eta) \sim F_{\{(N-1),(S-N)\}}$ for all η,
one can have

$$F_L \leq \frac{h(\eta)}{(N-1)MSW} \leq F_U, \qquad (2.47)$$

where F_L and F_U are lower and upper $\frac{\alpha}{2}$ limits of the F-distribution with (N-1) and (S-N) degrees of freedom, respectively.
From the above we have,

$$[(N-1)MSW \times F_L \ , \ (N-1)MSW \times F_U] \qquad (2.48)$$

as an approximate $100(1-\alpha)\%$ confidence interval for $h(\eta)$ and hence since $h(\eta)$ is an decreasing function of η then an approximate $100(1-\alpha)\%$ confidence interval for η is $(\hat{\eta}_L, \hat{\eta}_U)$ which are the solutions of

$$h(\eta) = [(N-1)MSW]F_U$$

and $$h(\eta) = [(N-1)MSW]F_L \ , \ \text{respectively,}$$

Hence a $100(1-\alpha)\%$ confidence interval for the measurement reproducibility, ρ will be

$$\left(\frac{\hat{\eta}_L}{1+\hat{\eta}_L} \ , \ \frac{\hat{\eta}_U}{1+\hat{\eta}_U}\right). \qquad (2.49)$$

2.3.2.4 A Specific Application

To illustrate the model with an order effect for the case of unequal number of observations per each individual, a sample of 12 individuals under exercise testing were chosen, where their Breathlessness

on a Visual Analogue Scale (VASB), were measured at distinct 2-minute intervals during the different visits. As an example where all individuals may not have the same number of replicates/visits, data from a specific time point (i.e. 18 minutes into the test when in fact some 'less fit' individuals will have dropped out) is considered here. In this case there is a significant effect of replicates/visits ($P < 0.05$).

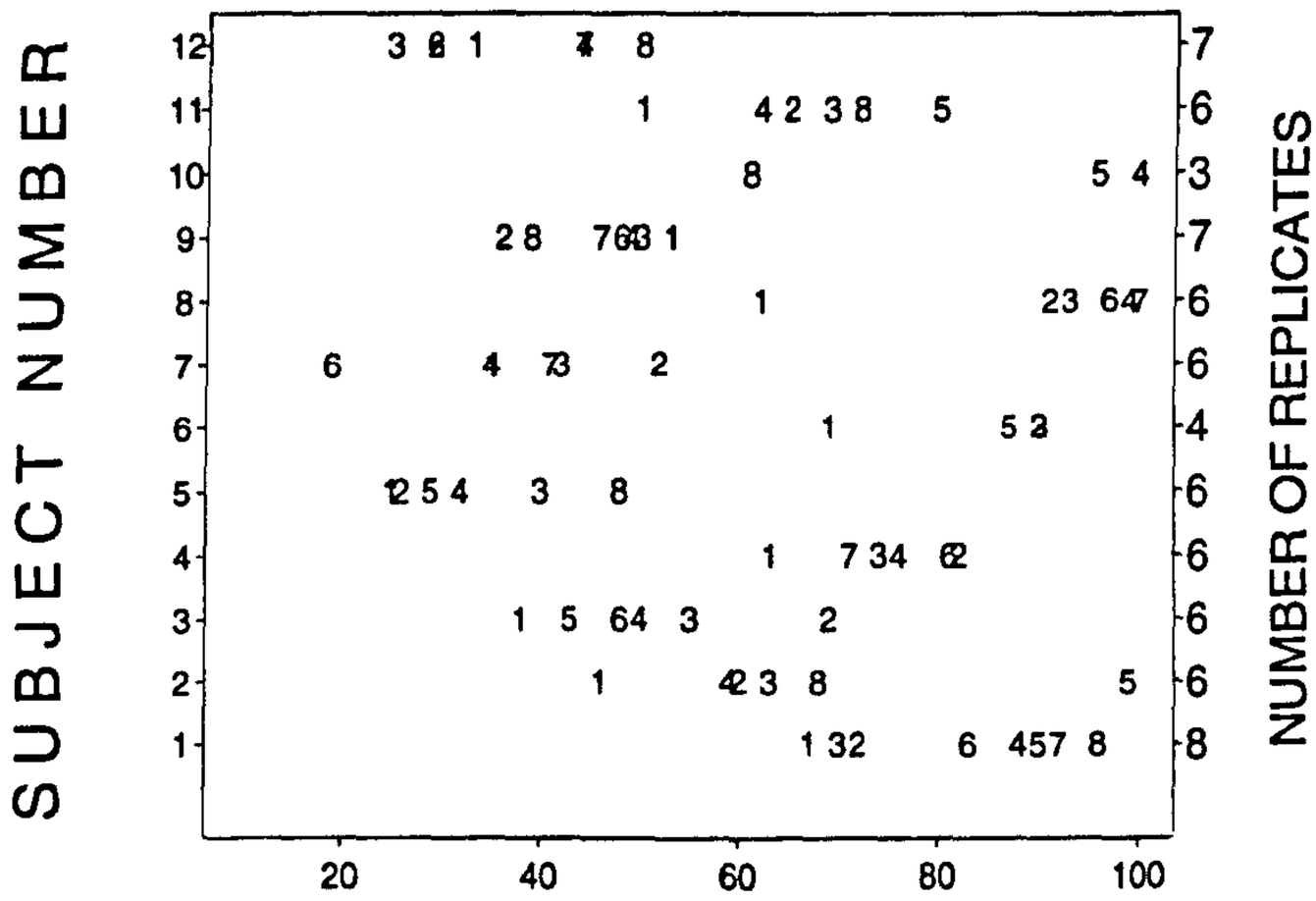

Figure 2.4: VAS for Breathlessness across each of the 8 visits (labelled in the plot by the order 'number' of the visit) for each of the 12 subjects.

Figure 2.4 shows a scatterplot of these data for each of the 8 visits and for all of the 12 individuals.

	$\hat{\sigma}_B$	$\hat{\sigma}_W$	$\hat{\rho}$	95% *C.I. of* ρ
No visit effect	24.01	12.94	0.77	(0.48 , 0.88)
Visit effect	23.74	11.57	0.81	(0.54, 0.91)

Table 2.4: Point estimates of components of variance and point and interval estimates of measurement reproducibility for the VASB data

Table 2.4 gives point estimates of within and between individuals variabilities as well as point and interval estimates of measurement reproducibility for the two situations of without and with taking into account the effect of learning/visit.

Obviously, in this case similar to that for an equal number of observations per individual (example 2.2.2.4), when the learning/visit effect is correctly included in the model, point estimate of the measurement reproducibility increases from 0.77 to 0.81 and the interval estimate gets slightly narrower.

B : Profile Likelihood Approach

2.4 Profile likelihood: a general definition

Suppose that $\underline{\theta}$ is the vector of unknown parameters in a model. We are usually not interested in all the components of $\underline{\theta}$, but rather a subset of them. If we define a $1-1$ mapping from $\underline{\theta} \rightarrow \underline{\phi}$ on such that $\phi_i = g_i(\underline{\theta}),\ i = 1, 2, \ldots, K$, we want to make inference about $\underline{\phi}_1$, where,

$$\underline{\phi} = \begin{pmatrix} \underline{\phi}_1 \\ \underline{\phi}_2 \end{pmatrix} \tag{2.50}$$

such that $\underline{\phi}_1$ is the vector of interest parameters and $\underline{\phi}_2$ is the vector of nuisance parameters. One method of eliminating the nuisance parameters is to maximize over them holding $\underline{\phi}_1$ as fixed, and define the profile likelihood for $\underline{\phi}_1$ as,

$$\begin{aligned} PLik(\underline{\phi}_1; \underline{X}) &= \max_{\underline{\phi}_2} Lik(\underline{\phi}_1, \underline{\phi}_2; \underline{X}) \\ &= Lik(\underline{\phi}_1, \hat{\underline{\phi}}_2(\underline{\phi}_1); \underline{X}) \end{aligned} \tag{2.51}$$

Then the log profile likelihood is

$$log_e PLik(\underline{\phi}_1; \underline{X}) = log_e Lik(\underline{\phi}_1, \hat{\underline{\phi}}_2(\underline{\phi}_1); \underline{X}) \tag{2.52}$$

and the relative log profile likelihood is,

$$rpl(\underline{\phi}_1; \underline{X}) = log_e Lik(\underline{\phi}_1, \hat{\underline{\phi}}_2(\underline{\phi}_1); \underline{X}) - log_e Lik(\hat{\underline{\phi}}_1, \hat{\underline{\phi}}_2; \underline{X}). \tag{2.53}$$

An $100Q\%\ \ (Q < 1)$ likelihood interval estimate for $\underline{\phi}_1$, based on profile likelihood, is the set of $\underline{\phi}_1$-values for which,

$$\left\{ \underline{\phi}_1 : rpl(\underline{\phi}_1; \underline{X}) \geq log_e Q \right\}. \tag{2.54}$$

If it is possible to make the standard likelihood assumption that

$$-2rpl(\underline{\phi}_1; \underline{X}) \sim \chi^2_{(1)} \tag{2.55}$$

then a likelihood interval with approximate $100(1-\alpha)\%$ confidence would be

$$\left[\underline{\phi}_1 : -2rpl(\underline{\phi}_1; \underline{X}) \leq \chi^2_{\{1;(1-\alpha)\}}\right]. \tag{2.56}$$

i.e. taking $log_e Q = -\frac{1}{2}\chi^2_{\{1;(1-\alpha)\}}$ give a $100Q\%$ likelihood interval as having roughly $100(1-\alpha)\%$ confidence.

2.5 Balanced Data

2.5.1 General Model

Here a more convenient way of writing the model used in section 2.1 is

$$\underline{X} = A\mu + U\underline{\tau} + \underline{\varepsilon} \tag{2.57}$$

where for N individuals and T, replicates per individual, also let $S = TN$ be the total number of observations,
Thus here

$\underline{X}$ is an $S \times 1$ vector of observations;
A is the $S \times (T+1)$ design matrix of zeros and ones;
$\underline{\mu}$ is an appropriate vector of one or more parameters depending on the particular model under consideration, e.g. $\underline{\mu} = (\mu)$ for the model 2.1 where

$$\underline{\mu}^t = (\mu, \beta_1, \ldots, \beta_T)$$

for the visit effect model of 2.2.2.1;
U is an $S \times N$ known vector of zeros and ones of the

form

$$U = \begin{pmatrix} \underline{1}_T & 0 & \dots & 0 \\ 0 & \underline{1}_T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \underline{1}_T \end{pmatrix}$$

where

$$\underline{1}^t = (1, 1, \dots, 1)$$

i.e. a t-vector of 1's;

and $\underline{\tau}$ is an $N \times 1$ vector of individual random effects

i.e.
$$\underline{\tau} = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_N \end{pmatrix}$$

Further

$\underline{\varepsilon}$ is an $S \times 1$ vector of error values, with the assumptions that $\underline{\tau}$ and $\underline{\varepsilon}$ are distributed independently with multivariate normal distributions of mean $\underline{0}$ and variances $\sigma_B^2 I_N$ and $\sigma_W^2 I_S$, respectively,
i.e.

$$\underline{\tau} \sim MN_N(\underline{0}\ ,\ \sigma_B^2 I_N), \quad \underline{\varepsilon} \sim MN_S(\underline{0}\ ,\ \sigma_W^2 I_S)$$

Applying these assumptions, one has

$$E(\underline{X}) = A\underline{\mu} \tag{2.58}$$

$$\begin{aligned} Var(X) &= Var(U\underline{\tau}) + Var(\underline{\varepsilon}) \\ &= U\sigma_B^2 I_N U^t + \sigma_W^2 I_S \\ &= \sigma_B^2 UU^t + \sigma_W^2 I_S \end{aligned} \tag{2.59}$$

For notational simplicity in future denote $\sigma_B^2 UU^t + \sigma_W^2 I_S$ by Σ.

2.5.1.1 Point Estimation of Measurement Reproducibility

Here there are 3 unknown parameters, $\underline{\mu}$, σ_B, and σ_W, the log likelihood function in this case is defined as,

$$logL = l = -\frac{S}{2}log(2\pi) - \frac{1}{2}log|\Sigma| - \frac{1}{2}(\underline{X} - A\underline{\mu})^t\Sigma^{-1}(\underline{X} - A\underline{\mu}) \quad (2.60)$$

Now, in fact, it is handier in this approach to reparameterise σ_B and σ_W as follows:

Remembering the definition of measurement reproducibility

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2},$$

and taking $\kappa = \sigma_B^2 + \sigma_W^2$, one has

$$\Sigma = \kappa V_\rho$$

where V_ρ is an $S \times S$ matrix of the form,

$$V_\rho = \begin{pmatrix} B_1 & 0 & \dots & 0 \\ 0 & B_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & B_N \end{pmatrix} \quad (2.61)$$

with each B_i a $T \times T$ matrix of the form

$$B_i = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}, \quad i = 1, 2, \dots, N$$

$$= (1-\rho)I_T + \rho\underline{1}_T\,\underline{1}_T^t$$

Hence using the standard form of the inverse of such a patterned matrix

$$B_i^{-1} = \frac{1}{(1-\rho)[1+(T-1)\rho]} \begin{pmatrix} \lambda_\rho & -\rho & \dots & -\rho \\ -\rho & \lambda_\rho & \dots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \dots & \lambda_\rho \end{pmatrix}$$

where

$$\lambda_\rho = 1 + (T-2)\rho$$

with T the number of replicates (visits) for each individual. Further

$$\Sigma^{-1} = (\kappa V_\rho)^{-1} = \frac{1}{\kappa} V_\rho^{-1} \tag{2.62}$$

and,

$$|\Sigma| = \kappa^S \left\{ (1-\rho)^{N(T-1)} \left[1 + (T-1)\rho\right]^N \right\} \quad \text{as long as} \quad |\rho| < 1 \tag{2.63}$$

Now returning to the full likelihood function reparameterised in terms of ρ then, after some algebra, one can write (2.60) as

$$\begin{aligned} l \;=\; & -\frac{S}{2} log(2\pi) - \frac{S}{2} log(\kappa) - \frac{N(T-1)}{2} log(1-\rho) - \frac{N}{2} log\left[1 + (T-1)\rho\right] \\ & -\frac{1}{2\kappa} (\underline{X} - A\underline{\mu})^t V_\rho^{-1} (\underline{X} - A\underline{\mu}). \end{aligned} \tag{2.64}$$

Hence, for known ρ, maximum likelihood estimators can be achieved by equating to zero the partial derivative of l with respect to $\underline{\mu}$ and κ to produce

$$\underline{\hat{\mu}}(\rho) = (A^t V_\rho^{-1} A)^{-1} A^t V_\rho^{-1} \underline{X} \tag{2.65}$$

and

$$\hat{\kappa}(\rho) = \frac{1}{S} \left(\underline{X} - A\underline{\hat{\mu}}(\rho)\right)^t V_\rho^{-1} \left(\underline{X} - A\underline{\hat{\mu}}(\rho)\right) \tag{2.66}$$

where

$$V_\rho^{-1} = \frac{1}{(1-\rho)\left[1 + (T-1)\rho\right]} \begin{pmatrix} C_1 & 0 & \dots & 0 \\ 0 & C_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & C_N \end{pmatrix}$$

with each C_i a $T \times T$ matrix of the form

$$C_i \;=\; \begin{pmatrix} \lambda_\rho & -\rho & \dots & -\rho \\ -\rho & \lambda_\rho & \dots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \dots & \lambda_\rho \end{pmatrix}, \quad i = 1, 2, \dots, N$$

$$= \lambda_\rho I_T - \rho \underline{1}_T \, \underline{1}_T^t$$

Replacing $\underline{\hat{\mu}}(\rho)$ and $\hat{\kappa}(\rho)$ in (2.64) gives the log profile likelihood for ρ as

$$logPLik(\rho) = -\frac{S}{2} log(\hat{\kappa}) - \frac{N}{2} log\left[1 + (T-1)\rho\right](1-\rho)^{T-1} \quad (2.67)$$

Now such a profile likelihood can be maximised by some iterative procedure to produce a point estimate for ρ or indeed an interval estimate.

2.5.1.2 Interval Estimation for Measurement Reproducibility

The relative profile likelihood required to provide a approximate $100Q\%$ likelihood interval for ρ is

$$\begin{aligned} rpl(\rho) &= logPLik(\rho) - logPLik(\hat{\rho}) \\ &= \frac{S}{2} log\left(\frac{\hat{\kappa}_{\hat{\rho}}}{\hat{\kappa}_\rho}\right) + \frac{N}{2} log\left\{\frac{\left[1 + (T-1)\hat{\rho}\right](1-\hat{\rho})^{T-1}}{\left[1 + (T-1)\rho\right](1-\rho)^{T-1}}\right\} \end{aligned} \quad (2.68)$$

where $\hat{\rho}$ is the maximum (profile) likelihood estimate for ρ. The approximate $100Q\%$ likelihood interval for ρ is

$$\{\rho : \; rpl(\rho) \geq log(Q)\} \quad (2.69)$$

Thus based on 2.55 and 2.56, a likelihood interval with approximate 95% confidence would be

$$\left\{\rho : \; rpl(\rho) \geq -\frac{1}{2}\chi^2_{(1,\ 0.95)}\right\}$$

Further, if one choose $Q = 0.147$, the likelihood interval with approximate 95% confidence will be the corresponding 14.7% likelihood interval.

Now these general results will be looked at for the cases considered previously, i.e. a simple replicate experiment by itself and then assuming visit/order effects both for balanced (i.e. same number of replicates per subject) and unbalanced (i.e. unequal number of replicates per subject) cases.

2.5.2 Simple Replication Model

2.5.2.1 Model

In this case, in the general model given by

$$X = A\underline{\mu} + U\underline{\alpha} + \underline{\varepsilon}$$

one has

$$A = \begin{pmatrix} \underline{1}_T \\ \underline{1}_T \\ \vdots \\ \underline{1}_T \end{pmatrix}$$

i.e. A a vector of 1's

and $\underline{\mu} = (\ \mu\ )$

i.e. μ is a scalar

2.5.2.2 Point Estimate of Measurement Reproducibility

From (2.66) one has

$$\hat{\kappa}(\rho) = \frac{1}{S} \frac{[1 + (T-2)\rho] \sum_{j=1}^{T} S_{jj} - 2\rho \sum_{j=1}^{T} \sum_{j^*=j+1}^{T} S_{jj^*}}{[1 + (T-1)\rho](1-\rho)} \qquad (2.70)$$

as long as $|\rho| < 1$,

where

$$S_{jj} = \sum_{i=1}^{N}(x_{ij} - \overline{x}_{..})^2, \quad \forall\ j = 1, 2, \ldots, T$$

$$S_{jj^*} = \sum_{i=1}^{N}(x_{ij} - \overline{x}_{..})(x_{ij^*} - \overline{x}_{..}), \quad \forall\ j \neq j^*$$

A maximum profile likelihood estimator of ρ, can be easily obtained from (2.67) as

$$\hat{\rho}_{PL} = \frac{2\sum_{j=1}^{T}\sum_{j^*=j+1}^{T} S_{jj^*}}{(T-1)\sum_{j=1}^{T} S_{jj}} \tag{2.71}$$

where T is the number of replicates per subject

2.5.2.3 Likelihood Interval for Measurement Reproducibility

From (2.5.1.2) a $100Q\%$ likelihood interval for this case is of the form

$$\{\rho : rpl(\rho) \geq log(Q)\}$$

where $rpl(\rho)$, as in (2.68), is

$$\begin{aligned} rpl(\rho) &= logPLik(\rho) - logPLik(\hat{\rho}_{PL}) \\ &= \frac{S}{2} log\left(\frac{\hat{\kappa}_{\hat{\rho}_{PL}}}{\hat{\kappa}_{\rho}}\right) + \frac{N}{2} log\left\{\frac{[1+(T-1)\hat{\rho}_{PL}]\,(1-\hat{\rho}_{PL})^{T-1}}{[1+(T-1)\rho]\,(1-\rho)^{T-1}}\right\} \end{aligned}$$

with approximate likelihood intervals based on section 2.5.1.2.

2.5.2.4 A specific Application

To illustrate the use of the above model and compare it with the model in section 2.2.1, the same data in exercise testing as used in the example 2.2.1.5 is used. Each of the 12 individuals underwent

8 separate exercise tests, where their Ventilation using a Douglas Bag, were measured at distinct 2-minute intervals.

In this example, the point estimate for measurement reproducibility is 0.86 and an approximate 95% confidence interval for this is (0.76 , 0.92).

Point and interval estimate for this approach and for the ANOVA-based approach (point estimate of 0.87 and interval estimate of 0.73 to 0.93) in section 2.2.1 are presented in Figure 2.5.

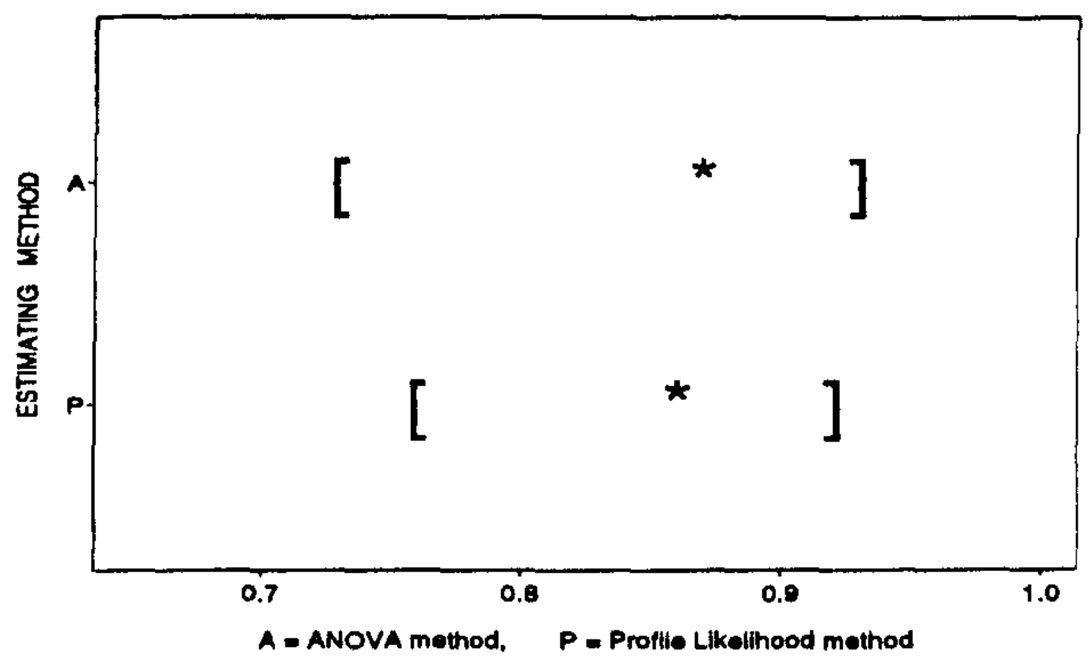

Figure 2.5: Point and interval estimates by each of the two methods of estimating measurement reproducibility

Here the Profile Likelihood approach provides a slightly smaller point estimate but, more importantly, reduces the width of the interval estimate from 0.20 to 0.16.

2.5.3 Replication Model with an Order Effect

2.5.3.1 Model

For the situation where possible learning/familiarisation effects of visits (replicates) are to be considered, the general model of

$$X = A\underline{\mu} + U\underline{\tau} + \underline{\varepsilon}$$

produces A to be an $S \times (T+1)$ matrix of the form

$$A = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix}$$

and $(T+1) \times 1$ vector of $\underline{\mu}$ is

$$\underline{\mu}^t = (\mu, \beta_1, \beta_2, \dots, \beta_T)$$

2.5.3.2 Point Estimation of Measurement Reproducibility

In this case $\hat{\kappa}$ can be written in the form

$$\hat{\kappa}(\rho) = \frac{1}{S} \frac{[1 + (T-2)\rho] \sum_{j=1}^{T} S^{\star}_{jj} - 2\rho \sum_{j=1}^{T} \sum_{j^*=j+1}^{T} S^{\star}_{jj^*}}{[1 + (T-1)\rho](1-\rho)} \qquad (2.72)$$

as long as $|\rho| < 1$,

but here

$$S^{\star}_{jj} = \sum_{i=1}^{N} (x_{ij} - \overline{x}_{.j})^2, \quad \forall\ j = 1, 2, \dots, T$$

$$S^{\star}_{jj^*} = \sum_{i=1}^{N} (x_{ij} - \overline{x}_{.j})(x_{ij^*} - \overline{x}_{.j^*}), \quad \forall\ j \neq j^*.$$

the maximum profile likelihood estimator of ρ can be obtained as

$$\hat{\rho}_{PLV} = \frac{2\sum_{j=1}^{T}\sum_{j=i+1}^{T} S^{\star}_{jj^{*}}}{(T-1)\sum_{j=1}^{T} S^{\star}_{jj}} \tag{2.73}$$

with T the number of replicates per subject.

2.5.3.3 Likelihood Interval for Measurement Reproducibility

As usual from section 2.5.1.2 a $100Q\%$ likelihood interval for this case is of the form

$$\{\rho : rpl(\rho) \geq log(Q)\}$$

where

$$\begin{aligned} rpl(\rho) &= logPLik(\rho) - logPLik(\hat{\rho}_{PLV}) \\ &= \frac{S}{2}log\left(\frac{\hat{\kappa}_{\hat{\rho}_{PLV}}}{\hat{\kappa}_{\rho}}\right) + \frac{N}{2}log\left\{\frac{[1+(T-1)\hat{\rho}_{PLV}](1-\hat{\rho}_{PLV})^{T-1}}{[1+(T-1)\rho](1-\rho)^{T-1}}\right\}. \end{aligned}$$

2.5.3.4 A specific Application

To illustrate the Profile Likelihood approach in the case of existence of an order effect and also compare it with the ANOVA approach, example 2.2.2.4 (section 2.2.1) is considered. For each of the 12 samples into Exercise Testing, Breathlessness on a Visual Analogue Scale(VAS), were measured at distinct 2-minute intervals during test. The test was repeated on 8 different visits, so, there are equal number of observations for each individual. Furthermore, as noted before, the learning/visit effect is significant ($P < 0.05$).

For this example, point estimate of measurement reproducibility without considering the learning/visit effect is 0.70 with a 95% like-

lihood interval of (0.54 , 0.87), whereas, after taking into account the learning/visit effect, point estimate would be 0.78 with a 95% confidence interval of (0.69 , 0.93).

To compare the estimates of measurement reproducibility using this approach with those based on the ANOVA-based approach (section 2.2.2), the results from the two methods are represented in Figure 2.6.

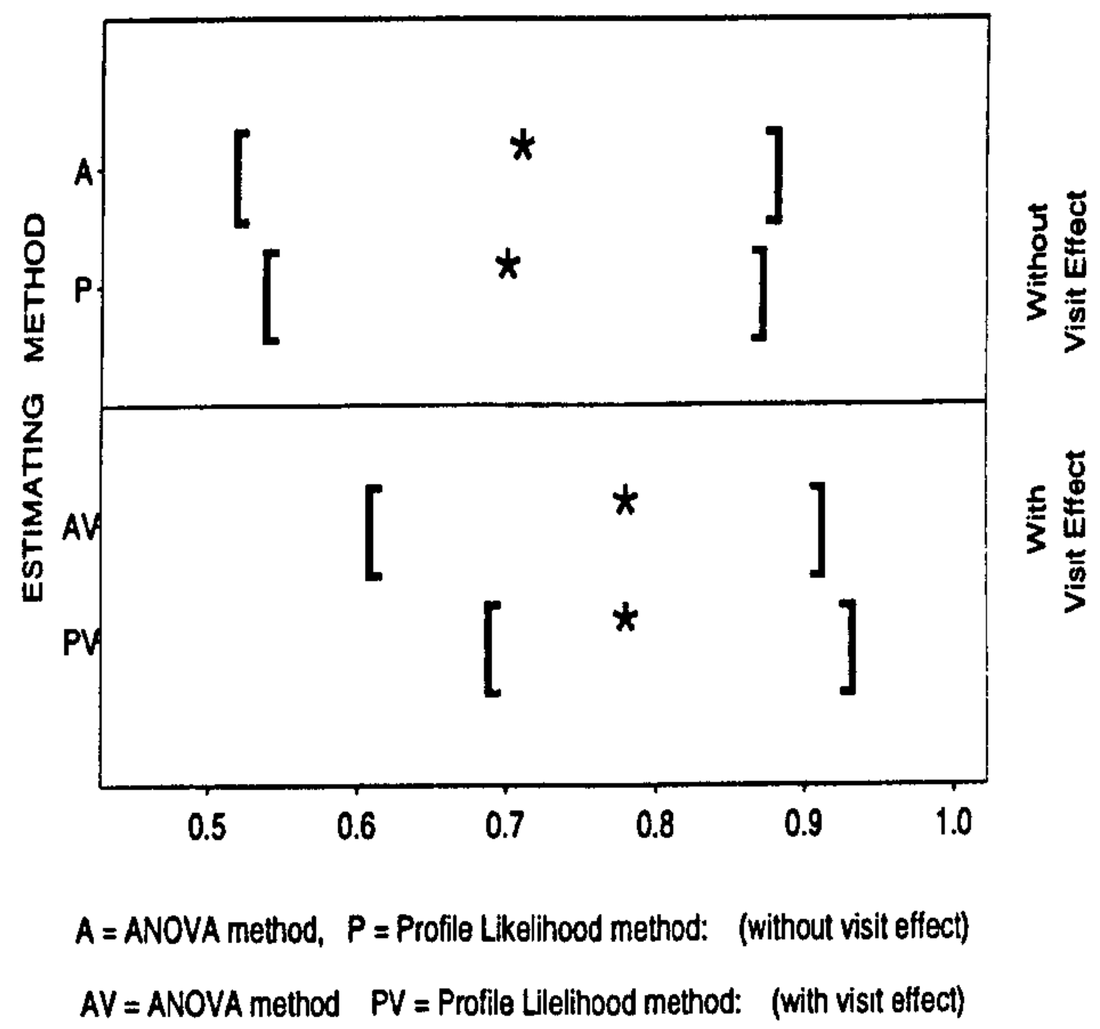

Figure 2.6: Point and interval estimates by each of the two methods of estimating measurement reproducibility both with and without fitting a visit effect

The results in this example and also in example 2.2.2.4 give an impression that including a significant learning/visit effect in the model results in 'improvement' (for both approaches) in the point and interval estimates of measurement reproducibility.

2.6 Unbalanced Data

2.6.1 General Model

In the case of unbalanced data in which there are not the same number of observations (replicates) for each individual, the general model can still have the form of

$$X = A\underline{\mu} + U\underline{\alpha} + \underline{\varepsilon} \tag{2.74}$$

Further let T_i be the number of replicates for individual i, $i = 1, 2, \ldots, N$, $S = \sum_{i=1}^{N} T_i$ total number of observations and

$$T = \max_{1 \le i \le N} T_i$$

2.6.1.1 Point Estimation of Measurement Reproducibility

In this case the likelihood function and log likelihood function have the same form as in the balanced case with the exception that in the matrix V_ρ (2.61) each B_i is a $T_i \times T_i$ matrix of the form

$$B_i = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}, \quad i = 1, 2, \ldots, N$$

$$= (1-\rho) I_{T_i} + \rho \underline{1}_{T_i} \underline{1}_{T_i}^t$$

so as before

$$B_i^{-1} = \frac{1}{(1-\rho)[1+(T_i-1)\rho]} \begin{pmatrix} \lambda_\rho & -\rho & \ldots & -\rho \\ -\rho & \lambda_\rho & \ldots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \ldots & \lambda_\rho \end{pmatrix}$$

with

$$\lambda_i = 1 + (T_i - 2)\rho$$

where T_i is the number of replicates for subject i.
Further

$$\Sigma^{-1} = (\kappa V_\rho)^{-1} = \frac{1}{\kappa} V_\rho^{-1} \tag{2.75}$$

and

$$\begin{aligned} |\Sigma| &= \kappa^S |\nu_\rho| \\ &= \kappa^S \prod_{j=1}^{T} \left\{(1-\rho)^{j-1}\left[1 + (j-1)\rho\right]\right\}^{n_j} \end{aligned}$$

where n_j, is the number of subjects with j replicates $j = 1, 2, \ldots, T$.
After some algebra, one can write (2.60) as

$$\begin{aligned} l &= -\frac{S}{2} log(2\pi) - \frac{S}{2} log\kappa - \frac{1}{2} log \prod_{j=1}^{T} \left\{(1-\rho)^{j-1}\left[1 + (j-1)\rho\right]\right\}^{n_j} \\ &\quad -\frac{1}{2\pi}(\underline{X} - A\underline{\mu})^t V_\rho^{-1}(\underline{X} - A\underline{\mu}) \end{aligned} \tag{2.76}$$

and by equating the partial derivatives with respect to $\underline{\mu}$ and κ to zero, for known ρ, $\underline{\hat{\mu}}(\rho)$ and $\hat{\kappa}(\rho)$ will be the same form as in section 2.5.1.1 with the exception that T changes to T_i

$$\underline{\hat{\mu}}(\rho) = (A^t V_\rho^{-1} A)^{-1} A^t V_\rho^{-1} \underline{X} \tag{2.77}$$

and

$$\hat{\kappa}(\rho) = \frac{1}{S}\left(\underline{X} - A\underline{\hat{\mu}}(\rho)\right)^t V_\rho^{-1}\left(\underline{X} - A\underline{\hat{\mu}}(\rho)\right) \tag{2.78}$$

where

$$V^{-1} = \begin{pmatrix} \frac{1}{\xi_1} C_1 & 0 & \ldots & 0 \\ 0 & \frac{1}{\xi_2} C_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{\xi_N} C_N \end{pmatrix}$$

in this matrix

$$\xi_i = (1-\rho)[1 + (T_i - 1)\rho]$$

and each C_i is a $T_i \times T_i$ matrix of the form

$$B_i = \begin{pmatrix} \lambda_i & -\rho & \dots & -\rho \\ -\rho & \lambda_i & \dots & -\rho \\ \vdots & \vdots & \ddots & \vdots \\ -\rho & -\rho & \dots & \lambda_i \end{pmatrix}, \quad i = 1, 2, \dots, N$$

$$= \lambda_i I_{T_i} - \rho \underline{1}_{T_i} \underline{1}_{T_i}^t$$

Now, after some algebra, the log profile likelihood for ρ will be

$$logPLik(\rho) = -\sum_{j=1}^{T} \frac{n_j}{2} log\left\{[1 + (j-1)\rho](1-\rho)^{j-1}\right\} - \frac{S}{2} log\hat{\kappa}_\rho \tag{2.79}$$

and such a log profile likelihood can be maximised to produce a point estimate for ρ.

2.6.1.2 Interval Estimation for Measurement Reproducibility

By using the relative profile log likelihood

$$\begin{aligned} rpl(\rho) &= logPLik(\rho) - logPLik(\hat{\rho}) \\ &= \frac{S}{2} log(\frac{\hat{\kappa}_{\hat{\rho}}}{\kappa_\rho}) + \sum_{j=1}^{T} \frac{n_j}{2} log \frac{\left\{[1 + (j-1)\hat{\rho}](1-\hat{\rho})^{j-1}\right\}}{\left\{[1 + (j-1)\rho](1-\rho)^{j-1}\right\}} \end{aligned} \tag{2.80}$$

a $100Q\%$ likelihood interval for ρ is

$$\{\rho : rpl(\rho) \geq log(Q)\} \tag{2.81}$$

and for $Q = 0.147$, a likelihood interval with approximate 95% confidence is achieved, since

$$-\frac{1}{2}\chi^2_{(1,0.95)} = log(0.147).$$

2.6.2 Simple Replicate Model

2.6.2.1 Model

For this case A and $\underline{\mu}$ in the general model have the following simple forms

$$A = \begin{pmatrix} \underline{1}_{T_1} \\ \underline{1}_{T_2} \\ \vdots \\ \underline{1}_{T_N} \end{pmatrix} \quad \text{and} \quad \underline{\mu} = (\mu)$$

2.6.2.2 Point Estimation of Measurement Reproducibility

From (2.66) after some algebra, $\hat{\kappa}_\rho$ is found to be

$$\hat{\kappa}_\rho = \frac{1}{S}\left\{ S_{11} + \sum_{i=2}^{T} \frac{1}{[1+(i-1)\rho](1-\rho)} \left[[1+(i-2)\rho] \sum_{j=1}^{i} S_{ij} - 2\rho \sum_{j=1}^{i} \sum_{j^*=j+1}^{i} S_{ijj^*} \right] \right\} \tag{2.82}$$

where

$$S_{ij} = \sum_{k=1}^{n_j} (x_{ijk} - \overline{x}_{...})^2$$

$$S_{ijj^*} = \sum_{k=1}^{n_j} (x_{ijk} - \overline{x}_{...})(x_{ij^*k} - \overline{x}_{...}),$$

$$\forall\ i = 1, 2, \ldots, T, \quad j \neq j^*, \quad j, j^* = 1, 2, \ldots, i.$$

A simple closed form for $\hat{\rho}$ is not obtainable in this context and hence numerical methods must be used to obtain both point and interval estimates. Either a simple search method or an iterative hill-climbing procedure such as Newton-Raphson can be used.

2.6.2.3 A specific Application

To illustrate the use of the above model and compare it with the model from section 2.3.1, the data from example 2.3.1.4 is used. Each of the 12 individuals underwent separate exercise tests, where their VO2 using a Douglas Bag, were measured at distinct 2-minute intervals. Since individuals had different number of replicates/visits, there are unequal number of observations per each individual. Moreover, the effect of learning/visits is not significant.

In this example, a point estimate of measurement reproducibility by the Profile Likelihood method is 0.85 and a 95% likelihood interval is (0.65 , 0.93).

To comparing the estimated values by this approach and the ANOVA-based approach, a graphical representation of the results by the two methods are shown in Figure 2.7.

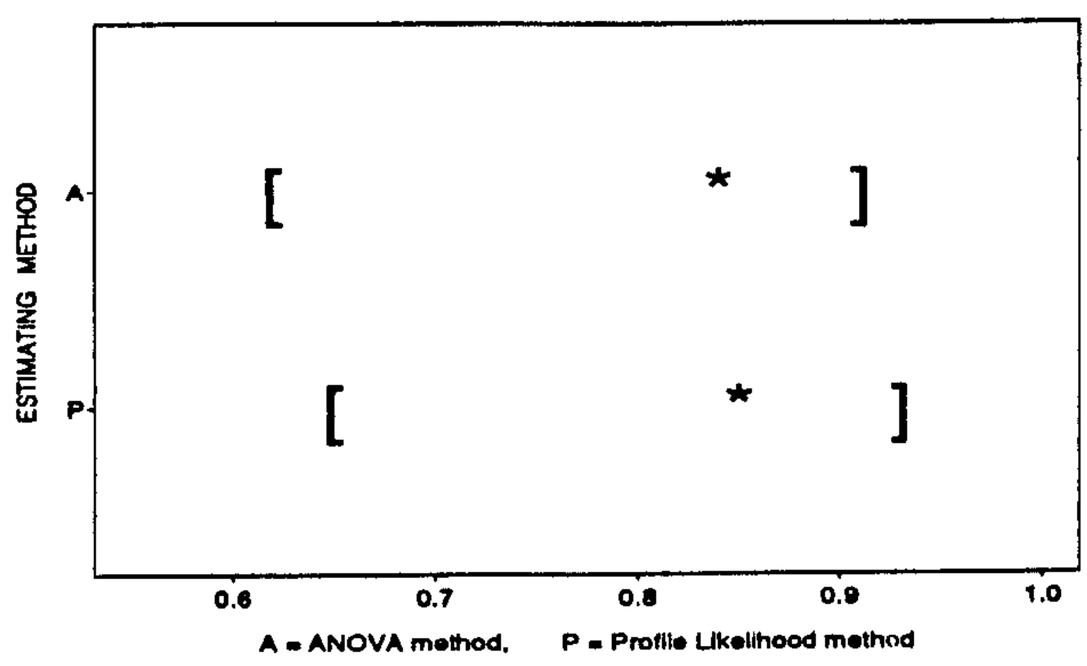

Figure 2.7: Point and interval estimates by each of the two methods of estimating measurement reproducibility

The point estimate by the Profile Likelihood approach is slightly higher than that by the ANOVA-based approach and the likelihood interval is narrower.

2.6.3 Measurement Reproducibility With an Order Effect

2.6.3.1 Model

In the case of unbalanced data where the effects of visits (replicates) per each individual must be taken into account, then model 2.57 should have

A as $S \times (T+1)$, matrix of the form

$$A = \begin{pmatrix} \underline{1}_{T_1} & I_{T_1} \\ \underline{1}_{T_2} & I_{T_2} \\ \vdots & \vdots \\ \underline{1}_{T_N} & I_{T_N} \end{pmatrix}$$

and the vector $\underline{\mu}$

$$\underline{\mu}^t = (\mu, \beta_1, \beta_2, \ldots, \beta_T) \quad \text{and} \quad \sum_{j=1}^{T} \beta_j = 0$$

where $T_i,\ (i = 1, 2, \ldots, N)$ is the number of replicates for subject i and

$$T = \max_{1 \le i \le N} T_i$$

2.6.3.2 Point Estimation of Measurement Reproducibility

In this case similar to (2.82) one finds that

$$\hat{\kappa}_\rho = \frac{1}{S}\left\{ S^*_{11} + \sum_{i=2}^{T} \frac{1}{[1+(i-1)\rho](1-\rho)} \left[[1+(i-2)\rho] \sum_{j=1}^{k} S^*_{ij} - 2\rho \sum_{j=1}^{i} \sum_{i=j+1}^{i} S^*_{ijj^*} \right] \right\}$$

with

$$S^*_{ij} = \sum_{k=1}^{n_j} (x_{ijk} - \overline{x}_{.j.})^2$$

and

$$S^*_{ijj^*} = \sum_{k=1}^{n_j} (x_{ijk} - \overline{x}_{.j.})(x_{ij^*k} - \overline{x}_{.j^*.}),$$

$$\forall i = 1, 2, \ldots, T, \quad j \neq j^*, j, j^* = 1, 2, \ldots, i$$

Again a simple closed form for $\hat{\rho}$ is not obtainable in this case and hence numerical methods must be used for both point and interval estimates.

2.6.3.3 A specific Application

To illustrate the above model and compare it with the model in section 2.3.2, the same data in exercise testing as used in the example 2.3.2.4 is used. The data were from a sample of 12 individuals who underwent separate exercise tests, where their Breathlessness on a Visual Analogue Scale (VAS), were measured at distinct 2-minute intervals. Since all individuals had not had the same number of repeats/visits, there were not equal number of observations per each individual. In addition, as noted, the learning/visit effect was significant ($P < 0.05$).

For this example, point estimate of measurement reproducibility without considering the visit effect is 0.78 with 95% likelihood interval of (0.51 , 0.89), while, point estimate of measurement reproducibility after correctly including the visit effect in the model, increases to 0.85 with an approximate 95% confidence interval of (0.61 , 0.93).

Graphical representation of the results by the two approaches of point and interval estimation of measurement reproducibility is shown in Figure 2.8.

Clearly, the point estimate by the Profile Likelihood approach in

the case of a significant learning/visit effect is, apparently, larger than that by the ANOVA-based approach, the likelihood interval is narrower.

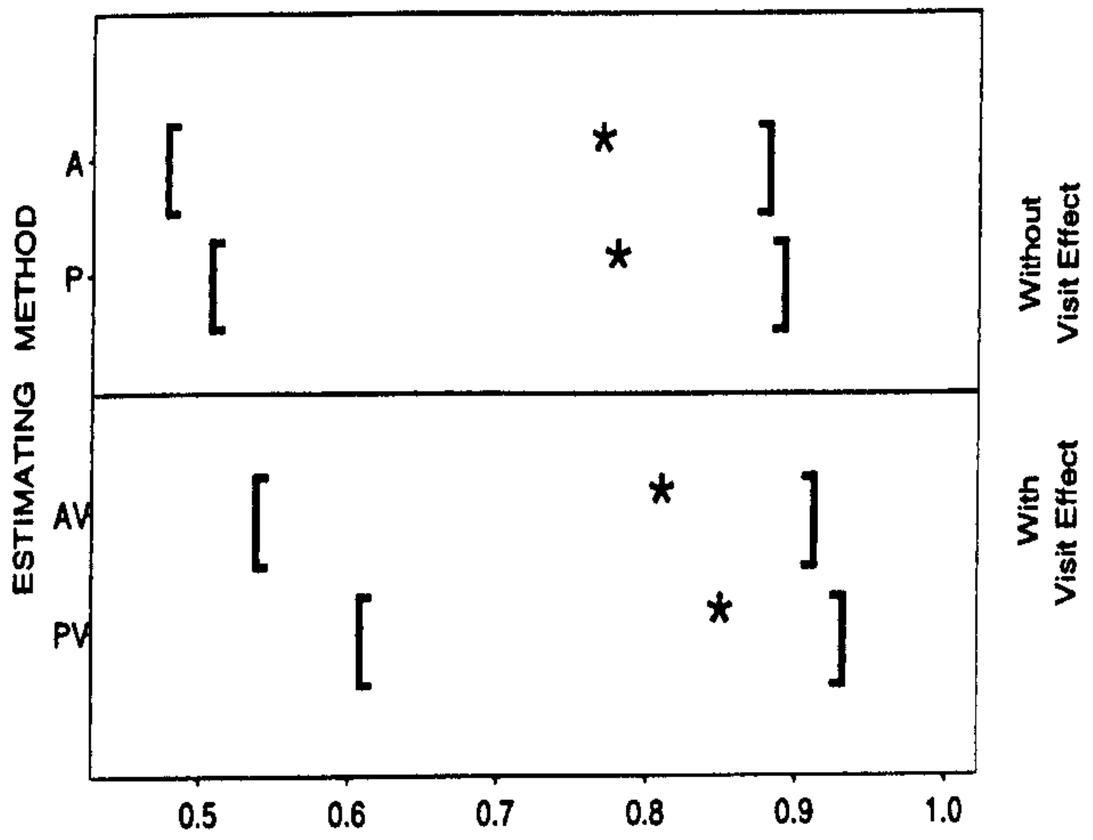

Figure 2.8: Point and interval estimates by each of the two methods of estimating measurement reproducibility both with and without fitting a significant visit effect

These results suggest that, inclusion of a significant effect of learning/visit in the model, will considerably 'improve' (for both approaches) point and interval estimation of measurement reproducibility.

C : Multivariate Approach

2.7 Introduction

A natural extension to the previous sections in the case of the Heart Failure Exercise Testing context is to estimate a pooled measurement reproducibility of a variable measured across the different time points during an exercise test. The multivariate approach deals with data containing observations from a fixed selection of time points which are measured on a set of individuals across a number of repeat tests.

