Assessment and Comparison of Continuous Measurement Systems

by

Nathaniel T. Stevens

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2014

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis we critically examine the assessment and comparison of continuous measurement systems. Measurement systems, defined to be the devices, people, and protocol used to make a measurement, are an important tool in a variety of contexts. In manufacturing contexts a measurement system may be used to monitor a manufacturing process; in healthcare contexts a measurement system may be used to evaluate the status of a patient. In all contexts it is desirable for the measurement system to be accurate and precise, so as to provide high-quality and reliable measurements.

A measurement system assessment (MSA) study is performed to assess the adequacy, and in particular the variability (precision), of the measurement system. The Automotive Industry Action Group (AIAG) recommends a standard design for such a study in which 10 subjects are measured multiple times by each individual who operates the measurement system. In this thesis we propose alternate study designs which, with little extra effort, provide more precise evaluations of the measurement system's performance.

Specifically, we propose the use of unbalanced augmented plans which, by strategically using more subjects and fewer replicate measurements, are substantially more efficient and more informative than the AIAG recommendation. We consider cases when the measurement system is operated by just one individual (or is automated), and when the measurement system is operated by multiple individuals, and in all cases, augmented plans are superior to the typical designs recommended by the AIAG.

In situations where the measurement system is used routinely, and records of these single measurements on many subjects are kept, we propose incorporating this additional 'baseline' information into the planning and analysis of an MSA study. Once again we consider the scenarios in which the measurement system is operated by a single individual, or multiple individuals. In all cases incorporating baseline information in the planning and analysis of an MSA study substantially increases the amount of information about subject-to-subject variation. This in turn allows for a much more precise assessment of the measurement system than is possible with the designs recommended by the AIAG.

Often new measurement systems that are less expensive, require less man-power, and are perhaps less time-consuming, are developed. In these cases, potential customers may wish to compare the new measurement system with their existing one, to ensure that the measurements by the new system agree suitably with the old. This comparison is typically done with a measurement system comparison (MSC) study, in which a number of randomly selected subjects are measured one or more times by each system. A variety of statistical techniques exist for analyzing MSC study data and quantifying the agreement between the two systems, but none are without challenges.

We propose the probability of agreement, a new method for analyzing MSC data, which more effectively and transparently quantifies the agreement between two measurement systems. The chief advantage of the probability of agreement is that it is intuitive and simple to interpret, and its interpretation is the same no matter how complicated the setting. We illustrate its applicability, and its superiority to existing techniques, in a variety of settings and we also make recommendations for a study design that facilitates precise estimation of this probability.

Acknowledgements

I would like to sincerely thank my supervisors Dr. Stefan Steiner and Dr. Jock MacKay for their guidance and support throughout my graduate studies. They have presented me with many wonderful opportunities that I would not have otherwise had, and for that I am extremely grateful. I would also like to thank my family for fostering a love for learning, and for teaching me the importance of hard work and dedication. And finally, I would like to thank Heather for her endless support, enthusiasm, and encouragement during this process.

Dedication

To my wife, Heather.

Table of Contents

4. Introduction: Measurement System Comparison 67

5. The Probability of Agreement: An Alternative to the Limits of Agreement Analysis 87

List of Figures

List of Tables

List of Abbreviations

ANOVA – Analysis of Variance

AP – Augmented Plan

CAD – Clinically Acceptable Difference

GPQ – Generalized Pivotal Quantity

GR&R – Gauge Repeatability & Reproducibility

LP – Leveraged Plan

LLA – Lower Limit of Agreement

LSL – Lower Specification Limit

ML – Maximum Likelihood

MLS – Modified Large Sample

MS – Measurement System

MSA – Measurement System Assessment

MSC – Measurement System Comparison

OLS – Ordinary Least Squares

PTR – Precision-to-Tolerance Ratio

REML – Restricted Maximum Likelihood

SNR – Signal-to-Noise Ratio

SP – Standard Plan

ULA – Upper Limit of Agreement

USL – Upper Specification Limit

WLS – Weighted Least Squares

Chapter 1

Introduction: Measurement System Assessment

The demand for high quality permeates all facets of society; for example consumers demand high quality goods, medical patients demand high quality healthcare, students demand high quality education and tax payers demand high quality services. This demand necessitates the ability to assess and monitor the quality of these outputs, and this in turn requires a method of measuring the output to ensure high quality is achieved.

Measurement systems (MS)- defined here to be the devices, people, and protocol used to make a measurement- are used to make such measurements, which are often used for decision making. For example, in a manufacturing process that produces golf balls, a device may be used to measure the diameter of each ball to ensure that it meets specification. If it does the product may be deemed suitable for shipment, but if not, the product may have to be scrapped or reworked. As another example, in healthcare diastolic and systolic blood pressure measurements are used to classify an individual's level of hypertension, and depending on the magnitude of these measurements medical intervention may be necessary. In instances like these, the measurement system is an invaluable component of the decision-making process, and so it is important that it provides high-quality measurements.

Unfortunately measurements are often subject to error. Generally speaking, we will say that a measurement system provides high-quality measurements if these measurements are both accurate and precise. To develop the concepts of accuracy and precision, let us define the measurand to be the characteristic of the object that is being measured, and define a measurement to be the assignment of a numerical value to the measurand. In the manufacturing context the object being measured may be a manufactured part for which a critical dimension must be determined. In the medical context the object may be a human subject for whom blood pressure must be assessed. Without loss of generality, within this thesis we refer to the objects whose characteristics we measure as subjects.

In general, we can define accuracy to be the “closeness” of measurements to the true value of the measurand, and precision to be the “closeness” of repeated measurements to one another [Automotive Industry Action Group, 2010]. In particular, an accurate measurement system will measure the true value of the measurand correctly on average, while a precise measurement system will produce repeated measurements which are very similar, but that may or may not be close to the true value.

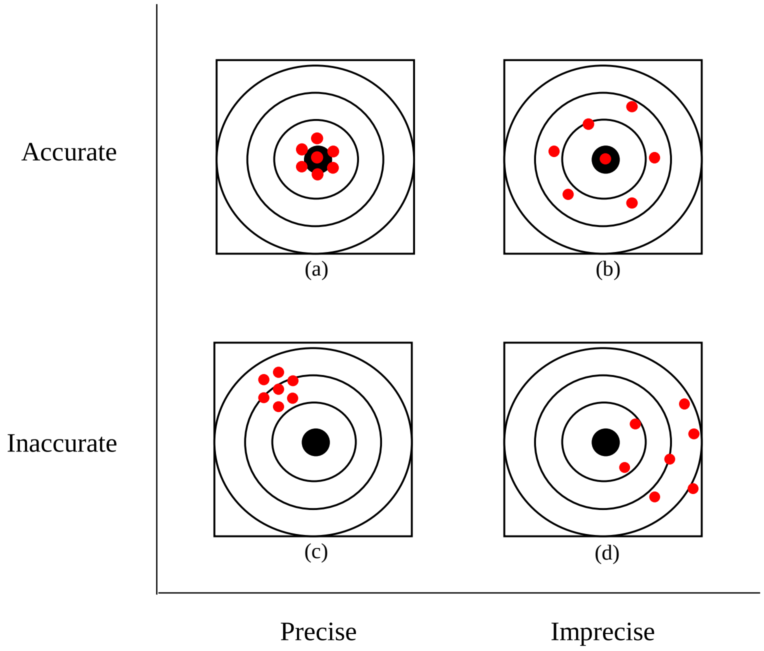

Figure 1.1: Dart Board Analogy Explaining Accuracy and Precision
(a) MS is accurate and precise (b) MS is accurate and imprecise
(c) MS is inaccurate and precise (d) MS is inaccurate and imprecise

To fully understand these concepts, a useful analogy can be drawn by considering a game of darts. Imagine that the “bull’s eye” is the true value of the measurand. A player who is accurate will hit the bull’s eye on average, and a player who is precise will not necessarily hit the bull’s eye but their darts will land fairly closely to one another. Figure 1.1 depicts the four cases that arise based on the level of a measurement system’s precision and accuracy. The optimal scenario is illustrated in (a) when the measurement system is precise and accurate and the least favourable scenario is shown in (d) when the measurement system is neither precise nor accurate. The situations corresponding to (b) and (c) are also undesirable.

Because a measurement system may not give accurate and precise measurements, a measurement system assessment (MSA) study may be undertaken to determine whether the measurement system is adequate. In fact, a periodic assessment of measurement systems is mandated within many quality systems such as ISO 9001:2008 [Myhrberg, 2009] and TS 16949:2009 [Automotive Industry Action Group, 2009], and checking the adequacy of the measurement system is an important step in many process improvement strategies [Steiner and MacKay, 2005; Breyfogle, 1999].

In such a study, the measurement system's accuracy and precision may be assessed. In statistical jargon we refer to a measurement system's accuracy as bias and its precision as variability. In particular, as bias increases accuracy decreases and as variability increases precision decreases. Thus we call a measurement system which is unbiased accurate, and one that is biased inaccurate. Similarly, we say that a measurement system is precise if it has little variability and we say that it is imprecise if it has excessive variability.

The Automotive Industry Action Group (AIAG) [2010] defines a measurement system to be linear if its bias and variability are constant over the distribution of the measurand's true values. That is, the bias and variability do not change as the true value of the measurand changes. They similarly define a measurement system to be stable if its bias and variability are constant over time [Automotive Industry Action Group, 2010]. Most often we assume that the measurement system being assessed is linear and stable, but it is worthwhile to check these assumptions. As well, a measurement system may be stable during the time of study, but this could change over time, and so it is important to periodically assess the measurement system.

Here we deal with the assessment of a non-destructive measurement system, which means the act of measuring does not alter the true value of the measurand, and multiple measurements of the measurand are possible. We also only consider continuous measurement systems (those that determine a single continuous measurand).