2.7.1 An Illustrative Example

To illustrate the multivariate approach of estimating a pooled measurement reproducibility, data from a set of separate Exercise Tests on a sample of 12 individuals are considered, where their VO_2, obtained from use of a Douglas Bag, were measured at distinct 2-minute intervals during the test. For this example data from 9 distinct time points during the exercise test (i.e. after 2 minutes, 4 minutes, 6 minutes, . . ., 18 minutes into the test) are used.

Figure 2.9 shows scatterplots of these data for each of the 9 distinct time points and at each time point for each of the 12 individuals across eight repeat tests. The estimated measurement reproducibilities along with their approximate 95% confidence interval estimates for each of the 9 time points are given in Table 2.5 with a graphical representation in Figure 2.10. Here, at least for the first 8 time points, the assumption of a common reproducibility appears plausible.

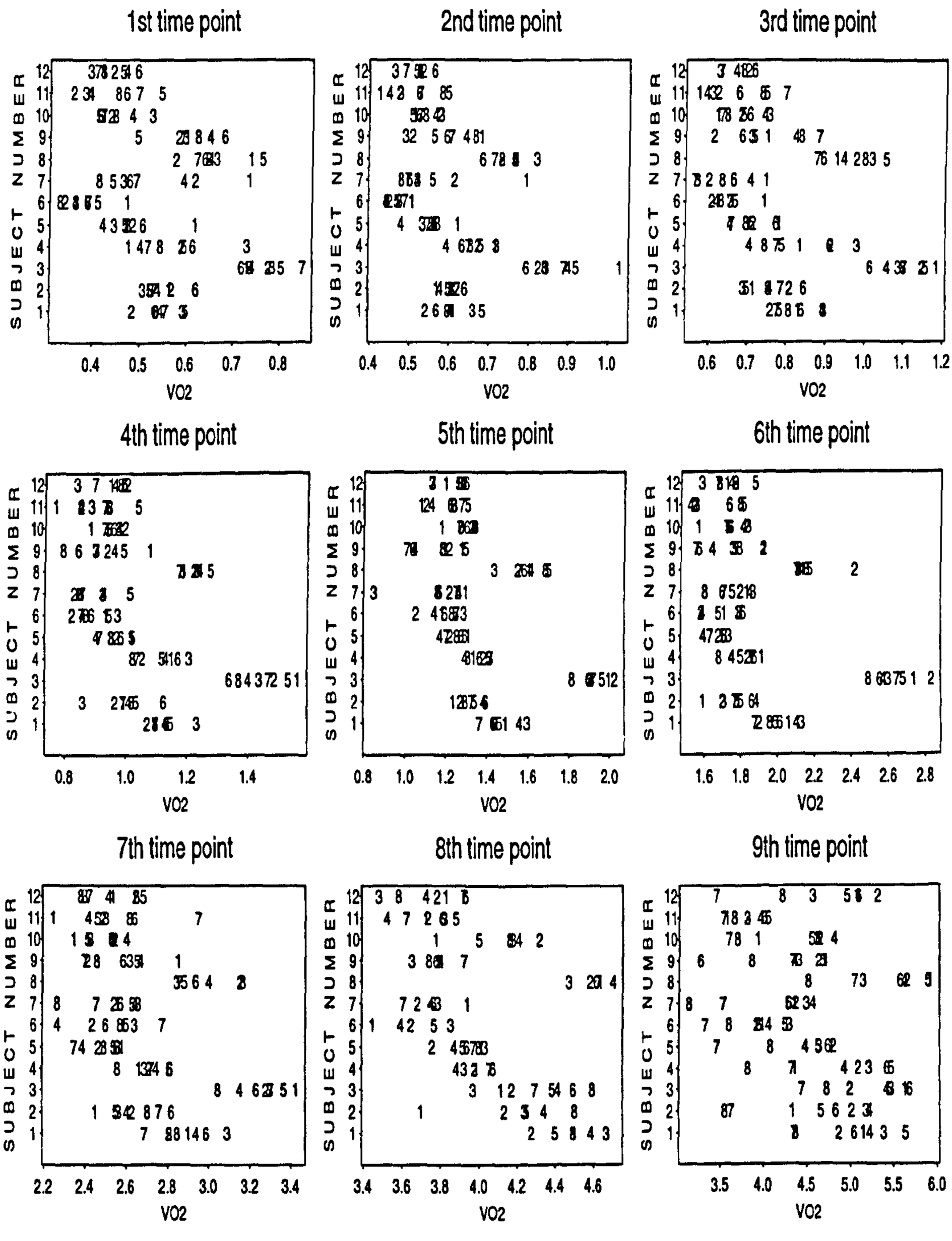

Figure 2.9: Scatterplots of VO_2 for each of 9 time points for each of the 12 individuals across repeat exercise tests (labelled in the plots by the order 'number' of the visit) for each individual.
Note: Different scales are used for each time point which are 2,4,6,...,18 minutes into the test

Time point	*Point Estimate*	95% *Interval Estimate*
1	0.77	(0.59 , 0.91)
2	0.82	(0.66 , 0.93)
3	0.80	(0.64 , 0.93)
4	0.87	(0.73 , 0.98)
5	0.91	(0.77 , 0.97)
6	0.89	(0.76 , 0.96)
7	0.88	(0.73 , 0.97)
8	0.85	(0.65 , 0.93)
9	0.53	(0.31 , 0.78)

Table 2.5: Points and interval estimates of measurement reproducibility for each of the 9 time points based on the Profile Likelihood approach

2.7.2 A pragmatic approach to pooling Measurement Reproducibility

One possible simple way of estimating a common measurement reproducibility, denoted by ρ_{com}, across time points is by assuming that the data from each time point is independent of that from all other times. Although this is clearly a false assumption, it is at least a pragmatic attempt at solving a difficult problem which would involve not only a more complicated mathematical solution but also require more detailed and perhaps too specific assumptions to the structure of such data.

Thus, effectively assuming a likelihood of the form

$$\prod_{k=1}^{K} PLik_k(\rho_{com})$$

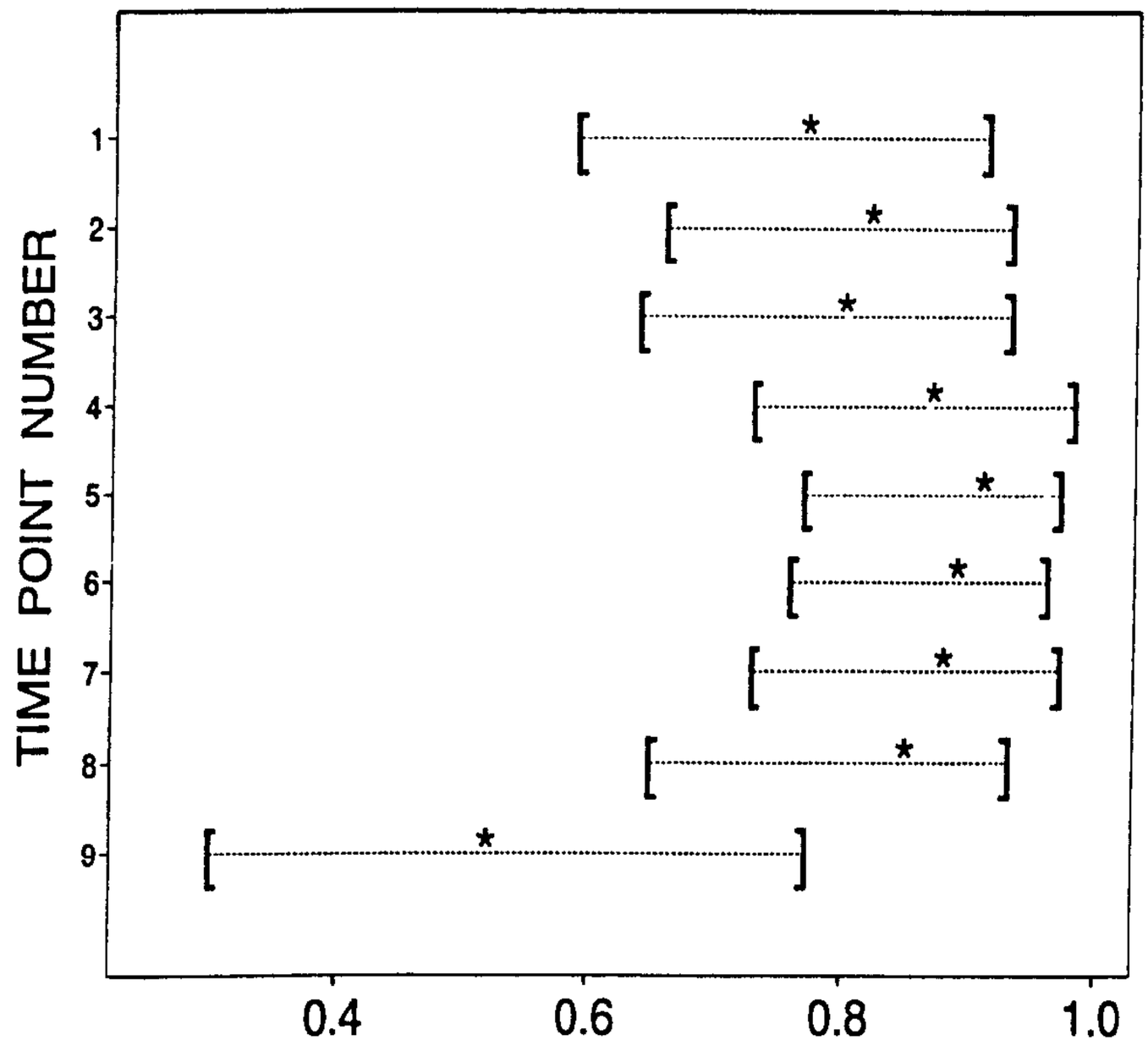

Figure 2.10: Points and interval estimates of separate measurement reproducibility for each of the 9 time points based on the Profile Likelihood approach

when $PLik_k(.)$ is the Profile Likelihood for the k^{th} time point whose resulting likelihood equation will be of the form given in (2.67) or (2.79) as appropriate.

The maximum Profile Likelihood estimate of ρ_{com} can thus clearly be evaluated by an appropriate iterative or search procedure.

Further, using the relative log profile likelihood

$$rpl(\rho_{com}) = \prod_{k=1}^{K} logPLik_k(\rho_{com}) - \prod_{k=1}^{K} logPLik_k(\hat{\rho}_{com})$$

a $100\%Q$ likelihood interval for ρ_{com} will be

$$\{\rho_{com} : rpl(\rho_{com}) \geq log(Q)\}$$

So, a likelihood interval with approximate 95% confidence is achievable for Q=.147, since

$$-\frac{1}{2}\chi^2_{(1,\ .95)} = log(.147).$$

Applying this process for the above example resulted in an (assumed common) point estimate of 0.80 for reproducibility of VO_2 in Exercise Testing with an interval estimate of (0.58 , 0.98).

2.7.3 Test of Equality of Measurement Reproducibility in Different Time Points

Before pooling the estimated measurement reproducibility across different time points to provide a common measurement reproducibility, it is important to test whether or not the measurement reproducibility at different time points is in fact the same. In other words, to test

$$H_0:\ \rho_1 = \rho_2 = \ldots = \rho_K \quad vs \quad H_1:\ \text{all } \rho_i \text{ not equal}$$

Suppose the appropriate models for the i^{th} individual on the j^{th} exercise test at the k^{th} time point, with N individuals and T_i replicates per individual, using the same notation as in sections 2.2.1 and 2.2.2 is

$$X_{kij} = \mu_k + \alpha_{ki} + e_{kj(i)} \tag{2.83}$$

$$i = 1, 2, \ldots, N, \quad J = 1, 2, \ldots, T_i \quad \text{and} \quad k = 1, 2, \ldots, K$$

for the Simple Replication Model and

$$X_{kij} = \mu_k + \alpha_{ki} + \beta_{kj} + e_{kj(i)}, \tag{2.84}$$

$$i = 1, 2, \ldots, N, \quad J = 1, 2, \ldots, T_i \quad \text{and} \quad k = 1, 2, \ldots, K$$

for the Replication Model with an Order Effect.
Note that when there are equal numbers of replicates/visits for each individual (balanced data), $T_i = T$ for each subject.

Now, by assuming that for each two time points k and k^* ($k \neq k^*; k, k^* = 1, 2, \ldots, K$), $e_{kj(i)}$ is independent of $e_{k^*j^*(i)}$ for any j and $j^* = 1, 2, \ldots, T_i$ and also assuming that $\alpha's$ and e's are independent,

$$\begin{aligned} Cov(X_{kij}, X_{k^*ij^*}) &= Cov(\alpha_{ki} + e_{kj(i)}\ ,\ \alpha_{k^*i} + e_{k^*j^*(i)}) \\ &= Cov(\alpha_{ki}, \alpha_{k^*i}) \\ &= \lambda\sigma_{B_k}\sigma_{B_{k^*}} \end{aligned} \tag{2.85}$$

i.e.

$$\Sigma_B = Var\begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \vdots \\ \alpha_{Ki} \end{pmatrix} = \begin{pmatrix} \sigma^2_{B_1} & \lambda\sigma_{B_1}\sigma_{B_2} & \ldots & \lambda\sigma_{B_1}\sigma_{B_K} \\ \lambda\sigma_{B_2}\sigma_{B_1} & \sigma^2_{B_2} & \ldots & \lambda\sigma_{B_2}\sigma_{B_K} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda\sigma_{B_K}\sigma_{B_1} & \lambda\sigma_{B_K}\sigma_{B_2} & \ldots & \sigma^2_{B_K} \end{pmatrix} \tag{2.86}$$

and

$$\Sigma_W = Var\begin{pmatrix} e_{1j(i)} \\ e_{2j^*(i)} \\ \vdots \\ e_{Kj(i)} \end{pmatrix} = \begin{pmatrix} \sigma^2_{W_1} & 0 & \ldots & 0 \\ 0 & \sigma^2_{W_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2_{W_K} \end{pmatrix} \tag{2.87}$$

Thus for any two time points k and k^*

$$Corr(X_{kij}, X_{k^*ij^*}) = \lambda\sqrt{\rho_k\rho_{k^*}} \quad \forall j, j^* \tag{2.88}$$

Using multivariate notation, for each time point k,
define $X_{ki} = (X_{kij})$, i.e. a T_i vector

$$X_{ki} \sim N_{T_i}\left\{\mu_k \underline{1}_{T_i}, (\sigma^2_{B_k} + \sigma^2_{W_k})V_{\rho_k}\right\}, \quad k = 1, 2, \ldots, K \tag{2.89}$$

with V_{ρ_k} a $T_i \times T_i$ matrix of the form

$$V_{\rho_k} = \begin{pmatrix} 1 & \rho_k & \ldots & \rho_k \\ \rho_k & 1 & \ldots & \rho_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k & \ldots & \rho_k & 1 \end{pmatrix}$$

where

$$\rho_k = \frac{\sigma^2_{B_k}}{\sigma^2_{B_k} + \sigma^2_{W_k}}$$

so by taking $\tau = \sigma^2_B + \sigma^2_W$

$$\underline{X_i} = \begin{pmatrix} \underline{X_{1i}} \\ \underline{X_{2i}} \\ \vdots \\ \underline{X_{Ki}} \end{pmatrix} \sim N_{KT_i} \left\{ \begin{pmatrix} \mu_1 \underline{1}_{T_i} \\ \mu_2 \underline{1}_{T_i} \\ \vdots \\ \mu_k \underline{1}_{T_i} \end{pmatrix}, \begin{pmatrix} \tau_1 V_{\rho_1} & W_{(\rho_1,\rho_2)} & \cdots & W_{(\rho_1,\rho_K)} \\ W_{(\rho_2,\rho_1)} & \tau_2 V_{\rho_1} & \cdots & W_{(\rho_2,\rho_K)} \\ \vdots & \vdots & \ddots & \vdots \\ W_{(\rho_K,\rho_1)} & W_{(\rho_K,\rho_2)} & \cdots & \tau_K V_{\rho_K} \end{pmatrix} \right\}$$

where $W_{(\rho_k,\rho_{k^*})}$ is a $T_i \times T_i$ matrix of the form

$$W_{(\rho_k,\rho_{k^*})} = \lambda \sqrt{\tau_k \tau_{k^*}} \sqrt{\rho_k \rho_{k^*}} \, \underline{1}_{T_i} \underline{1}^t_{T_i} \quad \forall \, k \neq k^* = 1, 2, \ldots, K \tag{2.90}$$

The likelihood function in this case is

$$Lik(\underline{\theta}, \underline{\rho}; \underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N) = \prod_{i=1}^{N} \frac{1}{|\Sigma|^{\frac{1}{2}}} exp \left\{ -\frac{1}{2} (\underline{X}_i - A\underline{\mu})^t \Sigma^{-1} (\underline{X}_i - A\underline{\mu}) \right\} \tag{2.91}$$

where

$$\underline{\theta} = \begin{pmatrix} \underline{\mu} \\ \tau_1 \\ \vdots \\ \tau_k \\ \lambda \end{pmatrix}, \quad \underline{\rho} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_K \end{pmatrix}, \quad A = \begin{pmatrix} \underline{1}_{T_i} & \underline{0}_{T_i} & \cdots & \underline{0}_{T_i} \\ \underline{0}_{T_i} & \underline{1}_{T_i} & \cdots & \underline{0}_{T_i} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{0}_{T_i} & \underline{0}_{T_i} & \cdots & \underline{1}_{T_i} \end{pmatrix}, \quad \underline{\mu} = \begin{pmatrix} \underline{\mu_1} \\ \underline{\mu_2} \\ \vdots \\ \underline{\mu_K} \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \tau_1 V_{\rho_1} & \lambda \sqrt{\tau_1 \rho_1} \sqrt{\tau_2 \rho_2} \underline{1}_{T_i} \underline{1}^t_{T_i} & \cdots & \lambda \sqrt{\tau_1 \rho_1} \sqrt{\tau_K \rho_K} \underline{1}_{T_i} \underline{1}^t_{T_i} \\ \lambda \sqrt{\tau_2 \rho_2} \sqrt{\tau_1 \rho_1} \underline{1}_{T_i} \underline{1}^t_{T_i} & \tau_2 V_{\rho_2} & \cdots & \lambda \sqrt{\tau_2 \rho_2} \sqrt{\tau_K \rho_K} \underline{1}_{T_i} \underline{1}^t_{T_i} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda \sqrt{\tau_K \rho_K} \sqrt{\tau_1 \rho_1} \underline{1}_{T_i} \underline{1}^t_{T_i} & \lambda \sqrt{\tau_K \rho_K} \sqrt{\tau_2 \rho_2} \underline{1}_{T_i} \underline{1}^t_{T_i} & \cdots & \tau_K V_{\rho_K} \end{pmatrix}$$

So, likelihood ratio statistic to test $H_0 : \rho_1 = \rho_2 = \ldots = \rho_K$ against $H_1 : \rho_k \neq \rho_{k^*}$ for at least one $k \neq k^*$ will be

$$\Lambda(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N) = \frac{\max Lik_{H_0}(\underline{\theta}, \underline{\rho}; \underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N)}{\max Lik_{H_1}(\underline{\theta}, \underline{\rho}; \underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N)} \tag{2.92}$$

where $Lik(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_N)$ must be estimated under H_0 and H_1.

Assuming that $-2log\Lambda$ will be approximately $\chi^2_{(K-1)}$ under H_0, the approximate test can be carried out.

2.7.4 Application to the Illustrative Example 2.7.1

In Example 2.7.1, Figure 2.10 showed a considerable difference between the point estimates of measurement reproducibility with that for time point 9 different from that of the other time points. For this example, $-2log\Lambda$ is 17.03 which compared to the upper 95 percentile of a Chi-Squared distribution with 8 degrees of freedom (i.e. 15.51), allows one to reject the assumption of equality of measurement reproducibilities of all time points. However, after removing time point number 9 from the analysis, $-2log\Lambda$ reduces to 13.14 and comparing this value with the upper 95 percentile of a Chi-Squared distribution with 7 degrees of freedom (i.e. 14.07), indicates that one should (just) not reject the assumption of equality of measurement reproducibilities across 9 time points. In this case an estimate of common measurement reproducibility will be 0.84 with 95% interval estimate of (0.65 , 0.98).

2.8 Summary

Two distinct methods based on a sums of squares (ANOVA) and Profile Likelihood approaches to point and interval estimation of measurement reproducibility of data were considered for the cases of balanced (equal number of observations per individual) and unbalanced (unequal number of observations per individual) situations of 'replicate' Exercise Tests under (allegedly) identical conditions. This was done for the situations of a replication experiment and also assuming an order effect to the replication.

Illustrative examples throughout the chapter were used to describe the performance of the two approaches on real data.

The illustrative examples implied that point estimates of measurement reproducibility by the Profile Likelihood approach were slightly higher and the interval estimates with this approach were narrower. Furthermore, the examples showed a considerable improvement (for both approaches) in point and interval estimates of measurement reproducibility if a (significant) visit effect existed and was allowed for in the analysis.

Overall, the Profile Likelihood approach, on the basis of the examples, appears to be a "better" method to provide point and interval estimates for measurement reproducibility. However, a full scale simulation study is required to assess this under a variety of different conditions. This is the basis of the following chapter.

The problem of combining measurement reproducibilities across a number of distinct time points during an Exercise Test was raised and a simple pragmatic solution offered.

Chapter 3

Estimating Measurement Reproducibility: A Simulation Study

3.1 Introduction

The previous chapter described the Analysis of Variance (ANOVA) and the Profile Likelihood approaches to point and interval estimation of measurement reproducibility.

To investigate performance of the two approaches, for the two situations of a simple replicate model and an order effect model, 1000 simulations of each of a number of configurations were carried out.

The configurations can be defined by these quantities:

i) Number of distinct subjects, N
ii) Number of repeats (visits) per subject, T
iii) The true measurement reproducibility, ρ_T

In the following simulations **all** combinations of the following were taken:

i) N, number of subjects, taken as 6 or 10 or 30;
ii) T, number of repeats (visits) per subjects, taken as 2 or 4 for balanced case and a maximum 2 or 4 replicates for the unbalanced case;
iii) ρ_T, true measurement reproducibility, taken as 0.75 or 0.85 or 0.95 .

3.1.1 Criteria Used to Judge the Approaches

For each simulation a set of data based on one of the above configurations was generated and the performance of the approaches was investigated based on a number of statistical criteria.

For each simulation the point and interval estimate of the measurement reproducibility for each approach was evaluated.

The performance of each approach was then assessed across all the simulations based on:

i) bias,
i.e. the long run average of the estimated minus the true reproducibility
ii) coverage rate,
i.e. the long run proportion of occasions when the interval estimate contains the true reproducibility
and
iii) the average width of the interval estimate.

This was carried out first for the simple case of replication with

no order effect. Then the simulation process was repeated in each of the following situations where an order effect to the replication process is relevant:

- an order effect was simulated in the generated data and was fitted in the model,
- an order effect was simulated in the generated data but was not fitted in the model,
- an order effect was not simulated in the generated data but was fitted in the model.

3.2 Balanced Data

Balanced data are defined as a replicate experiment where every subject has the same number of observations. This section is concerned with the simulation results from the case where subjects have an equal number of repeat Exercise Tests (visits).

3.2.1 Simple Replicate Model

The simulation study in this section is based on the *Simple Replicate Model* which was defined and illustrated in section 2.2.1 of the previous chapter.

For each approach (i.e. the ANOVA and the Profile Likelihood methods) the averages of the reproducibility estimates over 1000 simulations for each of the underlying configurations are shown in Table 3.1 with a graphical representation of the biases in Figure 3.1. Obviously, both the ANOVA and the Profile Likelihood ap-

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	0.68	0.69	0.79	0.80	0.92	0.93
	10	0.71	0.72	0.82	0.82	0.94	0.94
	30	0.74	0.74	0.84	0.84	0.95	0.95
Profile Likelihood Method	6	0.63	0.65	0.76	0.77	0.91	0.91
	10	0.69	0.70	0.80	0.81	0.93	0.93
	30	0.73	0.73	0.83	0.84	0.95	0.95

Table 3.1: Point Estimates of Measurement Reproducibility, outcome of 1000 simulations, for 2 and 4 replicates (visits) per subject in the case of simple replicate model (no order effect).

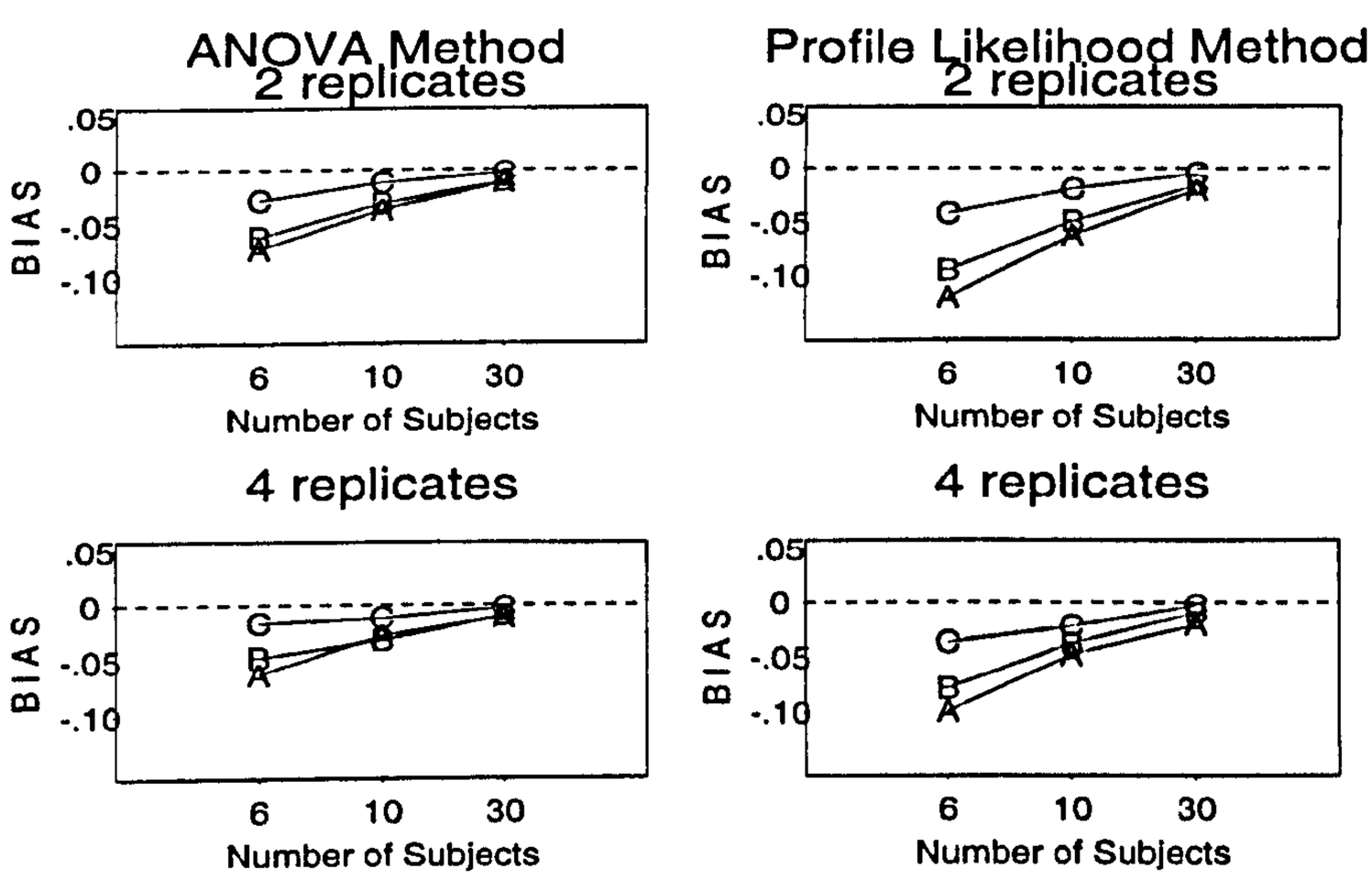

Figure 3.1: Bias from True Measurement Reproducibility for Simple Replicate Model (no order effect)

proaches, particularly for $\rho_T = 0.75$, underestimate the measurement reproducibility, although increasing the number of individuals in the study clearly reduces the bias. Furthermore, the point esti-

mates by the ANOVA approach are less biased than those of the other approach.

It seems that the increase in the number of replicates per subject from 2 to 4 has only a marginal effect on the estimates.

The coverage rate (which is the long run percentage of times that the true measurement reproducibility falls into the confidence interval and here estimated for 1000 simulations) is used as an index of measuring the performance of different methods of interval estimation.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	96	95	95	96	95	96
	10	96	96	95	95	95	96
	30	94	94	94	93	95	93
Profile Likelihood Method	6	93	99	94	95	93	96
	10	94	99	93	96	92	98
	30	93	97	93	98	91	99

Table 3.2: Estimated Coverage Rates for 2 and 4 replicates (visits) per subject in the case of simple replicate model (no order effect).

These rates are shown in Table 3.2 with a graphical presentation in Figure 3.2. The ANOVA approach provides consistent confidence in the range of 95% regardless of the number of replicates, although increasing the number of subjects slightly decreases this rate.

In contrast, in the case of 2 replicates per subjects, the Profile Likelihood approach provides slightly smaller coverage rates but the rates significantly increase as the number of replicates per subject increases from 2 to 4. Furthermore, unlike the ANOVA approach, in

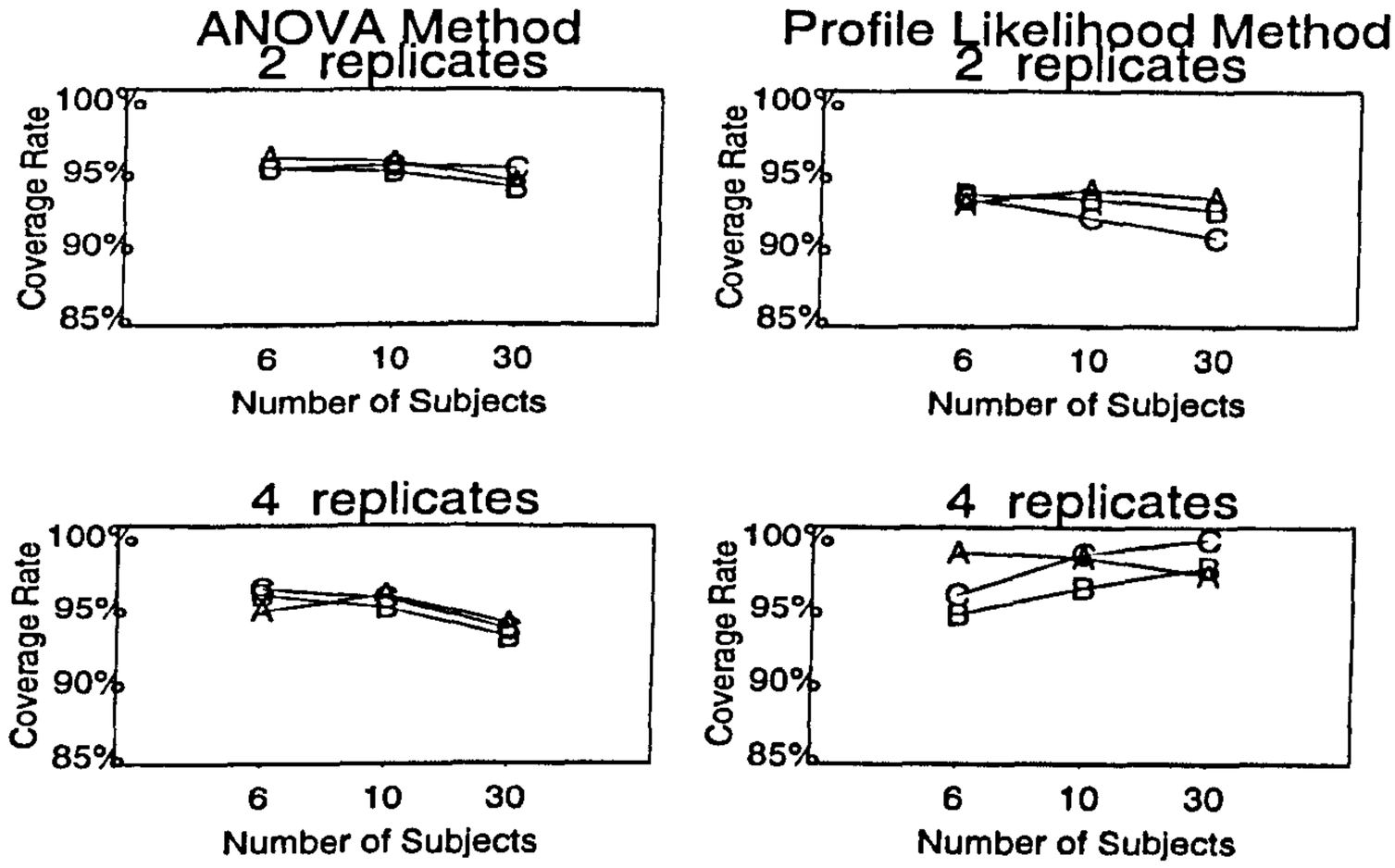

Figure 3.2: Coverage rates for Simple Replicate Model (no order effect)

the case of 4 replicates per subject and for higher true reproducibility (i.e. $\rho_T = 0.85$ or 0.95), the Profile Likelihood approach provides higher coverage rates for higher number of subjects (i.e. when the number of subjects increases from 6 to 10 or 30).

Plots of the bias against confidence interval width for each simulation across different simulation configurations are shown in Figure 3.3. Points inside the wedge shape (<) are intervals which capture the true value of measurement reproducibility while points outside fail to do so.

Both approaches not surprisingly provide wider intervals for small number of subjects which get narrower as the number of subjects increases (i.e. from 6 to 10 and 30). Moreover, it seems that confidence intervals for higher true reproducibility and larger number of subjects (i.e. ρ_T=0.95 and N=10 or 30) are narrower. However, the two approaches produce almost the same pattern of confidence intervals for a large number of subjects (i.e. N=30).

As far as the number of replicates per subject is concerned, an increase in the number of replicates significantly decreases the interval widths provided by the ANOVA approach but does not have a significant effect on the interval widths of the Profile Likelihood approach.

In general, the Profile Likelihood approach provides narrower intervals and this might well be preferred over the ANOVA approach.

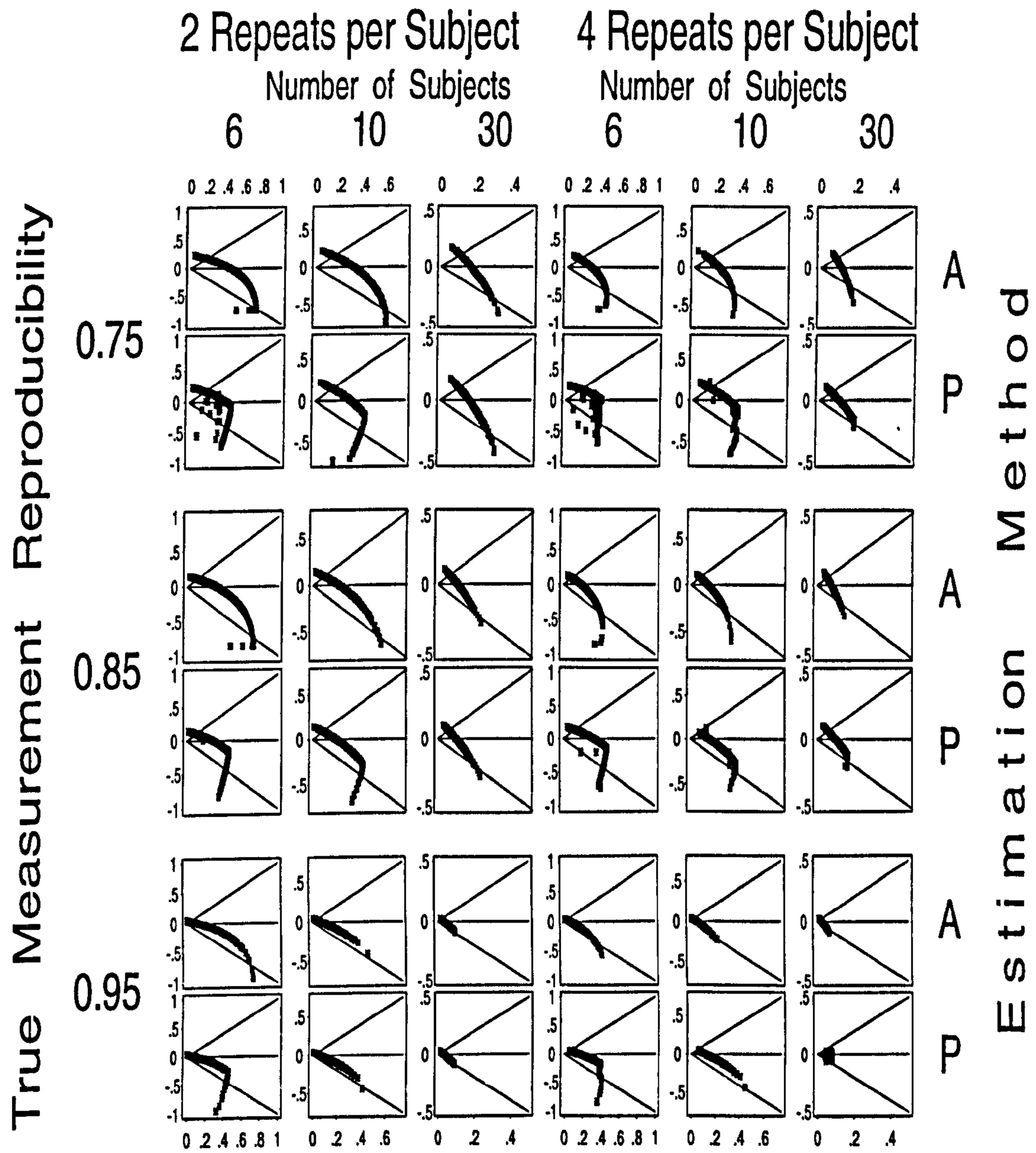

Figure 3.3: Confidence Diagrams with 2 and 4 replicates per subject (visits) for Simple Replicate Model (no order effect) for different combinations of number of subjects and true measurement reproducibility.
In each diagram Vertical Axis represents Bias and Horizontal Axis represents (Confidence Interval Width)/2.

3.2.2 Order Effect Model

In this section, the case where exercise tests are subject to order (i.e. learning or visit) effects are considered. The simulation results are based on the *Order Effect Model* which was illustrated in section 2.2.2 of the previous chapter.

In the following subsections, simulation results from three different situations are discussed. In the first situation, for each simulation, a data set with a significant learning or visit effect was generated and then the effect was fitted into the appropriate model. In the second situation, in spite of existence of a significant learning or visit effect in the generated data, an order effect was not fitted in the model. Finally in the third situation, in the absence of a significant learning or visit effect in the generated data, an order effect was fitted in the model. These three situations are summarized in the following table.

Situation	*Simulated*	*Fitted*	*Subsection*
1	Yes	Yes	3.2.2.1
2	Yes	No	3.2.2.2
3	No	Yes	3.2.2.3

3.2.2.1 Simulated and Fitted Order Effect

For each of the ANOVA and the Profile likelihood approaches, Tables 3.3 and 3.4 show the averages of the estimated measurement reproducibility and the coverage rates over 1000 simulations. Graphical presentation of the biases and the coverage rates are given in Figures 3.4 and 3.5, respectively.

Table 3.3 shows that on average there is 'good' agreement between the two approaches in providing point estimates of measurement reproducibility.

Estimation Method	*No. of Subjects (N)*	ρ_T *0.75* *no. of replicates* 2	ρ_T *0.75* *no. of replicates* 4	ρ_T *0.85* *no. of replicates* 2	ρ_T *0.85* *no. of replicates* 4	ρ_T *0.95* *no. of replicates* 2	ρ_T *0.95* *no. of replicates* 4
ANOVA Method	6	0.68	0.69	0.79	0.80	0.92	0.93
	10	0.71	0.72	0.82	0.82	0.94	0.94
	30	0.74	0.73	0.84	0.84	0.95	0.95
Profile Likelihood Method	6	0.67	0.68	0.78	0.80	0.92	0.93
	10	0.71	0.72	0.82	0.82	0.94	0.94
	30	0.74	0.74	0.84	0.84	0.95	0.95

Table 3.3: Estimated Measurement Reproducibility for 2 and 4 replicates (visits) per subject for the Order Effect Model (in the case of simulated and fitted order effect).

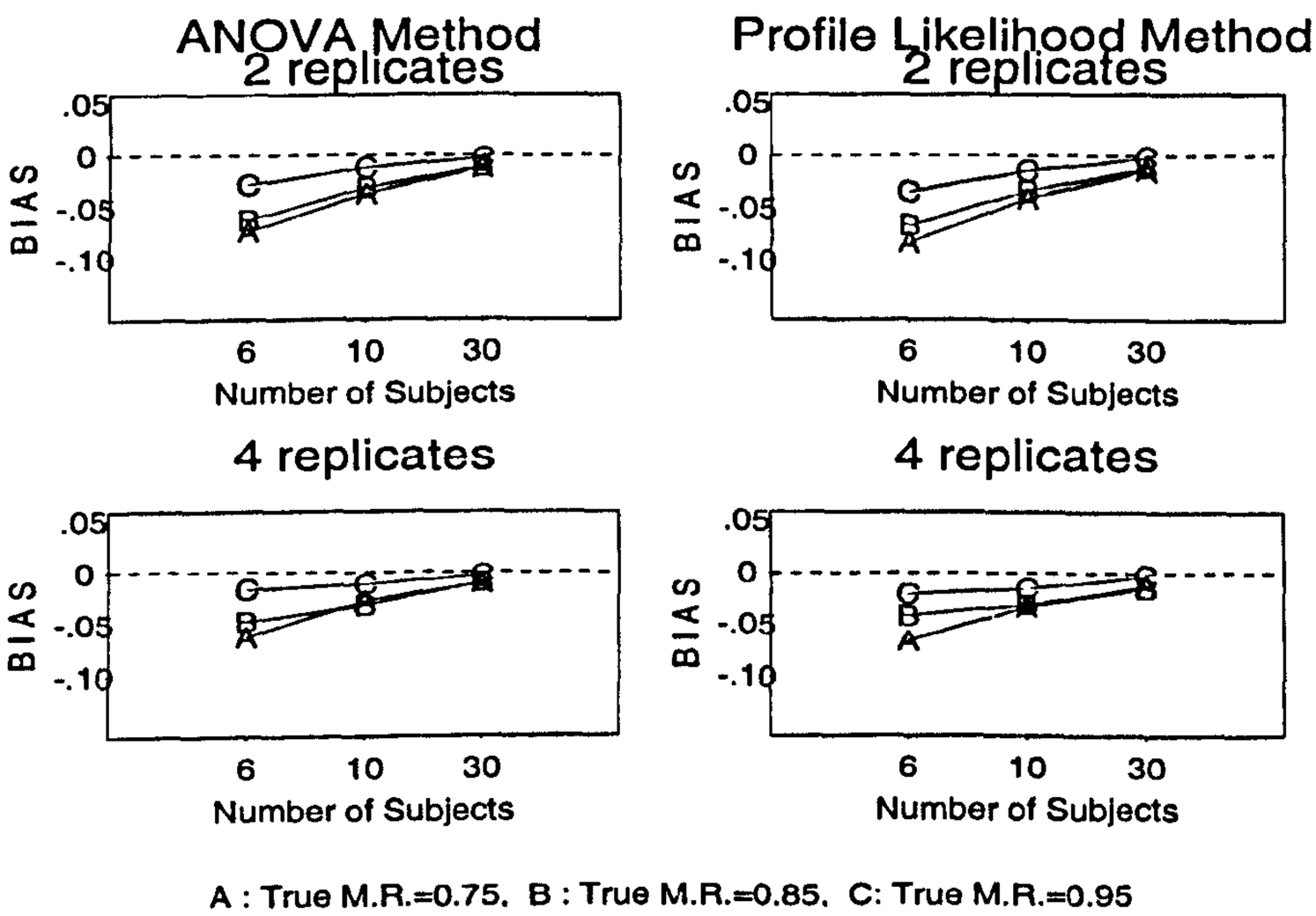

Figure 3.4: Bias from True Measurement Reproducibility for Order Effect Model (for the case simulated and fitted order effect)

As in the simple replicate model, both methods underestimate measurement reproducibility with a clear reduction in the biases as the number of subjects increase. It is clear that increase in the number of replicates per subject has no influence in the estimation of reproducibility.

In the case of high true measurement reproducibility (i.e. $\rho_T = 0.95$), both approaches produce point estimates with negligible bias.

The Coverage rates, displayed in Table 3.4 and Figure 3.5, suggest that the ANOVA approach provides consistent confidence in the range of 95%. Nevertheless, increase in the number of subjects (i.e. from 10 to 30), slightly reduces this rate. It seems that in this approach the number of replicates does not have a significant influence on the coverage rates.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	96	95	95	95	95	96
	10	95	95	95	95	95	95
	30	95	94	93	93	95	93
Profile Likelihood Method	6	93	99	92	96	92	97
	10	93	98	93	98	91	98
	30	94	98	93	98	90	99

Table 3.4: Estimated Coverage Rates for 2 and 4 replicates (visits) per subject for the Order Effect Model (for the case of simulated and fitted order effect).