In assessing such a measurement system it is standard practice to randomly sample n subjects and take r replicate measurements on each subject. If there are m observers included in the study, each one of these observers will take r replicate measurements of each subject, for an overall total of $N = nmr$ measurements. We call this a standard plan (SP). Common choices of n, m, and r are $n = 10$, $m = 2{,}3$ and $r = 2{,}3$ so $40 \leq N \leq 90$ [Tsai, 1988; Burdick et al., 2005; Automotive Industry Action Group, 2010]. If the measurement system is automated or has one observer, common choices are $n = 10$ and $r = 6$ [Automotive Industry Action Group, 2010]. We will return to the issue of planning a MSA study in Section 1.3.

Most often the primary concern of these studies is to investigate the variability of the measurement system. There is less emphasis placed on the bias because it is typically thought that “bias can be eliminated by proper calibration of the system” [Burdick et al., 2005, p. 3]. Addressing variability on the other hand, is not so straightforward. In the simplest case- when the measurement system is automated or has just one observer- we adopt the following random effects model:

$$Y_{ik} = S_i + M_{ik} \qquad [1.1]$$

where $i = 1,2,\dots,n$ indexes the subjects and $k = 1,2,\dots,r$ indexes the number of replicate measurements per subject. Thus Y_{ik} is a random variable which represents the observed response for the kth measurement on subject i. Here S_i is a random variable representing the unknown true value of the measurand for subject i, and M_{ik} is a random variable representing the measurement error associated with replicate measurements on subject i. We assume that for all i, $S_i \sim N(\mu, \sigma_s^2)$ where μ is the mean value of the measurand and σ_s^2 is the variance component which quantifies the variability in true values about μ. We additionally assume that for all i and k, $M_{ik} \sim N(\mu_m, \sigma_m^2)$ where the parameters μ_m and σ_m^2 represent the measurement bias and the measurement variation, respectively. However, because it is thought that bias can be eliminated through calibration, it is common to assume $\mu_m = 0$ [Burdick et al.,

2005]. Lastly we assume that the S_i and M_{ik} are mutually independent. As such, we find that the expectation and variance of the observed response is $E(Y_{ik}) = \mu$ and $Var(Y_{ik}) = \sigma_s^2 + \sigma_m^2$.

In healthcare and manufacturing measurement systems, there are many sources of variability that contribute to the size of σ_m^2. In particular, it is common to assume that the people taking the measurements (i.e. the observers or operators) can be an important source of variation.

Accordingly, we often include multiple observers in an MSA study and separate their effects from M_{ik} in model [1.1] so that we can assess their relative contribution to the overall measurement system variation. To do this we must adopt a different model which allows for the separate estimation of the observer effect. We give this model later in this section.

By separating the effect of observers we can partition the overall measurement variability into two pieces which we call the repeatability and the reproducibility of the measurement system. Reproducibility is the variability due to different observers using the measurement device and repeatability is the variability due to all other sources of variation inherent in the measurement system, and it reflects the precision of the system itself [Montgomery, 2005].

When considering the effect of the observers there is a critical statistical assumption that must be made: the observer effects can either be assumed to be fixed or random. When there are a large number of observers who regularly operate the measurement system, and only a sample of them is available for inclusion in the study, it is reasonable to assume the observer effects are random. However, when there are only a few observers operating the measurement system, all of whom participate in the study, it makes more sense to think of the observer effects as being fixed. In manufacturing contexts, where there is usually a small number of observers, it is more likely that a fixed effect approach would be warranted [Van den Heuvel and Trip, 2002; Burdick et al., 2005] and in a medical context when there are potentially many observers available, a random effects approach is appropriate [Steiner et al., 2011]. The work developed in this thesis primarily assumes fixed observer effects, with the random effect case briefly discussed in Section 1.3.

As noted, we must modify model [1.1] to account for observer variability. We do so by introducing a fixed effect for observer, resulting in the following two-factor mixed-effects model:

$$Y_{ijk} = S_i + o_j + SO_{ij} + M_{ijk} \qquad [1.2]$$

where $i = 1,2, \dots, n$ indexes the subject, $j = 1,2, \dots, m$ indexes the observer and $k = 1,2, \dots, r$ indexes replicate measurements. Thus Y_{ijk} represents the observed response for the k^{th} replicate measurement by observer j on subject i. As in [1.1] S_i is a random variable representing the unknown true value of the measurand for subject i, which has mean μ. M_{ijk} is a random variable representing the measurement error when the same observer takes replicate measurements of the same subject. Here o_j represents the fixed effect of observer j which is subject to the constraint

$$\sum_{j=1}^{m} o_j = 0$$

We also include the random variable SO_{ij} to allow the observer effect to change from subject to subject. Because the true values are described by a random effect, we adopt the traditional approach of also describing the possible subject-by-observer interaction with a random effect. The interaction is quantified parsimoniously by the single parameter σ_{so}^2. Under this model we make the distributional assumptions that $S_i \sim N(\mu, \sigma_s^2)$, $SO_{ij} \sim N(0, \sigma_{so}^2)$ and $M_{ijk} \sim N(0, \sigma_m^2)$ and that all of these random variables are mutually independent. As such, we find that the expectation and variance of the observed response is $E(Y_{ijk}) = \mu + o_j$ and $Var(Y_{ijk}) = \sigma_s^2 + \sigma_{so}^2 + \sigma_m^2$.

We use $\mu_j = \mu + o_j$ to denote the expected measurement by observer j, and with that define

$$\sigma_o^2 = \frac{1}{m}\sum_{j=1}^{m} \left(\mu_j - \mu\right)^2 \qquad [1.3]$$

where $\mu = \sum_{j=1}^{m} \mu_j / m$ is the overall mean value of the measurand, as in Burdick et al. [2005, p. 83]. Thus σ_o^2 quantifies the measurement variation due to the relative biases among observers. We then define and interpret the total variation in the observed measurements to be $\sigma_t^2 = \sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2$. Note that because σ_o^2 is not a variance in the usual sense under the fixed effect approach for observers (it is not the variance of a Normal distribution), neither is σ_t^2.

Montgomery [2005] notes that the primary goal of a MSA study is to estimate these individual variance components and hence quantify the measurement system variability in order to determine whether it is suitable for regular use.

1.1 MSA Metrics

A variety of quantities may be used for assessing the adequacy of a measurement system, many of which are defined in terms of the variance components associated with [1.2]. We will briefly discuss a number of them, but emphasis within this thesis will be placed on one in particular.

To begin, it is instructive to use these variance components to precisely define the repeatability and reproducibility of a measurement system. We define σ_m^2 to be the repeatability as it characterizes the variability among replicate measurements made by any particular observer on any particular subject, and we define the quantity $\sigma_o^2 + \sigma_{so}^2$ to be the reproducibility as it represents the variability among measurements made by many observers on the same subject [Vardeman and Valkenberg, 1999]. Based on these definitions, it is clear that $\sigma_o^2 + \sigma_{so}^2 + \sigma_m^2$ represents the variability due to the measurement system as a whole. Accordingly, let us define:

- $\sigma_{repeatability}^2 = \sigma_m^2$
- $\sigma_{reproducibility}^2 = \sigma_o^2 + \sigma_{so}^2$
- $\sigma_{MS}^2 = \sigma_{repeatability}^2 + \sigma_{reproducibility}^2$

Using this notation we will proceed with our discussion of MSA metrics.

One quantity that is often used to describe the adequacy of a measurement system in manufacturing contexts is the precision-to-tolerance ratio (PTR) [Automotive Industry Action Group, 2002; Burdick et al., 2005], which compares the width of the measurement error distribution to the width of the specification limits:

$$PTR = \frac{k\sigma_{MS}}{USL - LSL} \qquad [1.4]$$

where the upper specification limit (USL) and lower specification limit (LSL) respectively represent the largest and smallest allowable value for the quality characteristic being measured. The constant k typically takes on the values $k = 5.15$ or 6. The value $k = 6$ corresponds to the number of standard deviations between the natural tolerance limits that contain the middle 99.73% of a normal distribution, and $k = 5.15$ corresponds to limits that contain the middle 99.00% of a normal distribution [Burdick et al., 2005].

The size of PTR reflects the capability, or adequacy, of the measurement system. The guidelines suggested by the Automotive Industry Action Group [2002, p. 77] are as follows:

- $PTR \leq 0.1$: The measurement system is capable
- $0.1 \leq PTR \leq 0.3$: The measurement system may be capable

- $PTR \geq 0.3$: The measurement system is not capable

However, Wheeler and Lyday [1989] and Montgomery [2005] caution against using this metric as it does not involve the process variability σ_s^2. Note that in a manufacturing context, σ_s^2 quantifies part-to-part, or process, variability. When the process variability is small, the measurement system variability, σ_{MS}^2, must also be small (to adequately distinguish parts). However, suppose σ_s^2 is very small, and so the process produces parts well within specification, but the measurement system is highly variable (i.e. σ_{MS}^2 is large). Clearly we would like the metric to reflect the inadequacy of the measurement system, but σ_{MS}^2, although large, might still be small enough to ensure $PTR \leq 0.1$. As well, if the process variability is large, larger measurement variation can be tolerated [Mader et al., 1999]. This may cause $PTR \geq 0.3$, even though the measurement system is adequate.

Thus, because the precision-to-tolerance ratio does not quantify the measurement variation relative to the process variation, it does not always give a useful evaluation of a measurement system's capability.

More informative metrics typically compare the measurement variation to the between-subject variation σ_s^2, or the total variation σ_t^2. One such metric is the discrimination ratio which compares the relative sizes of the between-subject variation and the measurement system variation [Steiner and MacKay, 2005]:

$$D = \frac{\sigma_s}{\sigma_{MS}}$$

A scaling of this quantity is also used, and is referred to as the signal-to-noise ratio (SNR). See Burdick et al., [2005] and Montgomery [2005].