In contrast, the Profile Likelihood approach provides smaller coverage rates for the case of 2 replicates per subjects, but the rates significantly increase as the number of replicates increases. This in-

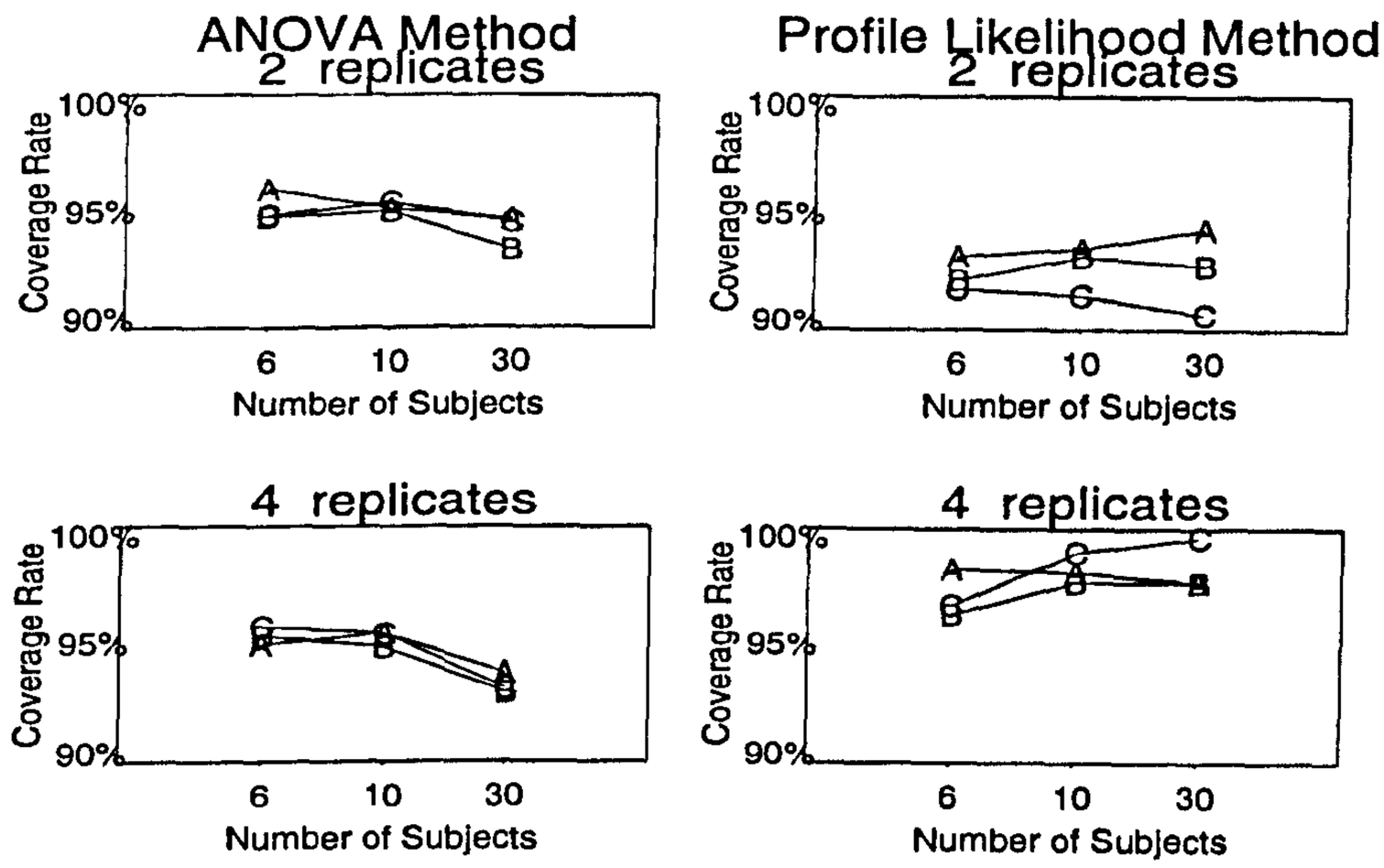

Figure 3.5: Coverage confidence for Order Effect Model (for the case simulated and fitted order effects)

crease is higher for higher true reproducibility (i.e. ρ_T=0.85 or 0.95) and larger number of subjects.

It appears that, in general, the Profile Likelihood approach with higher number of replicates per subject (i.e. T=4), provides the most consistent confidence.

Plots of bias against confidence interval width for each of the simulation configurations are presented in Figure 3.6. It appears that the pattern of confidence interval widths is almost the same as those in the previous section (section 3.2.1). Clearly, as in the previous case, the narrowest intervals are provided for the higher true reproducibility as well as the larger number of subjects.

Moreover, in the ANOVA approach the number of replicates per subject inversely affects the interval width, but it does not have a clear influence on the intervals which are produced by the Profile

Likelihood approach.

In general, in the case of small number of replicates per subject (e.g. 2 replicates) the Profile Likelihood approach provides narrower confidence intervals and may well be prefered, but as the number of replicates per subject increases, the two approaches have almost the same performance.

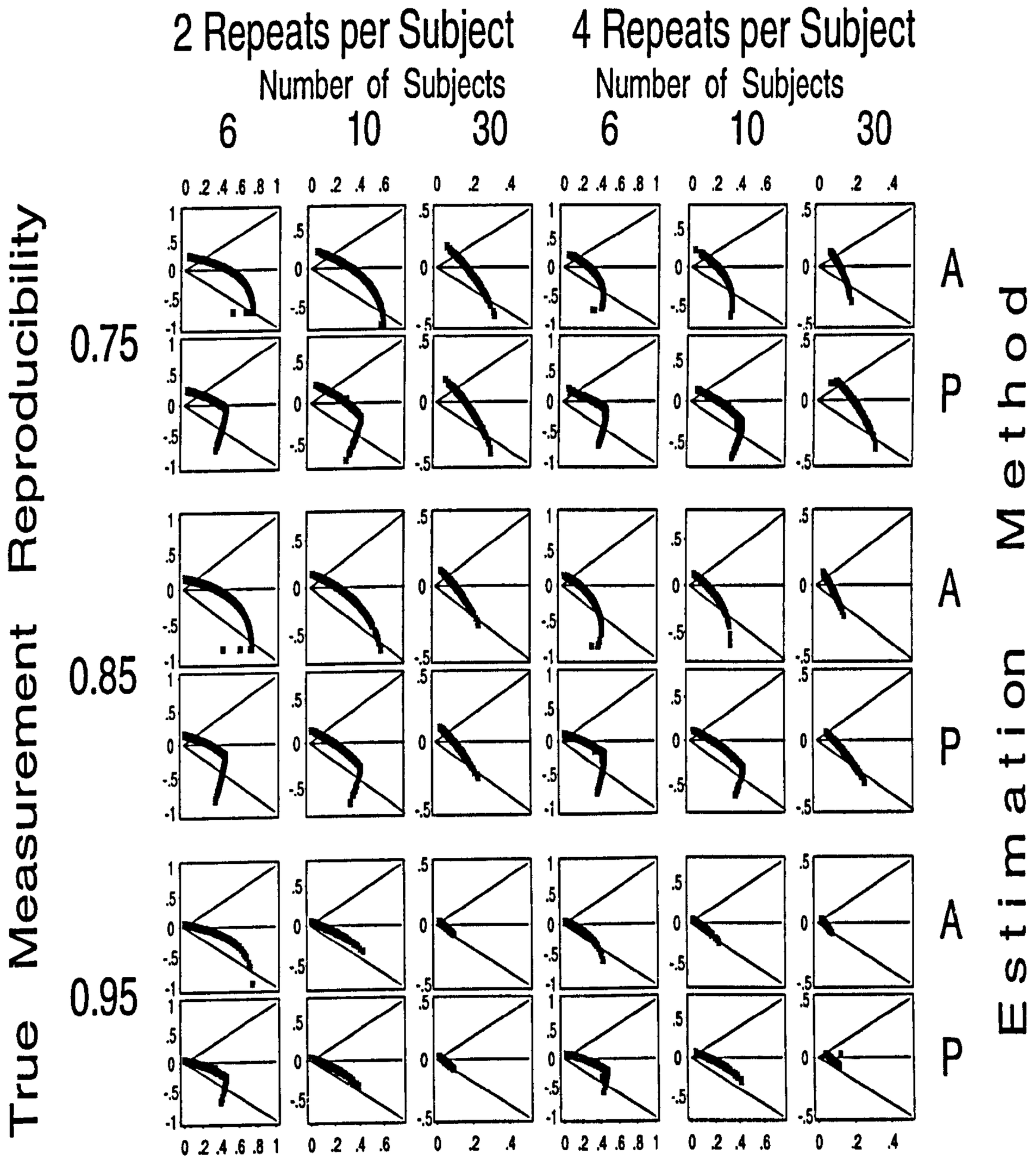

Figure 3.6: Confidence Diagrams with 2 and 4 replicates per subject (visits) for Order Effect Model (for the case simulated and fitted order effect) for different combinations of number of subjects and true measuement reproducibility.
In each diagram Vertical Axis represents Bias and Horizontal Axis represents (Confidence Interval Width)/2.

3.2.2.2 An Order Effect Simulated but not Fitted

Table 3.5 shows the averages of the point estimates over 1000 simulations for the underlying simulation configurations under the two approaches of ANOVA and Profile Likelihood. A graphical representation of the biases is given in Figure 3.7.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	0.66	0.66	0.77	0.76	0.88	0.88
	10	0.69	0.68	0.78	0.79	0.90	0.90
	30	0.71	0.71	0.81	0.80	0.91	0.91
Profile Likelihood Method	6	0.61	0.61	0.73	0.73	0.86	0.86
	10	0.66	0.66	0.77	0.77	0.89	0.89
	30	0.71	0.70	0.81	0.80	0.91	0.91

Table 3.5: Estimated Measurement Reproducibility for 2 and 4 replicates (visits) per subject for the Order Effect Model (for the case of an order effect simulated but not fitted).

Both approaches, as in the previous cases, underestimate measurement reproducibility with a significant decrease in the bias as the number of subjects increases. Furthermore, the number of replicates per subject, apparently, has minimum effect on bias. It appears that the estimated measurement reproducibilities provided by the Profile Likelihood approach in the case of a small number of subjects are more biased than those provided by the ANOVA approach.

Generally, in both approaches, and regardless of the number of subjects and replicates per subject, less bias is exhibited for high true measurement reproducibility (i.e. $\rho_T = 0.95$) than for lower true measurement reproducibilities (i.e. $\rho_T = 0.75$ or 0.85).

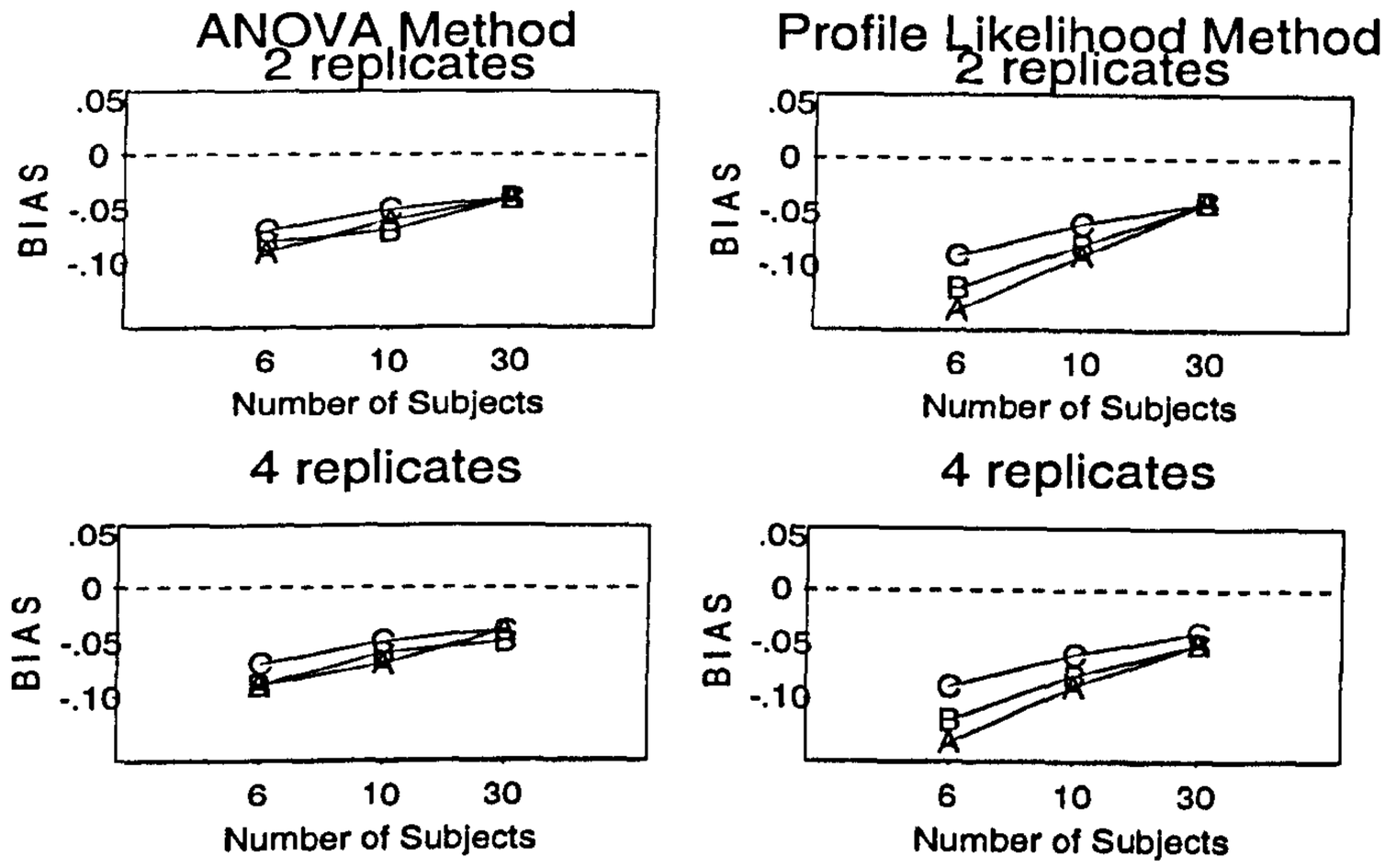

Figure 3.7: Bias from True Measurement Reproducibility for Order Effect Model (for the case of an order effect simulated but not fitted).

Comparing this situation with that where an order effect was simulated and fitted in the model (section 3.2.2.1), it is clear that a failure to fit the order effect in the model considerably increases the bias. This increase is more obvious for the Profile Likelihood approach in the case of a small number of subjects (i.e. N=6 or 10).

Coverage rates for each of the simulation configurations in Table 3.6 and their graphical representation in Figure 3.8 show that, based on the ANOVA approach, an increase in the number of subjects from 6 to 30 tends to a decrease in the coverage rates. This decrease is more obvious for higher number of replicates per subject (i.e. T=4).

In contrast, based on the Profile Likelihood approach, an increase in the number of the replicates slightly increases the coverage rates for all the cases, although this increase is larger for higher true measurement reproducibility and higher number of subjects (i.e. $\rho_T = .95$ and N=30).

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	94	96	95	96	94	94
	10	95	94	94	92	92	88
	30	93	90	91	84	85	85
Profile Likelihood Method	6	90	95	91	90	87	91
	10	92	97	91	90	87	93
	30	92	96	90	92	87	98

Table 3.6: Estimated Coverage Rates for 2 and 4 replicates (visits) per subject for the Order Effect Model (for the case of an order effect simulated but not fitted).

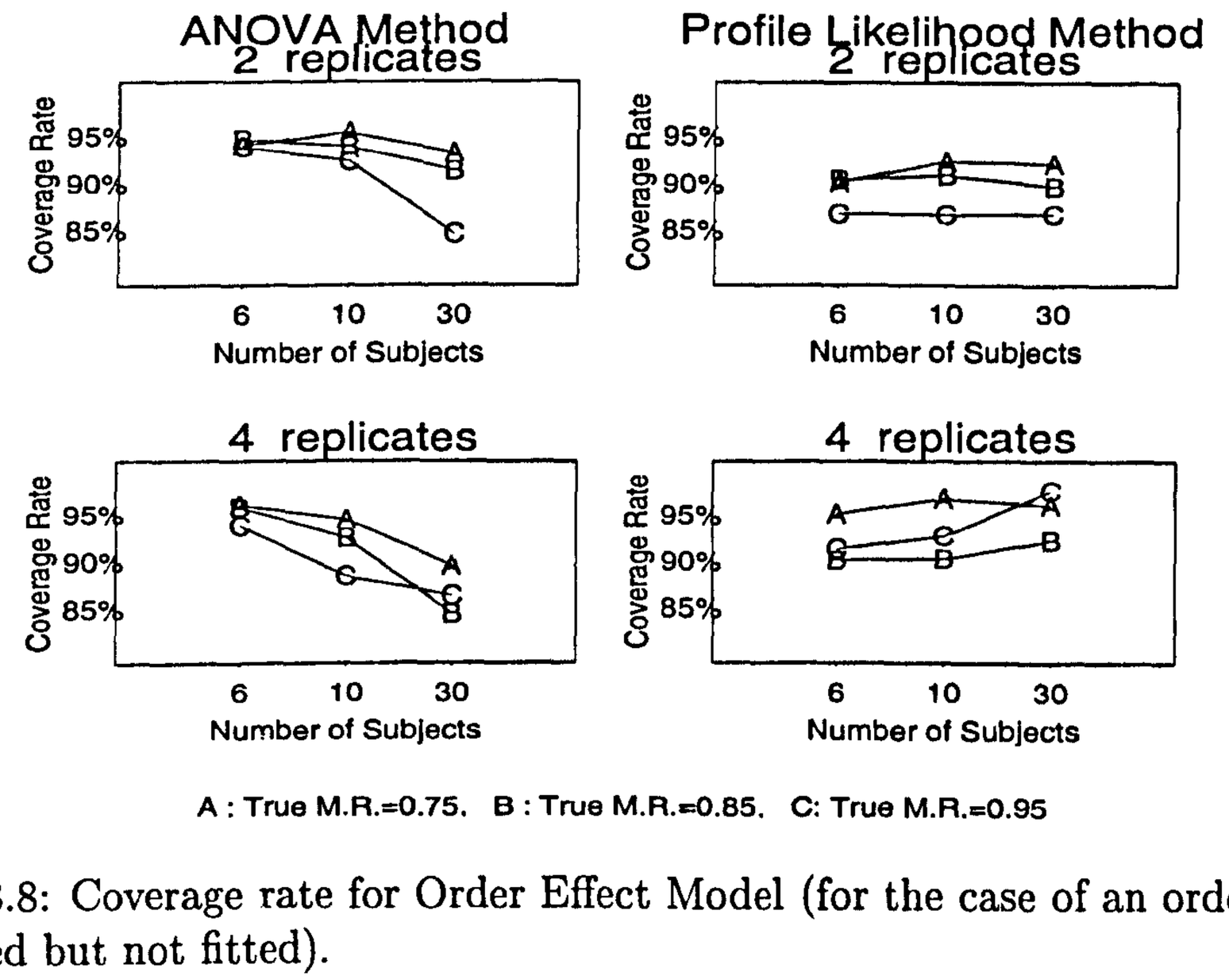

Figure 3.8: Coverage rate for Order Effect Model (for the case of an order effect simulated but not fitted).

Generally, in this case, the Profile Likelihood approach with 4 replicates per subject, particularly for higher true measurement reproducibility, performs better in the sense that it provides more consistent confidence intervals.

In comparison with the case of simulated and fitted order effect (section 3.2.2.1), it can be seen that a failure to fit the order effect in the model considerably decreases the coverage rates. This decrease is more obvious for the ANOVA approach with a large number of subjects (i.e. N=30).

Figure 3.9 shows plots of bias against confidence interval widths for two cases of 2 and 4 replicates per subject. As before, points inside the wedge shape (<) are the intervals which capture the true measurement reproducibility. The figure illustrates that although the ANOVA approach produces wide intervals, an increase in the number of subjects as well as replicates per subject decreases the widths.

On the other hand, confidence intervals provided by the Profile Likelihood approach get narrower as the number of subjects increases, but unlike the ANOVA approach, an increase in the number of replicates per subject does not have much effect on the interval width.

It seems that, for both approaches, the natural true value of the reproducibility does not significantly influence the confidence interval width.

Comparing these intervals with those from the situation where the order effect was simulated and also fitted in the model (section 3.2.2.1), it appears that, in general, when the order effect is not fitted in the model intervals are wider.

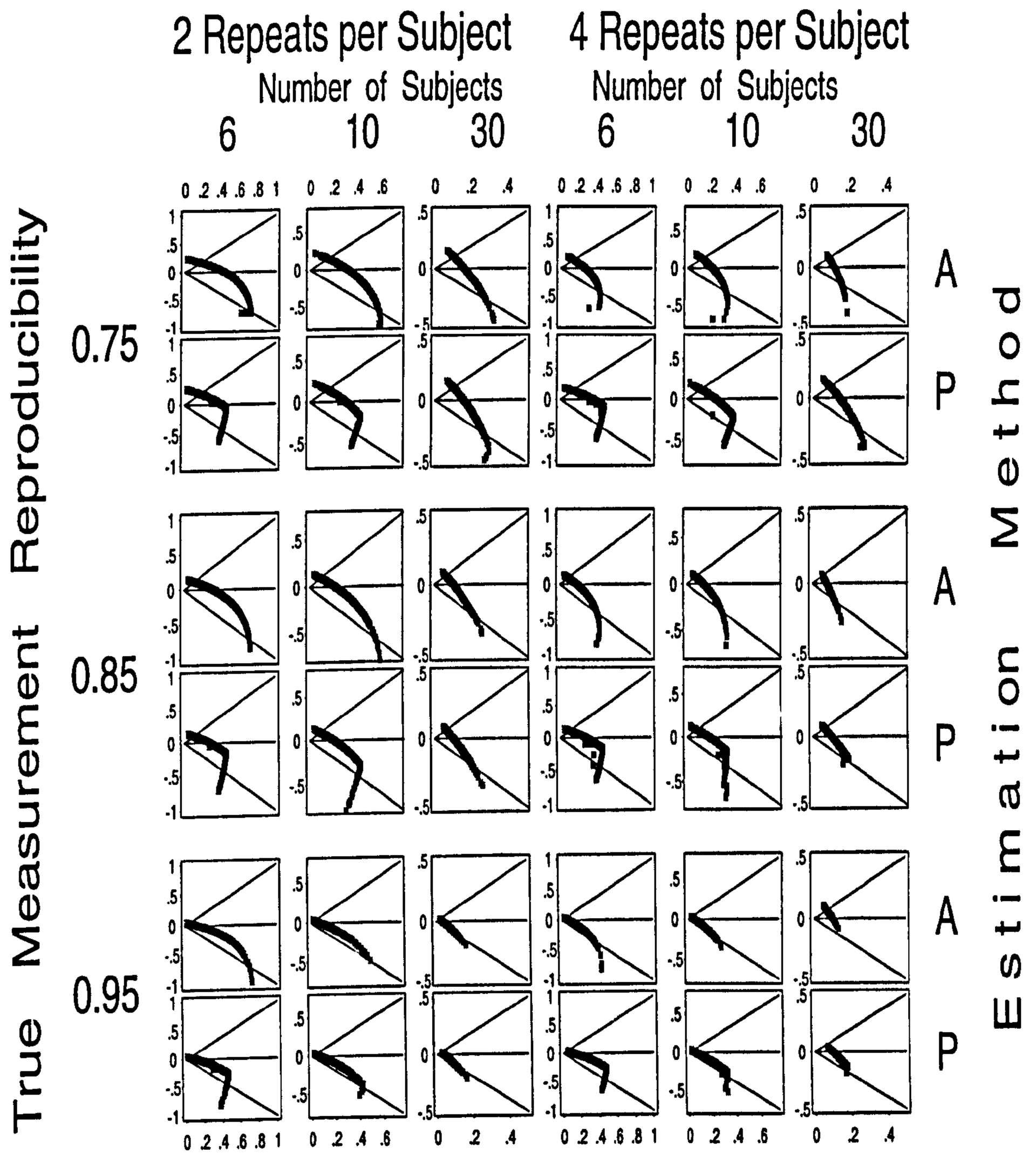

Figure 3.9: Confidence Diagrams with 2 and 4 replicates per subject (visits) for Order Effect Model (for the case of an order effect simulated but not fitted) for different combinations of number of subjects and true measuement reproducibility.
In each diagram Vertical Axis represents Bias and Horizontal Axis represents (Confidence Interval Width)/2.

3.2.2.3 An Order Effect not Simulated but still Fitted

In this section, the simulation results are concerned with the case where no order effect is simulated in the generated data but an order effect is included and fitted in the model.

For each of the two approaches, the averages of the estimated measurement reproducibility over 1000 simulations are shown in Table 3.7 and a graphical representation of the biases is given in Figure 3.10.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	0.69	0.69	0.80	0.80	0.92	0.93
	10	0.71	0.71	0.81	0.81	0.94	0.94
	30	0.74	0.73	0.84	0.84	0.95	0.95
Profile Likelihood Method	6	0.69	0.69	0.80	0.80	0.92	0.93
	10	0.71	0.71	0.81	0.82	0.94	0.94
	30	0.74	0.74	0.84	0.84	0.95	0.95

Table 3.7: Estimated Measurement Reproducibility for 2 and 4 replicates (visits) per subject for the Order Effect Model (in the case of an order effect not simulated but still fitted).

Figure 3.10 shows that there is a 'perfect' agreement between the two approaches to estimate measurement reproducibility. Furthermore, as in previous sections, both approaches underestimate the measurement reproducibility with a clear reduction in the bias as the number of subjects increases. It can be seen that the number of replicates per subject does not have a significant effect on the bias.

The point estimates of measurement reproducibility in this case

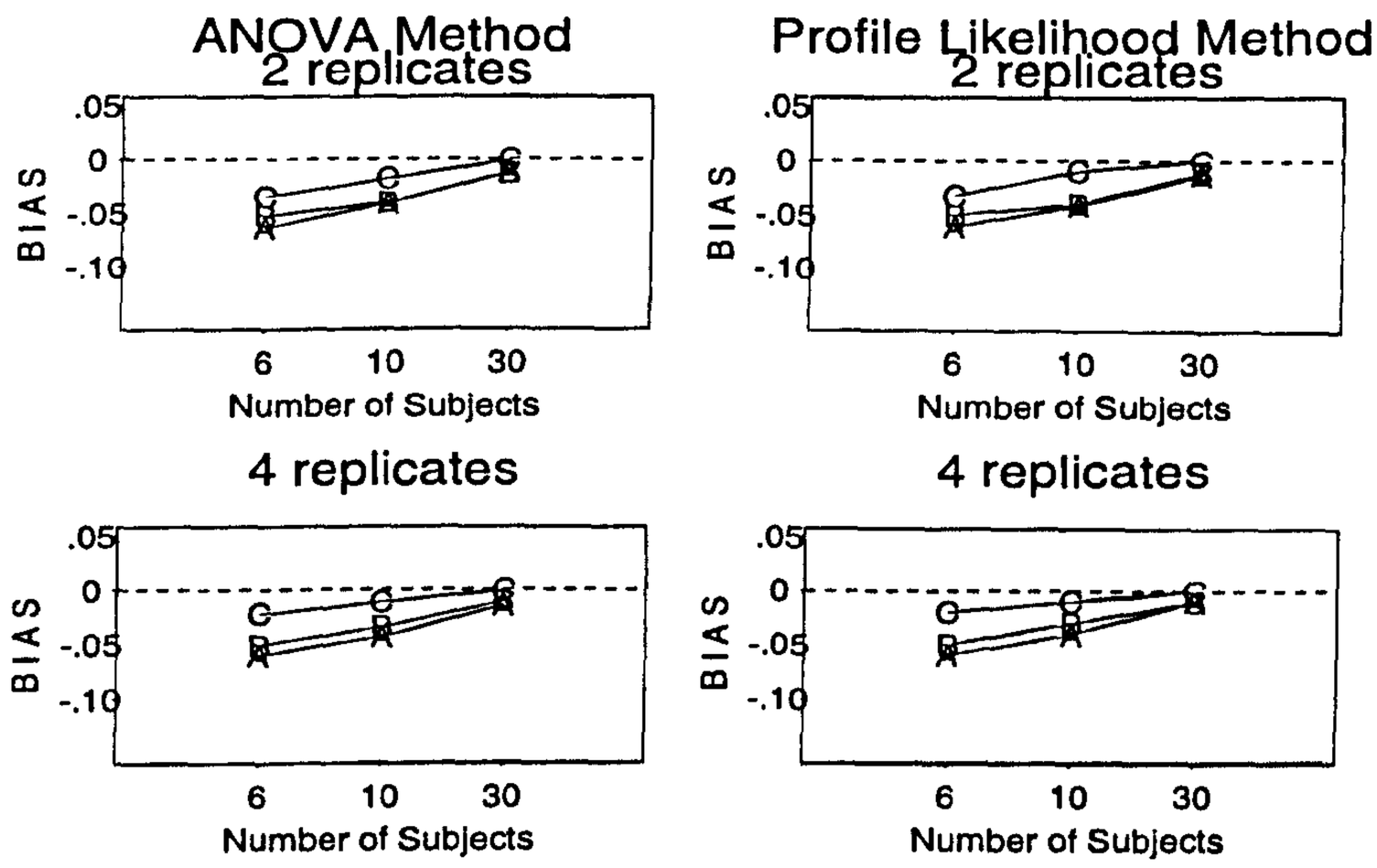

Figure 3.10: Bias from true Measurement Reproducibility for ordered effects model (for the case of an order effect is not simulated but still fitted in the model)

compare to those estimated in the previous two sections (i.e. 3.2.2.1 and 3.2.2.2), indicate that a failure to consider the order effect (even if it is not really significant) will increase the bias.

Coverage rates as an index for measuring the performance of the two approaches in producing confidence intervals are given in Table 3.8 and a graphical presentation is presented in Figure 3.11.

Apparently, the ANOVA approach, regardless of the number of subjects and repeats per subject, provides consistent confidence in the range of 95%. In contrast, the Profile Likelihood approach provides slightly lower coverage rates for the case of 2 replicates per subject (compare to corresponding rates by the ANOVA approach) but these rates significantly increase as the number of replicates increases from 2 to 4. Moreover, an increase in the number of subjects slightly increases the coverage rates.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	94	95	95	96	95	95
	10	96	95	93	95	96	95
	30	94	94	94	93	94	95
Profile Likelihood Method	6	90	97	91	95	93	98
	10	94	99	91	97	94	99
	30	94	99	94	97	94	99

Table 3.8: Estimated Coverage Rates for 2 and 4 replicates (visits) per subject for the Order Effect Model (for the case of an order effect not simulated but still fitted).

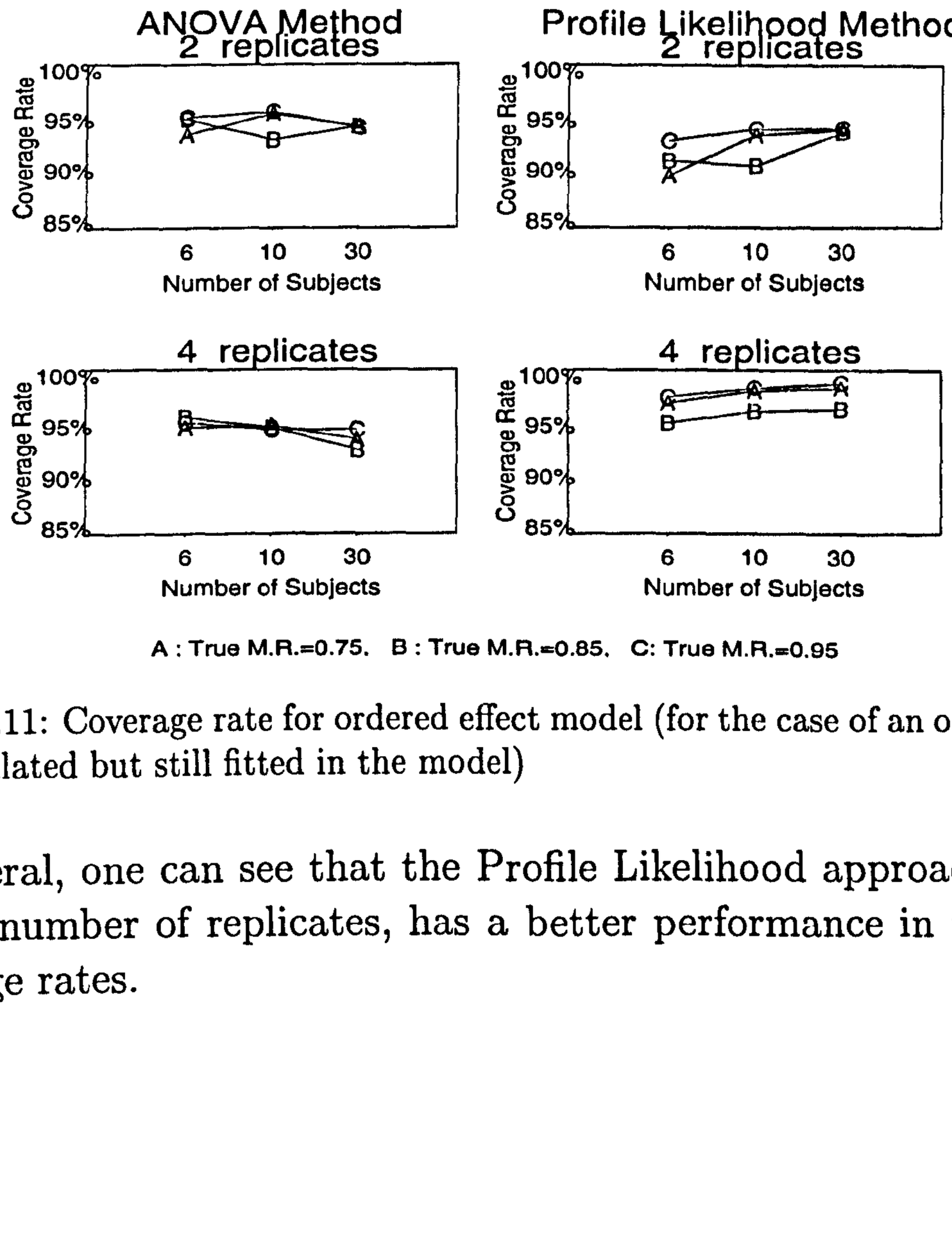

Figure 3.11: Coverage rate for ordered effect model (for the case of an order effect not simulated but still fitted in the model)

In general, one can see that the Profile Likelihood approach, for a higher number of replicates, has a better performance in terms of coverage rates.

Comparison of the coverage rates in this situation with those in the previous two situations suggest that, the coverage rates here are comparable with those in the situation where the order effect was simulated and fitted in the model (section 3.2.2.1). Generally they are also larger than those in the situation where the order effect was simulated but not fitted in the model (section 3.2.2.2). This again highlights the importance of correctly including an order effect in the model.

Figure 3.12 provides plots of bias against confidence interval width for different simulation configurations.

It is clear that, in the case of 2 replicates per subject, the ANOVA approach produces wider confidence intervals. However, the interval widths get narrower as the number of subjects as well as the number replicate observations per subject increases.

On the other hand, for the Profile Likelihood approach, the number of replicates does not have a clear effect on the interval widths whilst, as in the ANOVA approach, an increase in the number of subjects reduces the interval widths. This reduction is more obvious for the case of high true reproducibility. Nevertheless, it seems that, in the case of large number of subjects (i.e. N=30), both approaches produce almost the same pattern of intervals.

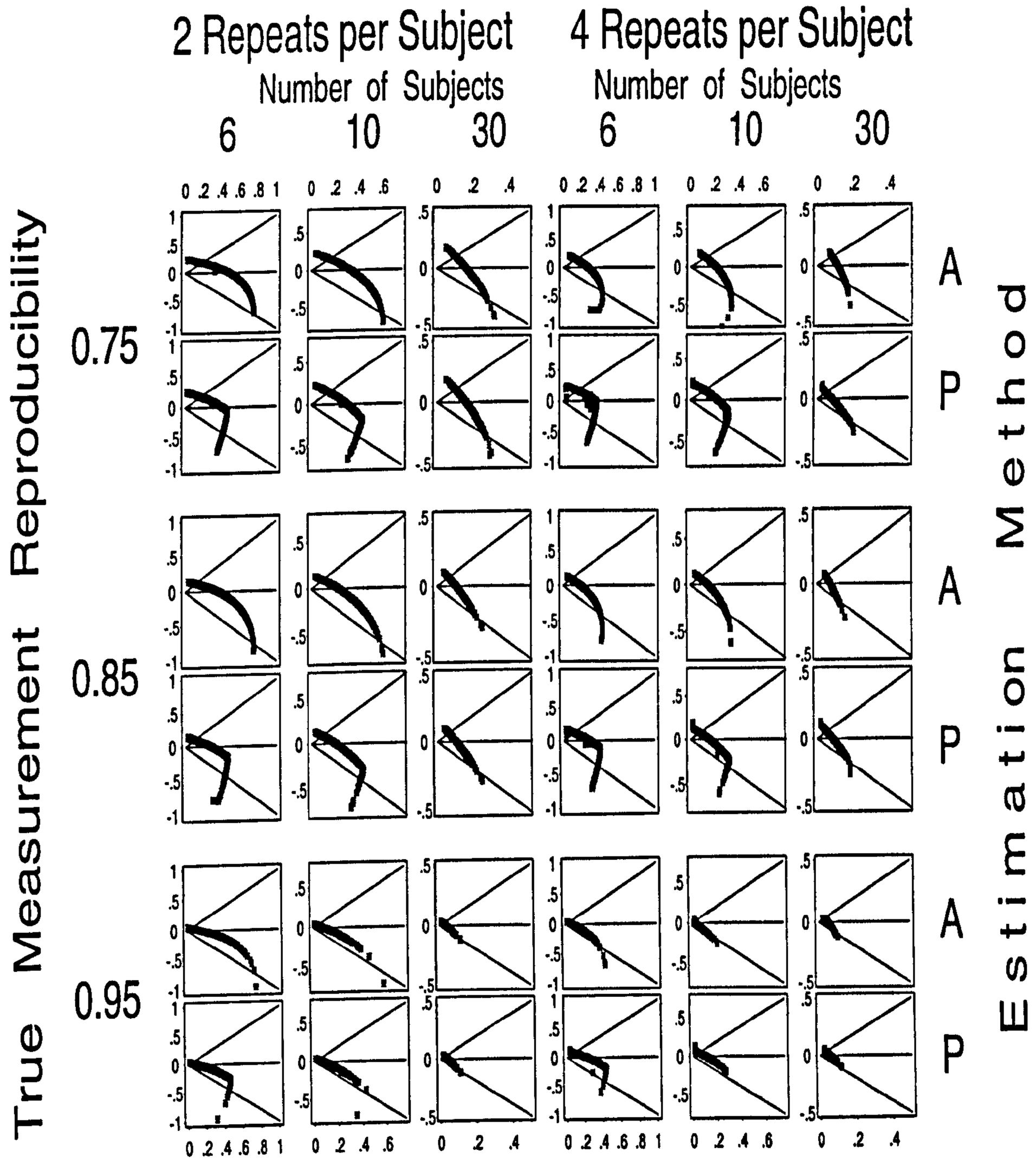

Figure 3.12: Confidence Diagrams with 2 and 4 replicates per subject (visits) for Order Effect Model (for the case where order effect is not simulated but is fitted in the model) for different combinations of number of subjects and true measuement reproducibility.
In each diagram Vertical Axis represents Bias and Horizontal Axis represents (Confidence Interval Width)/2.

3.3 Unbalanced Data

So far the data were simulated on the assumption that for each subject there were equal number of observations (visits or repeats). In this section the case where not all of the subjects have equal number of observations is considered.

The simulation configurations that were used are:

i) N, number of subjects, taken as 6 or 10 or 30;
ii) T, number of repeats (visits) per subjects, taken as maximum 2 or 4 replicates for the unbalanced* case;
iii) ρ_T, true measurement reproducibility, taken as 0.75 or 0.85 or 0.95 .

* The lack of balance was as follows:

a) For 2 replicates per subject,
50% of the cases were complete
and
50% of the cases had 1 missing replicate.

b) For 4 replicates per subject,
roughly 33.3% of the cases were complete
roughly 33.3% of the cases had 1 missing replicate
and
roughly 33.3% of the cases had 2 missing replicates

3.3.1 Simple Replicate Model

The simulation study in this section is based on the *Simple Replicate Model* for unbalanced data which was defined and illustrated in section 2.3.1 of the previous chapter.

Point estimates from the two approaches to estimating measurement reproducibility are given in Table 3.9 and the related biases are displayed in Figures 3.13.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	0.63	0.65	0.76	0.76	0.89	0.89
	10	0.64	0.68	0.79	0.79	0.91	0.92
	30	0.70	0.72	0.82	0.83	0.93	0.93
Profile Likelihood Method	6	0.64	0.66	0.76	0.78	0.88	0.84
	10	0.67	0.70	0.78	0.82	0.91	0.92
	30	0.72	0.73	0.83	0.84	0.94	0.95

Table 3.9: Estimated Measurement Reproducibility for 2 and 4 replicates (visits) per subject for Simple Replicate Model (no order effect).

It is clear that the two approaches behave similarly in terms of point estimation with respect to the number of subjects and replicates per subject. Both approaches underestimate the measurement reproducibility with a significant effect of increase in the number of subjects in reducing the bias, although more biased point estimates appear for lower true reproducibility. It seems that increasing the number of replicates per subject results in a significant decrease in the bias. This effect is more considerable for a smaller number of subjects (i.e. N=6 or 10).

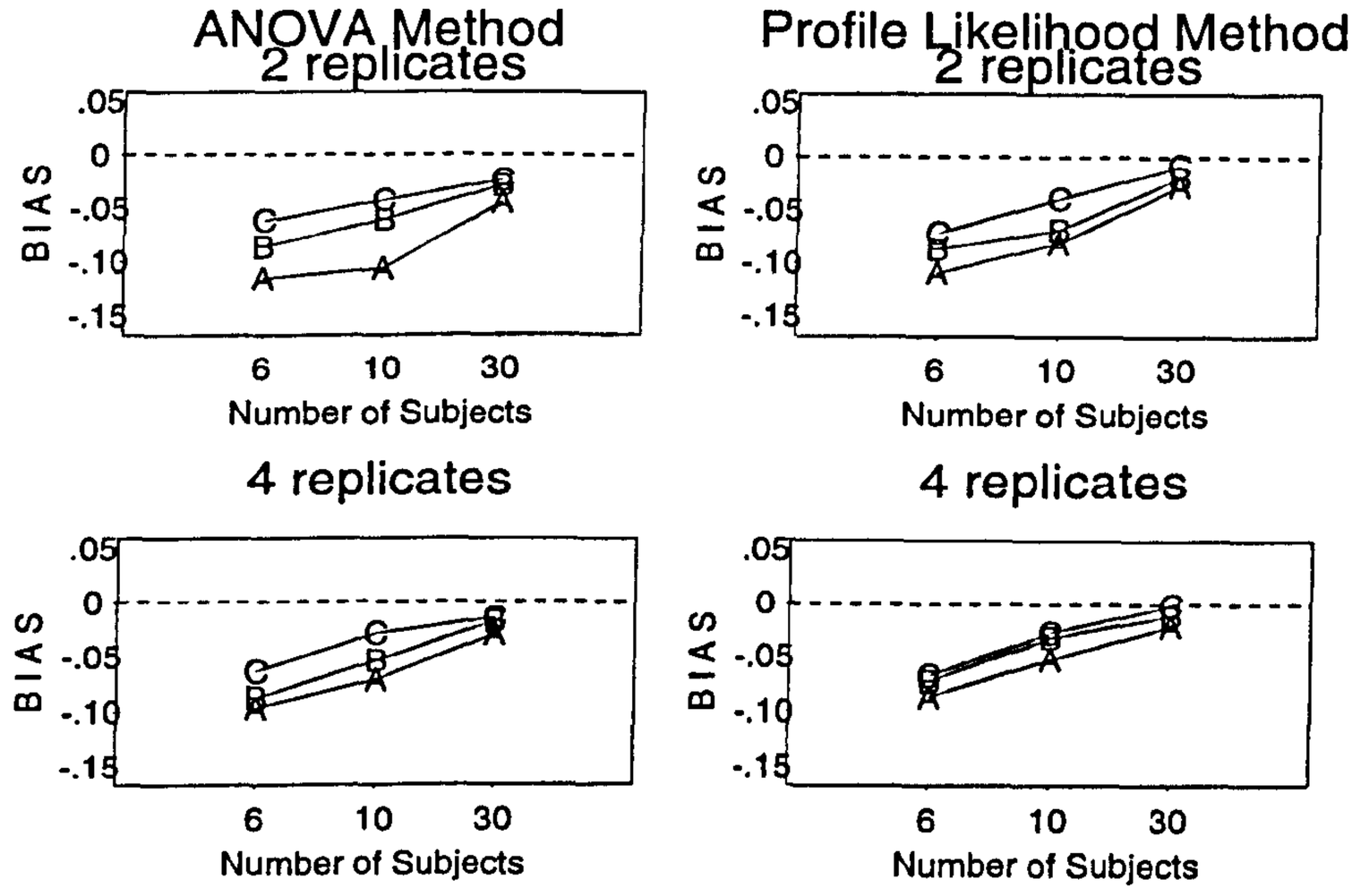

Figure 3.13: Bias from true measurement reproducibility for Simple Replicate Model(no order effect)

However, both approaches for a large number of subjects have almost the same performance in producing point estimates, but in general the Profile Likelihood approach, regardless of the number of subjects, produces less biased point estimates.

As far as the performance of the confidence intervals is concerned, as in the previous section, this is assessed by means of the *'Coverage Rate'* which is the percentage of times that the confidence interval captures the true measurement reproducibility.

Coverage rates based on the above model, are given in Table 3.10 and their graphical representation is displayed in Figure 3.14.

Both approaches, regardless of the number of replicates per subject, provide consistent confidence intervals in the range of around 95%. Nevertheless, the coverage rates produced by the ANOVA approach

are slightly poorer than those produced by the Profile likelihood approach. It is clear that the number of subjects as well as the true reproducibility does not have a significant effect on these rates.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	94	96	93	95	94	94
	10	94	94	94	95	94	93
	30	94	95	95	96	95	94
Profile Likelihood Method	6	93	94	93	95	96	95
	10	94	95	95	96	95	96
	30	94	96	95	97	95	96

Table 3.10: Estimated Coverage Rates for 2 and 4 replicates (visits) per subject for Simple Replicate Model (no order effect).

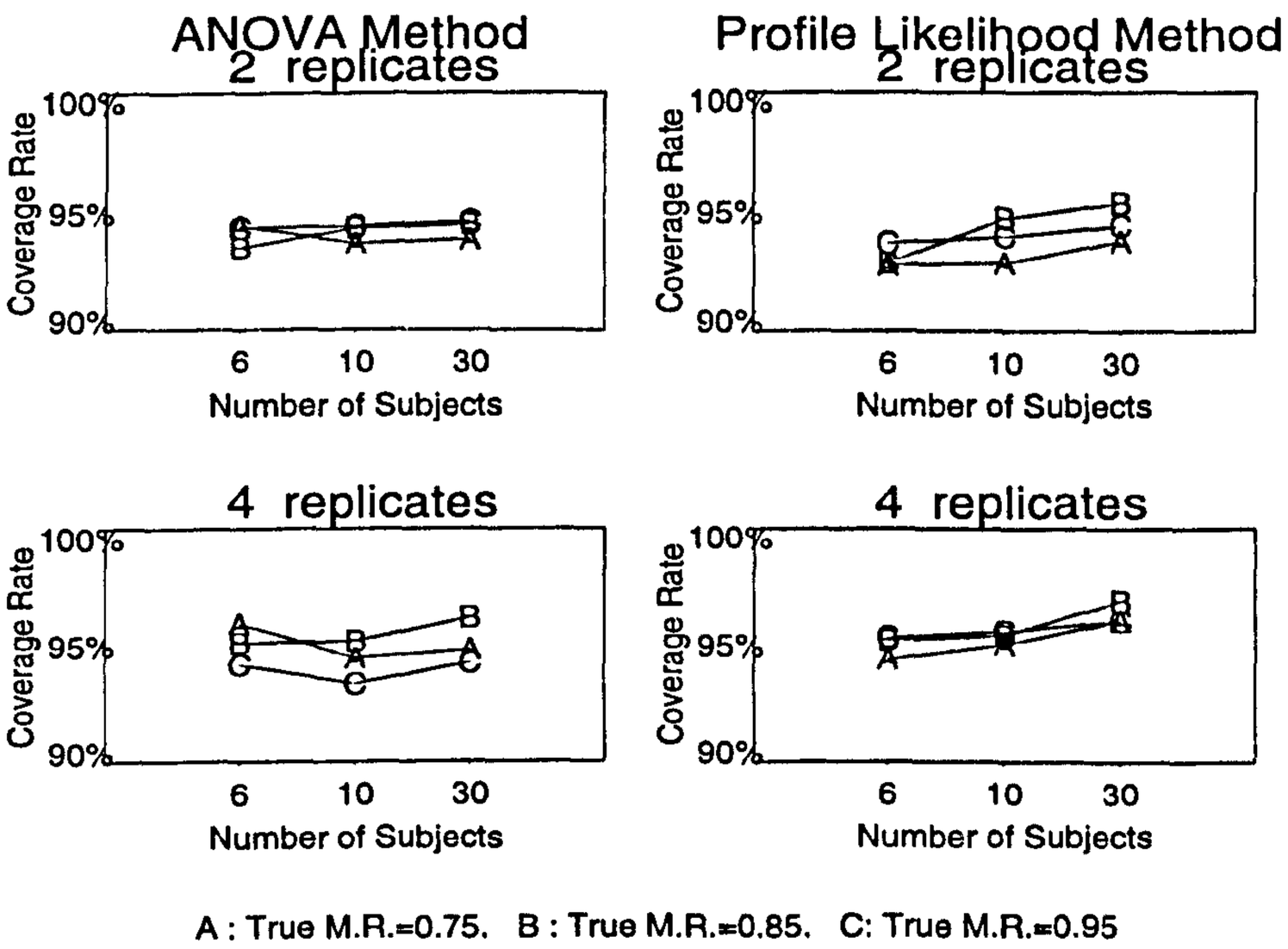

Figure 3.14: Coverage rates for Simple Replicate Model (no order effect)

Plots of bias against confidence interval widths are presented in

Figure 3.15. As noted before, points inside the wedge shape (<) are those intervals which capture the true measurement reproducibility.

Clearly, the number of subjects, regardless of the number of replicates per subject, inversely affects the interval widths. Furthermore, in different simulation configurations , the patterns of intervals produced by the two approaches are almost the same. It appears that in terms of confidence interval width, there is not a clear difference between the two approaches.

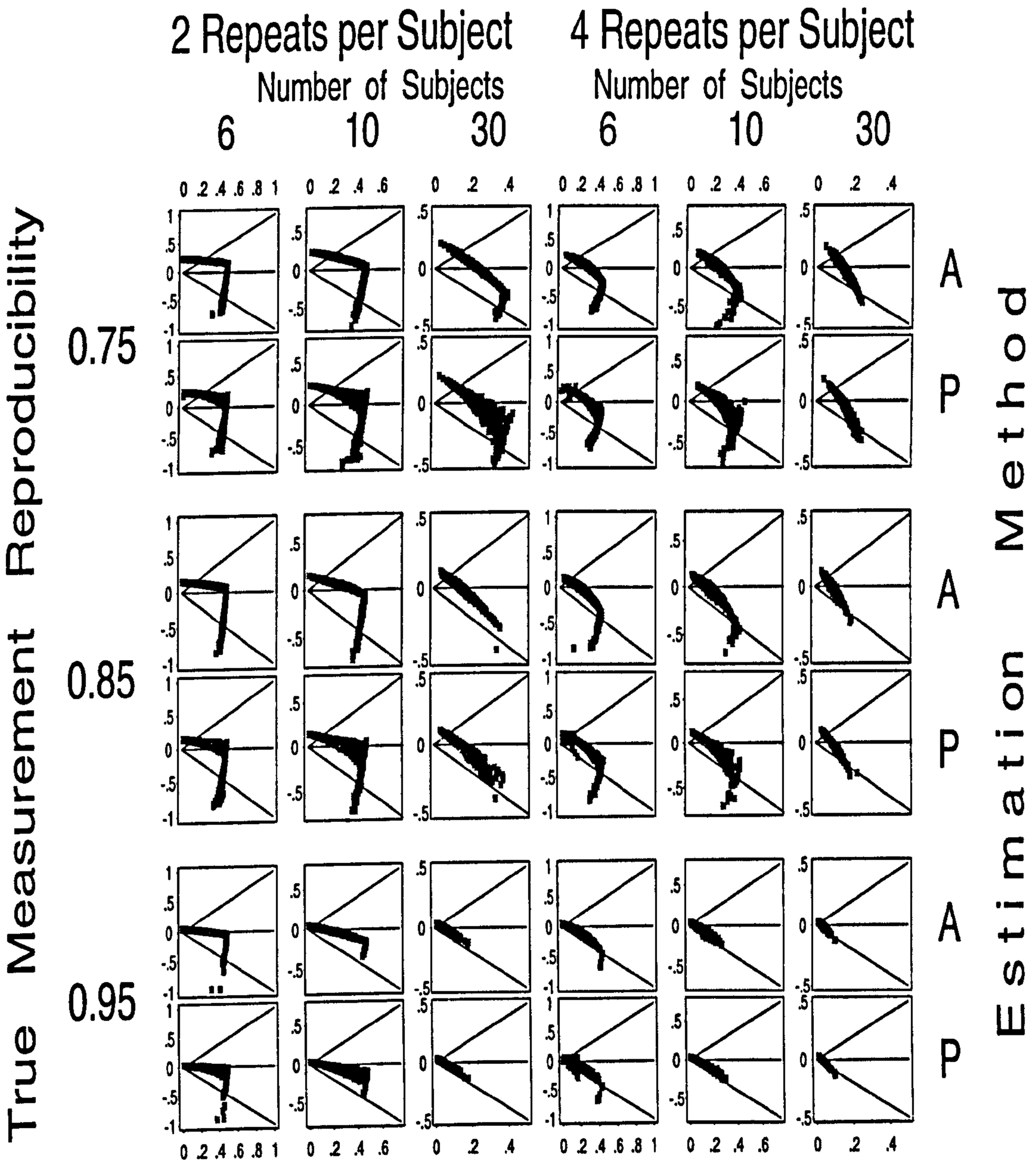

Figure 3.15: Confidence Diagrams with 2 and 4 replicates per subject (visits) for Simple Replicate Model (no order effect) for different combinations of number of subjects and true measuement reproducibility.
In each diagram Vertical Axis represents Bias and Horizontal Axis represents (Confidence Interval Width)/2.

3.3.2 Order Effect Model

In this section, the simulation study is based on considering an order effect in the simulation process. In exercise testing, for instance, individuals may be subject to a visit or learning effect. The model which is appropriate to such data is illustrated in the section 2.3.2 of previous chapter.

In the following three subsections, the same situations as in the balanced data are considered. Simulation results for these three situations are presented in subsections 3.3.2.1 to 3.3.2.3. These situations are summarized in the following table:

Situation	*Simulated*	*Fitted*	*Subsection*
1	Yes	Yes	3.3.2.1
2	Yes	No	3.3.2.2
3	No	No	3.3.2.3

3.3.2.1 Simulated and Fitted Order Effect

In this section results from the case where an order effect is simulated and fitted in the model are illustrated.

For each of the two approaches (i.e. ANOVA and Profile Likelihood) Tables 3.11 and 3.12 represent the averages of the estimated measurement reproducibilities and coverage rates over 1000 simulations and Figures 3.16 and 3.17 display the biases and the coverage rates, respectively.

As in previous cases, both approaches underestimate measurement reproducibility. Table 3.11 indicates that for the case of 2 repli-

cates per subject, highly unbiased point estimates of reproducibility rapidly improve as the number of subjects increases. The same trend exists for 4 replicates per subject but with a lesser influence of the number of subjects. In fact an increase in the number of replicates appears to have a large effect for the case of a small number of subjects (i.e. N=6).

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	0.54	0.63	0.64	0.74	0.75	0.85
	10	0.64	0.64	0.73	0.75	0.85	0.87
	30	0.71	0.72	0.82	0.82	0.92	0.92
Profile Likelihood Method	6	0.61	0.66	0.72	0.77	0.81	0.88
	10	0.68	0.67	0.77	0.78	0.88	0.90
	30	0.72	0.73	0.83	0.83	0.93	0.93

Table 3.11: Estimated Measurement Reproducibility for 2 and 4 replicates (visits) per subject for Order Effect Model(simulated and fitted order effect).

Apparently, estimated reproducibilities based on the ANOVA approach in the case of small number of subjects (i.e. N=6), in comparison with those from the Profile likelihood approach, are likely to be more biased. Further, it seems that true measurement reproducibility does not have a significant effect on the bias.

Generally, in this case, the Profile Likelihood approach shows a better performance throughout the underlying configurations in the sense that it produces less biased point estimates.

To examine the performance of confidence intervals from the two approaches, Figure 3.17 indicates that for lower true reproducibilities (i.e. ρ_T=0.75 or 0.85), both approaches provide consistent con-

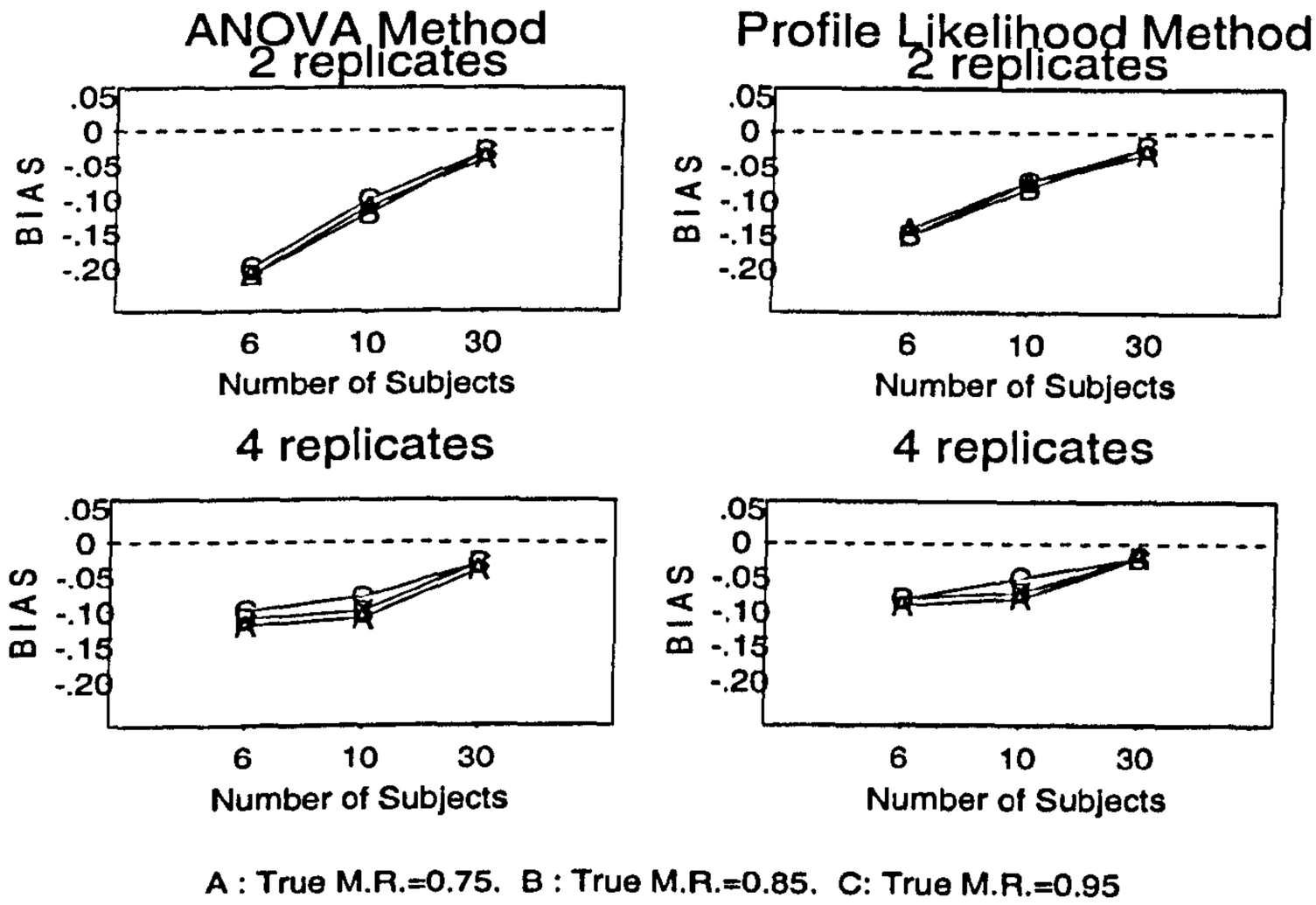

Figure 3.16: Bias from true Measurement Reproducibility for Ordered Effect Model (for the case where order effect is simulated and fitted in the model)

fidence in the range of 95%, while for high true measurement reproducibility (i.e. ρ_T=0.95), the two approaches behave differently with respect to an increase in the number of subjects. An increase in the number of subjects in the ANOVA approach inversely affects the coverage rate, whilst for the Profile Likelihood approach this slightly increases the rates. These changes are more obvious for a larger number of subjects. However, it seems that an increase in the number of replicates from 2 to 4 does not have a significant influence on the coverage rates.

Generally, it appears that the Profile Likelihood approach has a better performance in terms of coverage rate.

To compare performance of the two approaches to produce interval estimates, Figure 3.18 displays plots of bias against confidence interval width. In general, wider intervals in the case of a small number of subjects get narrower as the number of subjects increases. In

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	97	97	94	96	89	91
	10	96	96	95	94	88	89
	30	94	95	95	94	86	87
Profile Likelihood Method	6	96	97	95	95	90	92
	10	96	97	95	95	91	92
	30	95	96	96	96	93	94

Table 3.12: Estimated Coverage Rates for 2 and 4 replicates (visits) per subject for Simple Replicate Model (simulates and fitted order effect).

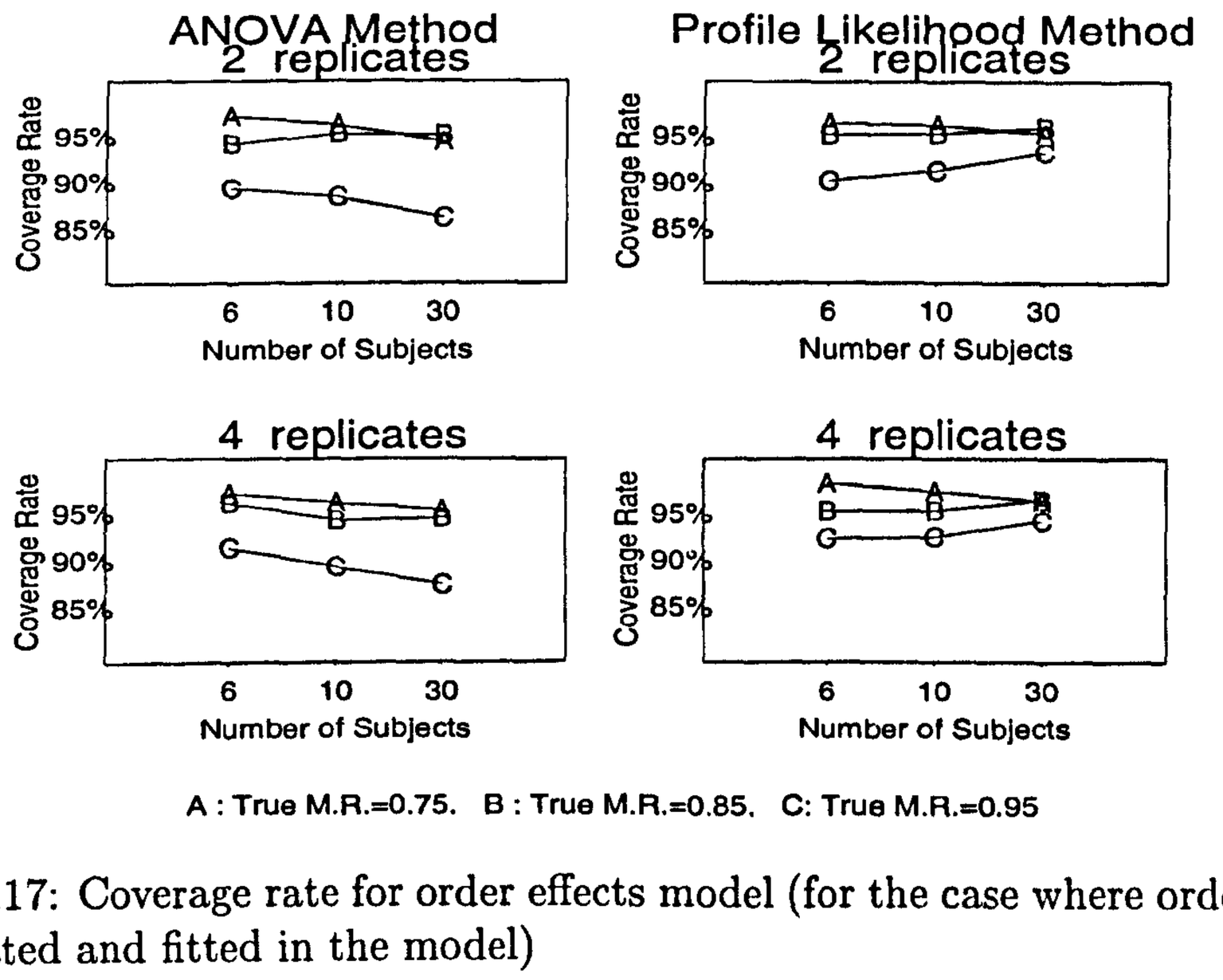

Figure 3.17: Coverage rate for order effects model (for the case where order effect is simulated and fitted in the model)

addition, in the case of 2 replicates per subject and small number of subjects, the intervals from the ANOVA approach are, on average, wider than those by the other approach. It seems that although an increase in the number of replicates changes the pattern of the intervals, it does not have a significant effect on the interval widths.

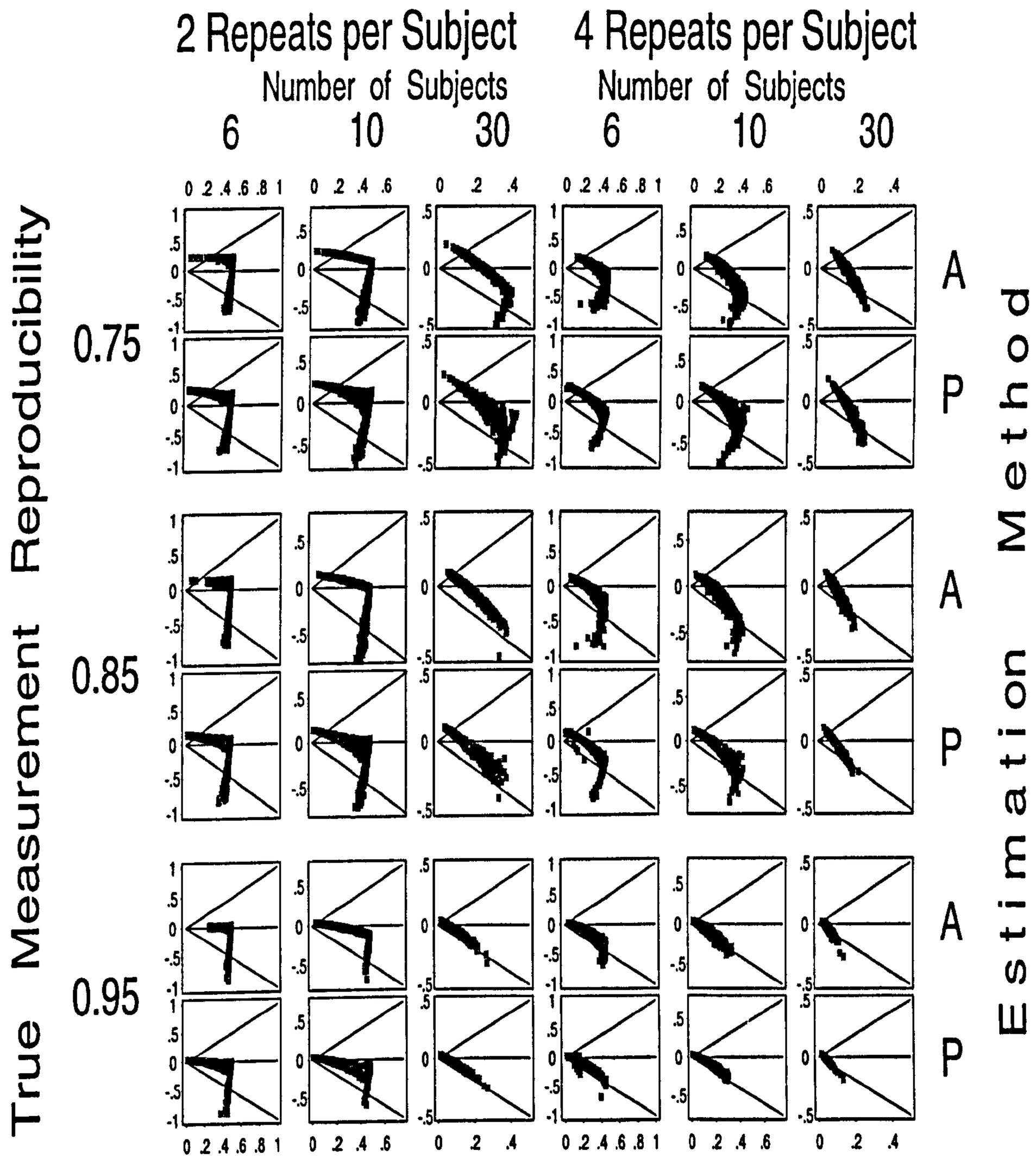

Figure 3.18: Confidence Diagrams with 2 and 4 replicates per subject (visits) for Order Effect Model (for the case where order effect is simulated and fitted in the model) for different combinations of number of subjects and true measuement reproducibility.
In each diagram Vertical Axis represents Bias and Horizontal Axis represents (Confidence Interval Width)/2.

3.3.2.2 An Order Effect Simulated but <u>not</u> Fitted

Simulation results from the situation where an order effect is simulated but not fitted in the model, are considered in this section.

Average of the estimated measurement reproducibilities and coverage rates over 1000 simulations are displayed in Tables 3.13 and 3.14 with graphical presentation of biases and coverage rates in Figures 3.19 and 3.20, respectively.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	0.60	0.62	0.72	0.72	0.83	0.84
	10	0.64	0.64	0.74	0.74	0.86	0.86
	30	0.68	0.69	0.77	0.80	0.90	0.91
Profile Likelihood Method	6	0.61	0.61	0.71	0.73	0.85	0.86
	10	0.64	0.65	0.75	0.76	0.89	0.89
	30	0.71	0.71	0.81	0.81	0.91	0.93

Table 3.13: Estimated Measurement Reproducibility for 2 and 4 replicates (visits) per subject for Order Effect Model(an order effect simulated but not fitted).

As in the previous sections, both the ANOVA and the Profile Likelihood approaches, regardless of the number of replicates, underestimate measurement reproducibility with a clear reduction in the bias for a larger number of subjects. Furthermore, point estimates for higher true reproducibility are slightly less biased than the estimates for other values of true reproducibility. In general, point estimates from the ANOVA approach look more biased than those from the Profile Likelihood approach.

Results from this section in comparison with those from the case

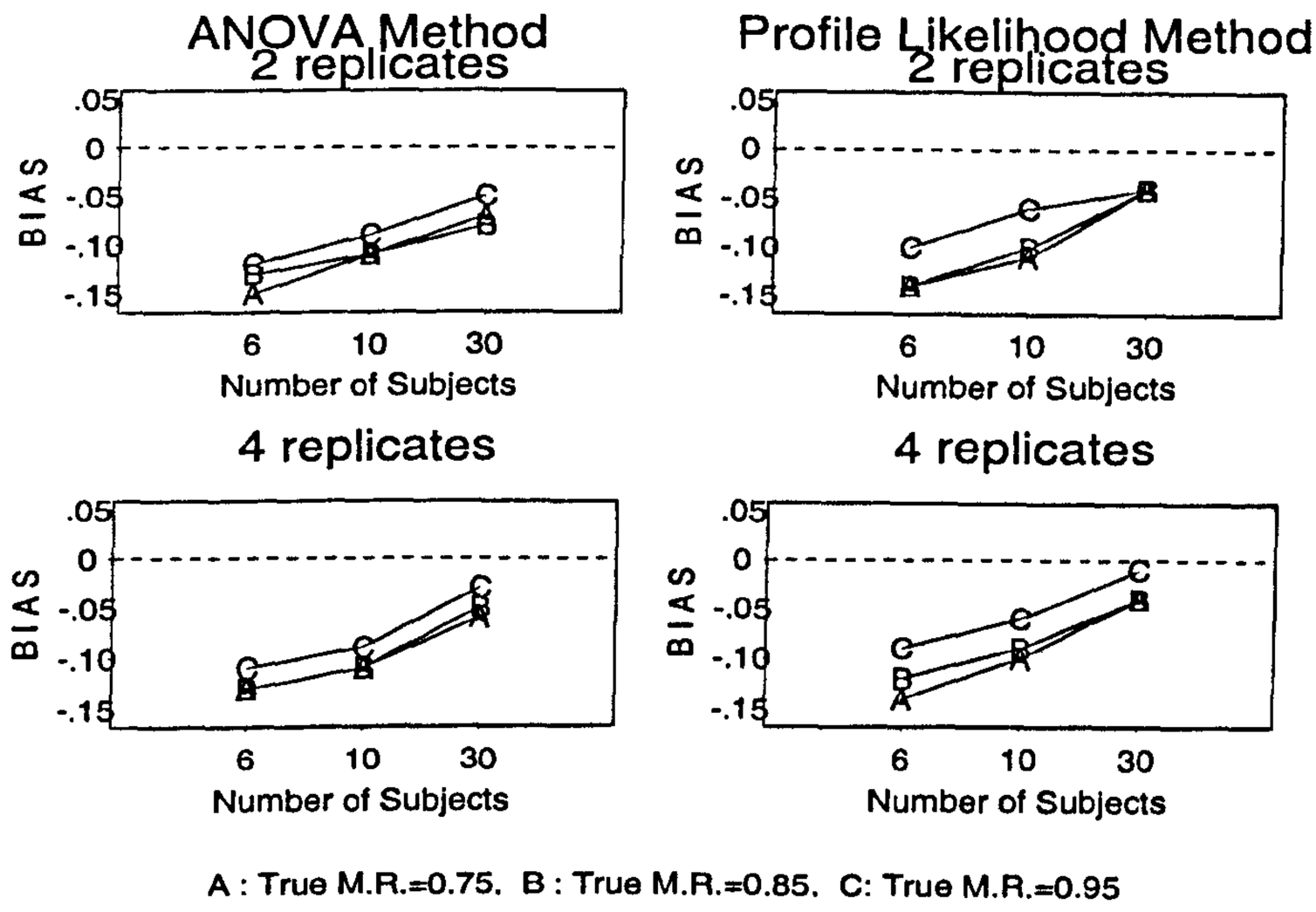

Figure 3.19: Bias from true Measurement Reproducibility for ordered effects model (for the case of an order effect simulated but not fitted in the model)

where the order effect was simulated and fitted in the model (section 3.3.2.1) show that, generally, a failure to consider the order effect in the model increases the bias except in the ANOVA approach for the cases of a small number of subjects with 2 replicates per subject.

As far as coverage rates, as a measure of performance of the two approaches to providing confidence intervals, are concerned, Table 3.14 and Figure 3.20 show that for lower values of measurement reproducibility (i.e. ρ_T=0.75 or 0.85), both approaches produce consistent confidence in the range of 95% but for high true reproducibility (i.e. $\rho_T = 0.95$), high coverage rates for the case of a small number of subjects significantly decrease as the number of subjects increases. Moreover, it seems that an increase in the number of replicates, does not affect the coverage rates for lower true reproducibility, but tends to a slight decrease in the coverage rates for high true reproducibility.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	94	94	94	95	92	92
	10	95	95	94	94	91	89
	30	94	94	93	92	84	83
Profile Likelihood Method	6	94	95	93	96	91	93
	10	94	96	94	95	91	92
	30	94	96	94	93	89	85

Table 3.14: Estimated Coverage Rates for 2 and 4 replicates (visits) per subject for Simple Replicate Model (an order effect simulated but not fitted).

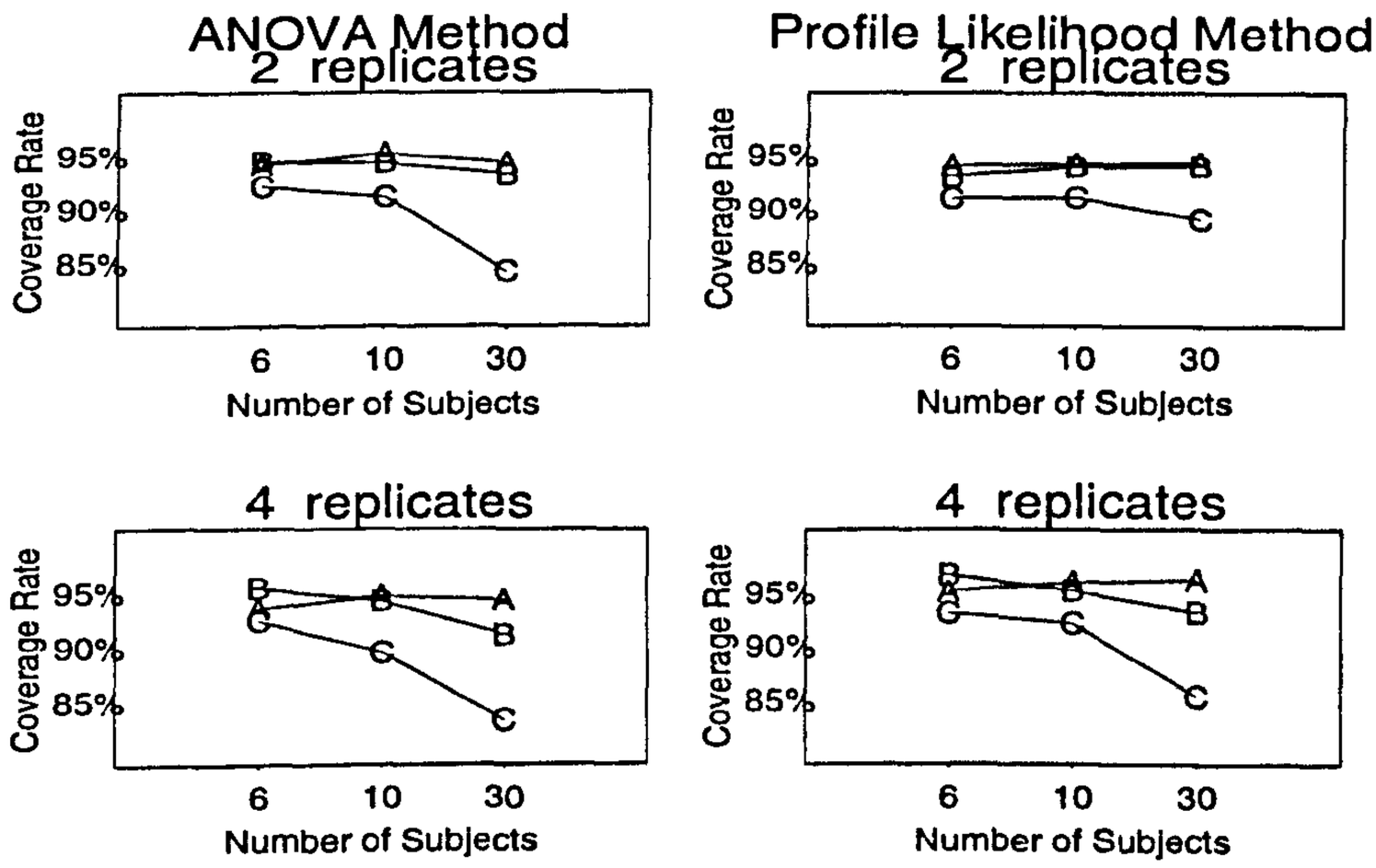

Figure 3.20: Coverage rate for ordered effects model (for the case of an order effect simulated but not fitted in the model)

In general, it appears that the Profile Likelihood approach has a better performance in terms of coverage rates.

To consider the effect of a failure to fit an order effect on the coverage

rates, Tables 3.12 and 3.14 show that this failure generally decreases the coverage rates. To examine the form of confidence intervals provided by the different approaches, Figure 3.21 represents plots of bias against confidence interval widths. As noted before, points inside the wedge shape (<) are those intervals that capture the true measurement reproducibility.

Both approaches provide wide intervals in the case of small number of subjects but get narrower as the number of subjects increases. This decrease in the interval widths is more obvious with 4 replicates per subjects. In addition, an increase in the number of replicates from 2 to 4, in most cases, does not have a significant effect on the intervals, although it slightly changes the pattern of the interval widths.

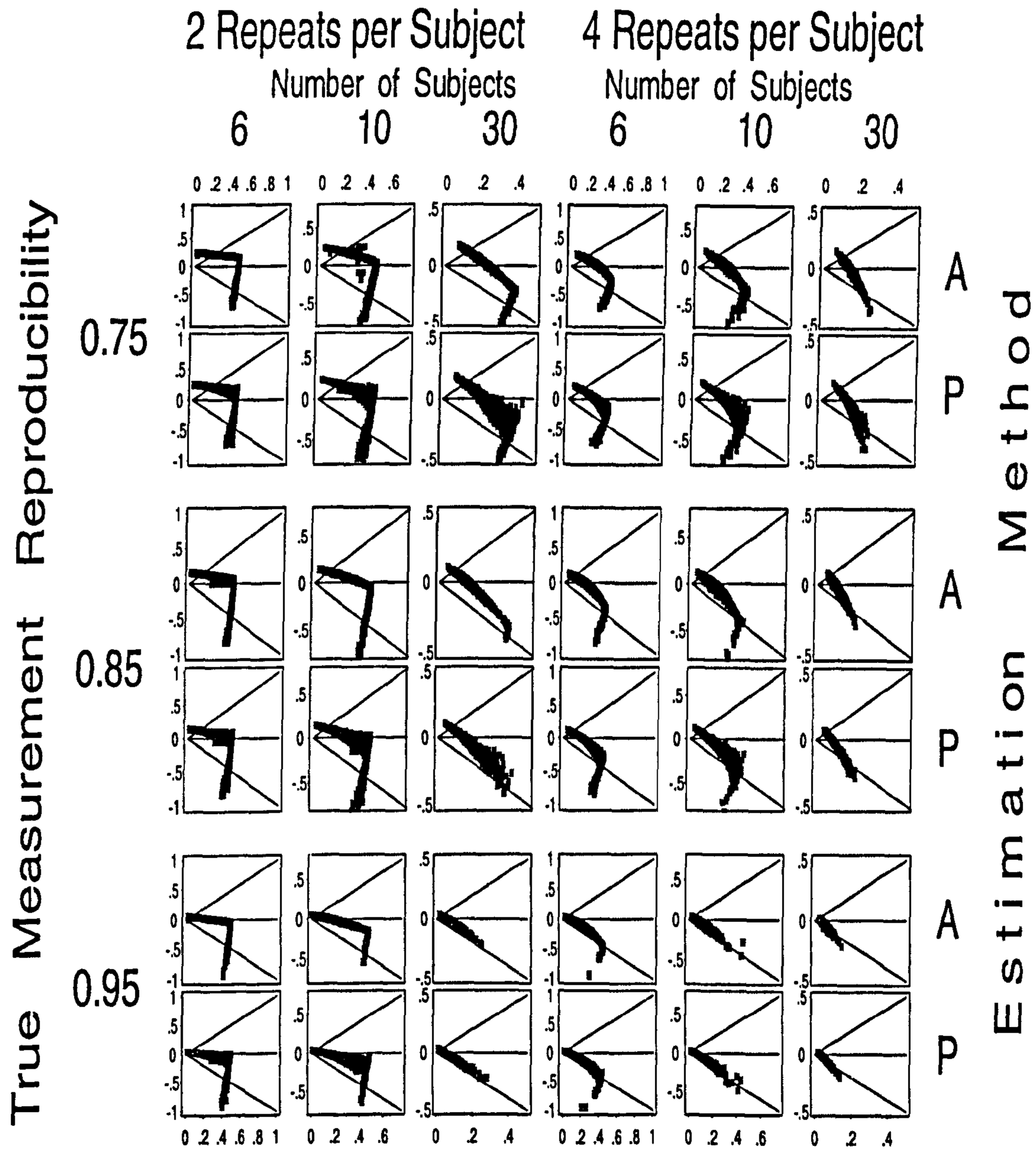

Figure 3.21: Confidence Diagrams with 2 and 4 replicates per subject (visits) for Order Effect Model (for the case where an order effect is simulated but not fitted in the model) for different combinations of number of subjects and true measuement reproducibility.
In each diagram Vertical Axis represents Bias and Horizontal Axis represents (Confidence Interval Width)/2.

3.3.2.3 An Order Effect not Simulated but still Fitted

Finally, the situation where the order (visit) effect is not simulated but actually is fitted in the model, is illustrated in this section.

Tables 3.15 and 3.16 present averages of estimated measurement reproducibility and coverage rates over 1000 simulations and bias values and coverage rates for different configurations are given in Figures 3.22 and 3.23.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	0.51	0.66	0.63	0.75	0.76	0.85
	10	0.62	0.64	0.73	0.75	0.87	0.88
	30	0.72	0.71	0.82	0.82	0.93	0.93
Profile Likelihood Method	6	0.62	0.65	0.71	0.74	0.80	0.86
	10	0.66	0.67	0.76	0.78	0.88	0.89
	30	0.73	0.73	0.83	0.83	0.93	0.93

Table 3.15: Estimated Measurement Reproducibility for 2 and 4 replicates (visits) per subject for Order Effect Model (an order effect not simulated but still fitted in the model).

As in the previous situations, both approaches underestimate the measurement reproducibility with a significant effect of the number of subjects. It is clear that for higher number of subjects (i.e. N=10 or 30), regardless of the number of replicates or true reproducibility, there is a 'relatively good' agreement between the two approaches to point estimation.

Furthermore, highly biased estimates by the ANOVA approach in the case of small number of subjects, dramatically improve with

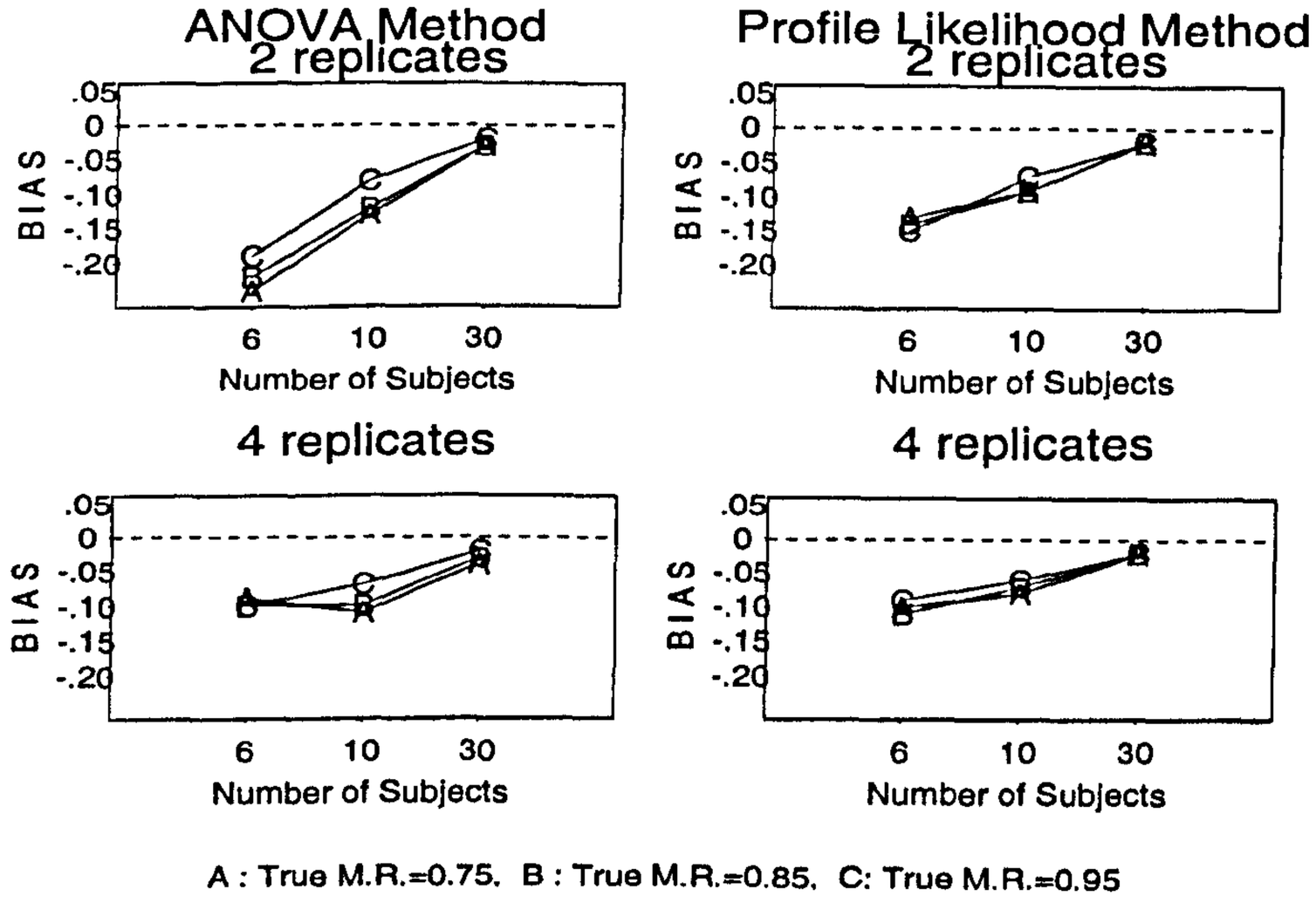

Figure 3.22: Bias from true Measurement Reproducibility for Order Effect Model (for the case where an order effect not simulated but still fitted in the model)

an increase in the number of replicates per subject. It seems that except for small number of subjects, for other simulation configurations, an increase in the number of replicates does not have a significant effect on the point estimates.

In general, the Profile Likelihood approach shows a better performance in estimating measurement reproducibility.

Comparing the point estimates from this case with those from the case where an order effect was simulated and fitted in the model (sections 3.3.2.1) gives an overall impression that fitting an order effect in the model, when in fact there is no evidence of a significant order effect, does not cause a significant change in the bias. However, when compared to estimates from the case where an order effect was simulated but not fitted in the model (section 3.3.2.2), it can be seen that fitting an order effect in the model, even if there is no evidence of a significant order effect , in general, reduces the bias

except for the ANOVA approach with a small number of subjects and 2 replicates per subject.

To assess the performance of the confidence intervals, the coverage rates, which are the percentage of times that the intervals contain the true measurement reproducibility, are given in Table 3.16 with a graphical representation in Figure 3.23.

High coverage rates (in the range of 95%) for lower values of true measurement reproducibilities (i.e. ρ_T=0.75 or 0.85), have been achieved regardless of the number of subjects or replicates per subject.

In the case of high true measurement reproducibility (i.e. ρ_T=0.95), both approaches provide lower coverage rates than 95%, but behave differently with an increase in the number of subjects and the number of replicates. For the ANOVA approach, an increase in the number of subjects tends to decrease the coverage rates whereas, for the Profile Likelihood approach, an increase in the number of subjects increases the coverage rates.

In general, the Profile Likelihood approach shows a better performance in terms of coverage rate.

Comparing the coverage rates in this section with those in the previous two sections, one can see that, in general, there does not appear to be a significant difference between the coverage rates in the case of not simulated but fitted order effects and those in the case where an order effect was simulated and fitted in the model (section 3.3.2.1). However, compared to the case where an order effect was simulated but not fitted in the model (section 3.3.2.2), it can be seen that, coverage rates significantly increase in almost all configurations except for high true reproducibility (i.e. $\rho_T = 0.95$). In this case a small number of subjects (i.e. N=6 or 10) results in a

decrease in the coverage rates, but with a large number of subjects there is an increase in the coverage rates. This shows that failure to considering a 'significant' order effect in the model decreases the coverage rates in almost all simulation configurations.

Estimation Method	*No. of Subjects (N)*	ρ_T					
		0.75		*0.85*		*0.95*	
		no. of replicates		*no. of replicates*		*no. of replicates*	
		2	4	2	4	2	4
ANOVA Method	6	96	96	96	96	84	87
	10	96	96	95	95	85	86
	30	96	96	94	95	88	84
Profile Likelihood Method	6	94	95	96	96	90	91
	10	96	96	97	97	91	93
	30	97	98	98	98	93	94

Table 3.16: Estimated Coverage Rates for 2 and 4 replicates (visits) per subject for Simple Replicate Model (an order effect not simulated but still fitted).