Another metric that is used frequently to assess the quality of a measurement system is the intraclass correlation coefficient [Shrout and Fleiss, 1979]:

$$\rho = \frac{\sigma_s^2}{\sigma_t^2} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2} \qquad [1.5]$$

This ratio examines the proportion of the overall variability (σ_t^2) that is attributable to variation in the true values (σ_s^2). Statistically, it is defined as the correlation between two measurements on the same subject by different observers. Note that $0 \leq \rho \leq 1$, and large values of ρ indicate that the measurement system contributes little to the overall variation, suggesting that it is acceptable.

A related ratio for assessing a measurement system compares the variation due to the measurement system (repeatability and/or reproducibility) to the overall variation of the measurements (i.e. due to differences in the true dimensions and the measurement system). The Automotive Industry Action Group [2010] defines the gauge repeatability and reproducibility (GR&R) ratio as:

$$\gamma = \frac{\sigma_{MS}}{\sigma_t} = \sqrt{\frac{\sigma_o^2 + \sigma_{so}^2 + \sigma_m^2}{\sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2}} \qquad [1.6]$$

When the measurement system is highly variable, σ_{MS} will be large relative to σ_t and so γ will be large. As such, smaller values of γ are desirable. The Automotive Industry Action Group [2010, p. 78] recommend the following acceptability criteria:

- $\gamma < 0.1$: The measurement system is acceptable
- $0.1 \leq \gamma \leq 0.3$: The measurement system needs improvement
- $\gamma > 0.3$: The measurement system is unacceptable

If the estimate of γ is large, we can examine the estimate of σ_o^2, σ_{so}^2, and σ_m^2 separately to identify the source of the large measurement system variation.

Note that the discrimination ratio D, the intraclass correlation coefficient ρ, and the gauge repeatability and reproducibility ratio γ are one-to-one functions of each other. In particular $D = \sqrt{\frac{1-\gamma^2}{\gamma^2}}$ and $\rho = 1 - \gamma^2$. As such, each of these metrics conveys the same information about the measurement system, and use of any of them is sufficient.

In Table 1.1 we translate the Automotive Industry Action Group's acceptability criteria for γ, for ρ. We also present the acceptability criteria suggested by Steiner and MacKay [2005] for D, which is less conservative than the AIAG recommendation in the sense that a measurement system judged as acceptable by this criteria may not be judged as acceptable by the AIAG criteria.

	Acceptable	Needs Improvement	Unacceptable
GR&R ratio	$\gamma \leq 0.1$	$0.1 < \gamma < 0.3$	$\gamma \geq 0.3$
Intraclass Correlation Coefficient	$\rho \geq 0.99$	$0.91 < \rho < 0.99$	$\rho \leq 0.91$
Discrimination Ratio	$D \geq 3$	$2 < D < 3$	$D \leq 2$

Table 1.1: Measurement System Acceptability Criteria

In manufacturing contexts it is common to use the GR&R ratio γ as the measure of measurement system adequacy, while in medical contexts it is typical to use the intraclass correlation coefficient ρ. However, because each of these metrics conveys the same information about the measurement system, and use of any one of them is sufficient, in this thesis we choose to focus attention on the repeatability and reproducibility ratio, γ, as the metric of interest. We do so for two reasons: the first is that convention in the industrial context dictates the use of γ to assess a measurement system when observer effects are assumed fixed. And the second reason, which is potentially the source of the first, is that the interpretation of γ, being the proportion of overall variation due to the measurement system, is appropriate given that the adequacy of a measurement system is judged based on how variable it is.

1.2 Estimation Techniques

In order to use the GR&R ratio in practice, we must obtain its estimate, denoted $\hat{\gamma}$. Doing so requires the estimation of the individual variance components. When multiple observers are considered, this means we must obtain $\hat{\sigma}_s^2$, $\hat{\sigma}_o^2$, $\hat{\sigma}_{so}^2$, and $\hat{\sigma}_m^2$. Note that we use Greek letters, for example θ, to denote parameters, and we use a Greek letter overscored by a circumflex ($\hat{\theta}$) to denote its estimate (a real number). When applicable we will use a Greek letter overscored by a tilde ($\tilde{\theta}$) to denote the corresponding estimator (a random variable).

A variety of estimation methods are used in practice. An early approach was a graphical technique that used control charts and an analysis of ranges to estimate $\sigma_{repeatability}^2$ and $\sigma_{reproducibility}^2$ [Wheeler and Lyday, 1989; Automotive Industry Action Group, 2010]. However, as Burdick et al. [2005] indicate, the range is an inefficient measure of measurement system variability and so this approach is not as commonly used as it once was.

A popular alternative is the analysis of variance (ANOVA). The philosophy of the ANOVA approach is to partition the total variability into its component sources. Recall that when the measurement system has multiple observers, and we separate the effect of the observers, then we adopt the two-factor mixed effect model [1.2]. In this case, the total variability is partitioned according to variability due to subjects, observer effects, subject-by-observer interaction effects, and the measurement system. Specifically, the total sum of squares is decomposed as follows:

$$SS_t = SS_s + SS_o + SS_{so} + SS_m$$

Table 1.2 displays the ANOVA table associated with this decomposition, where $\bar{Y}_{...} = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} Y_{ijk}/nmr$, $\bar{Y}_{ij\cdot} = \sum_{k=1}^{r} Y_{ijk}/r$, $\bar{Y}_{i\cdot\cdot} = \sum_{j=1}^{m}\sum_{k=1}^{r} Y_{ijk}/mr$, and $\bar{Y}_{\cdot j\cdot} = \sum_{i=1}^{n}\sum_{k=1}^{r} Y_{ijk}/nr$. Here we assume that the collected data were obtained from a study whose design followed the balanced standard plan (SP) described above. That is, each of m observers measures each of n subjects r times.

The estimates $\hat{\sigma}_s^2$, $\hat{\sigma}_o^2$, $\hat{\sigma}_{so}^2$, and $\hat{\sigma}_m^2$ are obtained by simultaneously solving the expected mean squares which are shown in Table 1.3 (reproduced from Burdick et al. [2005]). Solving these equations and substituting the observed mean squares for the expected gives:

$$\hat{\sigma}_s^2 = \frac{MS_s - MS_{so}}{mr} \qquad [1.7]$$

$$\hat{\sigma}_o^2 = \frac{(m-1)(MS_o - MS_{so})}{nmr} \qquad [1.8]$$

$$\hat{\sigma}_{so}^2 = \frac{MS_{so} - MS_m}{r} \qquad [1.9]$$

$$\hat{\sigma}_m^2 = MS_m \qquad [1.10]$$

Note that the mean squares used to calculate these estimates are constructed using the observed data and each of the corresponding estimators is unbiased for the true parameter. The estimates [1.7 – 1.10] are then substituted into [1.6] to obtain:

$$\hat{\gamma} = \sqrt{\frac{\hat{\sigma}_o^2 + \hat{\sigma}_{so}^2 + \hat{\sigma}_m^2}{\hat{\sigma}_s^2 + \hat{\sigma}_o^2 + \hat{\sigma}_{so}^2 + \hat{\sigma}_m^2}} \qquad [1.11]$$

Source	Sum of Squares	*df*	Mean Square	F-Statistic
S	SS_s	$n-1$	$MS_s = \frac{SS_s}{n-1}$	$\frac{MS_s}{MS_{so}}$
O	SS_o	$m-1$	$MS_o = \frac{SS_o}{m-1}$	$\frac{MS_o}{MS_{so}}$
SO	SS_{so}	$(n-1)(m-1)$	$MS_{so} = \frac{SS_{so}}{(n-1)(m-1)}$	$\frac{MS_{so}}{MS_m}$
M	SS_m	$nm(r-1)$	$MS_m = \frac{SS_m}{nm(r-1)}$	
Total	SS_t	$nmr-1$		

Table 1.2: ANOVA table for a two-factor mixed effect model

where

$$SS_s = mr\sum_{i=1}^{n}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\dots})^2$$

$$SS_o = nr\sum_{j=1}^{m}\left(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\dots}\right)^2$$

$$SS_{so} = r\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\dots}\right)^2$$

$$SS_m = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}\left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)^2$$

$$SS_t = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}\left(Y_{ijk} - \bar{Y}_{\dots}\right)^2$$

Expected Mean Squares
$E(MS_s) = \sigma_m^2 + r\sigma_{so}^2 + mr\sigma_s^2$
$E(MS_o) = \sigma_m^2 + r\sigma_{so}^2 + nmr\sigma_o^2/(m-1)$
$E(MS_{so}) = \sigma_m^2 + r\sigma_{so}^2$
$E(MS_m) = \sigma_m^2$

Table 1.3: Expected mean squares for two-factor mixed effect model

One major pitfall of the ANOVA approach to estimation is that it is possible that the variance component estimates $\hat{\sigma}_s^2$, $\hat{\sigma}_o^2$, $\hat{\sigma}_{so}^2$ could be negative [Montgomery and Runger, 1993a; Burdick et al., 2005]. Clearly this is a problem since variances are by definition non-negative. In the context of measurement system assessment, this issue arises most often when the subject-by-observer interaction effect is small. In particular, if $MS_{so} < MS_m$ then

$$\hat{\sigma}_{so}^2 = \frac{MS_{so} - MS_m}{r} < 0$$

One way to deal with this is to set this variance component equal to zero. However, in doing this, the estimates of the other parameters will be biased [Montgomery and Runger, 1993a].

If $MS_{so} < MS_m$ then it is likely that the subject-by-observer interaction is not significant. As such, Montgomery and Runger [1993a] suggest that we test this hypothesis using the F-statistic given on line 3 in Table 1.2. If the interaction is insignificant, then we can ignore its effects and perform an analysis of variance on the reduced model which does not include the SO_{ij} term.