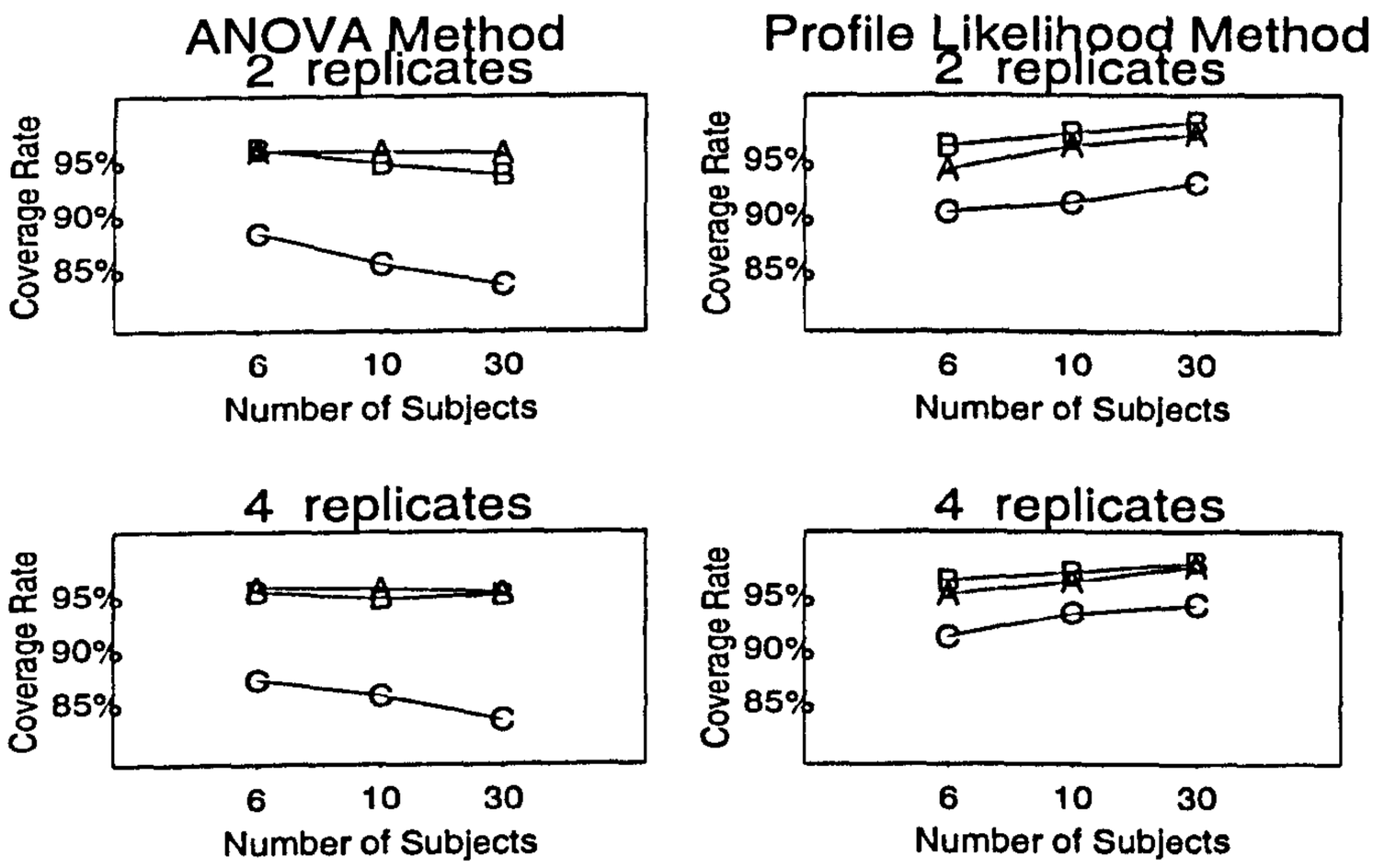

Figure 3.23: Coverage rate for ordered effects model (for the case where an order effect is not simulated but still fitted in the model)

To investigate the performance of the two approaches in terms of confidence intervals, plots of bias against confidence interval widths are presented in Figure 3.24.

As in the previous situation, for the case of 2 replicates per subject, an increase in the number of subjects does not have a significant effect on the interval widths, whereas for the case of 4 replicates, an increase in the number of subjects slightly reduces the interval widths.

Overall, it seems that there is not a significant difference between the pattern of interval widths from the two approaches.

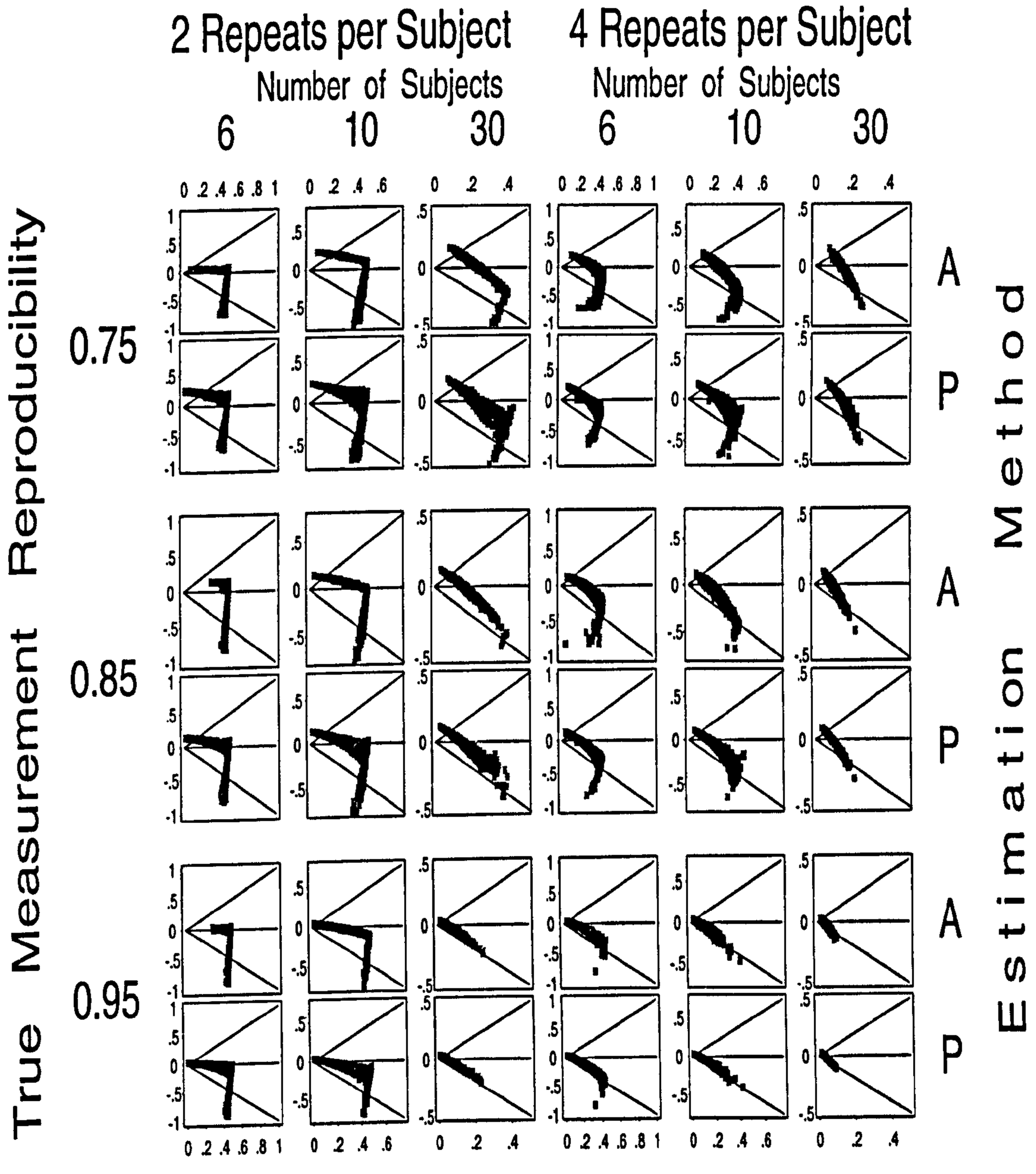

Figure 3.24: Confidence Diagrams with 2 and 4 replicates per subject (visits) for Order Effect Model (for the case where an order effect is not simulated but still fitted in the model) for different combinations of number of subjects and true measuement reproducibility.
In each diagram Vertical Axis represents Bias and Horizontal Axis represents (Confidence Interval Width)/2.

3.4 Summary

A simulation study was carried out to compare the performance of the ANOVA and the Profile Likelihood approaches to estimate measurement reproducibility and related confidence intervals. This was done for the cases of balanced and unbalanced data and based on a variety of underlying simulation configurations.

To investigate the effect of learning or familiarisation in the Exercise Tests, an Order Effect Model was also investigated across a variety of simulation configurations.

Both the ANOVA and the Profile Likelihood approaches underestimate measurement reproducibility with a clear reduction in the bias as the number of subjects increases. For the case of balanced data, more or less, there is a 'good' agreement between the two approaches to point estimation. This agreement is more obvious for a large number of subjects.

The ANOVA approach, in the case of balanced data and 2 replicates per subject, provides wider confidence intervals than the Profile Likelihood approach. In the other cases, there is not a clear difference between the estimated interval widths.

As an overall summary of the performance of both approaches across all the simulation situations considered the following table attempts to pull together the conclusions from each different situation considered:

	Estimation	*Model*			
		Simple R. model	*Order Effect Model*		
			SF	*SNF*	*NSF*
Balanced Data	*Bias*	A	N	A	N
	Coverage rate	P	P	P	P
	Interval Width	P	P	P	P
Unbalanced Data	*Bias*	P	P	P	P
	Coverage rate	P	P	P	P
	Interval width	N	P	N	N

Table 3.17: Summary results for all simulations.

Order Effect Structure

SF : Simulated and Fitted visit effect

SNF: Simulated but Not Fitted visit effect

SFN: Not Simulated but Fitted visit effect

Table Entries (on the basis of the specific simulation situation)

P : The Profile Likelihood approach performs better than the ANOVA approach

A: The ANOVA approach performs better than the Profile Likelihood approach

N : No clear difference between the approaches was appeared

Overall, the simulation results in this table give a clear impression that the Profile Likelihood approach is the better approach to producing point and interval estimates of measurement reproducibility.

Chapter 4

Estimating the Comparability of two distinct Variables: How to pool correlation coefficients

4.1 Introduction

In Sports Science, variables are often measured at a number of time points during an Exercise Test, and often such tests are repeated over a number of visits for each individual in a sample of individuals. Interest often focuses on the Comparability of variables e.g. whether the physiological variables such as frequency of breathing, ventilation, heart rate etc are at all well related to psychological variables such as the perceived rate of exertion or breathlessness across at least 'replicate visits' and/or indeed across individuals.

Clearly the strength of (linear) relationship between any two variables through one exercise test on one individual may be measured by a correlation coefficient. The *Comparability* between two vari-

ables is then defined as combining estimates of linear correlation coefficients from different replicate visits across individuals into either a "pooled estimate" of a common correlation or into an estimate of the typical correlation of the variables on a 'typical visit' for a typical subject. The first step in this process is the attempt in this chapter to investigate how best to provide point and interval estimates for an assumed common correlation by pooling across a number of visits/exercise tests for one individual.

4.2 Estimating a Common Correlation

4.2.1 Model

Assume that there are I independent estimates (e.g. I Exercise Tests on the same individual) of an assumed common correlation, ρ, i.e. I independent bivariate samples with, possibly, *different means* and *variances* but the *same population correlation* ρ.

Interest here is how to pool the separate estimates of correlation for each sample (i.e. each Exercise Test) to provide a point and interval estimate for this common correlation, ρ.

4.2.2 Data

Let $\{x_{ij}, y_{ij} : j = 1, 2, \ldots, n_i\}$ be the i^{th} random sample ($i = 1, 2, \ldots I$).
Now

$$\hat{\rho}_i = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

is the i^{th} sample estimate of correlation,

where

$$S_{ij} = \sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)(y_{ij} - \overline{y}_i) \quad \text{etc.}$$

Further let

$$N = n_1 + n_2 + \ldots + n_I$$

4.2.3 Methods of Point Estimation

Five distinct methods of producing a point estimate of ρ based on pooling the I sample correlations are investigated in this section.

4.2.3.1 Weighted Estimate

The simplest and most immediately appealing estimate is

$$\hat{\rho}_w = \frac{\sum_{i=1}^{I} n_i \hat{\rho}_i}{N} \tag{4.1}$$

i.e. an average of the individual sample correlations weighted by the appropriate sample sizes.

4.2.3.2 Unbiased Estimate

Since it is well known that $\hat{\rho}_i$ is a biased estimate of ρ_i,

i.e.

$$E(\hat{\rho}_i) \approx \rho_i + \frac{\rho_i(1-\rho_i)}{2n_i}$$

it has been suggested (Olkin and Pratt, 1958) that,

$$G(\hat{\rho}_i) = \hat{\rho}_i + \frac{\hat{\rho}_i(1-\hat{\rho}_i^2)}{2(n_i-3)} \tag{4.2}$$

will be approximately unbiased for ρ_i. Accordingly rather than use $\hat{\rho}_W$ above, one could use a weighted average of the individual unbiased estimates for each sample,
i.e.

$$\hat{\rho}_U = \frac{\sum_{i=1}^{I} n_i G(\hat{\rho}_i)}{N} \tag{4.3}$$

Obviously this will be approximately unbiased for ρ.

4.2.3.3 Fisher Estimate

Since any Normal approximation to the distribution of any estimator of ρ is clearly unsatisfactory (as ρ can only range between -1 and +1), Fisher(1921) suggested the use of the transformation,

$$F(\rho) = \frac{1}{2} log \frac{1+\rho}{1-\rho}$$

He then deduced that $F(\hat{\rho})$ was approximately normally distributed with mean $F(\rho)$ and variance $1/(n-3)$.

A possible point estimate of the common correlation investigated here based on the use such a transformation might be a weighted average of the form:

$$\hat{\rho}_F = F^{-1}\left\{\sum_{i=1}^{I}(n_i-3)F(\hat{\rho}_i)/(N-3I)\right\} \tag{4.4}$$

4.2.3.4 Hedges and Olkin Estimate

It is possible to write down a likelihood involving all I separate population means and variances as well as the common population correlation.

In general, the maximum likelihood estimate for ρ cannot be expressed in closed form. However, an approximate estimate of it can be obtained numerically as the solution of $g(\rho) = 0$,

where

$$g(\rho) = \frac{N\rho}{1-\rho^2} - \sum_{i=1}^{I} \frac{n_i \hat{\rho}_i}{1-\rho\hat{\rho}_i} \tag{4.5}$$

(Hedges and Olkin, 1985).

Denote such an estimate by $\hat{\rho}_{HO}$.

4.2.3.5 Profile Likelihood Estimate

Profile Likelihood may be used as a method of parameter estimation when many nuisance parameters are involved in a model (in this case the 'population' means and variances for each sample are nuisance parameters i.e. $\mu_{x_i}, \mu_{y_i}, \sigma^2_{x_i}, \sigma^2_{y_i}$, for $i = 1, 2, \ldots, I$). The likelihood contribution L_i from the i^{th} sample $(i = 1, 2, \ldots, I)$ can be written as

$$L_i = (\sigma^2_{x_i}\sigma^2_{y_i})^{-\frac{n_i}{2}} exp\left\{-\frac{1}{2(1-\rho^2)}\sum_{j=1}^{n_i}\left[\frac{(x_{ij}-\mu_{x_i})^2}{\sigma^2_{x_i}} - 2\rho\frac{(x_{ij}-\mu_{x_i})(y_{ij}-\mu_{y_i})}{\sigma_{x_i}\sigma_{y_i}} + \frac{(y_{ij}-\mu_{y_i})^2}{\sigma^2_{y_i}}\right]\right\}$$

with the full likelihood of $\prod_{i=1}^{I} L_i$.

To obtain the Profile Likelihood for ρ, maximise with respect to all

μ_{x_i}, σ_{x_i}, μ_{y_i} and $\sigma_{y_i}(i = 1, 2, \ldots, I)$ for fixed ρ to give

$$PLik(\rho) = \prod_{i=1}^{I} \left\{ (\hat{\sigma}_{x_i}^2 \hat{\sigma}_{y_i}^2)^{-\frac{n_i}{2}} (1-\rho^2)^{-\frac{n_i}{2}} exp\left\{ -\frac{n_i(1-\rho\hat{\rho}_i)}{1-\rho^2} \right\} \right\}$$

where

$$\hat{\sigma}_{x_i}^2 = \frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{n_i} \text{ and } \hat{\sigma}_{y_i}^2 = \frac{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{n_i}$$

Finally, after some algebra, the full log profile likelihood, could be written as (ignoring constants)

$$l(\rho) = -\frac{N}{2}\log(1-\rho^2) - \sum_{i=1}^{I} \frac{n_i(1-\rho\hat{\rho}_i)}{1-\rho^2} \tag{4.6}$$

The maximum profile likelihood estimator of ρ would be obtainable by equating to zero the partial derivative of $l(\rho)$ with respect to ρ, that is

$$\frac{\partial l(\rho)}{\partial \rho} = \frac{N\rho}{1-\rho^2} - \sum_{i=1}^{I} \frac{n_i(2\rho - \rho^2\hat{\rho}_i - \hat{\rho}_i)}{(1-\rho^2)^2} \tag{4.7}$$

Denote such an estimate by $\hat{\rho}_{PL}$.
Note how similar the form of this is to that in the previous section (i.e. $\hat{\rho}_{HO}$).

4.2.4 Methods of Interval Estimation

Rather than simply produce point estimates of a parameter, it is clearly more informative to produce interval estimates. Here, to provide approximate 95% confidence intervals for ρ, the following pivotal functions were used:

i) The Weighted, Unbiased and Hedges and Olkin estimates could all be used to provide interval estimates based

on the approximate result (Hedges and Olkins, 1985) that

$$\frac{\sqrt{N}(\hat{\rho}-\rho)}{1-\rho^2} \sim N(0,1). \tag{4.8}$$

Possible approximate 95% confidence intervals for ρ would be based on the solutions of

$$\left|\frac{\sqrt{N}(\hat{\rho}-\rho)}{1-\rho^2}\right|^2 = (1.96)^2 \tag{4.9}$$

which give intervals of the form

$$\left[\frac{\sqrt{N}-\sqrt{N-4\times 1.96(\sqrt{N}\hat{\rho}-1.96)}}{2\times 1.96}\,,\ \frac{-\sqrt{N}+\sqrt{N+4\times 1.96(\sqrt{N}\hat{\rho}+1.96)}}{2\times 1.96}\right] \tag{4.10}$$

where $\hat{\rho}$ in turn is replaced by $\hat{\rho}_W$, $\hat{\rho}_U$ and $\hat{\rho}_{HO}$ for the three methods.

ii) The Fisher estimate would use the following approximate result for the pivotal function,

$$\frac{F(\hat{\rho}_F)-F(\rho)}{\sqrt{1/(N-3I)}} \sim N(0,1)$$

to give approximate 95% confidence intervals for ρ of the form

$$F^{-1}\left[F(\hat{\rho}_F) \pm \frac{1.96}{\sqrt{N-3I}}\right]. \tag{4.11}$$

iii) For the Profile Likelihood method, the relative profile likelihood required to provide an approximate likelihood interval for ρ is

$$\begin{aligned} rl(\rho) &= l(\rho) - l(\hat{\rho}_{PL}) \\ &= -\frac{N}{2}\log\left(\frac{1-\rho^2}{1-\hat{\rho}_{PL}^2}\right) - \sum_{i=1}^{I}\frac{n_i(1-\rho\hat{\rho}_i)}{1-\rho^2} \\ &\quad + \sum_{i=1}^{I}\frac{n_i(1-\hat{\rho}_{PL}\hat{\rho}_i)}{1-\hat{\rho}_{PL}^2} \end{aligned} \tag{4.12}$$

Hence a likelihood interval with approximate 95% confidence would be

$$\{\rho : \ rlogPLik(\rho) \geq -1.92\}$$

since it is likely that

$$D = -2rlogPLik(\rho) \sim \chi^2_{(1)} \quad \text{approximately.}$$

4.2.5 Illustration of the Estimation methods

4.2.5.1 Data

In a study of the reliability of an Exercise Testing procedure, each of 12 subjects had their Breathlessness measured on a Visual Analogue Scale(VASB) and Ventilation measured using a Douglas Bag at two minute intervals through the test to give 8 or 9 time points per subject per test. The subjects exercised at progressive difficulties until voluntary withdrawal or 18 minutes. The Exercise Test was repeated on 8 different occasions for each subject. Scatterplots of these values for each of the 12 subjects are presented in Figure 4.1.

For each subject it is assumed that the correlation between VASB and Ventilation across time in any test is the same for all tests although the time profile may shift on one or both variables from test to test due to variations in temperature, fitness etc.

The correlations for each subject are displayed in Figure 4.2 and it appears they may be assumed constant across visits (i.e. in the notation of the previous sections I=8 and n_i=8 or 9, as appropriate, for each subject). It may even be reasonable to assume that the common correlation for the i^{th} subject (i=1,2,..., I) is the same for all subjects but this will be investigated later (see 4.2.6).

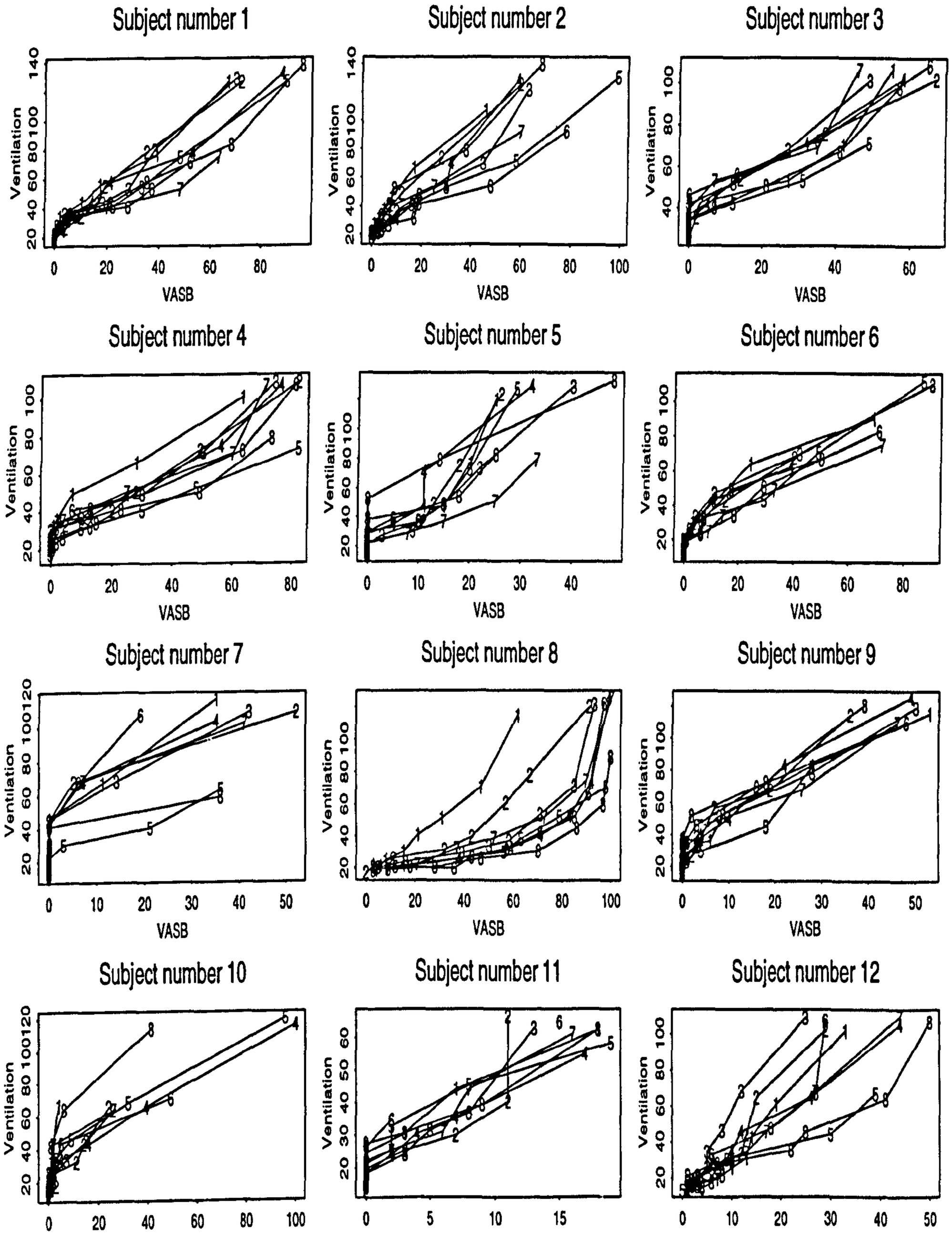

Figure 4.1: Scatterplots of the two variables Visual Analogue Scale for Breathlessness (VASB) and Ventilation for each of the 12 subjects.

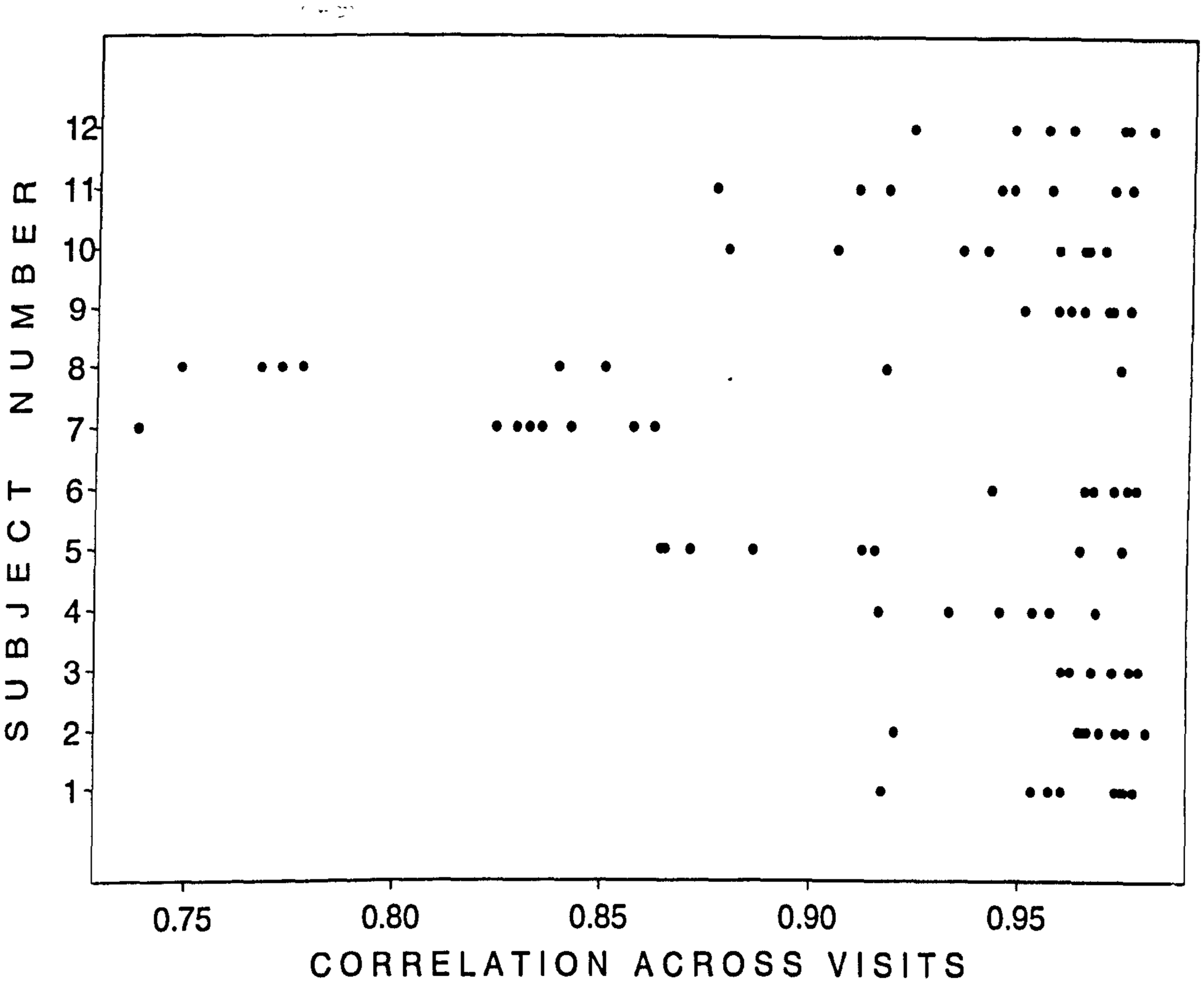

Figure 4.2: Sample Correlations between VASB and Ventilation across each of the 8 visits for each of the 12 subjects.

4.2.5.2 Results

Table 4.1 gives the point and interval estimates of the common correlations for each of the 12 subjects by each of the 5 methods of estimation and Figure 4.3 displays these results graphically.

Point estimates of the common correlation coefficients from the Weighted method are in general less than those estimates from the

other methods. The estimates from the Fisher method and Hedges and Olkin method appear very similar.

Figure 4.4 displays the appropriate approximate 95% confidence intervals. Except for the Profile Likelihood method which in all cases provides the narrowest intervals, there are no obvious general differences among the widths of the intervals from the other 4 methods, although, in cases with very narrow intervals, those based on the Hedges and Olkin estimate are narrower than those based on the other 3 methods.

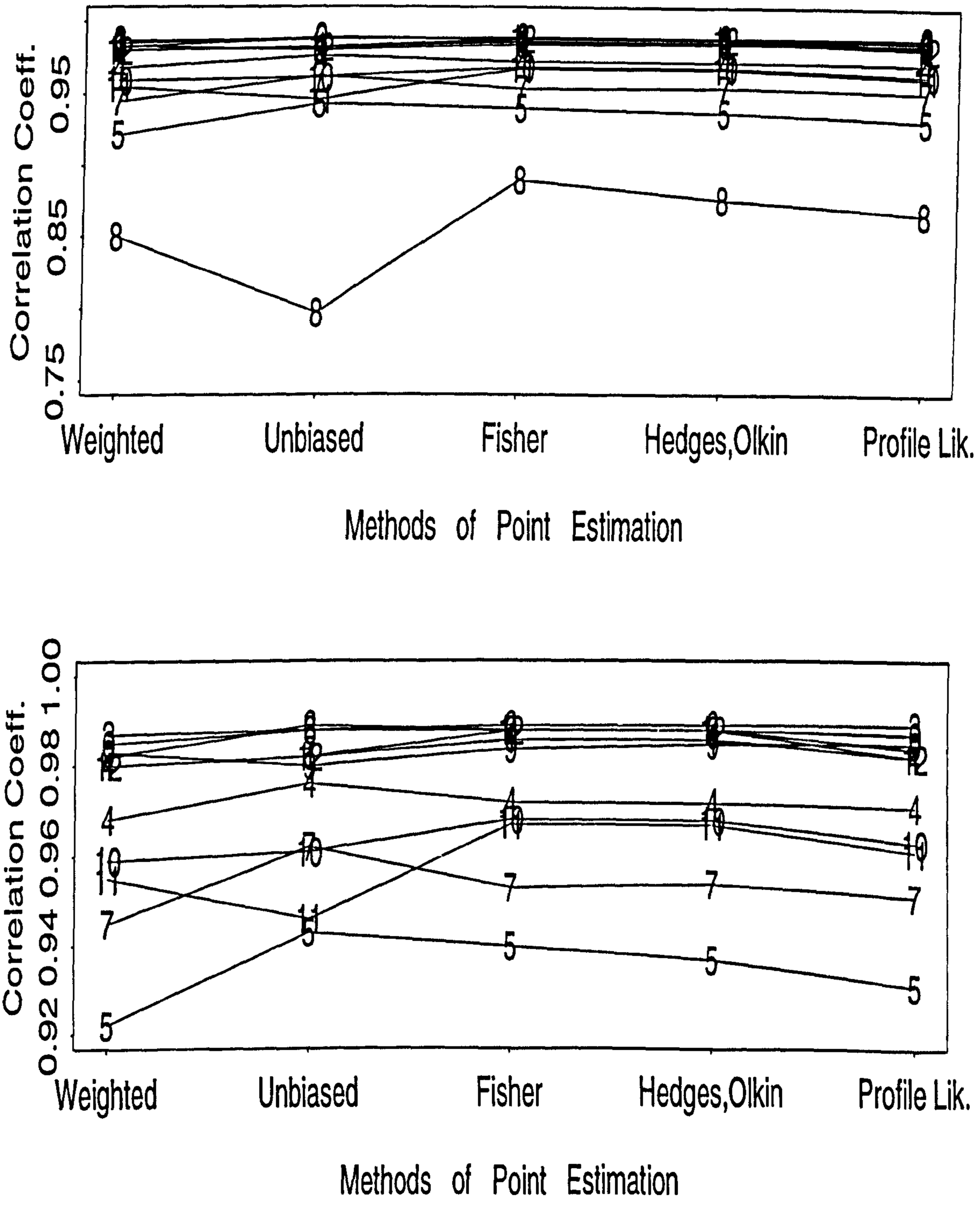

Figure 4.3: Estimates of Common Correlation for each subject between VAS for Breathlessness and Ventilation by 5 different methods for each of 12 subjects and the same Common Correlations after removing subject number 8 which has the lowest correlation coefficient.

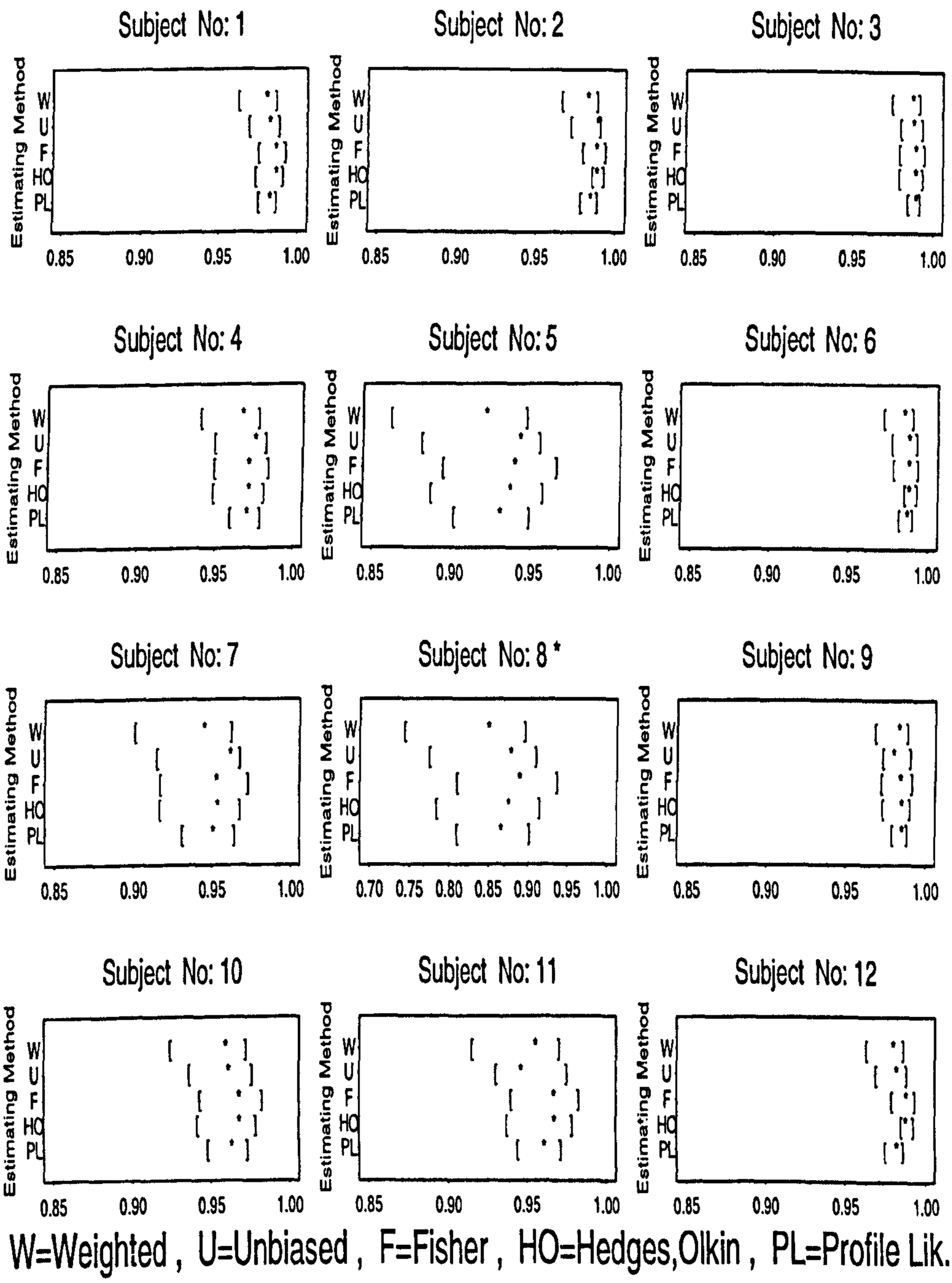

Figure 4.4: Point and Interval Estimates of Common Correlation for each subject between VAS for Breathlessness and Ventilation.
$*$: Different scale for subject number 8 with the widest confidence intervals is used.

Subject no.	*Weighted met.* (95% C.I.)	*Unbiased met.* (95% C.I.)	*Fisher met.* (95% C.I.)	*HO met.* (95% C.I.)	*PL met.* (95% C.I.)
1	0.980 **(0.962, 0.986)**	0.982 **(0.969, 0.988)**	0.986 **(0.975, 0.992)**	0.986 **(0.973, 0.990)**	0.982 **(0.975, 0.986)**
2	0.982 **(0.966, 0.988)**	0.989 **(0.972, 0.990)**	0.988 **(0.979, 0.993)**	0.988, **(0.985, 0.992)**	0.984 **(0.978, 0.988)**
3	0.987 **(0.974, 0.991)**	0.988 **(0.980, 0.993)**	0.989 **(0.979, 0.994)**	0.989 **(0.979, 0.993)**	0.989 **(0.984, 0.991)**
4	0.968 **(0.941, 0.978)**	0.976 **(0.950, 0.982)**	0.972 **(0.950, 0.984)**	0.972 **(0.949, 0.981)**	0.971 **(0.960, 0.979)**
5	0.922 **(0.862, 0.947)**	0.943 **(0.881, 0.955)**	0.940 **(0.895, 0.966)**	0.937 **(0.887, 0.957)**	0.931 **(0.902, 0.949)**
6	0.985 **(0.972, 0.990)**	0.988 **(0.977, 0.992)**	0.988 **(0.978, 0.993)**	0.988 **(0.985, 0.992)**	0.987 **(0.982, 0.990)**
7	0.945 **(0.901, 0.962)**	0.962 **(0.915, 0.968)**	0.953 **(0.917, 0.973)**	0.954 **(0.917, 0.968)**	0.951 **(0.931, 0.964)**
8	0.850 **(0.745, 0.896)**	0.879 **(0.776, 0.910)**	0.890 **(0.811, 0.937)**	0.876 **(0.785, 0.915)**	0.866 **(0.810, 0.902)**
9	0.983 **(0.968, 0.988)**	0.980 **(0.973, 0.990)**	0.984 **(0.972, 0.991)**	0.985 **(0.973, 0.990)**	0.985 **(0.979, 0.988)**
10	0.959 **(0.924, 0.972)**	0.961 **(0.936, 0.976)**	0.968 **(0.943, 0.982)**	0.968 **(0.941, 0.978)**	0.963 **(0.948, 0.973)**
11	0.955 **(0.915, 0.970)**	0.946 **(0.930, 0.975)**	0.967 **(0.940, 0.982)**	0.967 **(0.937, 0.978)**	0.961 **(0.944, 0.971)**
12	0.980 **(0.963, 0.986)**	0.982 **(0.969, 0.988)**	0.988 **(0.979, 0.993)**	0.988 **(0.985, 0.992)**	0.982 **(0.975, 0.986)**

Table 4.1: Estimated Common Correlations and approximate 95% confidence intervals for VAS for Breathlessness and Ventilation

4.2.6 Checking the Assumption of the Commonality of a sample of Estimated Correlation Coefficients

4.2.6.1 Introduction

Before pooling the sample correlations across different visits to provide an estimate of common correlation, it is important to test whether or not these correlations do in fact have the same common population correlation, ρ. In other words, to test the consistency of the estimated sample correlations.

4.2.6.2 Tests of Commonality

Suppose there are $\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_I$ independent sample correlation coefficient estimates, with each $\hat{\rho}_i$ based on a sample of n_i observations from a bivariate normal population having correlation coefficient ρ_i. Several approaches have been suggested to test the consistency of correlations (i.e. testing $H_0 : \rho_i = \rho$, vs $H_1 : \rho_i \neq \rho$ for all i, $i = 1, 2, \ldots, I$).
In the first approach which is based on the Fisher Transformation, let $F(\hat{\rho}_1), F(\hat{\rho}_2), \ldots, F(\hat{\rho}_I)$ be the Fisher transformed sample correlations and $F(\hat{\rho}_F)$ be the Fisher estimate of the common correlation (see section 1.2.3.3),
i.e.

$$F(\hat{\rho}_F) = \frac{\sum_{i=1}^{I}(n_i - 3)F(\hat{\rho}_i)}{N - 3I}$$

Now the commonality of the correlation coefficients is based on regarding

$$Q = \sum_{i=1}^{I}(n_i - 3)(F(\hat{\rho}_i) - F(\hat{\rho}_F))^2 \qquad (4.13)$$

as having approximately chi-square distribution with (I-1) degrees of freedom under H_0.
The second approach is based on a likelihood ratio test and the commonality of the correlations is based on

$$LRT = -\sum_{i=1}^{I}(n_i - 2)\log\left[1 - \left(\frac{\hat{\rho}_i - \hat{\rho}_{HO}}{1 - \rho_i\hat{\rho}_{HO}}\right)^2\right] \tag{4.14}$$

with $\hat{\rho}_{HO}$ as the Hedges and Olkin estimator of common correlation coefficient, also having approximately a χ^2_{I-1} distribution under H_0. For the example in section 4.2.5.1, for subject number 1 with 8 sample correlation coefficients from 8 assumed independent visits (i.e. I=8), the two observed values of the statistics (i.e. Q and LRT) are 7.64 and 8.60, respectively, neither of which proves significant to reject H_0 (i.e. both are less than, 14.07, the upper 95 percentile (critical value) of a Chi-Squared distribution with 7 degrees of freedom).

The appropriate values of the test statistics for each of the other 11 subjects are presented in Table 4.2.

Comparing these values with the upper 95 percentile of a chi-square distribution with 7 degrees of freedom (i.e. 14.07), indicates that one should not reject the assumption of the consistency of the correlation coefficients for any of the subjects, except for number 12.

There is considerable variability of sample correlations across visits for some of the 12 subjects. Figure 4.2 showed the wide range of these sample correlations for subject number 8 but only moderate variability of those for subject number 12 compared to other subjects. Thus it is somewhat surprising that since virtually all samples for every subject are based on the same number of observations (i.e. 7 or 8) the wide range of sample correlations for subject 8 proves non-significant while the more restricted but numerically

Subject	*Q*	*LRT*
2	10.03	10.61
3	3.05	3.62
4	3.06	3.53
5	9.60	10.35
6	3.86	4.47
7	3.76	4.25
8	13.36	13.18
9	2.60	2.99
10	5.53	6.34
11	7.91	8.93
12	16.90	17.99

Table 4.2: Fisher and Likelihood Ratio Test statistics for the other 11 subjects

higher values for subject 12 show apparently clear evidence of a lack of commonality. This is further investigated in the following section.

4.2.6.3 Suitability of χ^2 approximate for Q and LRT statistics under H_0

First of all, the assumption of the test statistics having Chi-Squared distributions under the null hypothesis of normality was considered. Estimated probability density functions for Q and LRT statistics under H_0 based on 1000 simulations for I=8 and different values of true common correlations (i.e. $\rho = 0.75, 0.85, 0.90, 0.95, 0.99$) were constructed and compared to the probability density function of a Chi-Squared distribution with 7 degrees of freedom. These are presented in Figure 4.5 and show that the assumption of underlying Chi-Squared approximations to the test statistics under H_0 is fairly

reasonable here especially for the LRT approach.

4.2.6.4 Deviation from bivariate normality

It has been shown that the distribution of Q will not be even approximately χ^2_{I-1} under deviation from bivariate normality (Duncan and Layard, 1973). For each of the 12 subjects, bivariate normality of the original data at each visit was assessed by using jacknifed mahalanobis probability distance plots and tests of skewness. These showed no particular evidence of a lack of bivariate normality for subject 12 or indeed any of the other 11 subjects.

4.2.6.5 Variability in the Sample Correlation

For each of the 12 subjects, the variation of the sample correlation coefficients across visits is another characteristic of the data which is considered in this subsection.

Table 4.3 gives the rank ordered standard deviations of the correlations across visits based on both raw values and on the Fisher transformed values. Clearly, due to the effect of the Fisher transformation at high (> 0.95) values of ρ, the rank values are far from similar. It can be seen that whereas, in the raw data space, subject 12 is only of modest variability compared to the other subjects, in the transformed space it is the most variable across visits. So this may well be the explanation as to why subject 12 has to be rejected for the assumption of common correlation. i.e. The Fisher transformation can provide large variability in the transformed space if the typical value of the correlation tends towards one (or indeed minus one).

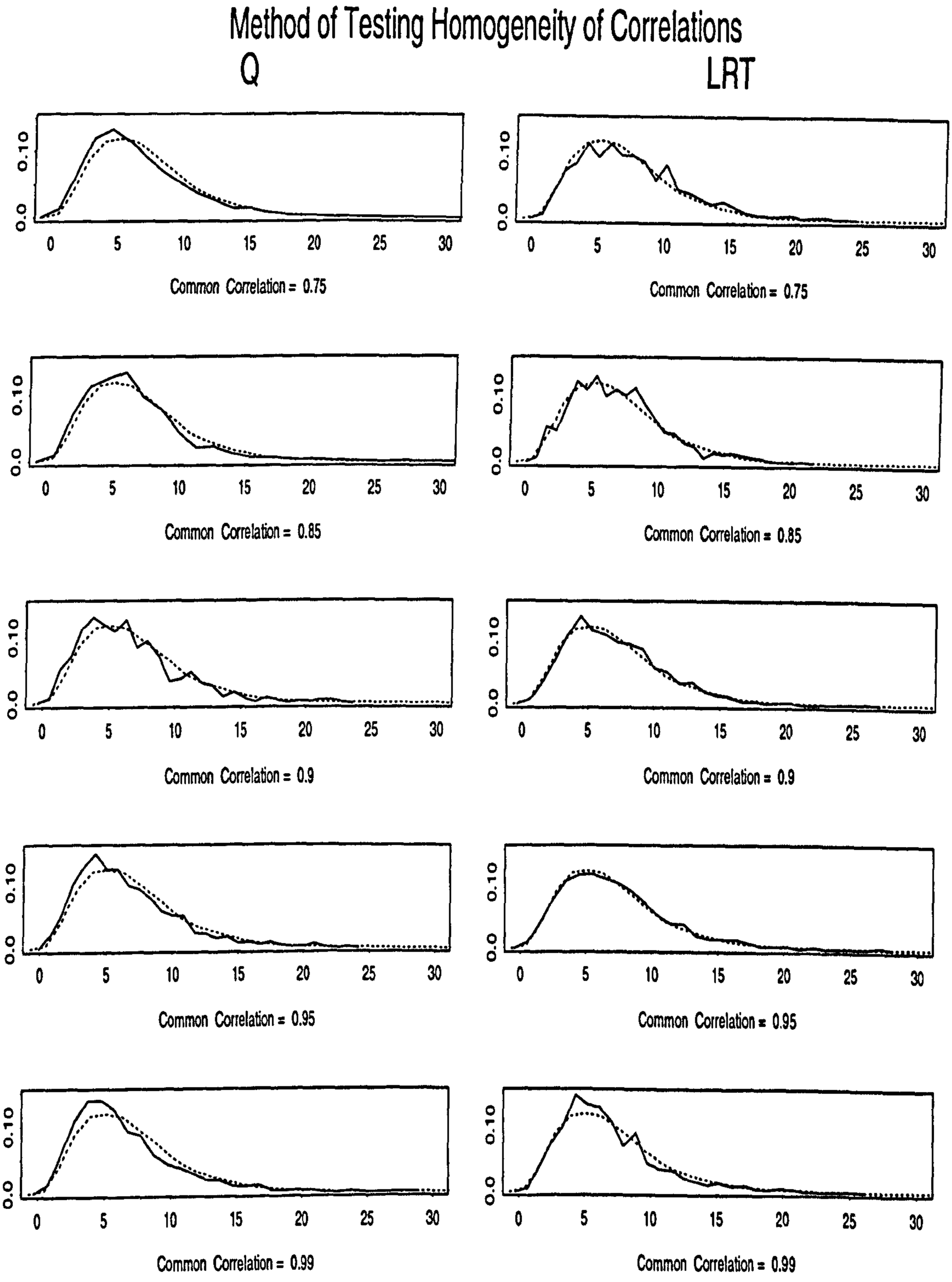

Figure 4.5: Estimated probability density functions,
——- : for Q and LRT statistics over 1000 simulations for different values of common correlation coefficient
..... : plot of a Chi-Square distribution with 7 degrees of freedom .

Order No.	*Subject No.*	*Std. Dev. of Raw Data*	*Subject No.*	*Std. Dev. of Fisher Trans. Data*
1	*3*	0.01	*9*	0.25
2	*9*	0.01	*4*	0.28
3	*6*	0.01	*3*	0.30
4	*4*	0.02	*7*	0.31
5	*2*	0.02	*6*	0.33
6	*1*	0.02	*10*	0.38
7	*12*	0.02	*1*	0.45
8	*10*	0.03	*2*	0.49
9	*11*	0.03	*11*	0.49
10	*7*	0.04	*5*	0.51
11	*5*	0.04	*8*	0.57
12	*8*	0.08	*12*	0.64

Table 4.3: Rank ordered Standard Deviation of Raw data and Fisher transformed data for each of the 12 subjects

4.2.6.6 Conclusions on the rejection of common correlation for subject 12

In previous subsections the suitability of the Chi-Squared approximation, the deviation from normality of the data and the variation of sample correlation coefficients in the Fisher transformed space were examined. The only characteristic of the data that might significantly influence the values of these statistics (i.e. Q and LRT) is the standard deviation of sample correlations across visits **in the Fisher transformed space**. In other words, these two statistics are very sensitive to the location of the transformed data. Subject number 12 with a comparable variability of sample correlations to other subjects but a relatively high range of values (i.e. 0.94 to 0.99) has the largest variability in the Fisher transformation space. This may indeed be the reason why both tests proved significant for this subject.

4.3 A Simulation Study

To compare and contrast the 5 distinct methods of point and interval estimation of a common correlation, 1000 simulations of each of a number of underlying configurations were carried out.

The configurations were defined by three quantities:

i) The number of distinct samples/exercise tests, I;
ii) The number of observations per sample, n_i;
iii) The true underlying common correlation, ρ_T.

In the simulations, all combinations of the following values were taken:

i) I=5 or 10;
ii) $n_i = n$ for all i and n=4, 8 or 12;
iii) ρ_T =0.1, 0.4, 0.7 or 0.95.

(i.e. $2 \times 3 \times 4 = 24$ separate configurations were investigated)

4.3.1 Summary of Results of the Simulations

4.3.1.1 Point Estimation

For each method of estimation of a common correlation, the average estimate over the 1000 simulations for each of the above configurations are presented in Table 4.4 and the estimated biases (i.e. the average estimate for 1000 simulations $- \rho_T$) are displayed in Figure 4.6.
Obviously as the number of observations per sample increases, all

methods tend towards minimal bias. Generally there is more bias at moderate ρ (i.e. 0.4 and 0.7) than at 'extreme' $\rho = 0.1$ or 0.95.

The Weighted and Profile Likelihood methods are negatively biased while the Hedges and Olkin as well as the Fisher methods are positively biased.

For small number of observations per sample (i.e. n=4) the Unbiased method is slightly (positively) biased.

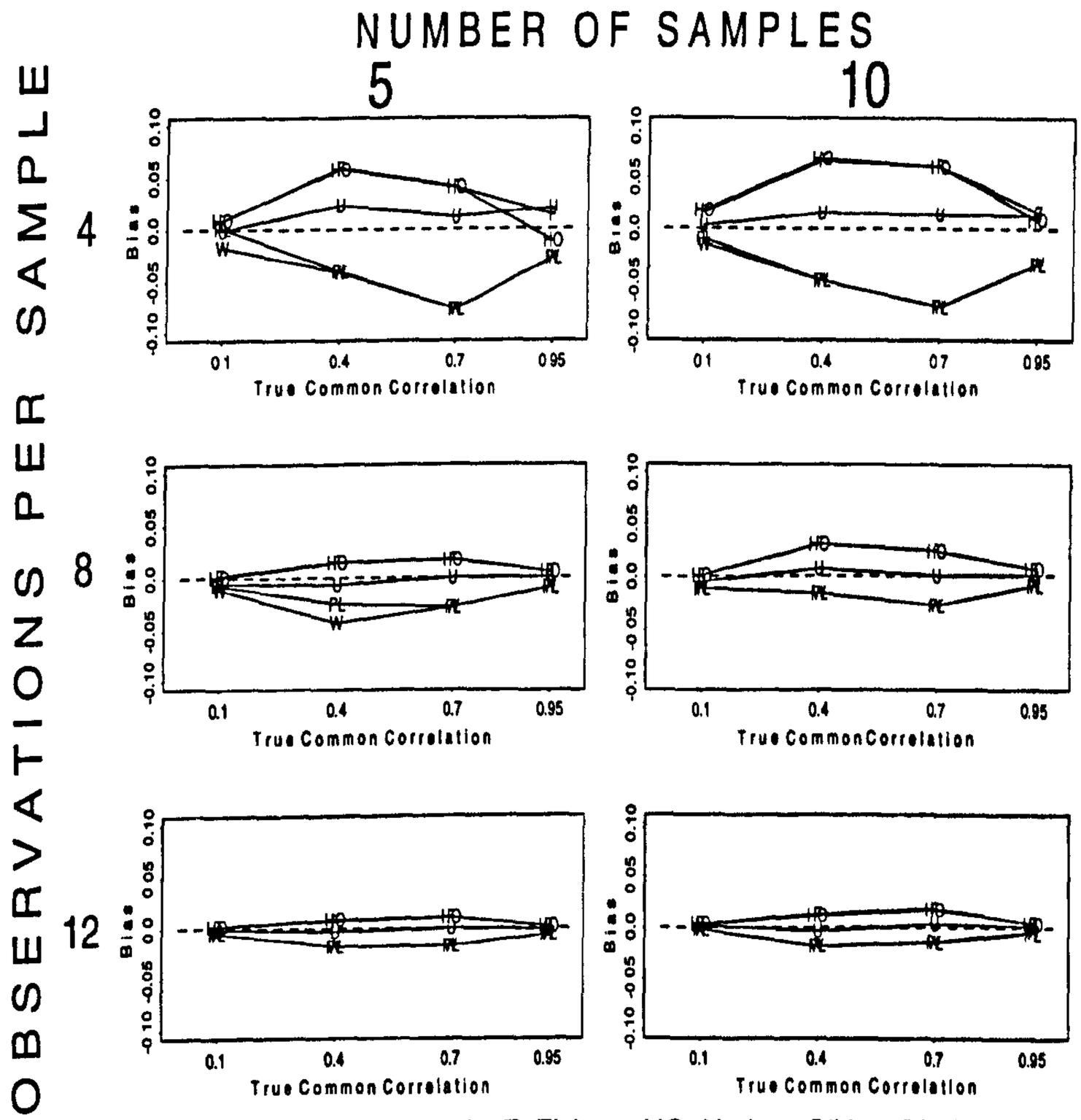

Figure 4.6: Estimates of Bias for each method of common correlation estimation based on the results of 1000 simulations

ρ_T	Method	I=5			I=10		
		n=4	n=8	n=12	n=4	n=8	n=12
0.1	Weighted	0.083	0.089	0.095	0.085	0.089	0.098
	Unbiased	0.099	0.095	0.099	0.103	0.095	0.102
	Fisher	0.108	0.102	0.102	0.116	0.101	0.103
	H.O.	0.109	0.101	0.102	0.118	0.101	0.103
	P.L.	0.101	0.093	0.095	0.090	0.089	0.098
0.4	Weighted	0.359	0.375	0.382	0.350	0.384	0.382
	Unbiased	0.423	0.397	0.397	0.415	0.407	0.396
	Fisher	0.459	0.415	0.407	0.467	0.431	0.412
	H.O.	0.458	0.415	0.408	0.469	0.431	0.411
	P.L.	0.360	0.375	0.382	0.350	0.384	0.382
0.7	Weighted	0.622	0.672	0.683	0.624	0.672	0.686
	Unbiased	0.711	0.700	0.700	0.714	0.700	0.703
	Fisher	0.740	0.717	0.710	0.762	0.724	0.718
	H.O.	0.741	0.717	0.710	0.761	0.724	0.717
	P.L.	0.622	0.672	0.683	0.624	0.672	0.686
0.95	Weighted	0.923	0.941	0.944	0.916	0.941	0.946
	Unbiased	0.970	0.951	0.949	0.963	0.951	0.951
	Fisher	0.963	0.955	0.952	0.966	0.956	0.954
	H.O.	0.938	0.955	0.952	0.959	0.956	0.954
	P.L.	0.922	0.941	0.944	0.915	0.941	0.945

Table 4.4: Average values for each method of common correlation estimation based on the results of 1000 simulations

It appeares that increasing the number of samples from 5 to 10 has no major influence on the pattern or magnitude of such biases.

4.3.1.2 Interval Estimation

To compare the performance of the methods of interval estimation, the estimated coverage rate (i.e. the percentage of times in the 1000 simulations that ρ_T lay inside the alleged 95% confidence intervals) for each of the above configurations was evaluated. These are presented in Table 4.5 with a graphical presentation in Figure 4.7.

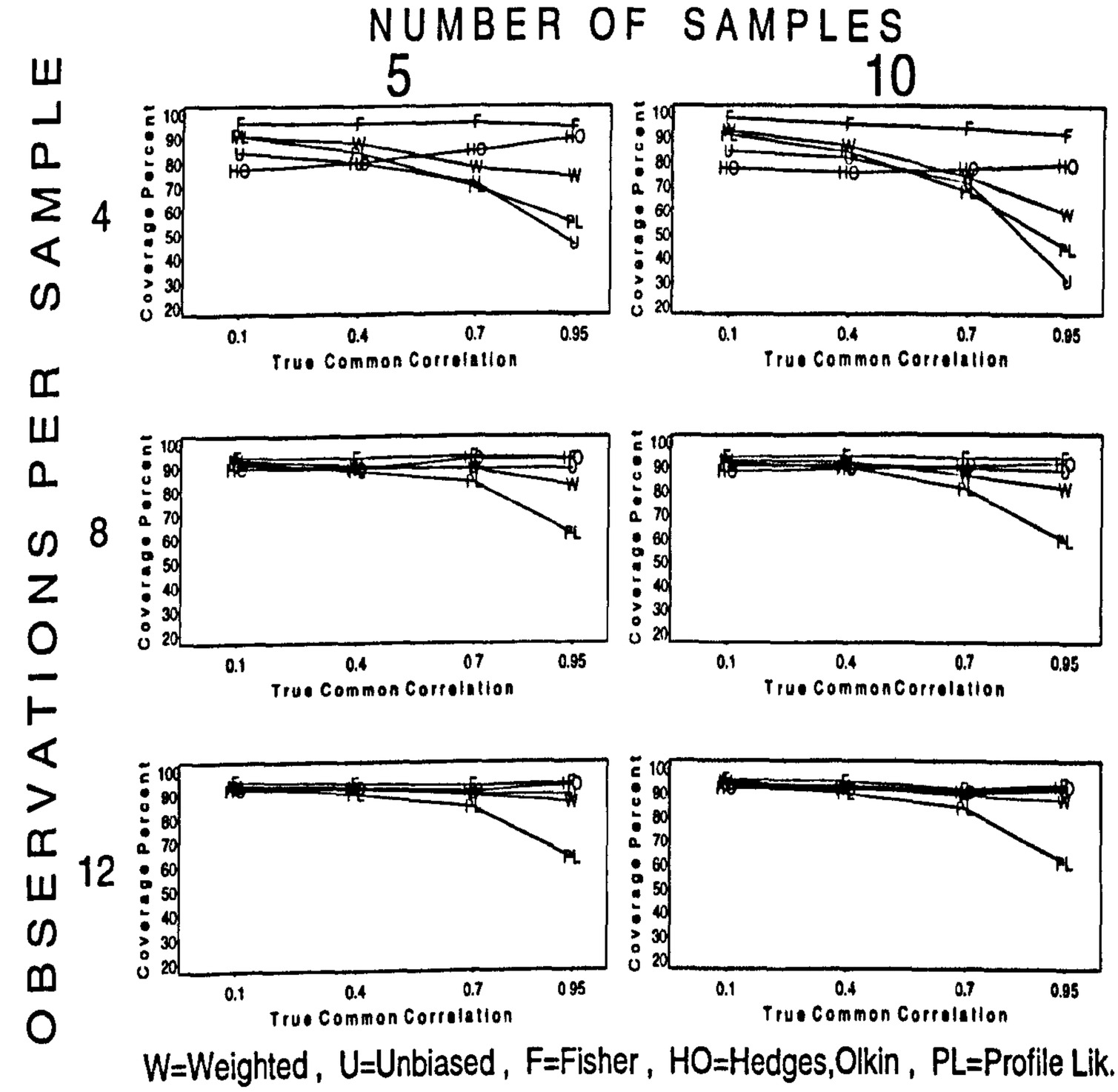

Figure 4.7: Coverage Rates for each of the methods

ρ_T	Method	I=5			I=10		
		n=4	n=8	n=12	n=4	n=8	n=12
0.1	Weighted	91	93	94	93	93	95
	Unbiased	84	91	93	85	90	94
	Fisher	96	94	95	98	94	96
	H.O.	77	89	92	77	88	92
	P.L.	91	92	93	91	92	94
0.4	Weighted	88	91	93	87	92	93
	Unbiased	80	90	92	82	91	93
	Fisher	96	94	95	95	94	95
	H.O.	80	89	93	76	89	92
	P.L.	84	89	90	84	90	90
0.7	Weighted	78	89	90	74	86	89
	Unbiased	72	90	90	72	89	91
	Fisher	96	95	94	94	93	92
	H.O.	85	94	92	77	90	91
	P.L.	70	84	85	68	81	84
0.95	Weighted	74	83	87	59	81	87
	Unbiased	46	90	90	31	88	91
	Fisher	94	94	95	92	93	94
	H.O.	90	94	94	79	91	93
	P.L.	55	62	63	44	59	61

Table 4.5: Proportion of times in 1000 simulations that the interval estimate captured the true common correlation

The Hedges and Olkin method provides poor confidence for low numbers of observations per sample (i.e. $n_i = 4$) and low true common correlations, but this improves as the number of observations per sample and the true value of common correlation increases, while the Profile likelihood method produces higher confidence for low values of true common correlations but decreases as the value of the true common correlation increases.

Poor confidence is obtained using the Unbiased method for low values of observations per sample and high values of true common correlation. This rate dramatically improves with an increase in the number of observations per sample.

Only the Fisher method provides consistent confidence in the range of 95% regardless of the number of subjects, observations per sample and the true value of the common correlation.

All the other four methods do not perform adequately throughout the simulations, particularly for a low number of observations per sample (i.e. $n_i = 4$) as well as for a true common correlation of 0.95.

It seems that increasing the number of samples from 5 to 10 does not radically improve the pattern of coverage rates.

4.3.2 Comparison of Confidence Interval Widths and Coverage Rates in Different Methods

To investigate the differences in the estimated confidences for the underlying simulation configurations, scatterplots of the bias against confidence interval width for each simulation are used. These plots across different simulation configurations are shown in Figures 4.8

(I=5) and 4.9 (I=10). To provide a clearer presentation for the true common correlation of 0.95, the same form of plots with a larger scale is presented in Figure 4.10.

Points inside the wedge (<) shape are interval estimates which capture the true value of the common correlation while points outside fail to do so.

Apparently increasing in the number of samples/exercises as well as increase in the number of observations per sample/exercise decreases length of the estimated confidence intervals. While the Fisher approach provides the most stable coverage rates irrespective of the length of estimated confidence intervals, in the other approaches coverage rates increase as the confidence interval widths decrease.

The Fisher approach obviously provides the widest confidence interval with the best estimated confidence based on coverage rates. In contrast, the Profile Likelihood approach provides the narrowest intervals and in all situations, except in the case of small number of observations per sample (i.e. $n_i = 4$), gives the smallest confidence. The skewness of the intervals as the true common correlation increases is clearly obvious.

Unsurprisingly, the Weighted, Unbiased and Hedges and Olkin approaches, with the same pivotal functions to produce confidence intervals, give the same pattern of confidence intervals. In these three approaches, plots of bias against width overlap each other and so it is impossible to distinguish the performance of each approach.

To have a clearer presentation of confidence interval width for each of these three approaches, Figure 4.11 shows the distributions of the produced confidence interval widths by each of the three mentioned methods for the case of $\rho_T = 0.95$. The Unbiased approach,

obviously, produces the narrowest confidence intervals for the case of 4 observations per sample irrespective of the number of samples. It appears that in other situations these three approaches produce almost the same pattern of confidence interval widths.

4.3.3 Conclusion of the simulations

All five approaches of estimating a common correlation, provide biased point estimates. Generally the bias rate is more at moderate true common correlations (i.e. ρ_T=0.4 or 0.7) and tends to decrease as the number of observations per sample increases.

Regarding the effect of different factors on the attained confidence of the different approaches of interval estimation, it is clear that a combination of factors influence the coverage rates.

To illustrate the influence of different factors on the estimated confidence, the five different approaches to estimate common correaltion can be summarised as follows:

1) The Fisher approach
This approach provides a moderate, positive bias, but has the most stable coverage rates. It seems that number of observations per sample and true common correlation, which affect the width of the intervals, have no significant effect on the achieved confidence.

2) The Hedges and Olkin approach
This approach like the Fisher approach provides positively biased point estimates and it appears that the number of observations per sample is the most important factor influencing the coverage rate. Obviously in the case of a small number of samples with a small

number of observations per sample (i.e. I=5 and n=4), higher true common correlation significantly increases the achieved confidence.

3) The Weighted, the Unbiased and the Profile Likelihood approaches

In these three approaches, the Unbiased approach provides slightly positively biased point estimates and the Weighted and the Profile Likelihood approaches provide negatively biased point estimates. Generally an increase in the true common correlation results in narrower confidence intervals and smaller coverage rates, but an increase in the number of observations per sample (e.g. from 4 to 8) improves the coverage rate. With the Unbiased method, for example, increase in the true common correlation continuously decreases the provided coverage rates with its minimum value at $\rho_T = 0.95$, but the higher number of observations per each sample quickly recovers this rate.

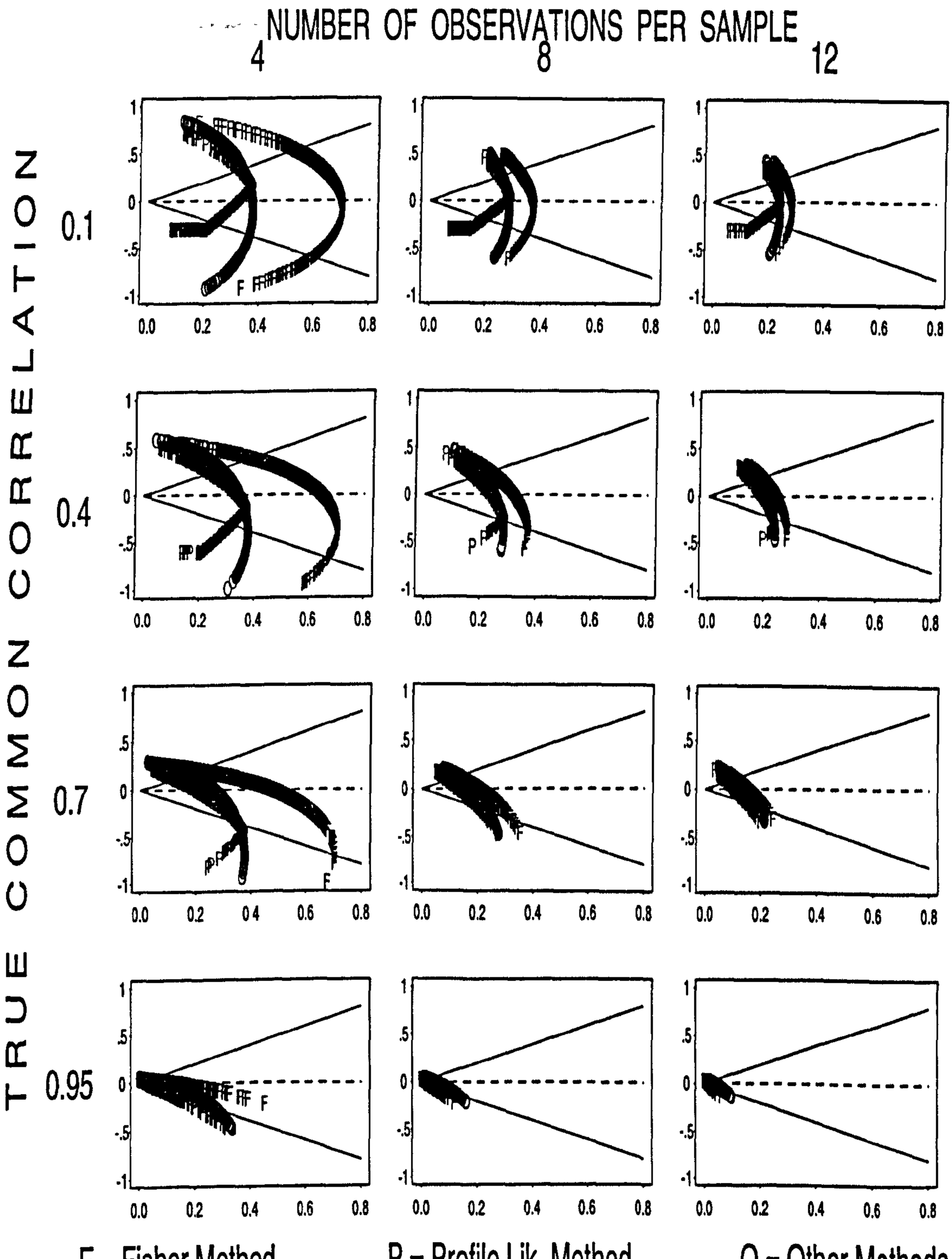

Figure 4.8: Plots of bias against width for the case I=5 samples and for different values of 4,8 and 12 observations per sample.
In each diagram vertical axis represents Bias and horizontal axis represents (Confidence Interval Width)/2.

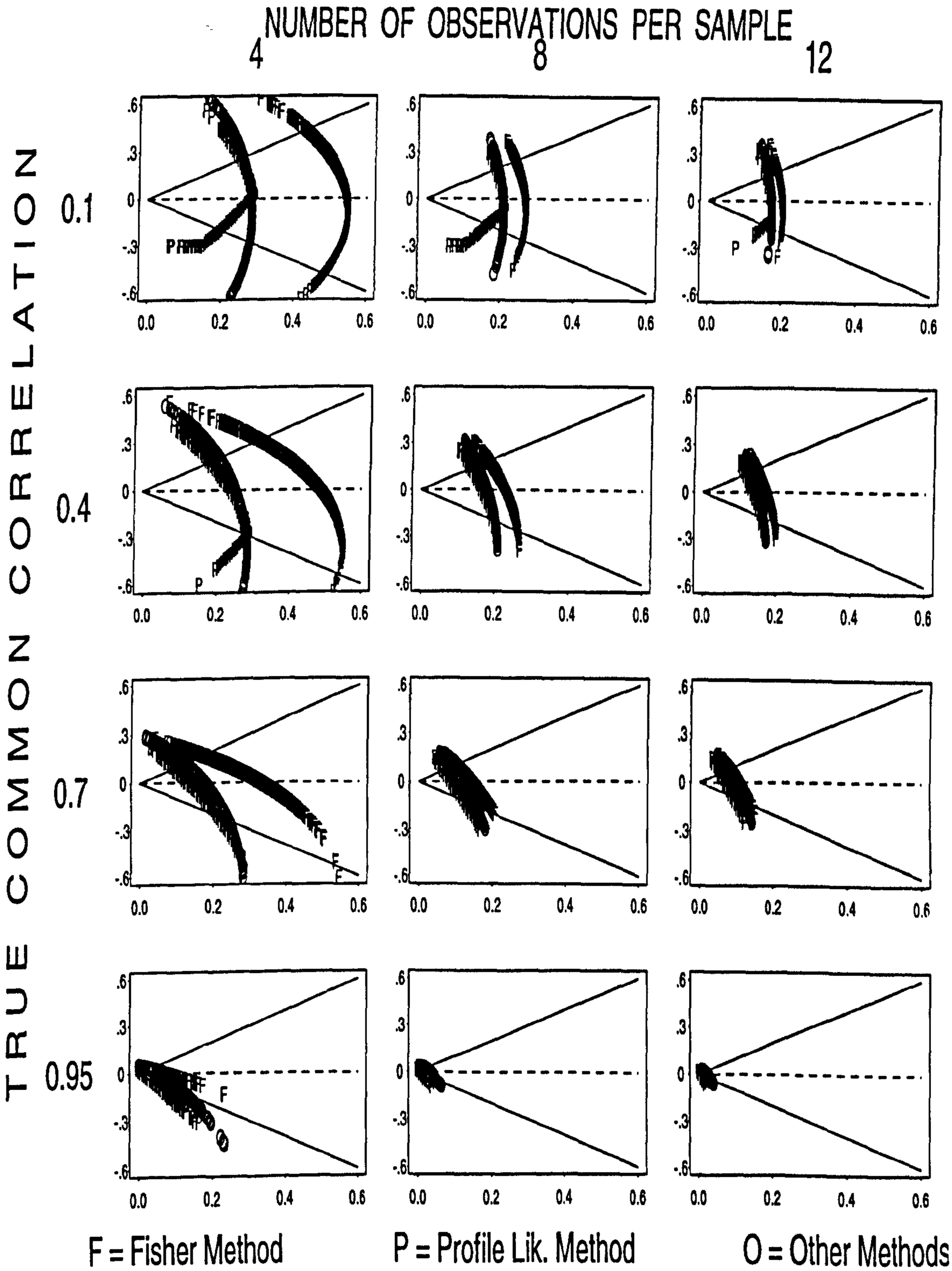

Figure 4.9: Plots of bias against width for the case of I=10 samples and for different values of 4,8 and 12 observations per sample.
In each diagram vertical axis represents Bias and horizontal axis represents (Confidence Interval Width)/2.

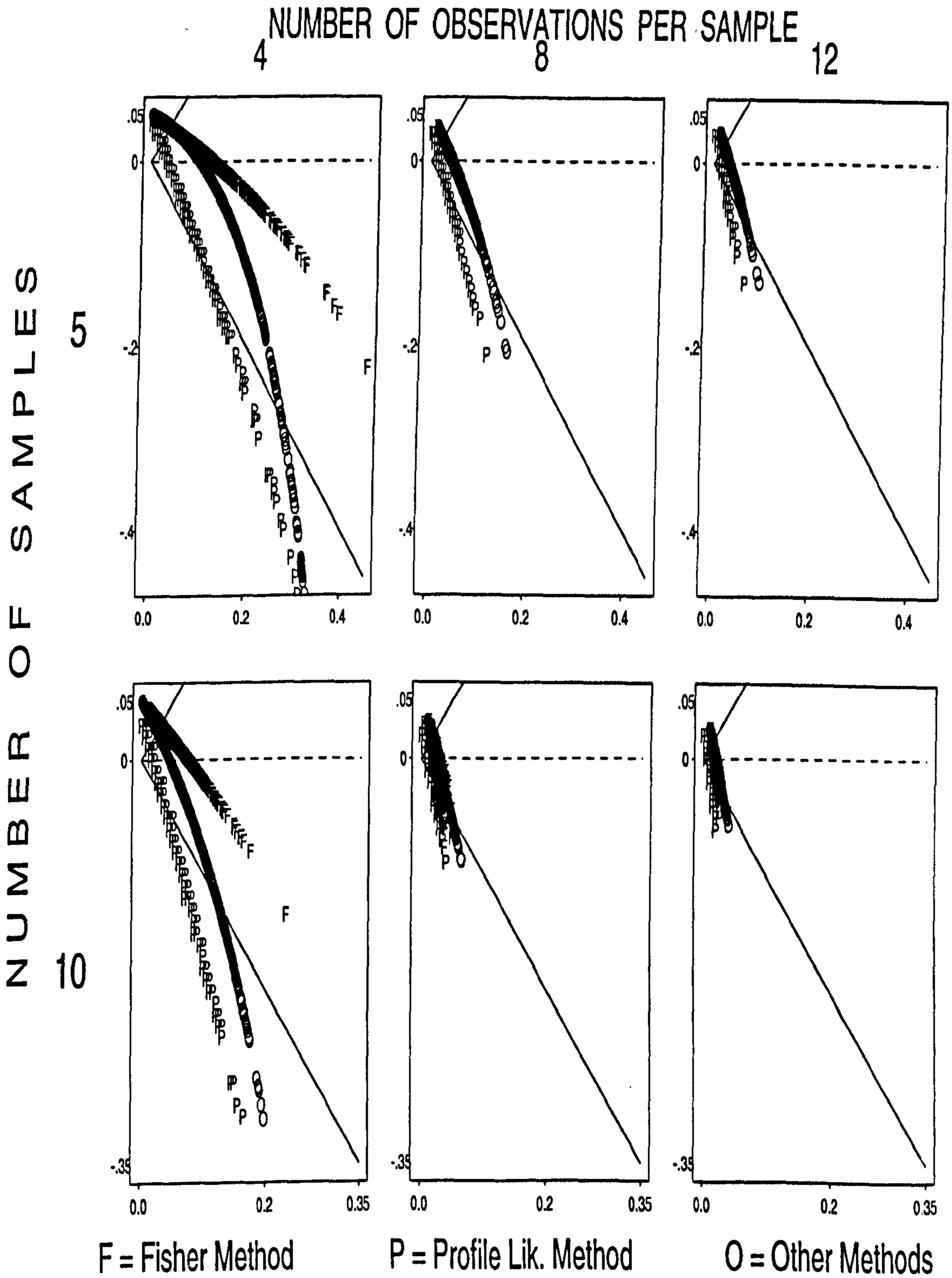

Figure 4.10: Plots of bias against width for the true common correlation of 0.95 and for different values of 4,8 and 12 observations per subject.
In each diagram vertical axis represents Bias and horizontal axis represents (Confidence Interval Width)/2.

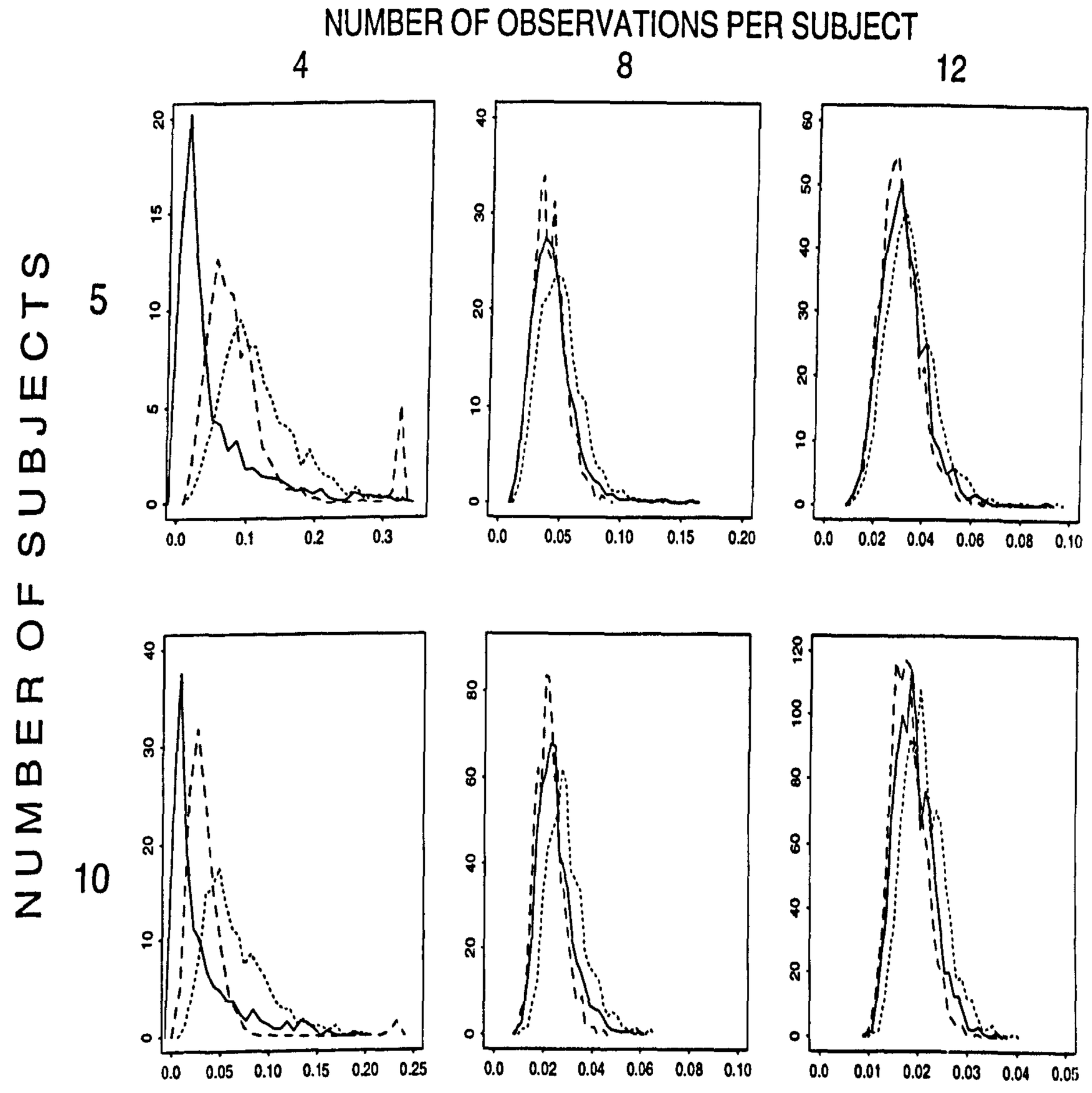

Figure 4.11: Distribution of confidence interval widths produced by the three approaches of Weighted, Unbiased and Hedges & Olkin in the case of high common correlation (i.e. $\rho = 0.95$).
..... : Weighted, —— : Unbiased, - - - - : Hedges and Olkin

4.4 Summary

Five distinct methods of point and interval estimation were considered for the case of estimating a common correlation coefficient.

A particular example suggested that both the Fisher and Hedges and Olkin methods would provide similar results for high (i.e. greater than 0.95) correlations.

A simulation study supported the Fisher method as the only method to achieve near the required 95% confidence for interval estimation across a variety of underlying situations.

Overall the Fisher method appears to be, on the basis of this study, the best method to estimate a common correlation.

Chapter 5

Estimating the Comparability of two distinct Variables: How to model across individuals and repeat Exercise Tests

5.1 Introduction

Essentially, a correlation coefficient measures the linear *relationship* between two variables of interest. The *'Comparability'* between variables refers to combining estimates of simple correlation coefficients from different replicates across individuals into either a pooled estimate of a common correlation if this is an appropriate assumption or into an estimate of the "typical/average" correlation from the variables on a 'typical replicate' exercise test of a 'typical individual'. It is thus latter interpretation of Comparability that is considered in this chapter.

Various methods of pooling estimates of a common correlation coefficient from different samples were considered in the previous chap-

ter. In this chapter it is intended to investigate how to model the correlation between a pair of variables (e.g. physiological and psychological assessments of physical effort) across replicate visits for a number of individuals and then to model these across all individuals in a sample from an appropriate population. First one models how one can estimate the 'typical or average' correlation for an individual (one-stage modelling) and then extend this idea to obtain an overall estimate of the 'typical' correlation of the 'typical' individual (two-stage modelling).

Throughout this chapter the modelling of the data will be in the *'Fisher transformation space'*. For instance if an individual has a true correlation ρ for a specific visit/retest, then the corresponding Fisher value will be

$$F(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho} \tag{5.1}$$

The choice of such a space for modelling in this chapter is clearly dictated by the results in the previous chapter where the Fisher transformation was found to be 'best' with respect to the estimation of a common correlation coefficient.

5.2 One-Stage Modelling Process

Consider an individual with a set of J bivariate samples of size n_j (j=1,2,..., J), which arise from bivariate (X,Y) normal distributions and let F_T and $\hat{F}_j = F(\hat{\rho}_j)$, (j=1,2,...,J), be the transformed (Fisher space) 'average' population correlation coefficient (or Comparability) and sample correlation coefficients between the two variables of interest, respectively. In Exercise Testing, for example, an individual may be tested on each of J different visits where a physiological variable X and a psychological variable Y are measured at

n_j specific time intervals during an exercise test on that visit with resulting observations

$$\{x_{jk}, y_{jk}; k = 1, 2, \ldots, n_j, j = 1, 1, \ldots, J\}$$

The sample correlation for the j^{th} test will be

$$\hat{\rho}_j = \frac{S^j_{xy}}{\sqrt{S^j_{xx} S^j_{yy}}}$$

where

$$S^j_{xy} = \sum_{k=1}^{n_j} (x_{jk} - \overline{x}_{j.})(y_{jk} - \overline{y}_{j.}) \quad \text{etc.}$$

Further denote

$$\hat{F}_j = F(\hat{\rho}_j) = \frac{1}{2} \log \frac{1 + \hat{\rho}_j}{1 - \hat{\rho}_j}$$

This section is concerned with two possible models whereby the $\hat{F}_j$ (j=1,2,...,J) may be used to estimate F_T.

5.2.1 The Basic (Fisher) Model

Since, for each individual, there are J underlying 'true correlations' $\rho_j, (j = 1, 2, \ldots, J)$, it is reasonable to suppose that

> *All the J tests should measure exactly the same true underlying correlation ρ_T.*

In the Fisher space this could give rise to the following models

$$F_j = F_T, \quad \text{for all } j = 1, 2, \ldots, J$$
$$\text{where } F_T = F(\rho_T)$$

Now, of course, what is actually observed is $\hat{F}_j$ which, on the basis of the Fisher transformation, one could assume

$$\hat{F}_j = F_j + \lambda_j \quad \text{for all} \quad j = 1, 2, \ldots, J \tag{5.2}$$

and where it might be further assumed that

$$\lambda_j \sim N\left(0, \frac{1}{n_j - 3}\right) \qquad j = 1, 2, \ldots, J(n_j > 3)$$

with these λ_j assumed independent of each other.
Standard Normal Theory then provides a maximum likelihood estimate

$$\hat{F}_T = \frac{\sum_{j=1}^{J}(n_j - 3)\hat{F}_j}{\sum_{j=1}^{J}(n_j - 3)} \tag{5.3}$$

with

$$Var(\hat{F}_T) = \frac{1}{\sum_{j=1}^{J}(n_j - 3)} \tag{5.4}$$

Further a 95% confidence interval for ρ_T, the assumed common correlation, can be produced in the form

$$\left[F^{-1}\left(\hat{F}_T \pm 1.96 \times \frac{1}{\sqrt{\sum_{j=1}^{J}(n_j - 3)}}\right)\right] \tag{5.5}$$

as long as all $n_j > 3$.

(c.f. section 4.2.3.3)

5.2.2 Multiplicative Fisher Model

A natural extension to the model which is found to be applicable in practice is that the λ_j do not adequately mirror the variability seen in the $\hat{F}_j$.
Accordingly one could hypothesise the model:

$$\hat{F}_j = F_T + \gamma_j \quad \text{for all } j = 1, 2, \ldots, J \tag{5.6}$$

where one assumes that

$$\gamma_j \sim N\left(0,\ \frac{\sigma_T^2}{n_j - 3}\right)$$

independently of other $\gamma_j's$

i.e. $$\hat{F}_j \sim N\left(F_T,\ \frac{\sigma_T^2}{n_j - 3}\right)$$

Thus if $\sigma_T^2 = 1$ the model of the previous section is appropriate whereas if $\sigma_T^2 < 1$ then the standard Fisher model *overestimates* the variability in the $\hat{F}_j$.

Standard Maximum Likelihood Estimators for F_T and σ_T^2 are obtained as

$$\hat{F}_{ML} = \frac{\Sigma_{j=1}^{J}(n_j - 3)\hat{F}_j}{\Sigma_{j=1}^{J}(n_j - 3)} \tag{5.7}$$

and

$$\hat{\sigma}_T^2 = \frac{1}{J}\sum_{j=1}^{J}(n_j - 3)(\hat{F}_j - \hat{F}_{ML})^2. \tag{5.8}$$

Further, it can be shown that

$$Var(\hat{F}_{ML}) = \frac{\hat{\sigma}_T^2}{\Sigma_{j=1}^{J}(n_j - 3)} \tag{5.9}$$

and a 95% confidence interval for ρ_T, the estimated common correlation, is of the form

$$\left[F^{-1}\left(\hat{F}_{ML}\ \pm\ t_{(J-1,0.975)} \times \frac{\hat{\sigma}_T}{\sqrt{\frac{J-1}{J}\Sigma_{j=1}^{J}(n_j - 3)}}\right)\right] \tag{5.10}$$

as long as all $n_j > 3$.

Note: An Alternative (Additive) Fisher Model

Another potential model that might be considered is the model based on additive rather than multiplicative errors

$$\hat{F}_j = F_T + \varepsilon_j + \lambda_j \quad \text{for all } j = 1, 2, \ldots, J \tag{5.11}$$

with

$$\varepsilon_j \sim N(0, \sigma_T^2)$$

independently of

$$\lambda_j \sim N[0, \frac{1}{(n_j - 3)}]$$

i.e.

$$\hat{F}_j \sim N[F_T, \sigma_T^2 + 1/(n_j - 3)] \tag{5.12}$$

where Maximum Likelihood estimators for F_T and σ_T^2 can be produced by equating to zero the partial derivative of log-likelihood equation with respect to F_T and σ_T^2 and simultaneously solving the two equations:

$$\frac{\partial l}{\partial F_T} = \sum_{j=1}^{J} \frac{F_j - F_T}{\sigma_T^2 + \frac{1}{n_j - 3}} = 0 \tag{5.13}$$

and

$$\frac{\partial l}{\partial \sigma_T} = \sum_{j=1}^{J} \frac{1}{\sigma_T^2 + \frac{1}{n_j - 3}} - \sum_{j=1}^{J} \frac{(F_j - F_T)^2}{(\sigma_T^2 + \frac{1}{n_j - 3})^2} = 0 \tag{5.14}$$

to obtain point estimates for F_T and σ_T^2.

However in the remainder of this chapter the Multiplicative Fisher model is used as it is certainly simpler to fit and, on the grounds of practical experience, is a substantially better fit for physiological data than the above Additive model.

5.2.3 A Specific Application

To illustrate the use of the above models, data from 8 separate exercise tests of a specific individual are considered, where his Fatigue

on a Visual Analogue Scale (VAS), and Ventilation using a Douglas Bag, were measured at 6 or 7 or 8, as appropriate [1] , distinct 2-minute intervals during the test (i.e. J=8 while n_j=6 or 7 or 8). Figure 5.1 shows a scatterplot of these data for each of the 8 visits of the individual.

On the basis of this part it seems reasonable to assume that the correlation between Ventilation and VAS for Fatigue across time in any test is the same for all tests for that individual.

Figure 5.2 gives a plot of the estimated sample correlation coefficients for each test for this individual as well as interval estimates for the common correlation coefficients under both the Fisher and Multiplicative Fisher models.

While the two point estimates are obviously the same, the appropriate confidence interval produced by the Multiplicative Fisher model is considerably narrower than that under the Fisher model.

In fact the 95% confidence interval for σ_T^2 is (0.09 , 0.81), so in this example the Multiplicative Fisher model is clearly appropriate since the interval is 'completely' less than one, i.e. $\sigma_T^2 < 1$.

In this example, an approximate 95% confidence interval for the common correlation between two variables, VAS for Fatigue and Ventilation across time during an exercise test, is 0.95 to 0.98.

---

[1] Note that the number of time points was determined by how long the subject managed to run on the treadmill for each of the 8 visits.