Another drawback to the ANOVA method of estimation is that it cannot be applied to all study designs. Although Burdick et al. [2005] develop the ANOVA-based estimates for mixed and random effect models within many atypical designs, there are still some unbalanced designs to which the procedure cannot be applied. We will address this point further in Chapter 2.

To avoid the problem of negative variance component estimates, the use of maximum likelihood (ML) estimation has been proposed as an alternative [Montgomery and Runger, 1993b]. As well, it is flexible enough to accommodate non-standard study designs. However, there are two main drawbacks associated with the ML approach to variance component estimation. The first is the well-known problem that maximum likelihood estimates are biased when sample sizes are small [Swallow and Monahan, 1984], and the second is that it is computationally more intensive than the ANOVA approach (closed form expressions for ML estimates rarely exist). However, with the powerful computational resources and statistical software available to practitioners, this procedure is not nearly as cumbersome as it once was.

Because of its flexibility with regard to study design and the nice asymptotic properties of ML estimates, we adopt maximum likelihood estimation as the preferred estimation technique throughout this thesis. We note that the issue of biased estimates with small sample sizes is of little concern because the sample sizes in MSA studies are typically large enough to avoid any substantial bias. When necessary we check this assumption using simulation.

We close this subsection with a discussion of estimation uncertainty. One common criticism associated with using an estimate such as [1.11] to judge the acceptability of a measurement system is that $\hat{\gamma}$ is a point estimate, and on its own does not provide any information regarding the uncertainty of its estimation [Burdick et al., 2005]. As such, there has been much effort directed at calculating confidence intervals for the parameters of an MSA study. In particular, Burdick et al. [2005] provide techniques for calculating approximate confidence intervals for γ (and many other metrics) within the context of one-factor random effects models and two-factor mixed effect models for both balanced and unbalanced designs. The approximate intervals presented in this reference are calculated using modified large sample (MLS) and generalized pivotal quantity (GPQ) techniques which are based on the distributions of the sums of squares shown in Table 1.2. As such, these intervals are applicable when using the ANOVA method to estimate variance components. Shrout and Fleiss [1979] similarly present confidence intervals for the intraclass correlation coefficient defined within one-factor random effects models and two-factor mixed effect models.

We note that approximate confidence intervals for γ can also be calculated when employing a maximum likelihood approach. Using the inverse of the expected Fisher information matrix we can obtain the asymptotic variances of the maximum likelihood estimators $\tilde{\sigma}_s^2$, $\tilde{\sigma}_{so}^2$, $\tilde{\sigma}_m^2$ and $\tilde{\mu}_j$ for $j = 1,2, \ldots, m$. To determine the asymptotic variance of functions of these estimators, say $\tilde{\sigma}_o^2$ or $\tilde{\gamma}$, we can apply the delta method and pre- and post-multiply the information matrix by a change-of-variable matrix of suitable partial derivatives. Because maximum likelihood estimators have asymptotic normal distributions and are asymptotically unbiased, these asymptotic variances can be used to construct approximate confidence intervals for the parameters of interest using the critical values of the standard normal distribution [Casella and Berger, 2002]. The performance of any such interval (regardless of the estimation method) will depend on the sample size and design of the study.

1.3 Designing MSA Studies

The design of MSA studies is an extremely important, albeit under-emphasized, topic in measurement system assessment literature. In fact Mazu [2006, p. 305] remarks that "if there is an area in gauge R&R studies that needs more extensive information and guidelines for the practitioner, it is the planning of gauge studies". Recall that the most common design, which we refer to as the standard plan (SP), has $m = 2,3$ observers measure each of $n = 10$ subjects $r = 2,3$ times each.

Regardless of the specific allocation of measurements, every practitioner planning an MSA study should consider the three fundamental experimental design principles replication, randomization, and blocking- doing so improves the validity and efficiency of the study [Montgomery and Runger, 1993a]. In the context of MSA studies, replication refers to an observer making multiple measurements on a subject, where each measurement is an independent execution of the measurement system. Doing this allows for estimation of σ_m^2, the repeatability of the measurement system, and the more replicated measurements there are, the more precise this estimate will be. It is important to distinguish between replicate measurements and repeated measurements. Repeated measurements refer to consecutive measurements that an observer takes on a subject without changing the set-up of the instrumentation. In this way repeated measurements may not be independent. A study which uses repeated measurements and not replicate measurements will likely under estimate σ_m^2 [Burdick et al., 2005].

The second principle of experimental design is randomization. In the context of MSA studies, there can be two levels of randomization; the first level of randomization refers to the manner in which subjects are selected for inclusion in the study and the second level of randomization refers to the order in which observers measure the selected subjects. It is important to note that in practice, subjects included in the study are often not randomly selected, however the validity of many statistical methods relies on the assumption of randomization, and so it is important that it be incorporated in the design of MSA studies where possible.

The final principle experimental design principle to consider is blocking. Blocking is a method for eliminating the effect of nuisance factors (extraneous sources of variation whose effect we are not concerned with). This is done by taking measurements at fixed levels of these nuisance factors so that they cannot introduce any additional variability into the measurements. However, we would prefer to investigate and quantify these possible sources of variation rather than remove them from the analysis, and so blocking is not particularly useful in an MSA setting.

The next major consideration in planning a MSA study is the number of subjects n, observers m and replicate measurements r. When the measurement system is automated or has a single observer, there has been considerable activity in choosing n and r, and hence altering the allocation of resources compared to the common $n = 10,\ r = 6$ SP recommended by the Automotive Industry Action Group [2010]. For example, Shainin and others [Shainin, 1992;

Traver, 1995] recommend an Isoplot™ study, where $n = 30$ subjects are selected and each is measured twice, i.e. $r = 2$. The Shainin plan provides better balance between the number of degrees of freedom available for estimating the measurement and subject variation, and as we shall see in Chapter 2, is the optimal SP for estimating γ in this case. Hamada and Borror [2012] consider repeatability and reproducibility estimation within unreplicated designs, i.e. when $r = 1$, and Vardeman and Valkenberg [1999] point out that if the only goal is to estimate the variation due to the measurement system, σ_m^2, then it is best to take $n = 1$ and take replicate measurements of this single subject.

There has also been considerable activity for choosing sample sizes in medical contexts when ρ, the intraclass correlation coefficient, is used as measure of adequacy. Specifically there are two main methods to choose the optimal number of subjects n and replicate measurements r. The first is based on the power of a hypothesis test regarding ρ such as

$$H_0: \rho = \rho_0 \text{ versus } H_A: \rho > \rho_0 \qquad [1.12]$$

where ρ_0 corresponds to a minimum value of ρ that investigators deem acceptable [Donner and Eliasziw, 1987]. Donner and Eliasziw [1987] and Walter et al. [1998] use such a power analysis. The former authors present power contours for varying values of n and r when $\alpha = 0.5$, $1 - \beta = 0.8$ and $\rho_0 = 0.2, 0.4, 0.6, 0.8$ with which a practitioner can choose optimal values of n and r. The latter authors provide a functional approximation to these exact results which allows for more flexibility in terms of investigating other plans with different values of $1 - \beta$. Note that α and β here respectively represent the probability of Type I and Type II error.

In both papers, the general recommendation when the true value of ρ is relatively large (as it should be in such studies) is to maximize the number of subjects n and minimize the number of replicate measurements r for fixed $N = nr$. In particular, $r = 2$ or 3 is sufficient. This agrees with the Shainin plan described above.

Within this framework Eliasziw and Donner [1987] also suggest that the number of subjects and replicates in the study can be determined by the costs associated with sampling subjects and performing replicate measurements; if it is expensive to sample subjects, then it might be beneficial to choose a design that achieves the same power but uses fewer subjects and more replicate measurements. Accordingly, they present a method which minimizes a linear cost function subject to the statistical power constraint $(1-\beta)$ using Legrange multipliers. The conclusions they draw are rather intuitive: if sampling subjects is expensive, reduce n and increase r, and if replicate measurements are expensive, increase n and reduce r. In Eliasziw and Donner [1987] tables of possible (n,r) pairings and an example are given to help choose n and r in practice.

The second method of choosing n and r in medical contexts is done by finding the values of n and r that maximize the precision with which ρ is estimated. This method is suggested as an alternative to the power analysis technique because the power analysis requires an arbitrary choice of ρ_0 in [1.12] and a choice of the significant difference one wants to be able to identify. Giraudeau and Mary [2001] note that these assumptions may be difficult ones to make, and once they are made, they may be questionable. Giraudeau and Mary [2001] and Bonnett [2002] choose values of n and r that minimize the width of confidence intervals for ρ. In both cases they arrive at similar recommendations as those given above: for large ρ, the size typical of ρ, it is best to increase the number of subjects and reduce the number of replicates. For fixed $N = nr$ they similarly suggest taking $r = 2$ or 3, and choosing n accordingly.

Taken together, these results suggest that when the MSA study design follows the standard plan, and estimating γ (or ρ) is of interest, unless it is cost-prohibitive, it is best to maximize the number of subjects and minimize the number of replicate measurements. However, these results ignore the potential effect of observers. The optimal allocation of measurements when multiple observers are involved in the study remains unclear.

As well, these results assume the MSA study design adheres to the standard plan. There has been a recent development of alternative study designs that may be superior to the standard plan in their ability to efficiently estimate parameters of interest.

To motivate these alternative designs, recall that the Automotive Industry Action Group [2010] typically recommends $n = 10$ subjects be included in the study. But with so few subjects, the resulting estimate of σ_s^2 (the between-subject variability) may be poor; $n = 10$ subjects does not allow for precise estimation of σ_s^2.