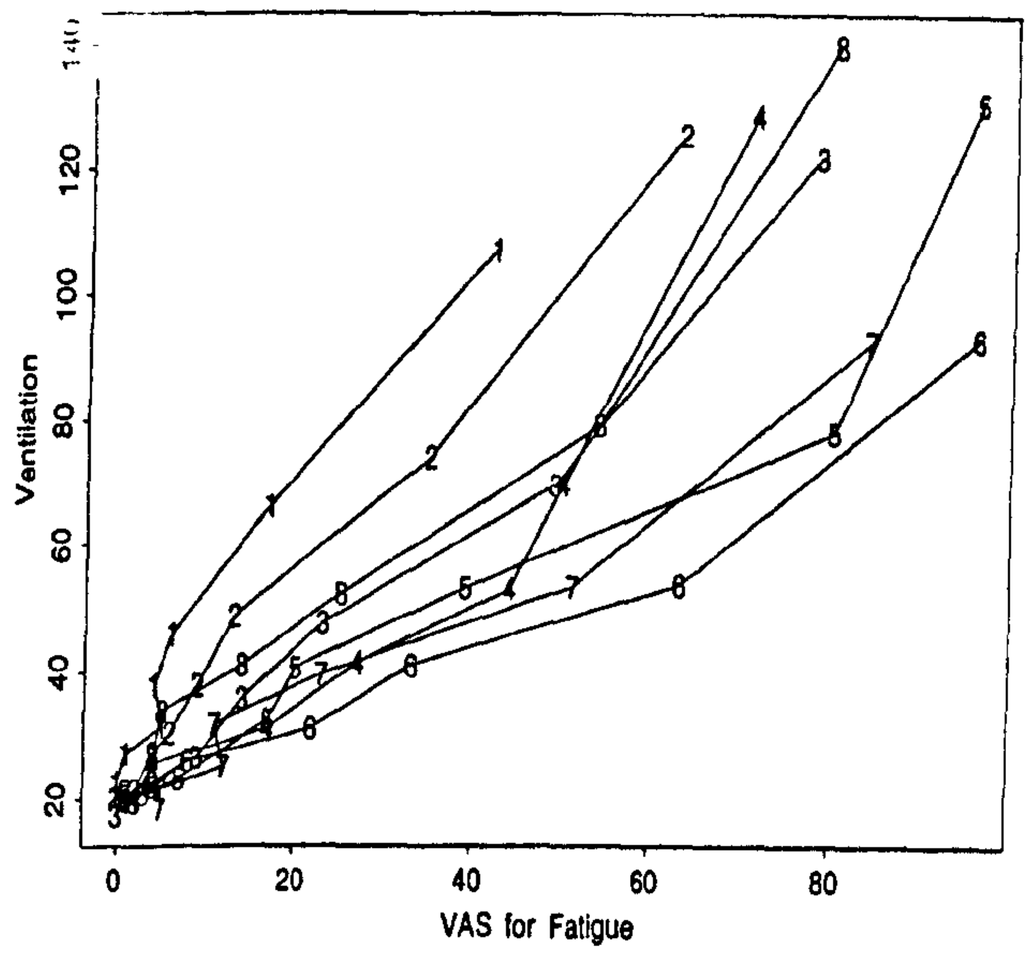

Figure 5.1: Scatterplot of the two variables VAS for Fatigue and Ventilation for each of the 8 visits

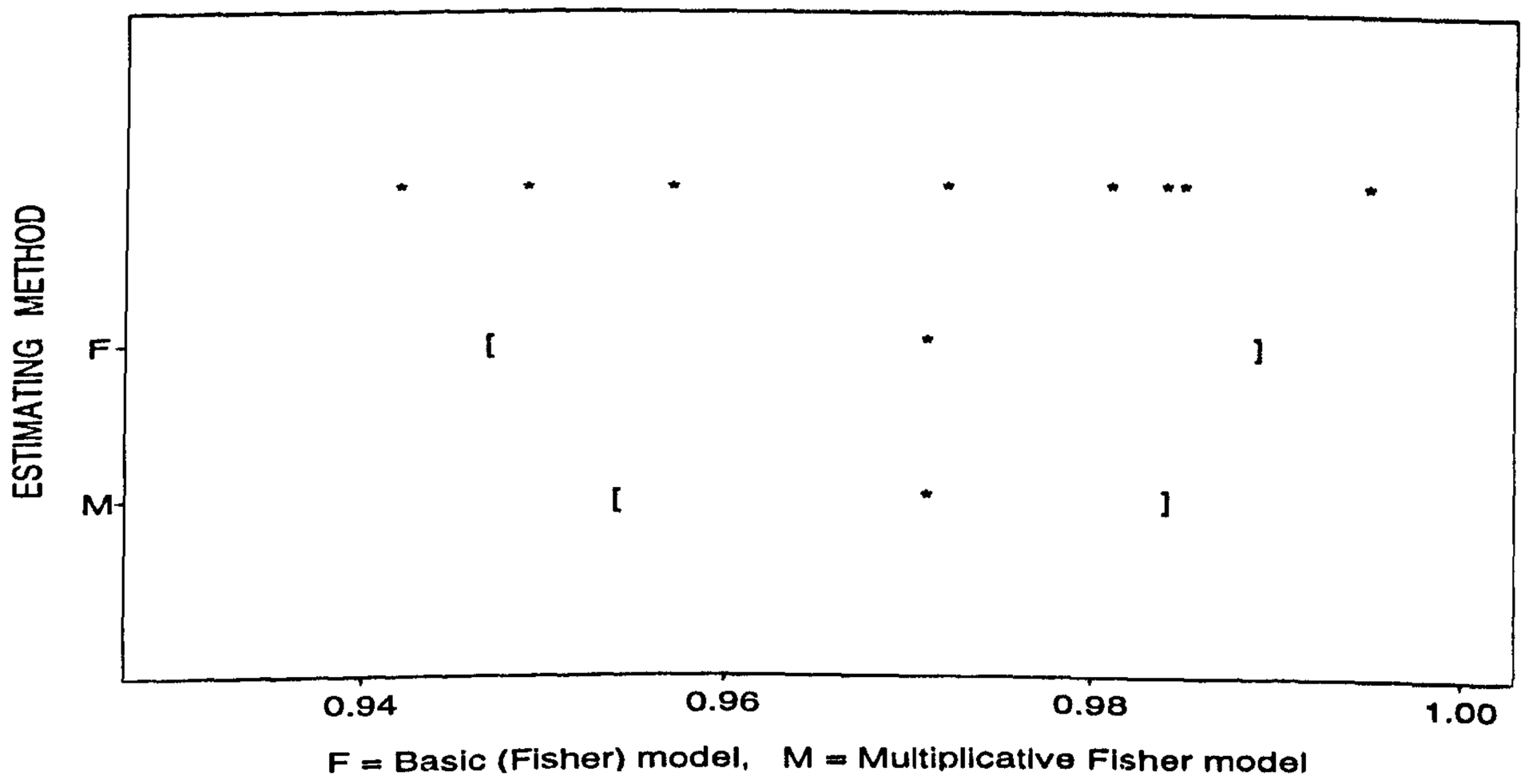

Figure 5.2: Point and interval estimates of the Comparability between VAS for Fatigue and Ventilation from 8 tests on the same individual under the Fisher and Multiplicative Fisher model

5.3 Two-Stage Modelling Process

In this section the case considered is where a sample of I individuals is taken, and each of these provides a set of J bivariate samples of size n_{ij}, which are assumed to arise from bivariate (X, Y) normal distributions with an 'average' correlation, ρ_T, across all visits from all individuals (and corresponding Fisher value of $F_T = F(\rho_T)$).

For example, in Exercise Testing, each of I subjects was tested on each of J different visits, where the two variables, VAS for Breathlessness and VO_2, were measured at n_{ij} different time points into each test at 2 minute intervals (n_{ij} usually 6 or 7 or 8).

Let $\hat{\rho}_{ij}$ be the sample correlation coefficient for the j^{th} sample $[(x_{ijk}, y_{ijk}), k = 1, 2, \ldots, n_{ij}]$ from the i^{th} individual, $(j = 1, 2, \ldots, J,$ $i = 1, 2, \ldots, I)$, and $N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ be the total number of observations.
i.e.

$$\hat{\rho}_{ij} = \frac{S_{xy}^{ij}}{\sqrt{S_{xx}^{ij} S_{yy}^{ij}}}$$

where

$$S_{xy}^{ij} = \sum_{k=1}^{n_{ij}} (x_{ijk} - \overline{x}_{ij.})(y_{ijk} - \overline{y}_{ij.}) \qquad \text{etc.}$$

Further let $\hat{F}_{ij}$ denotes the Fisher transformed sample correlation coefficient on the j^{th} test of the i^{th} individual.
The major aim in this section is to model the relationship between the set of $\hat{F}_{ij}$ for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$ and the true underlying 'population correlation', F_T.

5.3.1 The Potential Models

Since there are I individuals and, for each individual, through n_{ij} time points, J underlying 'true correlations' ρ_{ij}, $(i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J)$, there are three models which might arise naturally when considering data from a Two-Stage context:

i) Basic Model: Each of the true correlation coefficients between the variables of interest across all replicates of all individuals can assumed to be the same with the Fisher model explaining all the variability across the $\hat{\rho}_{ij}$.

ii) Multiplicative Model: As **i)** but with the Fisher model not fully explaining variability across $\hat{\rho}_{ij}$. In fact the true correlation coefficients between the variables of interest are assumed to be the same, on average, for all individuals but there are variations in the actual correlation across different replicates, i.e. there is no systematic difference in the average correlations across individuals.

iii) Components of Variance Model: Two sources of variability of the 'common correlation' may be considered, one of which is the variability *within* each individual across different replicates, i.e. the variability of correlations between replicates on the same individual, and the other is the variability in the 'average' correlation per individual across individuals, i.e. the variability of correlations *between* individuals.

What is actually observed of course are the $\hat{F}_{ij}$ which, based on the Fisher Transformation, could be assumed to be

$$\hat{F}_{ij} = F_{ij} + \lambda_{j(i)} \tag{5.15}$$

where the $\lambda_{j(i)}$ are independently distributed as

$$\lambda_{j(i)} \sim N\left(0,\ \frac{1}{n_{ij}-3}\right) \quad \text{as long as } n_{ij} > 3$$

The next three subsections describe in turn each of the three above models.

5.3.2 Basic Model

Basically all sets of observations $[(x_{ijk}, y_{ijk}), k = 1, 2, \ldots, n_{ij}]$, for i=1,2,...,I, j=1,2,...,J, through different visits and across all individuals are sampled from populations with the same correlation between the two variables, i.e. a constant correlation structure between the two variables. Thus the model can be written

$$\hat{F}_{ij} = F_T + \lambda_{j(i)}, \quad i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J. \tag{5.16}$$

where

$$\lambda_{j(i)} \sim N[0\ , \frac{1}{n_{ij}-3}]$$

and all the $\lambda_{j(i)}$ are independent.

The natural point estimate of F_T would be:

$$\hat{F}_T = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-3)\hat{F}_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-3)} \tag{5.17}$$

as long as $n_{ij} > 3$ for all i, j, with

$$Var(\hat{F}_T) = \frac{1}{N-3IJ}$$

Finally a 95% confidence interval for ρ_T would be of the form:

$$\left[F^{-1}(\ \hat{F}_T\ \pm\ 1.96/\sqrt{N-3IJ}\ )\right]. \tag{5.18}$$

5.3.3 Multiplicative Model

In the case where the assumed variance in the Basic Model does not adequately describe the variability of $\hat{F}_{ij}$, the above model could be extended to

$$\hat{F}_{ij} = F_T + \lambda_{j(i)} \tag{5.19}$$

where it might be assumed that $\lambda_{j(i)}s$ are independently distributed as

$$\lambda_{j(i)} \sim N\left(0\ , \frac{\sigma_T^2}{n_{ij}-3}\right)$$

i.e.

$$\hat{F}_{ij} \sim N\left(F_T,\ \frac{\sigma_T^2}{n_{ij}-3}\right)$$

Maximum Likelihood estimators for F_T and σ_T^2 are

$$\hat{F}_{ML} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-3)\hat{F}_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-3)} \tag{5.20}$$

and

$$\hat{\sigma}_T^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-3)(\hat{F}_{ij}-\hat{F}_{ML})^2}{IJ} \tag{5.21}$$

Further

$$Var(\hat{F}_{ML}) = \frac{\hat{\sigma}_T^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-3)} \tag{5.22}$$

and a 95% confidence interval for ρ_T, the true Comparability, is of the form

$$\left[F^{-1}\left(\hat{F}_{ML}\ \pm t_{(IJ-1\ ,\ 0.975)}\frac{\hat{\sigma}_T}{\sqrt{\frac{IJ-1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij}-3)}}\right)\right] \tag{5.23}$$

5.3.4 Components of Variance Model

A more natural model arises where, apart from the variability mentioned in the previous models, there also exists variability across individuals.

In this case

$$\hat{F}_{ij} = F_T + \alpha_i + \lambda_{j(i)} \tag{5.24}$$

where one can assume that

$$\alpha_i \sim N(0\ , \sigma_B^2), \quad i = 1, 2, \ldots, I$$

independently of

$$\lambda_{j(i)} \sim N(0\ , \frac{\sigma_T^2}{n_{ij} - 3}), \quad j = 1, 2, \ldots, J$$

and all α_i and $\lambda_{j(i)}$ are mutually independent.

i.e.

$$Var(\hat{F}_{ij}) = Var(\alpha_i + \lambda_{j(i)}) = \sigma_B^2 + \frac{\sigma_T^2}{n_{ij} - 3}$$

and

$$Cov(\hat{F}_{ij}, \hat{F}_{ij^*}) = \sigma_B^2$$

Thus the full model here is

$$\hat{F}_{ij} \sim N\left(F_T\ ,\ \sigma_B^2 + \frac{\sigma_T^2}{n_{ij} - 3}\right)$$

Maximum Likelihood estimation of all the three parameters is not possible so a numerical solution has to be sought. Thus an Iterative Generalized Least Squares method may be used to estimate these parameters.

Suppose, for instance, for each individual, say individual i, with J replicates of correlation coefficients, the $J \times J$ matrix V_i contains

the variance of F_{ij} in the diagonal positions and the covariance between two correlation coefficients, σ_B^2, in each of the other positions. Since the covariance between each of two individuals is supposed to be zero, it is possible to form the full covariance matrix V for all observations, which is a 'block diagonal structure' matrix.

i.e.

$$V_i = \begin{pmatrix} \sigma_B^2 + \frac{\sigma_T^2}{n_{ij}-3} & \sigma_B^2 & \dots & \sigma_B^2 \\ \sigma_B^2 & \sigma_B^2 + \frac{\sigma_T^2}{n_{ij}-3} & \dots & \sigma_B^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_B^2 & \sigma_B^2 & \dots & \sigma_B^2 + \frac{\sigma_T^2}{n_{ij}-3} \end{pmatrix}, \quad i = 1, 2, \dots, I$$

and

$$V = \begin{pmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V_I \end{pmatrix}$$

so, applying Generalized Least Squares estimation for known V, an estimator for F_T would be

$$\hat{F}_T = (A^t V^{-1} A)^{-1} A^t V^{-1} \underline{F} \tag{5.25}$$

with

$$Var(\hat{F}_T) = (A^t V^{-1} A)^{-1} \tag{5.26}$$

where

A is a JI design vector of ones

and $\underline{\text{F}}$ is the JI vector of observations (i.e. F_{ij})

When in the above model α_i and $\lambda_{j(i)}$ are assumed to be normally distributed, this estimator gives maximum likelihood estimates (Goldstein, 1987), but the matrix V is unknown so an iterative procedure should be applied. Suppose V_0 is a sensible matrix

close to $\hat{V}$ chosen as a starting point (it may be reasonable to use the results from the multiplicative model as starting vectors, i.e. assuming $\sigma_B^2 = 0$, to form the matrix V_0). Using GLS estimation (equation 5.25), an initial estimator for F_T would be

$$\hat{F}_{T(1)} = (A^t V_0^{-1} A)^{-1} A^t V_0^{-1} \underline{F} \tag{5.27}$$

then the raw residuals are

$$\tilde{F}_{(1)} = \hat{F} - \hat{F}_{T(1)} \tag{5.28}$$

i.e. $\tilde{F}_{(1)}$ is a IJ vector of residuals.

If one forms the cross-product matrix $\tilde{F}_{(1)}\tilde{F}_{(1)}^t$, it is straightforward to show that $E(\tilde{F}_{(1)}\tilde{F}_{(1)}^t) = V$.

In the first stage of the iterative procedure, and after rearranging the two matrixes $\tilde{F_{(1)}}\tilde{F_{(1)}}^t$ and V as vectors, the relationship between these two vectors can be regarded as a linear model with the elements of $\tilde{F_{(1)}}\tilde{F_{(1)}}^t$ as responses and V contains two explanatory variables with coefficients σ_B^2 and σ_T^2 to be estimated. Again, one may use Generalized Least Squares to estimate them.

With these two estimates, say $\sigma_{B(1)}^2$ and $\sigma_{T(1)}^2$, a new estimate of the matrix V is obtainable that may again be applied in (5.25) to obtain a 'new' estimate of F_T.

The same procedure should be continued for sufficient times until estimates from one iteration to the next do not change, i.e. for iteration l, (l=1,2,...) $\sigma_{B(l+1)}^2 = \sigma_{B(l)}^2$, $\sigma_{T(l+1)}^2 = \sigma_{T(l)}^2$ and $\hat{F}_{T(l+1)} = \hat{F}_{T(l)}$ to some required degree of accuracy.

5.3.4.1 Interval Estimation of Comparability

Remembering that $(A^tV^{-1}A)^{-1}$ is the variance of $\hat{F}_T$ and letting $\hat{F}_T$ and $\hat{V}$ be the final estimates of F_T and V from the iterative procedure, $\{(F_T - \hat{F}_T)^2(A^tV^{-1}A)^{-1}\}$ has approximately a χ^2 distribution with 1 degree of freedom (Goldstein, 1987). So an approximate 95% confidence interval for F_T is obtainable by equating

$$(F_T - \hat{F}_T)^2(A^t\hat{V}^{-1}A)^{-1}$$

to the 95% upper tail of a Chi-squared distribution with 1 degree of freedom, that is

$$\hat{F}_T \pm \left[(A^t\hat{V}^{-1}A)^{-1}\chi^2_{1,.95}\right]^{1/2}$$

i.e.

$$\hat{F}_T \pm 1.96\sqrt{(A^t\hat{V}^{-1}A)^{-1}}$$

and then an approximate 95% confidence interval for ρ_T would be of the form

$$\left[F^{-1}\left(\hat{F}_T \pm 1.96\sqrt{(A^t\hat{V}^{-1}A)^{-1}}\right)\right]. \tag{5.29}$$

5.3.5 A specific application

This procedure is illustrated using data from a study where each of 12 individuals underwent 8 separate exercise tests, (*i.e.* $J = 8$), with for each individual Breathlessness on a Visual Analogue Scale(VAS), and VO_2 using a Douglas Bag, being measured at 6 or 7 or 8, as appropriate, distinct 2-minute intervals during the test (i.e. at 2,4,6,... or 16 minutes, k=1,2,...,n_{ij} and j=1,2,...,8). Scatterplots of these values for each of the 12 subjects are presented in Figure 5.3.

Sample correlation coefficients between these two variables for each of the 12 subjects across different visits are given in Table 5.1 with a graphical presentation in Figure 5.4.

Figure 5.5 shows the estimated point estimation for overall correlation coefficient and approximate 95% confidence intervals for each of the three models.

While the point estimates of F_T from the Fisher and the Multiplicative models are obviously the same, that from the Components of Variance model is slightly bigger. The confidence interval provided by the Multiplicative model is slightly narrower than that provided by the Fisher model while that provided by the Components of Variance model is considerably wider than either of those obtained using the other two models.

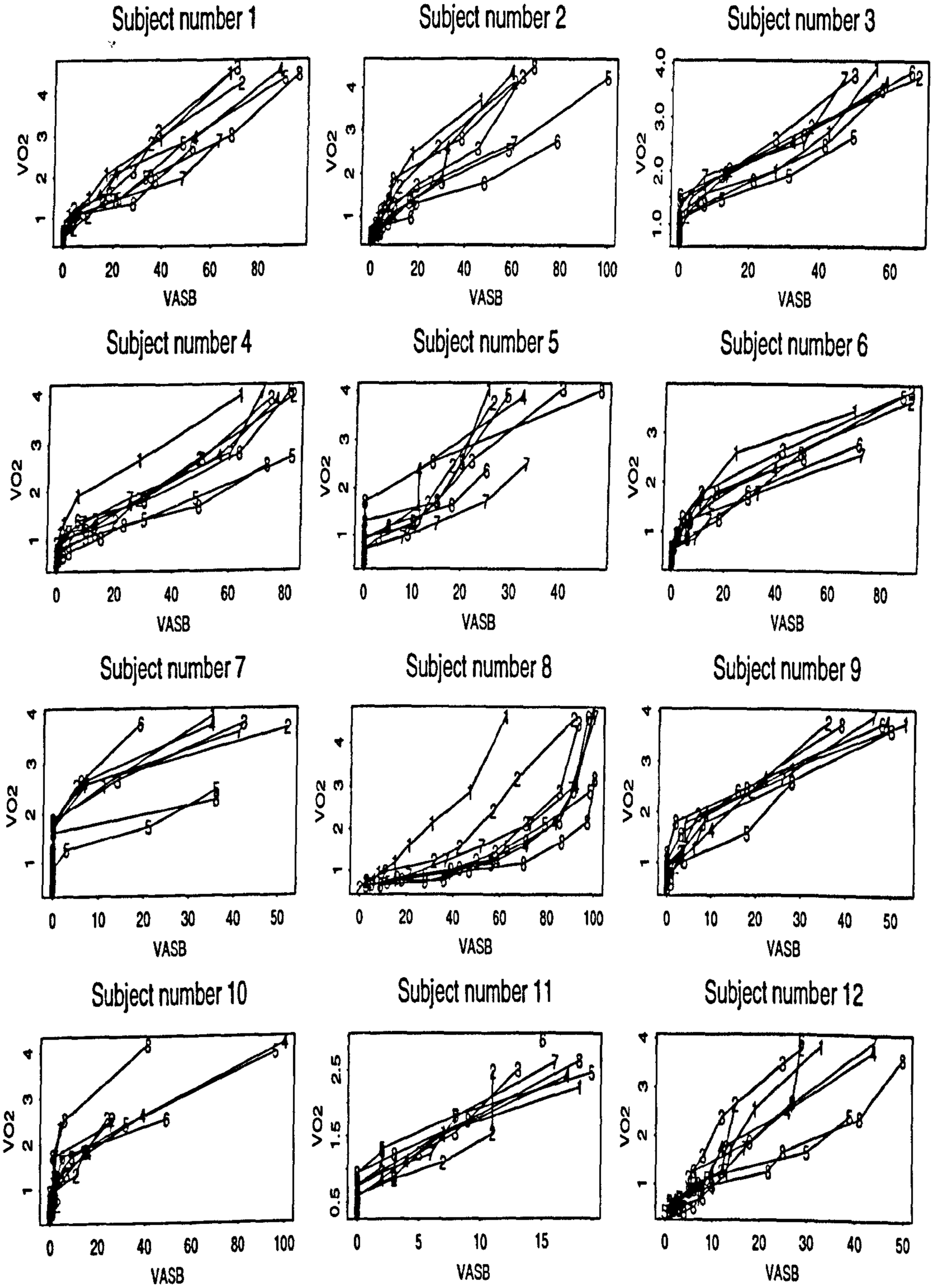

Figure 5.3: Scatterplot of the two variables VAS for Breathlessness and VO_2 for each of 12 subjects

Subject number	*1st visit* (n_{ij})	*2nd visit* (n_{ij})	*3rd visit* (n_{ij})	*4th visit* (n_{ij})	*5th visit* (n_{ij})	*6th visit* (n_{ij})	*7th visit* (n_{ij})	*8th visit* (n_{ij})
1	0.966 (8)	0.995 (8)	0.966 (8)	0.965 (8)	0.990 (8)	0.971 (7)	0.977 (7)	0.986 (8)
2	0.974 (8)	0.993 (8)	0.985 (8)	0.972 (8)	0.982 (8)	0.984 (7)	0.985 (7)	0.948 (8)
3	0.957 (7)	0.980 (7)	0.983 (7)	0.990 (7)	0.990 (6)	0.977 (7)	0.953 (7)	0.982 (7)
4	0.896 (8)	0.982 (8)	0.987 (8)	0.975 (8)	0.993 (7)	0.964 (8)	0.966 (8)	0.983 (8)
5	0.912 (8)	0.931 (8)	0.984 (8)	0.913 (8)	0.962 (8)	0.968 (7)	0.971 (7)	0.835 (8)
6	0.981 (8)	0.971 (8)	0.925 (8)	0.929 (8)	0.970 (8)	0.990 (7)	0.984 (7)	0.945 (7)
7	0.869 (8)	0.858 (8)	0.837 (8)	0.858 (8)	0.967 (7)	0.947 (8)	0.863 (8)	0.806 (7)
8	0.996 (8)	0.954 (8)	0.928 (8)	0.907 (8)	0.946 (7)	0.942 (8)	0.977 (7)	0.848 (7)
9	0.979 (8)	0.929 (8)	0.957 (8)	0.982 (8)	0.958 (7)	0.916 (8)	0.911 (8)	0.891 (8)
10	0.950 (7)	0.981 (7)	0.990 (7)	0.967 (8)	0.927 (8)	0.903 (7)	0.864 (7)	0.913 (8)
11	0.973 (7)	0.882 (7)	0.983 (7)	0.986 (7)	0.958 (7)	0.964 (6)	0.984 (7)	0.973 (7)
12	0.936 (8)	0.921 (8)	0.965 (8)	0.979 (8)	0.974 (7)	0.970 (8)	0.991 (8)	0.979 (8)

Table 5.1: Simple correlation coefficients between two variables VASB and VO_2 and in brackets the number of observations per visit for each of the 8 visits

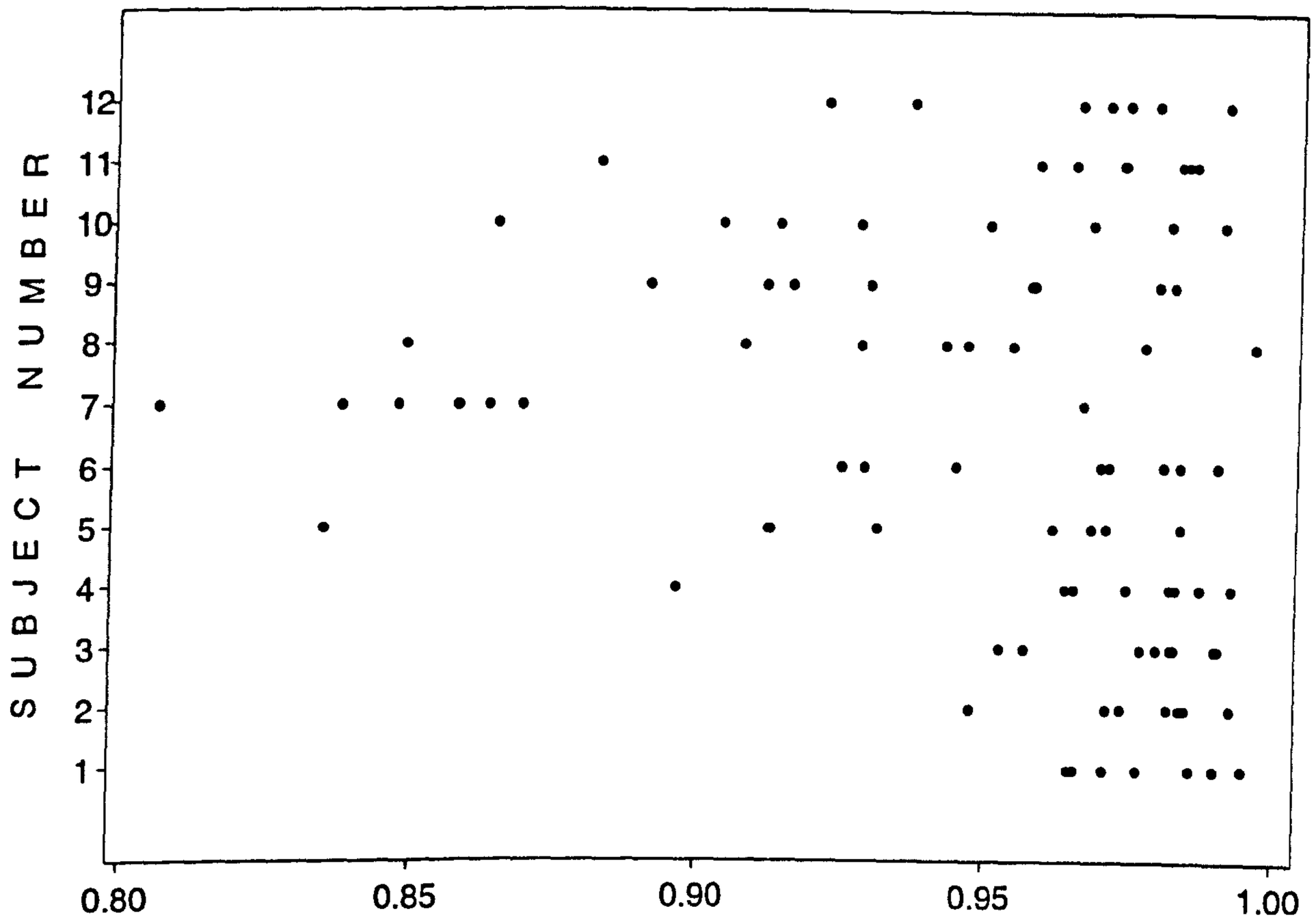

Figure 5.4: Sample correlation coefficient between VAS for Breathlessness and VO_2 for 8 tests on each of 12 individuals

In fact, in this example, an approximate 95% confidence interval for σ_B^2 is [0.01 , 0.20] and for σ_T^2 is [0.50 , 0.82]. The former result suggesting that the between individual variance is significantly greater than zero (but perhaps only marginally). The latter interval clearly demonstrates that $\sigma_T^2 < 1$ and so the basic Fisher model is again inadequate to describe this data. So, in this example the Components of Variance is clearly the most appropriate model since σ_B^2 is 'apparently' greater than zero and σ_T^2 is 'significantly' less than 1.

In this example, an approximate 95% confidence interval for the Comparability (i.e. 'average' correlation, F_T) between two variables

VAS for Breathlessness and VO_2 across time during an exercise test and over the sample of 12 individuals is 0.951 to 0.976.

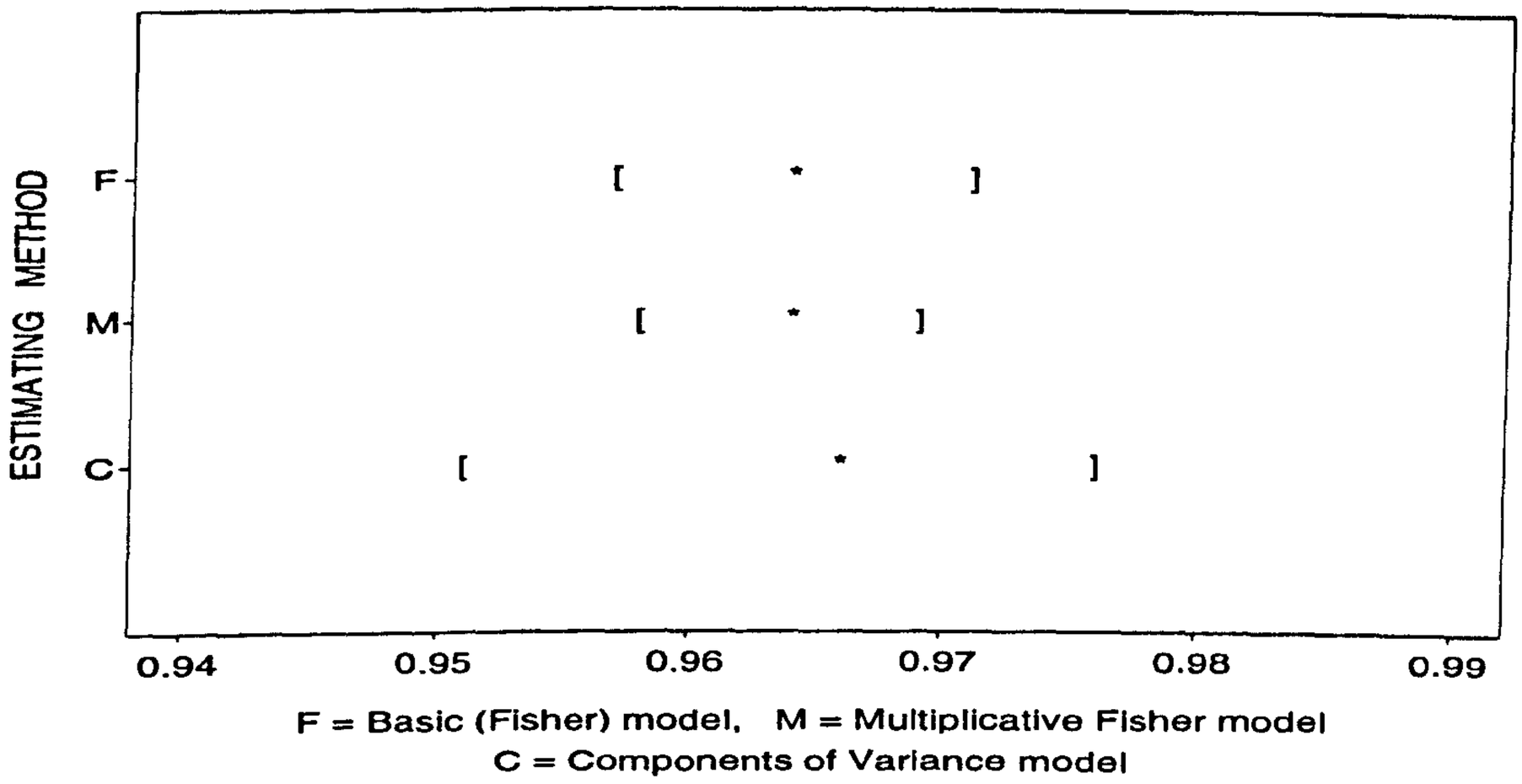

Figure 5.5: Point and interval estimates of the Comparability between VAS for Breathlessness and VO_2 from 8 tests on 12 individuals under three different models

5.4 A Simulation Study

To compare and contrast the different methods of modelling a population common correlation in both the One-Stage and Two-Stage processes, a simulation study was carried over a number of underlying configurations. For each configuration, 1000 simulations were undertaken.

5.4.1 One-Stage Modelling Simulation

In One-Stage modelling the configurations were defined by three quantities:

i) The number of distinct samples/exercise tests, J;
ii) The number of observations per sample, n_j;
iii) The true underlying Comparability, ρ_T.

In the simulations all combinations of the following values were taken:

i) J=4 or 8 or 12;
ii) $n_j = n$ for all j and n=6, 10 or 15;
iii) $\rho_T = 0.4$, 0.7 or 0.95.

i.e. $3 \times 3 \times 3 = 27$ separate configurations

5.4.1.1 Summary of Results of the Simulations in One-Stage Modelling

The mean of the estimated correlations over 1000 simulations as well as the coverage rate (the number of times in 1000 simulations that

the interval estimate captures the true Comparability) for different values of $\sigma_T^2 = 1$, $\sigma_T^2 = 0.5$ or $\sigma_T^2 = 0.1$ are reported in Tables 5.2 and 5.3. A graphical representation of estimated biases (i.e. the mean estimate for 1000 simulations $- \rho_T$), coverage rates and also average confidence interval widths are shown in Figures 5.6 to 5.8, respectively.

Both models produce identical and positively biased point estimates of Comparability with a sharp decrease in the bias as the number of observations per sample increases from 6 to 15. The decrease in the bias is more obvious for lower and intermediate true values of Comparability. Generally, in the case of a smaller true Comparability, the bias is higher and it decreases as the true Comparability increases from 0.4 to 0.95. Furthermore, an increase in the number of samples tests from 4 to 12 slightly decreases the bias.

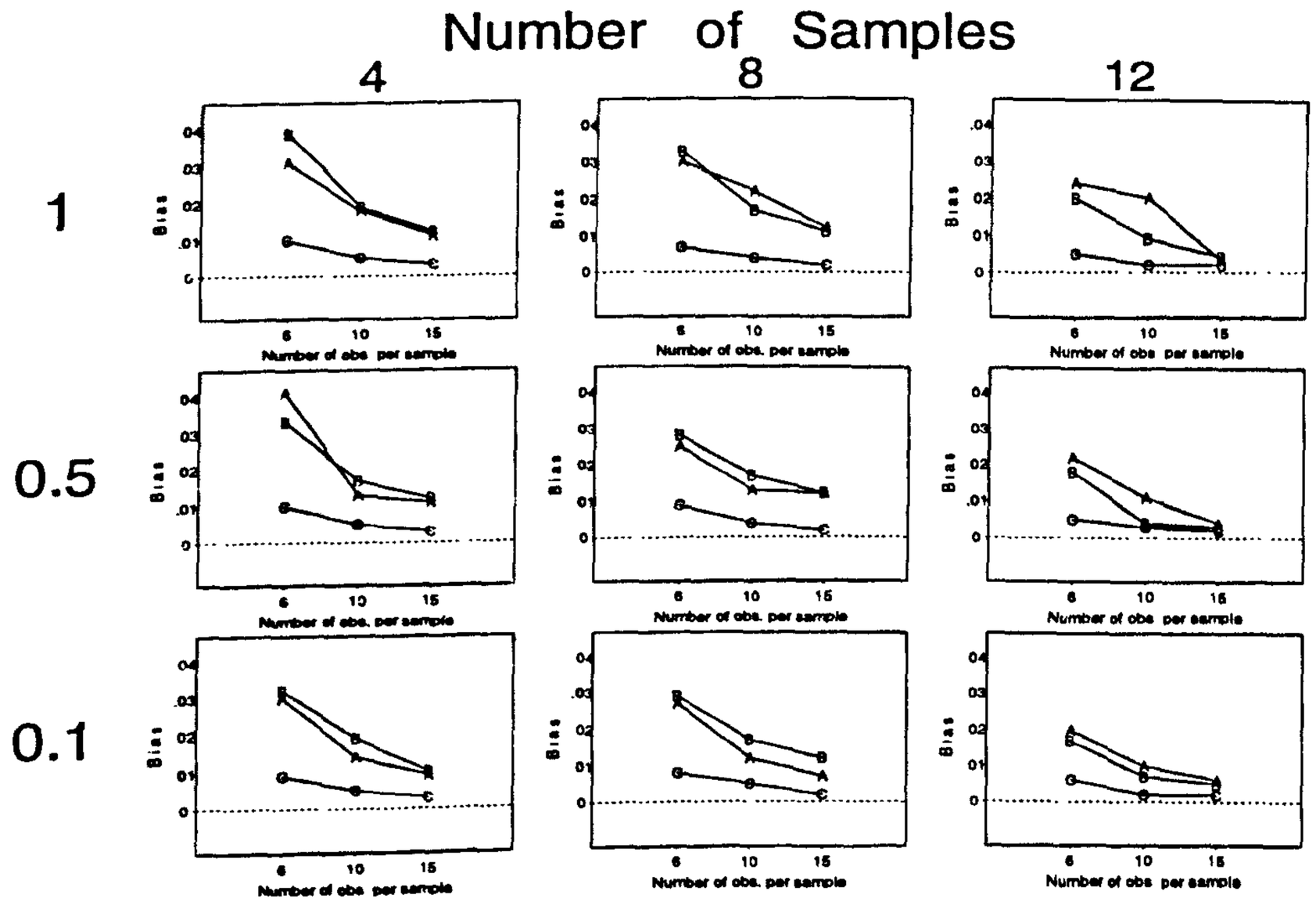

Figure 5.6: Plots of Biases with respect to different number of distinct samples, observations per samples and σ_T^2.

σ_T^2	*Obs. per Sample*	ρ_T								
		0.4			**0.7**			**0.95**		
		No. of samples, J			*No. of samples, J*			*No. of samples, J*		
		4	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	**6**	0.43	0.43	0.42	0.74	0.73	0.72	0.96	0.96	0.96
	10	0.42	0.42	0.42	0.72	0.72	0.71	0.95	0.95	0.95
	15	0.41	0.41	0.40	0.71	0.71	0.70	0.95	0.95	0.95
0.5	**6**	0.44	0.42	0.42	0.73	0.73	0.72	0.96	0.96	0.95
	10	0.41	0.41	0.41	0.72	0.72	0.70	0.95	0.95	0.95
	15	0.41	0.41	0.40	0.71	0.71	0.70	0.95	0.95	0.95
0.1	**6**	0.43	0.43	0.42	0.74	0.73	0.72	0.96	0.96	0.96
	10	0.41	0.41	0.41	0.72	0.71	0.71	0.95	0.95	0.95
	15	0.41	0.41	0.41	0.71	0.71	0.71	0.95	0.95	0.95

Table 5.2: Mean of the estimated Comparability across 1000 simulations for the One-Stage model with different numbers of samples and observations per sample.

The coverage rates which are used as an index of measuring performance of the models in reaching the desired confidence level, show that when $\sigma_T^2 = 1$ both the Basic Fisher and the Multiplicative models provide consistent confidence in the range of 95% although increase in the number of samples slightly reduces this rate. In contrast, when $\sigma_T^2 < 1$ coverage rates differ with respect to the size of σ_T^2 and the number of samples. Decrease in the size of σ_T^2 from 0.5 to 0.1 appears to reduce the coverage rate. In this case the coverage rates provided by the Multiplicative Fisher model are generally higher than those provided by the Fisher model. It seems that the number of samples inversely influences the coverage rate whilst the number of observations per sample does not have a considerable effect on it.

Figures of average confidence interval widths show that, in general, an increase in the number of samples or observations per sample significantly reduces the average confidence interval widths in all cases and especially for smaller true Comparability.

Regarding the effect of σ_T^2 on the confidence , when $\sigma_T^2 = 1$ the intervals on average are slightly wider for the Multiplicative model, whereas for $\sigma_T^2 < 1$, the Multiplicative model, especially for smaller σ_T^2 (i.e. $\sigma_T^2 = .1$), provides considerably narrower intervals.

Comparing coverage rates and average confidence interval widths in Figures 5.7 and 5.8, it is clear that when $\sigma_T^2 = 1$, the Multiplicative model, has provided wider intervals on average but with a more stable confidence level. While, for $\sigma_T^2 < 1$ significantly higher coverage rates were produced by the Multiplicative model with considerably narrower confidence intervals. This shows that capturing the true Comparability may not always be related to the width of confidence interval!

Overall, the Multiplicative model seems to perform better with respect to narrower interval on average and a 'more consistent' confidence level across the configurations covered in this simulation study.

σ_T^2	Obs. per Sample	Model M	ρ_T = 0.4			ρ_T = 0.7			ρ_T = 0.95		
			No. of samples, J			No. of samples, J			No. of samples, J		
			4	8	12	4	8	12	4	8	12
1	6	F	96	96	95	96	96	95	96	96	95
		M	98	97	94	98	97	95	97	96	94
	10	F	96	96	95	96	96	95	97	97	95
		M	98	96	95	97	96	94	97	96	95
	15	F	95	94	96	97	97	96	97	96	96
		M	97	94	94	97	96	94	96	94	94
0.5	6	F	92	87	89	95	86	87	91	85	84
		M	95	92	91	96	91	87	96	87	86
	10	F	93	88	90	95	88	90	92	86	87
		M	96	92	91	96	92	90	97	91	88
	15	F	93	87	86	94	86	83	94	87	86
		M	96	95	95	96	94	94	96	93	94
0.1	6	F	83	81	79	84	81	77	82	78	74
		M	89	86	82	89	84	79	86	81	76
	10	F	82	81	80	83	81	78	82	79	78
		M	87	86	82	90	85	79	89	83	80
	15	F	83	83	79	82	82	77	82	80	77
		M	89	88	83	85	85	82	87	86	78

Table 5.3: Percentage of cases, over 1000 simulations, where the estimated confidence interval captures the true Comparability based on different number of samples and observations per sample. In this table:

F = Fisher model

M = Multiplicative Fisher model.