Steiner and MacKay [2005] suggest incorporating baseline (historical) data in a measurement system analysis. Doing so increases the number of degrees of freedom for estimating σ_s^2, yielding more precise estimates. Browne et al. [2009a] and Danila et al. [2008, 2010] in the binary situation demonstrate the considerable value of this extra information.

Another example comes from a series of papers by Browne et al. [2009a, 2009b, 2010] that consider the use of leveraging to increase the efficiency of standard plans. The proposed leveraged plan (LP) is completed in two stages. In Stage 1, a baseline is collected: b subjects are randomly selected from the usual process over a long enough time frame to obtain a good estimate of σ_t^2, and each of these subjects is measured once. In Stage 2, n subjects are collected from the baseline sample and re-measured $r = 2{,}3$ times each. The term 'leverage' arises because of the way in which these n subjects are selected; the n selected subjects are considered extreme in relation to the baseline mean. In particular, they suggest selecting $n/2$ subjects with the smallest and $n/2$ subjects with the largest (non-outlier) baseline measurements.

If the study includes $m > 1$ observers, then each observer measures b/m different subjects in Stage 1, and in Stage 2, n/m extreme subjects are selected from each observer's Stage 1 sample for replicated measurements. The authors suggest that if a total of N measurements can be made in the study, we should allocate roughly $N/2$ measurements to Stage 1 and $N/2$ measurements to Stage 2. If $N = 60$ and $m = 3$, Browne et al. [2010] suggest $b = 11, n = 3, r = 3$ and if $N = 90$ and $m = 3$ then they suggest $b = 18$, $n = 6$, $r = 2$.

Browne et al. demonstrate the benefit in using a leveraged plan as opposed to a standard plan within the framework of models [1.1] [2009a, 2009b, 2009c] and [1.2] [2010]. In particular, an LP with the same total number of measurement as an SP will provide more precise estimates of the gauge R&R ratio γ [1.6] and the intraclass correlation coefficient ρ [1.5], and with fewer overall measurements an LP can achieve the same power that an SP can when testing hypotheses regarding ρ such as [1.12] or hypotheses involving γ:

$$H_0\colon \gamma \geq \gamma_0 \text{ versus } H_A\colon \gamma < \gamma_0 \qquad [1.13]$$

where common choices for γ_0 are the acceptability/unacceptability criteria 0.1 and 0.3 suggested by the Automotive Industry Action group [2010].

Steiner et al. [2011] investigate measurement system assessment when the observer effects are assumed to be random. In this case they incorporate the random effect $O_j \sim N(0, \sigma_O^2)$ (instead of the fixed effect μ_j) into their two-way model. Here, because the observer effect is random, many more observers are needed to obtain a precise estimate of σ_O^2 and hence γ. In fact Burdick et al. [2005, p. 58] suggest that in this case $m \geq 6$ observers be included in the study. Steiner et al. [2011] propose that, rather than performing one usual standard plan, it is beneficial to perform replicates of a smaller standard plan. In particular they propose using the plan in which two subjects are measured once by each of two observers (for a total of four measurements), and then they suggest replicating this plan with different subjects and observers. Doing so balances the degrees of freedom necessary for estimating σ_S^2 and σ_O^2, resulting in better estimates. This in turn facilitates more precise estimation of γ.

1.4 Looking Ahead

To this point we have discussed the goals and importance of a measurement system assessment study, the metrics used to judge acceptability of a measurement system and methods used to estimate these metrics. We have also highlighted the importance of the study design and have discussed recent developments in this area. In Chapter 2 and 3 we present new contributions to this field resulting from work on this thesis. In particular we investigate and develop alternative MSA study designs that continue to improve upon the standard plan in their ability to precisely estimate γ.

In Chapter 2 we investigate the performance of two types of (unbalanced) assessment plans, collectively referred to as Augmented Plans [Stevens et al., 2010]. In each type we use a standard plan, in which m observers measures n subjects, r times each, and we augment this with additional measurements by each observer. In type A augmentation, each observer measures a different set of subjects once each. In type B augmentation, each observer measures the same set of subjects once each. The goal of these designs is to supplement the information gained from the standard plan, so as to more precisely estimate γ, the gauge repeatability and reproducibility ratio. When a measurement system is used by multiple observers we show that use of an appropriate augmented plan can produce substantial gains in precision for estimating γ compared to the best standard plan with the same total number of measurements.

In Chapter 3 we focus on measurement systems that are used routinely and that have a record of the single measurements made during regular use. We consider incorporating these baseline measurements in both the planning and analysis of an MSA study [Stevens et al., 2013]. Specifically we quantify the substantial benefits of incorporating baseline data with regard to precisely estimating γ, and we search for good standard plans with a fixed total number of measurements that take into account available baseline data. The benefit of incorporating baseline data into the analysis is significant and most of the gains in precision can be obtained with small baseline sample sizes. In general, depending on the baseline sample size, the number of observers and whether we wish to estimate a subject-by-observer interaction, the standard plan with either the minimum or maximum number of subjects is recommended.

In the second part of this thesis (Chapters 4, 5 and 6) we consider the comparison of two measurement systems. Often new measurement systems are developed that are cheaper and perhaps easier to use. In these cases a potential buyer may want to compare the performance of this new measurement system with their existing one, and decide whether the new one can be used in place of the old one. This comparison is typically done with a measurement system comparison (MSC) study. In Chapter 4 we discuss, in more detail, the MSC study and we review several existing techniques for analyzing MSC data. In Chapter 5 we propose the probability of agreement, a new method for analyzing MSC data, which more effectively and transparently quantifies the agreement between two measurement systems [Stevens et al., 2014 (*under revision*)]. We also make recommendations for a study design that facilitates precise estimation of this probability. In Chapter 6 we relax various model assumptions made in the previous chapter and consider the application of the probability of agreement when comparing two measurement systems in these more general settings. Specifically, we consider the situation in which the true values of the measurand do not follow a normal distribution, and when the measurement variation of one or both systems depends on this unknown true value.

Chapter 2

Augmented Measurement System Assessment

In Chapter 1 we introduced the idea of a measurement system assessment (MSA) study, the purpose of which is primarily to assess the variability of the measurement system. We also described the typical balanced design of such a study, which we refer to as the standard plan (SP), in which a random sample of n subjects is measured r times by each of m observers, for a total of $N = nmr$ measurements. For a fixed number of observers, we denote this plan $SP(n, r)$. In this chapter we propose the use of unbalanced assessment plans that we refer to as augmented plans (AP), as an alternative to the standard plan [Stevens et al., 2010].

To begin we will re-state the two-way mixed effects model that is typically used to describe MSA study data:

$$Y_{ijk} = S_i + o_j + SO_{ij} + M_{ijk} \qquad [2.1]$$

Recall that $i = 1,2, \dots, n$ indexes subjects, $j = 1,2, \dots, m$ indexes observers and $k = 1,2, \dots, r$ indexes replicate measurements. As such, Y_{ijk} represents the observed response for the k^{th} replicate measurement by observer j on subject i. As well, $S_i \sim N(\mu, \sigma_s^2)$ is a random variable representing the unknown true value of the measurand and $M_{ijk} \sim N(0, \sigma_m^2)$ is a random variable representing the measurement error when the same observer takes replicate measurements of the same subject. Recall also that we choose to describe the effect of observer j with the fixed effect o_j, as opposed to a random effect because there is interest in understanding each observer's potential bias. We also include the random variable $SO_{ij} \sim N(0, \sigma_{so}^2)$ to allow the observer effect to change from subject to subject. Under this model we further assume that S_i, SO_{ij} and M_{ijk} are mutually independent.

With this model we also define σ_o^2, which quantifies the measurement variation due to the relative biases among observers:

$$\sigma_o^2 = \frac{1}{m}\sum\nolimits_{j=1}^{m}\left(\mu_j - \mu\right)^2$$

where $\mu_j = \mu + o_j$ is the expected measurement by observer j and μ is the overall mean value of the measurand. We then define and interpret the total variation in the observed measurements to be $\sigma_t^2 = \sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2$.

In Chapter 1 we also discussed a variety of metrics that are used in practice to quantify the adequacy of the measurement system. In this chapter we will focus on the gauge repeatability and reproducibility (GR&R) ratio γ, which compares the variability due to the measurement system to the total variability attributable to both the subjects and the measurement system. For convenience we re-state its definition here:

$$\gamma = \sqrt{\frac{\sigma_o^2 + \sigma_{so}^2 + \sigma_m^2}{\sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2}} \qquad [2.2]$$

In this chapter we consider three different cases depending on the number of observers included in the study, and whether a subject-by-observer interaction is suspected. Model [2.1] and the definition for γ given in [2.2] correspond to the multiple observer case, when estimation of a subject-by-observer interaction is of interest. In the other two cases, the model and the definition of γ both simplify.

When we assume the system is automated with no observer effects, or has a single observer ($m = 1$), we have $\sigma_o^2 = 0$ and we cannot estimate σ_{so}^2 separately from σ_m^2, so we let $\sigma_{so}^2 = 0$. In this case model simplifies to the random effects model [1.1], and γ reduces to [2.3]:

$$\gamma = \sqrt{\frac{\sigma_m^2}{\sigma_s^2 + \sigma_m^2}} \qquad [2.3]$$

When we assume the measurement system is operated by multiple observers, but assume their effects are the same for every subject (i.e. no subject-by-observer interaction), we set $\sigma_{so}^2 = 0$. Knowing whether a subject-by-observer interaction exists may be evidenced from previous MSA studies or prior knowledge of the observer effects. In this case, when we set $\sigma_{so}^2 = 0$, the SO_{ij} term is dropped from model [2.1] and γ reduces to [2.4]:

$$\gamma = \sqrt{\frac{\sigma_o^2 + \sigma_m^2}{\sigma_s^2 + \sigma_o^2 + \sigma_m^2}} \qquad [2.4]$$

To estimate γ, we need a plan that provides an estimate of the between-subject variation σ_s^2. Alternate metrics, such as the precision-to-tolerance ratio (PTR) [1.4], depend only on the measurement system variation, and do not require an estimate of σ_s^2. The optimal design of an assessment study to estimate PTR and other such metrics will be different from what we propose. In other situations, the goal may be to estimate the individual variance components. This change of goal will also lead to different assessment plans. In this chapter, we focus on finding good plans for estimating γ while preserving some information about the separate variance components.

Specifically, we compare standard plans with two new plans, which we call augmented plans, in which not all subjects are measured the same number of times [Stevens et al., 2010]. In all cases, we assume the number of observers m is fixed. The augmented plans have two components: one component is a standard plan using n subjects with r replicate measurements by each observer, and there are two possibilities for the other component:

Type A: Randomly sample n_A subjects (different from those selected in the SP component) where n_A is a multiple of the number of observers, m. Each observer then measures n_A/m different subjects once. We call this an A plan, and for a fixed number of observers we denote it by $A(n, r, n_A)$.

Type B: Randomly sample n_B subjects (different from those selected in the SP component). Each observer then measures each of these subjects once. We call this a B plan, and for a fixed number of observers we denote it by $B(n, r, n_B)$.

Plan $A(n, r, n_A)$ has a total of $N = nmr + n_A$ measurements using $n + n_A$ subjects, and Plan $B(n, r, n_B)$ has a total of $N = nmr + mn_B = m(nr + n_B)$ measurements using $n + n_B$ subjects. Within each component, every subject is measured the same number of times.

If we set n_A or n_B to zero, the corresponding augmented plan is an SP. And note that the second component of plan B corresponds to a SP with $r = 1$. We also point out that the two components of an augmented plan can be conducted simultaneously or in any order.

The goal of this work is to identify augmented and standard plans that precisely estimate γ when N, the total number of measurements available, is fixed. As noted, estimation of γ requires estimation of σ_s^2, the between-subject variability. And so precise estimation of γ requires precise estimation of σ_s^2. In order to precisely estimate the subject-to-subject variation, an assessment study must include many subjects. In Chapter 1 we stated that the Automotive Industry Action Group [2010] typically recommends that $n = 10$ subjects be included in the study, which is not necessarily enough to provide a good estimate of σ_s^2. The idea of augmentation is to increase the number of subjects in the study, allowing for a more precise estimate of σ_s^2 and hence γ.

We measure the efficiency of any augmented plan at a particular set of parameter values and fixed N by comparing the asymptotic standard deviations (not variance) of the maximum likelihood estimates of γ from the augmented plan relative to the best SP. Specifically, we define the efficiency of an augmented plan relative to the best standard plan by dividing the asymptotic standard deviation of $\tilde{\gamma}$ associated with the standard plan by that associated with the augmented plan. Here the "best" standard plan is the one with the smallest asymptotic standard deviation at the given parameter values. Thus we search for augmented plans that have efficiency greater than 1 over a range of values for the unknown parameters.

If we set $r = 1$ in an SP or in the SP component of either plan A or B, then no subject is measured more than once by any observer. Looking at the model [2.1], we see that in this case, σ_{so} and σ_m are not separately identifiable or estimable but $\sigma_{so}^2 + \sigma_m^2$, and hence γ, can be estimated. Also, any plan B is now an SP with each of the m observers measuring $n + n_B$ subjects once each. If we suspect that there is subject-by-observer interaction that we want to identify separately, then we include only plans with $r > 1$ in the comparisons.

The outline of this chapter is as follows. In Section 2.1, we derive the likelihood function and the Fisher information of the standard and augmented plans. Then in Section 2.2, we use the marginal asymptotic standard deviation of γ to rank various possible plans. We consider separately the special cases when there is no subject-by-observer interaction (i.e. $\sigma_{so}^2 = 0$) and when there are no observer effects ($m = 1$). In each case, we recommend specific plans and, when appropriate, calculate the efficiency of the recommended plans relative to the best standard plan. Because our choice is based on an asymptotic criterion, we also check the performance of the recommended plans using simulation. We conclude the chapter in Section 2.3 with a brief discussion and summary.

2.1 Likelihood, Fisher Information, and Asymptotic Standard Deviations

We rank standard and augmented plans using the asymptotic precision as given by the inverse of the Fisher information matrix. We focus on estimation of γ with the variance components considered secondary. To derive the information matrix, we need the likelihood. In this section we sketch the derivation and avoid the tedious details by using Maple [Maplesoft, 2014] and Matlab [The MathWorks Inc., 2013] to carry out the symbolic and numerical calculations.

In model [2.1], we assume that measurements made on different subjects are independent; thus we can write the log-likelihood and Fisher information for each subject and then add over all subjects. Consider the distribution of all measurements on a randomly selected subject i that is measured by m observers r times each, as in the SP (or the SP component of plan A or B). We order the random variables by observer so that $\vec{Y}_i = \left(\vec{Y}_{i1}^T, \vec{Y}_{i2}^T, \dots, \vec{Y}_{im}^T\right)^T$ where $\vec{Y}_{ij} = \left(Y_{ij1}, Y_{ij2}, \dots, Y_{ijr}\right)^T$ corresponds to the r measurements by observer j on subject i. We let J_a be a column vector of a 1's, $J_{a\times b}$ be an $a \times b$ matrix of 1's and I_a be the $a \times a$ identity matrix. From model [2.1] we have for subject i, $\vec{Y}_i \sim MVN(\vec{\mu}, \sum)$ with

$$\vec{\mu} = (\vec{\mu}_1, \vec{\mu}_2, \dots, \vec{\mu}_m)^T$$

where

$$\vec{\mu}_j = \mu_j(1,1, \dots, 1)^T = \mu_j J_r$$

and

$$\sum = \sigma_s^2 J_{mr\times mr} + \sigma_{so}^2 \begin{bmatrix} J_{r\times r} & 0 & \cdots & 0 \\ 0 & J_{r\times r} & 0 & \ddots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & J_{r\times r} \end{bmatrix} + \sigma_m^2 I_{mr}$$

The matrix $\sum$ has a special form that allows us to write its inverse and determinant explicitly as

$$\sum^{-1} = a_1 I_{mr} + a_2 I_m \otimes J_{r\times r} + a_3 J_{mr\times mr}$$

and

$$|\sum| = (\sigma_m^2 + r\sigma_{so}^2 + mr\sigma_s^2)(\sigma_m^2 + r\sigma_{so}^2)^{m-1}(\sigma_m^2)^{m(r-1)}$$

where the Kronecker product $\otimes$ creates the appropriate block diagonal matrix, and

$$a_1 = \frac{1}{\sigma_m^2}$$

$$a_2 = \frac{-\sigma_{so}^2}{\sigma_m^2(\sigma_m^2 + r\sigma_{so}^2)}$$

$$a_3 = \frac{-\sigma_s^2}{(\sigma_m^2 + r\sigma_{so}^2)(\sigma_m^2 + r\sigma_{so}^2 + mr\sigma_s^2)}$$

See Appendix A for more details on calculating the inverse and determinant of the variance covariance matrix $\sum$.

Denoting the observed data by y_{ijk} $i = 1,2,\ldots,n$, $j = 1,2,\ldots,m$ and $k = 1,2,\ldots,r$, and using a lower-case y to denote the observed data vectors, the log-likelihood contribution from subject i with r replicate measurements by m observers is

$$-mrln(2\pi) - \frac{1}{2}\ln|\sum| - \frac{1}{2}(\vec{y}_i - \vec{\mu})^T \sum^{-1}(\vec{y}_i - \vec{\mu})$$

Expanding and adding over all subjects gives the full log-likelihood contribution from a standard plan or the SP component of an augmented plan:

$$\begin{aligned} & l_{SP}(\vec{\mu},\sigma_s^2,\sigma_{so}^2,\sigma_m^2) = \\ & -nmrln(2\pi) - \frac{n}{2}\ln\left[(\sigma_m^2 + r\sigma_{so}^2 + mr\sigma_s^2)(\sigma_m^2 + r\sigma_{so}^2)^{m-1}(\sigma_m^2)^{m(r-1)}\right] \\ & -\frac{1}{2}\left\{ a_1 \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}\left(y_{ijk}-\mu_j\right)^2 \right. \\ & \qquad + a_2 \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\sum_{k=1}^{r}\left(y_{ijk}-\mu_j\right)^2\right] \\ & \qquad \left. + a_3 \sum_{i=1}^{n}\left[\sum_{j=1}^{m}\sum_{k=1}^{r}\left(y_{ijk}-\mu_j\right)\right]^2 \right\} \end{aligned} \qquad [2.5]$$

We can find the log-likelihood contribution for the data from the augmented component of plan B by setting $r = 1$ in equation [2.5]. If we denote the observed data in this component by z_{ij}, $i = 1,2,\ldots,n_B$, $j = 1,2,\ldots,m$ we have:

$$\begin{aligned} & l_B(\vec{\mu},\sigma_s^2,\sigma_{so}^2,\sigma_m^2) = \\ & -n_B mln(2\pi) - \frac{n_B}{2}\ln[(\sigma_m^2 + \sigma_{so}^2 + m\sigma_s^2)(\sigma_m^2 + \sigma_{so}^2)^{m-1}] \\ & \qquad -\frac{1}{2}\left\{ (a_1 + a_2)\sum_{i=1}^{n_B}\sum_{j=1}^{m}\left(z_{ij}-\mu_j\right)^2 \right. \\ & \qquad \left. + a_3 \sum_{i=1}^{n_B}\left[\sum_{j=1}^{m}\left(z_{ij}-\mu_j\right)\right]^2 \right\} \end{aligned} \qquad [2.6]$$

For an A plan, each observer measures different subjects once, and measurements on all subjects are independent. From model [2.1], we have, for any measurement made by observer j, $Z_{jl} \sim N\left(\mu_j, \sigma_s^2 + \sigma_{so}^2 + \sigma_m^2\right)$ and so, denoting the observed measurements by

$$z_{jl}, j = 1,2,\ldots,m,\ l = 1,2,\ldots,n_A/m,$$

the log-likelihood contribution from the augmented component of plan A is:

$$l_A(\vec{\mu}, \sigma_s^2, \sigma_{so}^2, \sigma_m^2) = \\ -n_A ln(2\pi) - \frac{n_A}{2}\ln(\sigma_s^2 + \sigma_{so}^2 + \sigma_m^2) \\ - \frac{1}{2(\sigma_s^2 + \sigma_{so}^2 + \sigma_m^2)} \sum_{j=1}^{m} \sum_{l=1}^{\frac{n_A}{m}} \left(z_{ij} - \mu_j\right)^2 \qquad [2.7]$$

Because we assume that the subjects used in the two components of the augmented plans are different (and hence independent), the overall log-likelihood for an augmented plan is the sum of the SP log-likelihood and the log-likelihood corresponding to the augmented component. Thus the overall log-likelihood $l(\mu_1, \dots, \mu_m, \sigma_s^2, \sigma_{so}^2, \sigma_m^2)$ is the sum of the contributions from equations [2.5] and [2.7] for plan A and equations [2.5] and [2.6] for plan B.