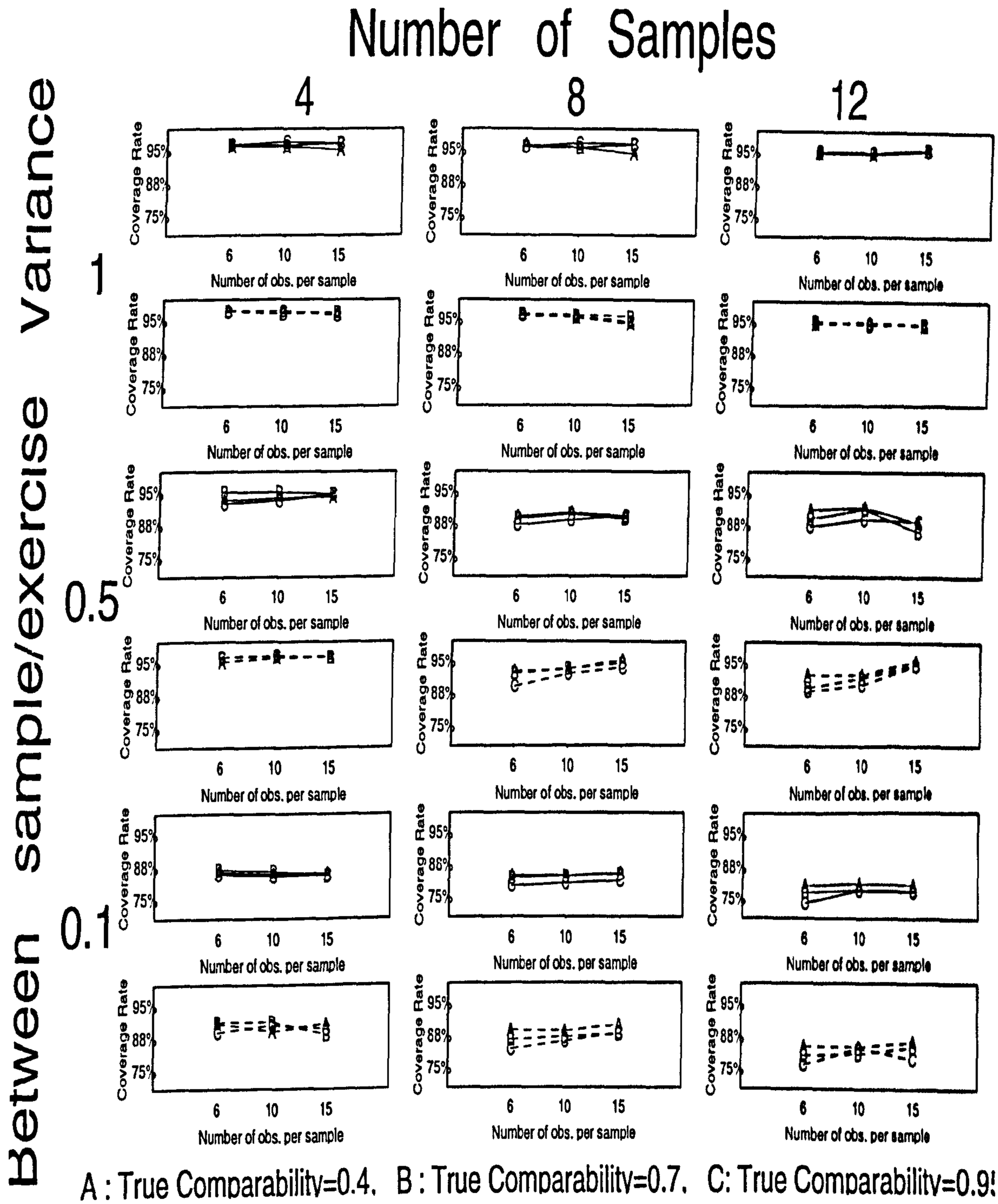

Figure 5.7: Plots of coverage rates for both Basic Fisher model and Multiplicative Fisher model with respect to different number of distinct samples, observations per samples and σ_T^2.
——— : Basic Fisher model,
..... : Multiplicative Fisher model

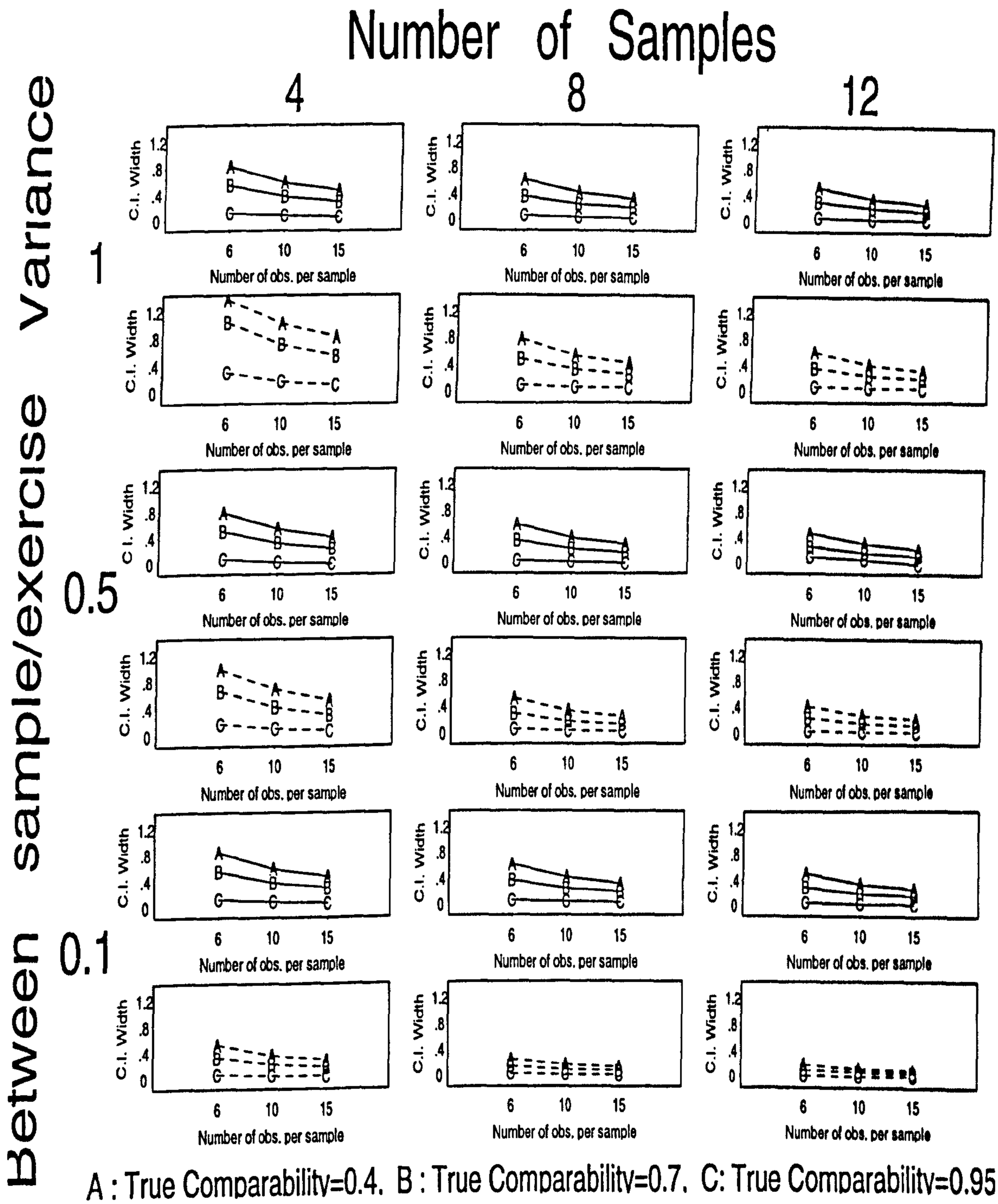

Figure 5.8: Plots of average confidence interval widths with respect to different number of distinct samples, observations per samples and σ_T^2 for both Basic Fisher model and Multiplicative model

——— : Basic Fisher model

- - - - : Multiplicative model

5.4.2 Two-Stage Modelling Simulation

In Two-stage modelling the underlying configurations were based on the following four quantities:

i) The number of individuals into exercise testing, I;
ii) The number of distinct samples/exercise tests per individual, J;
iii) The number of observations per sample, n_{ij};
iv) The true underlying Comparability, ρ_T.

with combinations of all of the following values:

i) I=6 or 15;
ii) J=4 or 8;
iii) $n_{ij} = n$ for all j and n=6 or 12;
iv) $\rho_T = 0.4$, 0.7 or 0.95.

i.e. $2 \times 2 \times 2 \times 3 = 24$ separate configurations

5.4.2.1 Summary of Results of the Simulations in Two-Stage Modelling

For each of the three methods of modelling Comparability, the average point estimates of the true Comparability, ρ_T, coverage rates and average interval estimate widths were produced over 1000 simulations. These are reported in Tables 5.4 and 5.5. Graphical presentation of coverage rates as well as average biases and average interval estimate widths are provided in Figures 5.9 to 5.11, respectively.

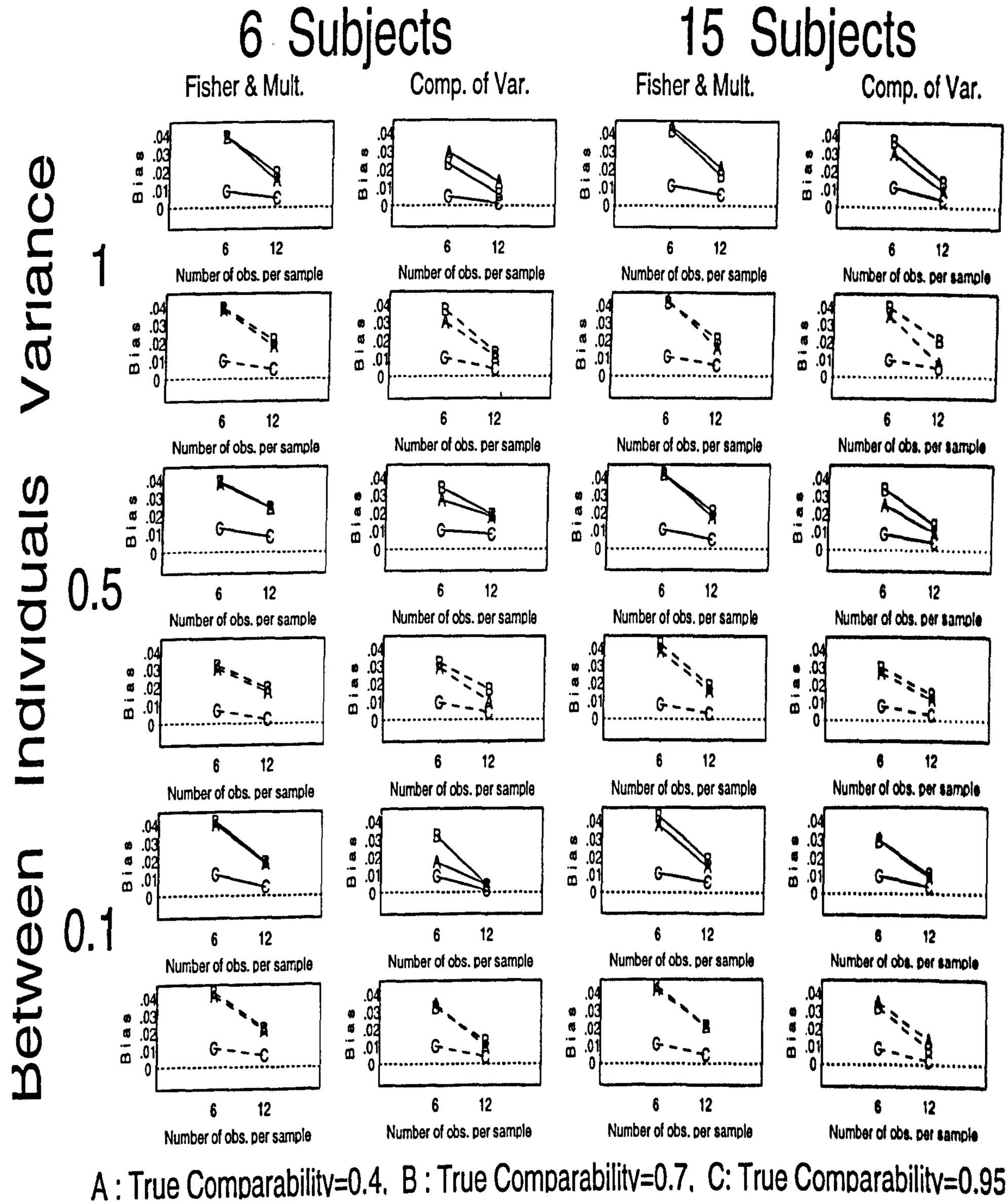

Figure 5.9: Plots of Biases for each of the three models with respect to different number of subjects, distinct samples, observations per samples and σ_τ^2.
——— : 4 distinct samples,
- - - - : 8 distinct samples

σ_T^2	*Number of Subjects*	*Sample Per Subject*	*Obs. Per Sample*	*True Comparability,* ρ_T					
				0.4		**0.7**		**0.95**	
				Model		*Model*		*Model*	
				F & M	**C**	**F & M**	**C**	**F & M**	**C**
1	*6*	**4**	**6**	0.44	0.44	0.74	0.74	0.96	0.96
			12	0.42	0.42	0.72	0.72	0.95	0.95
		8	**6**	0.43	0.44	0.73	0.74	0.96	0.96
			12	0.41	0.41	0.71	0.71	0.95	0.95
	15	**4**	**6**	0.44	0.44	0.74	0.74	0.96	0.96
			12	0.42	0.42	0.72	0.72	0.96	0.96
		8	**6**	0.43	0.43	0.73	0.74	0.96	0.96
			12	0.41	0.42	0.71	0.72	0.95	0.95
0.5	*6*	**4**	**6**	0.43	0.43	0.72	0.72	0.96	0.96
			12	0.42	0.41	0.72	0.72	0.96	0.95
		8	**6**	0.43	0.43	0.73	0.73	0.96	0.96
			12	0.42	0.41	0.72	0.72	0.96	0.95
	15	**4**	**6**	0.43	0.43	0.72	0.72	0.96	0.96
			12	0.42	0.41	0.72	0.72	0.95	0.95
		8	**6**	0.43	0.42	0.73	0.72	0.96	0.96
			12	0.41	0.41	0.71	0.71	0.95	0.95
0.1	*6*	**4**	**6**	0.43	0.43	0.73	0.73	0.96	0.96
			12	0.42	0.41	0.72	0.71	0.95	0.95
		8	**6**	0.43	0.42	0.73	0.73	0.96	0.96
			12	0.41	0.41	0.71	0.71	0.95	0.95
	15	**4**	**6**	0.43	0.43	0.73	0.73	0.96	0.96
			12	0.41	0.41	0.72	0.71	0.95	0.95
		8	**6**	0.43	0.42	0.73	0.72	0.96	0.96
			12	0.41	0.41	0.71	0.71	0.95	0.95

Table 5.4: Mean estimate of Comparability over 1000 simulations based on different number of subjects, samples per subject and observations per sample, by each of the three models. In this table:

F = Basic Fisher Model
M = Multiplicative Fisher Model
C = Components of Variance Model

The point estimates from the Fisher model and the Multiplicative Fisher model are naturally the same whereas those from the Component of Variance are somewhat different. The point estimators generally overestimate the Comparability with the larger the number of observations per sample the smaller the bias. This improvement is more obvious for lower values of the true Comparability (i.e. ρ_T=0.4 or 0.7) but unlike the One-Stage modelling, the number of samples per subject has minimal effect on the point estimates. It seems that an increase in the number of subjects from 6 to 15 slightly decreases the bias.

Estimated coverage rates show that there are, generally, higher coverage rates for a larger number of subjects (i.e. I=15) as well as for a larger number of observations per sample (i.e. n=12). This simply indicates the positive effect of these two factors in increasing the performance of the models in providing consistent confidence levels. Clearly, in the case of a small number of observations per sample, an increase in the number of samples from 4 to 8 reduces the coverage rates for higher values of the true Comparability (i.e. ρ_T=0.7 or 1).

When $\sigma_T^2 = 1$, the estimated coverage rates under the Fisher and the Multiplicative models are almost the same, while the Components of Variance model, especially in the case of 4 samples, provides slightly higher and more stable coverage rates. For $\sigma_T^2 < 1$, coverage rates under the Fisher and the Multiplicative models are smaller than those for the Components of Variance model. This reduction applies to the cases of higher true Comparability (i.e. ρ_T=0.7 or 0.95) and is more obvious when the Fisher model is used.

σ_T^2	*Number of Subjects*	*Sample Per Subject*	*Obs. Per Sample*	*True Comparability,* ρ_T								
				0.4			**0.7**			**0.95**		
				Model			*Model*			*Model*		
				F	**M**	**C**	**F**	**M**	**C**	**F**	**M**	**C**
1	*6*	**4**	**6**	84	84	85	77	78	78	67	70	79
			12	87	90	88	84	85	85	79	82	82
		8	**6**	85	87	88	76	76	79	69	71	74
			12	93	95	92	89	90	90	80	81	83
	15	**4**	**6**	85	86	88	82	87	86	77	84	85
			12	86	88	88	84	87	86	83	88	87
		8	**6**	87	88	88	82	85	86	74	76	79
			12	89	90	88	86	89	87	84	86	87
0.5	*6*	**4**	**6**	77	93	93	72	82	87	56	70	85
			12	83	93	94	79	85	87	70	82	89
		8	**6**	79	92	94	64	75	84	50	60	74
			12	83	93	95	76	87	85	59	70	88
	15	**4**	**6**	79	91	91	77	87	90	69	81	87
			12	82	92	92	77	89	91	75	87	90
		8	**6**	83	93	94	75	88	90	64	77	85
			12	84	95	95	78	90	91	75	89	91
0.1	*6*	**4**	**6**	84	89	86	70	75	86	53	61	64
			12	85	92	95	82	85	91	70	77	89
		8	**6**	88	89	92	66	71	84	48	57	77
			12	93	94	94	82	84	88	68	71	89
	15	**4**	**6**	81	91	92	79	87	91	71	78	89
			12	84	93	94	82	91	93	79	89	92
		8	**6**	89	91	93	80	81	85	65	68	81
			12	93	94	94	88	90	91	77	82	89

Table 5.5: Percentage of the cases, over 1000 simulations, where the confidence interval captures the true Comparability based on different numbers of subjects, samples per subject and observations per sample, for each of the three models. In this table:

F = Basic Fisher Model
M = Multiplicative Fisher Model
C = Components of Variance Model

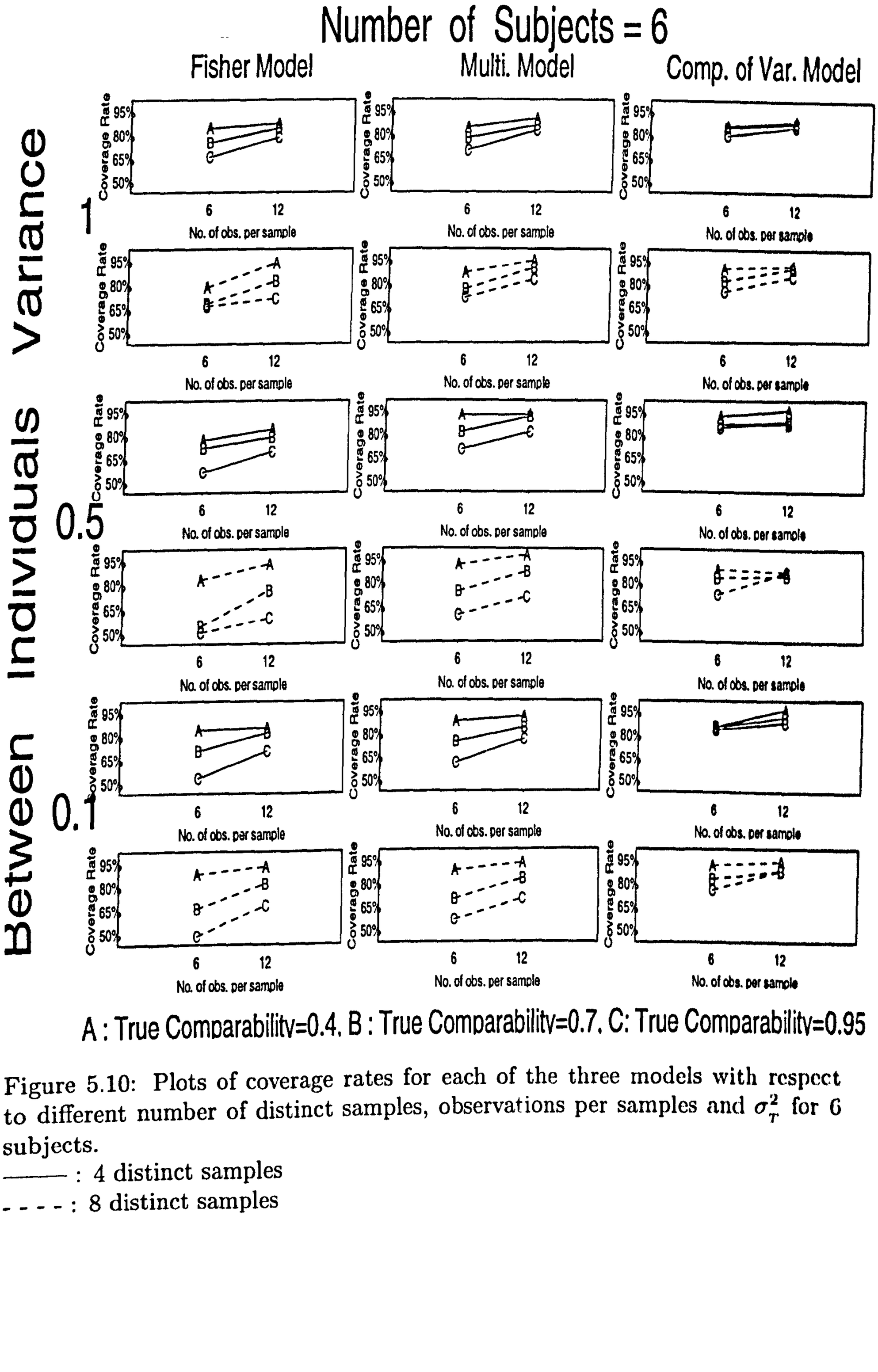

Figure 5.10: Plots of coverage rates for each of the three models with respect to different number of distinct samples, observations per samples and σ_T^2 for 6 subjects.

——— : 4 distinct samples

- - - - : 8 distinct samples

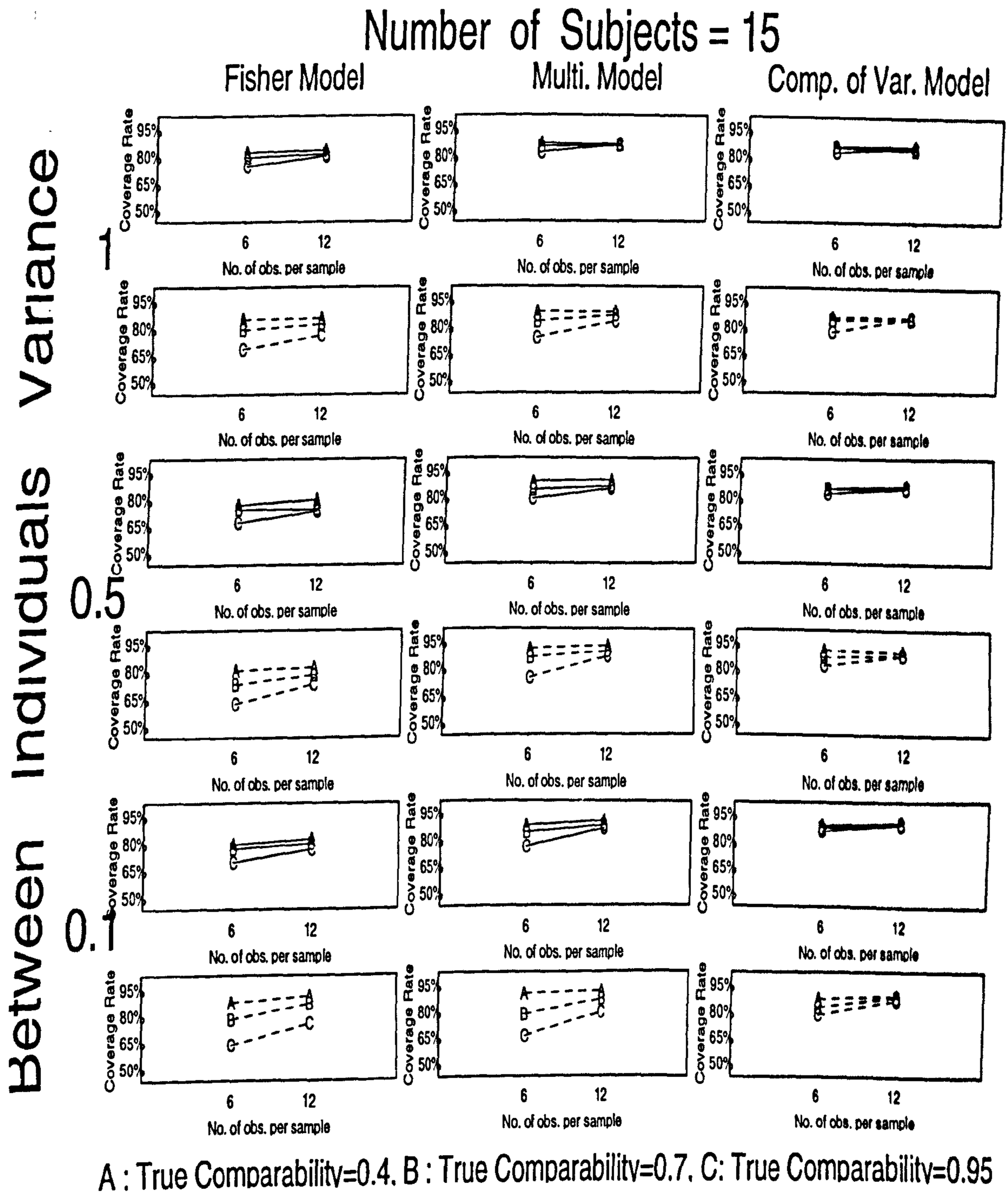

Figure 5.11: Plots of coverage rates for each of the three models with respect to different numbers of distinct samples, observations per samples and σ_T^2 for 15 subjects.
——— : 4 distinct samples,
- - - - : 8 distinct samples

Plots of average confidence interval width (Figures 5.12 and 5.13) show that the case of a larger number of subjects tends, not surprisingly, to produce narrower confidence intervals. For lower true Comparability (i.e. ρ_T=0.4 or 0.7), confidence intervals under the Fisher and the Multiplicative models are, on average, narrower than the Components of Variance model as the number of observations per samples increase. When $\sigma_T^2 = 1$, confidence intervals from the Multiplicative model are wider than those from the Fisher model, whereas in the cases of $\sigma_T^2 < 1$ the intervals from the Multiplicative model are, on average, slightly narrower than those from the Fisher model. Further, for lower true Comparability (i.e. ρ_T=0.4 or 0.7), the typical confidence interval from the Components of Variance model is wider than those provided by the other models, with an increase in the number of samples decreasing the width in general. In these cases it seems that the number of observations per sample has no significant effect on average interval width.

Overall, the Components of Variance model seems to provide higher and more stable coverage rates, irrespective of the size of σ_T^2 and the true Comparability, at least over these simulation configurations.

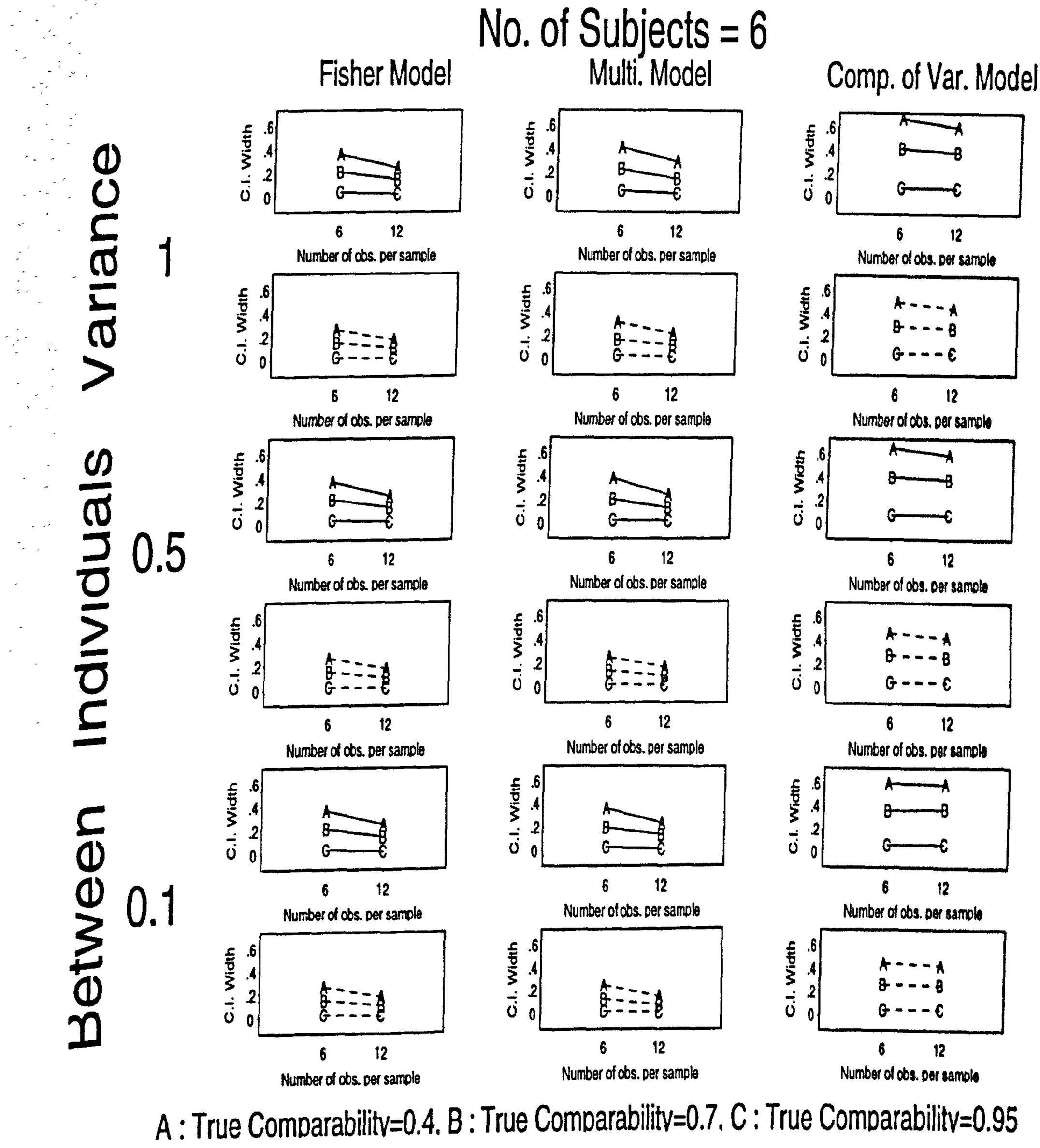

Figure 5.12: Plots of confidence interval widths for each of the three models with respect to different number of distinct samples, observations per samples and σ_τ^2 for 6 subjects.
——— : 4 distinct samples,
- - - - : 8 distinct samples

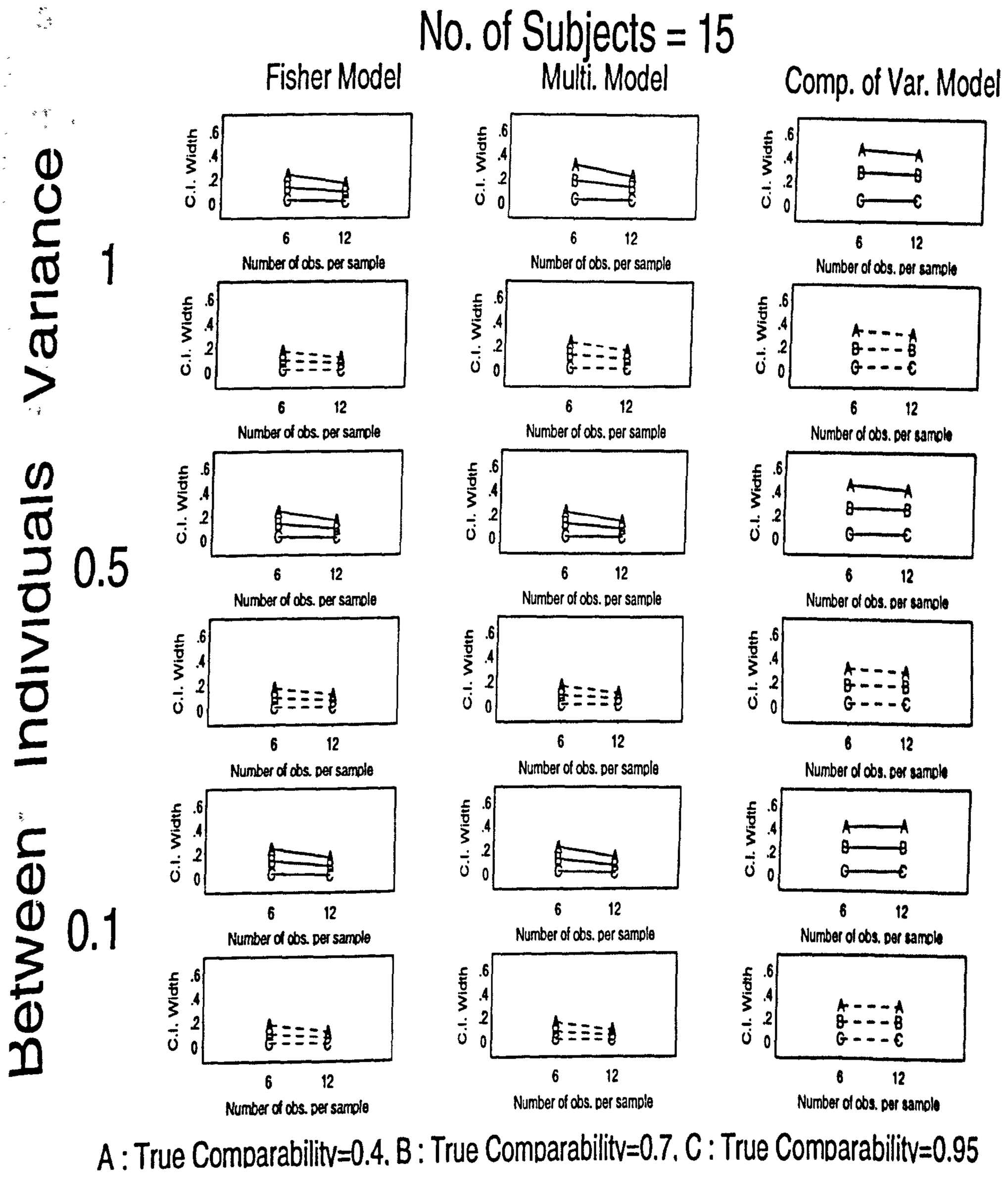

Figure 5.13: Plots of confidence interval widths for each of the three models with respect to different number of distinct samples, observations per samples and σ_τ^2 for 15 subjects.
——— : 4 distinct samples,
- - - - : 8 distinct samples

5.4.3 Summary

Two competing approaches for a One-Stage model to estimating a population Comparability and three distinct approaches for a Two-Stage model were introduced and assessed.

A specific illustration of One-Stage modelling suggested that, while the two point estimates based on the two models are always the same, the confidence interval under the Multiplicative model tends to be narrower than that under the Basic Fisher model.

A simulation study strongly suggested narrower confidence intervals and better performance in achieving higher confidence for interval estimation under the Multiplicative model across a variety of simulated configurations.

A specific illustration of Two-Stage modelling showed that while the two point estimates based on the Fisher model and the Multiplicative model are always the same, that based on the Components of Variance model is slightly different. Confidence intervals under the Fisher model and the Multiplicative model are almost the same in this instance, while the confidence interval under the Component of Variance model is wider that those from the other two models.

A simulation study showed a tendency for wider confidence intervals under the Components of Variance model but with better performance in achieving the required 95% confidence across the range of simulated configurations.

Chapter 6

Conclusions and Further Work

6.1 Conclusions

Estimating the Reproducibility of variables and the Comparability between two variables (based on common correlation coefficients or modelling a population of correlation coefficients) were the main topics covered in this thesis. Chapters 2 and 3 were concerned with the estimation of measurement reproducibility of data and its application in Exercise Testing data. In chapter 4 five different statistical approaches to estimating a common correlation coefficient were examined and finally in chapter 5 modelling a population of correlation coefficients was introduced and developed to allow the estimation of Comparability between two variables.

Chapter 2 dealt with the estimation of measurement Reproducibility of data from mixed effects models involving two variance components. Two models, one based on the idea of using sums of squares (ANOVA) and the other based on a Profile Likelihood approach, were set forth for both the cases of balanced (equal number of observations per individual) and unbalanced (unequal number of

observations per individual) data to provide point and interval estimates of the measurement Reproducibility. This was carried out for two different models, one for simple replication and the other one assuming an order effect to the replications.

Throughout the chapter, illustrative examples using exercise testing data, each with different features, were used to describe the performance and the applicability of the two approaches. It has been found that while the point estimates from both approaches are almost the same, interval estimates from the Profile Likelihood approach tend to be narrower. The basic recommendation here is to use the Profile Likelihood approach in both the point and interval estimation of measurement Reproducibility.

Performances of the two approaches were subject to further investigation and comparison in chapter 3, where a simulation study, with a variety of underlying configurations for both a simple replicate model and for a replicate model with an order effect, was carried out. The performances then were compared according to three basic criteria (bias, coverage rate and interval estimate width).

The simulation study indicated that both approaches on average tend to underestimate measurement Reproducibility; however, an increase in the number of subjects significantly reduces this bias. It was concluded that failure to fit a (significant) visit effect in the model considerably increases the bias as well as decreasing the consistency of the interval estimation in terms of coverage rate. For balanced data, the ANOVA-based approach showed a better performance in terms of bias although, in comparison with the Profile Likelihood approach, lower coverage rates as well as wider interval estimates were revealed. For unbalanced data, the Profile Likelihood approach provided a better overall performance in the sense that it produces less biased point estimates with narrower confidence

intervals in general.

In the choice of a better approach, the results of the simulation study over the different configurations seemed on balance to favour the Profile Likelihood approach.

The second aspect of the thesis in chapter 4 involved, as a first step in estimating the Comparability of two variables, the estimation of a common correlation coefficient from a sample of correlation coefficients. Five different methods of point and interval estimation of a common correlation coefficient were examined. These were the Weighted method, the Unbiased method based on an approximately unbiased estimate for a common correlation coefficient suggested by Olkin and Pratt, the Fisher method based on a Fisher transformation of the simple correlation coefficients, a method proposed by Hodges and Olkin, and finally a Profile Likelihood based method.

An illustrative example with data from an Exercise Testing study was used to compare the performance of the five methods. The point estimates from the Weighted method were, in general, less than those from the other methods while the estimates from the Fisher and the Hedges and Olkin methods appeared similar. On the other hand, except for the Profile Likelihood method which provided the narrowest interval estimates, there was no obvious differences among the widths of the interval estimates from the other 4 methods.

For this example, the assumption of commonality of correlation coefficients was validated and the problem with inconsistency of correlation coefficients for a specific subject was investigated. It was shown that, the major characteristic of the data that may have a significant effect on statistics for testing commonality of correlation is the standard deviation of sample correlations in the Fisher Transformation space.

Further investigations of the performances of the five methods were carried out by means of a simulation study with a variety of underlying configurations. Some methods were better than others on certain simulation configurations but the overall results suggested that the Fisher method was the best in the sense that it provides the "most stable" confidence levels irrespective of the number of subjects, replicates per subject or true common correlation.

The Comparability of two variables was finally modelled, in chapter 5, by developing structures for 'pooling' correlation coefficient across replicate visits for individuals. In this chapter, estimation of a 'typical or average' correlation coefficient between two variables of interest for an individual (One-Stage modelling) was developed and then extended to the "overall" estimation of a 'typical' correlation for a 'typical' individual (Two-Stage modelling). Based on the results from the previous chapter, the modelling process here was carried out in the Fisher Transformation space.

Two distinct approaches were developed for the One-Stage modelling (i.e. the Basic Fisher model and the Multiplicative Fisher model) and three distinct approaches (the Basic model, the Multiplicative model and the Components of Variance model) were adopted for Two-Stage modelling. Illustrative examples from Exercise Testing were used to investigate the performance of these models on real data.

The example for the models in One-Stage modelling showed identical point estimates of Comparability for both models, but a considerably narrower interval under the Multiplicative model. The real data example for the models in Two-Stage modelling showed that basically the point and interval estimates under the Fisher model and the Multiplicative model were similar, but a slightly different point estimate and a wider interval estimate were found under the

Components of Variance model.

The performance and applicability of the models developed in this chapter was considered by means of a simulation study based on a wide variety of configurations. The main conclusion drawn from these for One-Stage modelling was that the Multiplicative Fisher model is better in the sense that it produces narrower interval estimates and more consistent confidence levels.

The results from the simulation study for Two-Stage modelling suggested that the Components of Variance model, in spite of the fact that it tends to produce wider interval estimates than the other models, has the advantage of providing higher and more stable coverage rates.

6.2 Possible Further Work

Methods for estimating Reproducibility of variables and Comparability between two variables using different techniques were described in this thesis. In this section, some points are given that can be the subject of further work:

- Data in medical sciences often has a hierarchical organisation in which units at one level are grouped in units at another level. In Exercise Testing, for example, measurements/replicates of one individuals are nested in the individual and individuals are nested in time points. One type of popular analysis for this kind of data is known as "Multilevel Modelling". As use of Multilevel techniques and related software becomes widespread in different areas of statistical researches, one possible approach is

to apply standard Multilevel Modelling techniques to estimate components of variance and hence measurement reproducibility.

- A pragmatic solution to the multivariate approach of estimating a pooled Measurement Reproducibility was described in chapter 2. The case was restricted to the situation where the time points are 'independent' from each other. In reality, time points will not be independent in any sense and estimation of a pooled measurement reproducibility, in spite of the fact that it may be computationally difficult, should be a subject of further investigation.

- From a practical point of view, the effect of medication changes on Exercise Testing results may be considered in the analysis of such data. For instance, individuals with a heart failure problem may take a 'Beta-blocker' as medication. The effect of such a medication changes an individual's response and its effect (i.e. the Sensitivity of the measurement) should be estimated within a reproducibility study.

- Validity of Measurement Reproducibility estimates can be increased by considering the results from independent studies. In Exercise Testing, for instance, exercise testing may be carried out in "exercise" centres in different hospitals or different countries with the same exercise device (i.e. treadmill or bicycle). Such data has a hierarchical structure. Measurements of one individual are nested in the individual, which in turn are nested in the exercise centre in hospitals, hospitals in countries, and so on. One might wish to consider these centres as fixed but it is often of more interest to consider the exercise centres as a random sample drawn from a potentially much bigger population of possible hospital and/or countries. In this case the estimation of three or four variance components (for subject effects, for hospital effects and country effects as well as the 'natural

variability') would be involved. This would be extended to the situation of unequal number of individuals per each centre as well as unequal number of tests per individual.

- Modelling a population of correlation coefficients to estimate the Comparability between two variables across individuals was described in chapter 5. The idea can be extended to the case where variables of interest are measured in different centres, e.g. different hospitals or countries. In such a situation it may be possible to extending the modelling techniques for estimating a "pooled" Comparability between two variables. In this case another sources of variability i.e. those within and between different centres, e.g. across hospitals and/or countries, would be involved.

References

Bartko, J.J., Carpenter, W.T. (1976) On the Methods and Theory of Reliability. *The Journal of Nervous and Mental Disease*, **163**, No. 5, 307-317.

Bartlett, R.F. (1993) Linear Modelling of Pearson's Product Moment Correlation Coefficient: An Application of Fisher's z-transformation. *The Statistician*, **42**, 45-53.

Burdic, R.K., Birch, N.J., Graybill, F.A. (1986) Confidence Intervals on Measures of Variability in an Unbalanced Two-Fold Nested Design with Equal Subsampling. *J. Statist. Comput. Simul.*, **25**, 259-272.

Bushman, J.B., Wang, C.M. (1995) A procedure for Combining Sample Correlation Coefficients and Vote Counting to Obtain an Estimate and a Confidence Interval for the Population Correlation Coefficient. *Psychological Bulletin*, **117**, No. 3, 530-546.

Donner, A. (1986) A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. *International Statistical Review*, **54**, No. 1, 67-82.

Donner, A., Eliasziw, M. (1987) Sample Size Requirements for Reliability Studies. *Statistics in Medicine*, **6**, 441-448.

Duncan, G.T., Layard, M.W. (1973) A Monte-Carlo study of Asymptotically Robust Tests for Correlation Coefficients. *Biometrika*, **60**, No. 3, 551-558.

Dunn, G. (1989) Design and Analysis of Reliability studies; The Statistical evaluation of Measurement Errors. Oxford University Press, New York.

Dunn, G. (1992) Design and Analysis of Reliability Studies. *Statistical Methods in Medical Research*, **1**, 123-157.

Goldberg, D.P. (1975) The Detection of Psychologic Illness by Questionnaire. Oxford University Press. London.

Goldman, L., Hashimoto, B., Cook, E.F., Loscalzo, A. (1981) Comparative Reproducibility and Validity of Systems for Assessing Cardiovascular Functional Class: Advantages of a New Specific Activity Scale. *Circulation*, **64**, No. 6, 1227-1233.

Goldstein, H. (1987) Multilevel Models in Education and Social Research. Oxford University Press, New York.

Goldstein, H. (1995) Multilevel Statistical Models. Edward Arnold, London.

Graybill, F. A. (1969) Introduction to Matrices with Application in Statistics. Wadsworth Publishing Company, Inc. Belmont, California.

Grubbs, F.E. (1973) Errors of Measurement, Precision, Accuracy and the Statistical Comparison of Measuring Instruments. *Technometrics*, **15**, No. 1, 53-66.

Hartley, H.O., Rao, J.N.K. (1967) Maximum Likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93-108.

Hedges, L.V., Olkin, I (1985) Statistical methods for Meta-Analysis. Academic Press, London.

Hemmerle, W.J., Hartley, H.O. (1973) Computing Maximum Likelihood estimates for the mixed model using the W-transformation. *Technometrics*, **15**, 819-831.

Henderson, C.R. (1953) Estimation of Variance and Covariance Components. *Biometrics*, **9**, 226-252.

Kelly, R.J., Mathew, T. (1993) Improved Estimation of Variance components with Smaller Probability of Negativity. *Journal of Royal Statistics Society*, **55**, No. 4, 897-911.

Kelly, R.J., Mathew, T. (1994) Improved Nonnegative Estimation of Variance Components in Some Mixed Models With Unbalanced Data. *Technometrics*, **36**, No. 2, 177-181.

Khuri. A.I., Sahai, H., (1985) Variance Components analysis: A Selective Literature Survey. *International Statistical Review*, **53**, 279-300.

References

Kowalski, C.J. (1972) On the Effects of Non-normality on the Distribution of the Sample Product-moment correlation Coefficient. *Applied Statistics*, **21**, 1-12.

Kraemer, H.C. (1975) On Estimation and Hypothesis Testing Problems for Correlation Coefficients. *Psychometrika*, **40**, No. 4, 473-485.

Kraemer, H.C. (1979) Tests of Homogeneity of Independent Correlation Coefficients. *Psychometrika*, **44**, No. 3, 329-335.

Miller, J.J. (1977) Maximum Likelihood estimation of variance components; a Monte Carlo study. *Journal of Statist. Comput. Simul.*, **8**, 175-190.

Naughton, J. P., Hellerstein, H K. (1973) Exercise Testing and Exercise Training in Coronary Heart Disease. Academic Press, New York, London.

Olkin, I., Pratt, J.W. (1958) Unbiased estimation of certain Correlation Coefficients. *Annals of Mathematical Statistics*, **29**, 201-211.

Olkin, I. (1995) Meta-Analysis: Reconciling the Results of Independent Studies. *Statistics in Medicine*, **14**, 457-472.

Rosenthal, R. (1978) Combining Results of Independent Studies. *Psychological Bulletin*, **85**, No. 1, 185-193.

Searle, S. R., Casella, G., McCulloch, C E. (1992) Variance Components. John Wiley & Sons, Inc., New York.

Searle S.R. (1987) Linear Models for Unbalanced Data. John Wiley & Sons, New York.

Sharkey, B. J. (1991) New Dimensions in Aerobic Fitness. Human Kinetics Books, Champaign, Illinois.

Shrout, P.E., Fleiss, J.L. (1979) Intraclass Correlation: Uses in Assessing Rater Reliability. *Psychological Bulletin*, **66**, No. 2, 420-428.

Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Skinner, J.S. (1987) Exercise Testing and Exercise Prescription for Special Cases; Theoretical Basis and Clinical Application. Lea and Febiger, Philadelphia.

Smith, D.W., Murray, L.W. (1984) An Alternative to Eisenhart's Model II and Mixed Model in the Case of Negative Variance Estimates. *Journal of American Statistical Association*, **79**, 145-151.

Thompson, W.A. (1962) The Problem of Negative Estimates of Variance Components. *Annals of Mathematical Statistics*, **33**, 273-289.

Tippett, L.H.C. (1931) The Methods of Statistics. Williams and Norgate; An Introduction Mainly for Workers in Biological Sciences, London.

Venzon, D.J., Moolgavkar, S.H. (1988) A Method for Computing Profile-Likelihood-Based Confidence Intervals. *Applied Statistics*, **37**, No. 1, 87-94.

Viana, M.A.G. (1980) Statistical Methods for sumarizing independent correlational results. *Journal of Educational Statistics*, **5**, 83-104.