To calculate the asymptotic standard deviations for any assumed values for the unknown parameters, we find the Fisher information matrix symbolically using Maple [Maplesoft, 2014] to calculate the appropriate second partial derivatives of the overall log-likelihood function. We then take expected values of the sums of squares involving the data in equations [2.5], [2.6] and [2.7], as given below.

$$E\left[\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{r} \left(y_{ijk} - \mu_j\right)^2\right] = nmr(\sigma_s^2 + \sigma_{so}^2 + \sigma_m^2) \qquad [2.8]$$

$$E\left[\sum_{i=1}^{n} \sum_{j=1}^{m} \left\{\sum_{k=1}^{r} \left(y_{ijk} - \mu_j\right)\right\}^2\right] = nmr(r\sigma_s^2 + r\sigma_{so}^2 + \sigma_m^2) \qquad [2.9]$$

$$E\left[\sum_{i=1}^{n} \left\{\sum_{j=1}^{m} \sum_{k=1}^{r} \left(y_{ijk} - \mu_j\right)\right\}^2\right] = nmr(mr\sigma_s^2 + r\sigma_{so}^2 + \sigma_m^2) \qquad [2.10]$$

and for plan A,

$$E\left[\sum_{i=1}^{\frac{n_A}{m}} \sum_{j=1}^{m} \left(z_{ij} - \mu_j\right)^2\right] = n_A(\sigma_s^2 + \sigma_{so}^2 + \sigma_m^2) \qquad [2.11]$$

and for plan B,

$$E\left[\sum_{i=1}^{n_B}\sum_{j=1}^{m}\left(z_{ij}-\mu_j\right)^2\right]=n_B m(\sigma_s^2+\sigma_{so}^2+\sigma_m^2) \qquad [2.12]$$

$$E\left[\sum_{i=1}^{n_B}\left\{\sum_{j=1}^{m}\left(z_{ij}-\mu_j\right)\right\}^2\right]=n_B m(m\sigma_s^2+r\sigma_{so}^2+\sigma_m^2) \qquad [2.13]$$

We invert the Fisher Information matrix numerically using Matlab [The MathWorks Inc., 2013]. This gives the asymptotic variances of $\tilde{\mu}_j$, $j=1,2,\dots,m$, and $\tilde{\sigma}_s^2$, $\tilde{\sigma}_{so}^2$ and $\tilde{\sigma}_m^2$. But because we are interested only in γ, σ_s^2, σ_{so}^2 and σ_m^2 we apply the Delta Method [Lehmann and Casella, 1998] and pre- and post-multiply the inverse of the information matrix by a change-of-variables matrix of suitable partial derivatives:

$$D=\frac{\partial(\gamma,\sigma_s^2,\sigma_{so}^2,\sigma_m^2)}{\partial(\mu_1,\dots,\mu_m,\sigma_s^2,\sigma_{so}^2,\sigma_m^2)}$$

Again, we use Maple [Maplesoft, 2014] to calculate these partial derivatives and Matlab [The MathWorks Inc., 2013] to find their numerical values. The square root of the (1, 1) element of the resulting matrix gives the asymptotic standard deviation of $\tilde{\gamma}$, which we use to rank plans.

When we consider the special case where we assume there is no subject-by-observer interaction we set $\sigma_{so}^2=0$ in the above calculations up to the stage of finding the partial derivatives of the log-likelihood function. We then proceed as before, except we purge the appropriate row and column corresponding to σ_{so}^2 from the information matrix and the change-of-variables matrix D. We also use this calculation for the case $r=1$ in the SP component for both type A and B plans, where σ_m^2 is now the sum of the repeatability and subject-by-observer components of the variation. When we consider the case with a single observer (or no observer effects), we set $m=1$ and $\sigma_{so}^2=0$ and $\sigma_o^2=0$ and alter the information matrix and the matrix D accordingly. Recall that in these two cases the definition of γ simplifies to [2.4] and [2.3], respectively.

2.1.1 An Important Property of the Fisher Information Matrix

Without displaying the information matrix explicitly, we note one of its properties that has important consequences. Because we can calculate the overall information by summing the components for each subject, a scale change in the number of subjects in a standard plan or in an augmented plan (A or B) produces the same scale change in the information that then acts inversely on the asymptotic variance. As such the efficiency of an SP with n subjects compared with an AP with $n + n_A$ subjects is the same as the efficiency of an SP with λn subjects compared with an AP of $\lambda(n + n_A)$ subjects.

To see this invariance property let FI_{SP} denote the Fisher Information matrix for a standard plan with n subjects, and let FI_A denote the Fisher Information matrix for an augmented plan A with $n + n_A$ subjects. For simplicity this demonstration compares an SP to a type A augmented plan, but we could replace A's by B's in what follows and the result applies equivalently to the comparison of an SP and a type B augmented plan.

Because we assume subjects are independent, $FI_{SP} = \sum_{i=1}^{n} FI_{SP_i}$ where FI_{SP_i} is the Fisher Information matrix for subject $i = 1, \dots, n$. From equation [2.5] and the expected sums of squares [2.8-2.10] we can see that FI_{SP_i} is the same for all i, and so we have $FI_{SP} = nFI_{SP_i}$.

Similarly, the Fisher Information matrix for a type A augmented plan is $FI_A = \sum_{i=1}^{n} FI_{A1_i} + \sum_{i=1}^{n_A} FI_{A2_i}$, where FI_{A1_i} is the Fisher Information matrix for one of the n subjects in the SP component, and FI_{A2_i} is the Fisher Information matrix for one of the n_A subjects in the augmented component. As before, the Fisher Information matrix is the same for all subjects in the SP component, and from equation [2.7] and the expected sums of squares in [2.11] we see that the Fisher Information matrix will be the same for all subjects in the augmented component as well. Thus we have $FI_A = nFI_{A1_i} + n_A FI_{A2_i}$.

The asymptotic variance of $\tilde{\gamma}$ associated with an SP is found by pre- and post-multiplying the inverse ${FI_{SP}}^{-1}$, by the vector of partial derivatives:

$$D = \frac{\partial \gamma}{\partial(\mu_1, \dots, \mu_m, \sigma_s^2, \sigma_{so}^2, \sigma_m^2)}$$

yielding

$$D{FI_{SP}}^{-1}D^T = D\left[nFI_{SP_i}\right]^{-1}D^T$$

Similarly, the asymptotic variance of $\tilde{\gamma}$ associated with an augmented plan A is:

$$DFI_A{}^{-1}D^T = D\left[nFI_{A1_i} + n_A FI_{A2_i}\right]^{-1}D^T$$

As such, the efficiency, which we have defined as the asymptotic standard deviation of $\tilde{\gamma}$ associated with the SP divided by the asymptotic standard deviation from the AP, is given by:

$$\sqrt{\frac{D\left[nFI_{SP_i}\right]^{-1}D^T}{D\left[nFI_{A1_i} + n_A FI_{A2_i}\right]^{-1}D^T}}$$

We can see explicitly that increasing the number of subjects in both plans by a factor of λ, i.e. from n to λn for the SP, and from $n + n_A$ to $\lambda(n + n_A)$ for the AP, does not change the efficiency. To see this, consider substituting λn for n in the numerator, and λn for n and λn_A for n_A in the denominator of the expression above. In doing this, there is a common factor of the scalar λ in both the numerator and denominator which cancels out, resulting in the same efficiency as when the SP has n subjects and the AP has $n + n_A$ subjects.

So, for example, if we have $m = 3$ observers, the relative efficiency of $A(5,2,30)$ to $SP(10,2)$ (two plans each with 60 measurements) is the same as the relative efficiency of $A(7,2,42)$ compared with $SP(14,2)$ (two plans each with 84 measurements). Here, the number of subjects has increased by a factor of $\lambda = 1.4$ from one set of plans to the next, but the relative efficiency within a set does not change. We illustrate the benefits of this result in Section 2.2.

2.2 Comparison of Plans

In this section we compare augmented and standard plans on the basis of their ability to estimate γ precisely in three situations characterized by the number of observers and the incorporation of subject-by-observer interaction. First we compare plans in the context of an automated measurement system, or equivalently, a measurement system with only one observer. In this case there is no way to examine an observer effect and no possibility of a subject-by-observer interaction. Next we compare plans for measurement systems with multiple observers, but no subject-by-observer interaction. And lastly we compare plans for a measurement system with multiple observers and we include the possibility of a subject-by-observer interaction effect.

To compare plans, we suppose that the total number of measurements N and the number of observers m are fixed. Here we consider values of N between 60 and 100 with $1 \leq m \leq 4$. With each combination, we examine each possible SP (integer values for n, r such that $N = nmr$), plan A (integer values for n, r, n_A such that $N = nmr + n_A$), and plan B (integer values for n, r, n_B such that $N = m(nr + n_B)$). Note that for a given AP and SP with N measurements, the number of subjects n, and the number of replicate measurements r, will not be the same.

We then substitute a range of possible values for the unknown parameters and rank all possible plans according to the asymptotic standard deviation of $\tilde{\gamma}$. Because γ is defined as the square root of a ratio of variances, with no loss of generality we can set $\sigma_t^2 = \sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2 = 1$. Note throughout the comparisons of augmented plans, we use the asymptotic standard deviation of the estimator $\tilde{\gamma}$ (not its variance) from the best standard plan at the particular parameter values as the basis to calculate relative efficiency.

2.2.1 Plans with One or No Observer

Many measurement systems are automated with no observer effects. This also corresponds to a system with a single observer. In our formulation of the problem, we then have $m = 1$, $\sigma_o = 0$, $\sigma_{so} = 0$, and so γ, the parameter of interest, simplifies to [2.3]. Also with $m = 1$, augmented plans A and B are equivalent. Both augmented plan types start with a standard plan with n subject each measured r times. Then, in the augmented component, we measure an additional n_A (or n_A) subjects once.

Suppose $N = 60$ and the true value of γ equals 0.3. In Table 2.1, we list the best four plans in increasing order of the asymptotic standard deviation of γ. For purposes of comparison, we also include the standard plan with $n = 10$ subjects, each measured $r = 6$ times, which is the recommendation by the Automotive Industry Action Group [2010] in this situation.

Plan	SE($\hat{\gamma}$)	SE($\hat{\sigma}_m$)	Relative Efficiency
$SP(30,2)$	0.0523	0.0387	1.00
$A(29,2,2)$	0.0525	0.0394	1.00
$A(28,2,4)$	0.0527	0.0401	0.99
$A(16,3,12)$	0.0529	0.0375	0.99
$SP(10,6)$	0.0680	0.0300	0.77

Table 2.1: Five plans for estimating γ when $m = 1$, $N = 60$ and $\gamma = 0.3$

The best plan is $SP(30{,}2)$, the Shainin proposal for an Isoplot$^{\text{TM}}$ study [Shainin, 1992]. Not surprisingly, we see similar results (not presented here) for other values of N and γ. In general, when there are no observer effects, it is best to use a standard plan that balances the degrees of freedom for estimating σ_s and σ_m by minimizing the number of replicate measurements, i.e., choosing $r = 2$. There is a substantial improvement (23%) in precision when estimating γ by the Shainin plan versus the default AIAG standard plan with 10 subjects. Because of the scaling property discussed in Section 2.1.1, for example, we also have a 23% improvement using $SP(45{,}2)$ over $SP(15{,}6)$ when we increase the number of subjects by a factor of $\lambda = 1.5$ and $N = 90$.

Thus we recommend that when a measurement system is automated, or has just one observer, and a total of N measurements can be taken in the study, $n = N/2$ subjects be randomly sampled and measured $r = 2$ times each. It is important to remark that this is very different from the AIAG recommendation.

Augmentation provides no benefit here. As such we can use standard ANOVA methods to analyze the data from the recommended standard plan, instead of maximum likelihood estimation. In Section 1.2 we described how to estimate γ with the ANOVA method when the measurement system had multiple observers and a subject-by-observer interaction was present. The ANOVA method in the single-observer scenario, which is a special case of this, is described below.

In this case, the total variability is partitioned into variability due to subjects and variability due to the measurement system, and the total sum of squares is decomposed as follows:

$$SS_t = SS_s + SS_m$$

Table 2.2 displays the ANOVA table associated with this decomposition, where $\bar{Y}_{\cdot\cdot} = \sum_{i=1}^{n}\sum_{k=1}^{r} Y_{ik}/nr$, and $\bar{Y}_{i\cdot} = \sum_{k=1}^{r} Y_{ik}/r$. The expected mean squares in this case are given by $E(MS_s) = \sigma_m^2 + r\sigma_s^2$ and $E(MS_m) = \sigma_m^2$. The estimates $\hat{\sigma}_s^2$ and $\hat{\sigma}_m^2$ are obtained by simultaneously solving these equations and substituting the observed means squares in place of the expected, which gives

$$\hat{\sigma}_s^2 = \frac{MS_s - MS_m}{r} \qquad \hat{\sigma}_m^2 = MS_m$$

Note that the mean squares used to calculate these estimates are constructed using the observed data. These estimates are then substituted into [2.3] to obtain $\hat{\gamma}$.

Source	Sum of Squares	df	Mean Square	F-Statistic
S	$SS_s = r\sum_{i=1}^{n}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$n-1$	$MS_s = \dfrac{SS_s}{n-1}$	MS_s/MS_m
M	$SS_m = \sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{ik} - \bar{Y}_{i\cdot})^2$	$n(r-1)$	$MS_m = \dfrac{SS_m}{n(r-1)}$	
Total	$SS_t = \sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{ik} - \bar{Y}_{\cdot\cdot})^2$	$nr-1$		

Table 2.2: ANOVA table for a one-factor random effect model

2.2.2 Plans with More than One Observer and No Subject-by-Observer Interaction

Now suppose we have more than one observer and we assume that there is no subject-by-observer interaction. We consider three cases where the number of observers is $m = 2$, 3, or 4, each with two values of N close to 60 and 96. We use $N = 64$ for $m = 4$ so that there is a large number of possible augmented plans. We specify three values of $\gamma =$ 0.5, 0.3, and 0.1, corresponding to a poor, acceptable, and good measurement system. Because γ is given by [2.4] in this case, we set $\sigma_s^2 + \sigma_o^2 + \sigma_m^2 = 1$ (without loss of generality because γ is not changed by a scale change) so that $\sigma_o^2 + \sigma_m^2 = \gamma^2$. We then specify

$$\delta = \frac{\sigma_m^2}{\sigma_o^2 + \sigma_m^2} \qquad [2.14]$$

where $\delta =$ 0.1, 0.5, 0.9, to look at situations when the repeatability (σ_m^2) makes up a small, medium, or large proportion of the overall measurement system variation as captured by γ. Algebraically we have $\sigma_m^2 = \delta\gamma^2$.

For each value of m, N, and the nine pairs of values for γ and δ, we rank all possible SP, A, and B plans using the asymptotic standard deviation associated with the ML estimator of γ as described in Section 2.1. For example, with $m = 2$ and $N = 60$, there are 103 plans of type A and B and 8 standard plans. Table 2.3 presents a comparison of the best plans of each type when $m = 2$, $N = 60$, $\gamma = 0.3$, and $\delta = 0.1$ (σ_m is relatively small compared with σ_o). We also include the widely used standard plan $SP(10,3)$, recommended by the Automotive Industry Action Group [2010].

Plan	SE($\hat{\gamma}$)	SE($\hat{\sigma}_m$)	SE($\hat{\sigma}_o$)	Relative Efficiency
$A(5,2,40)$	0.0347	0.0173	0.0210	1.07
$B(2,2,26)$	0.0383	0.0119	0.0122	0.97
$SP(30,1)$	0.0371	0.0122	0.0122	1.00
$SP(10,3)$	0.0621	0.0095	0.0122	0.60

Table 2.3: Four plans with $m = 2$, $N = 60$, $\gamma = 0.3$ and $\delta = 0.1$

Here there is a small gain over the best SP (about 7% reduction in standard error) in estimating γ with an augmented plan A with 5 subjects measured twice by both observers and then two sets of 20 subjects measured once by each by each observer, i.e., the $A(5,2,40)$ plan. As well, all of the best plans are substantially better than the SP with $n = 10$. We also note that the plan $A(5,2,40)$ has the same 7% gain over the best SP when $\delta = 0.1$ for other values of γ. However, for $\delta = 0.5$ or 0.9, the plan $A(5,2,40)$ is 12% to 25% less efficient than the best SP. In this case, we recommend the $SP(30,1)$ plan.

We see a similar pattern for $m = 3$. For example, when $N = 60$, in all cases except when $\gamma = 0.5$ and $\delta = 0.9$, there is a plan A that is superior to the best SP. Unfortunately, the best plan A varies as the parameters are changed. When $\delta = 0.1$, the plan $A(3,2,42)$ has a relative efficiency of about 1.26. However for larger values of δ, this plan is 2% to 12% less efficient than the best SP. Accordingly, with $m = 3$, we recommend the $SP(20,1)$ plan.

Plan	SE($\hat{\gamma}$)	SE($\hat{\sigma}_m$)	SE($\hat{\sigma}_o$)	Relative Efficiency
$A(4,2,32)$	0.0456	0.0283	0.0366	1.18
$B(2,2,12)$	0.0567	0.0212	0.0265	0.95
$SP(16,1)$	0.0537	0.0217	0.0265	1.00
$SP(8,2)$	0.0720	0.0200	0.0265	0.75

Table 2.4: Comparison of plans with $m = 4$, $N = 64$, $\gamma = 0.3$ and $\delta = 0.5$

In Table 2.4, we present a second example with $m = 4$, $N = 64$, $\gamma = 0.3$, $\delta = 0.5$. Here there is an 18% gain in estimating γ by using the best plan A rather than the best SP. With 4 observers, as shown by the right-most column of Table 2.5, the plan $A(4,2,32)$ does well over the entire parameter space compared with the best standard plan $SP(16,1)$. There is significant improvement relative to the best SP in all cases except when $\gamma = 0.5$ (the measurement system is highly variable) and $\delta = 0.9$ (most of the measurement variability is due to repeatability, i.e.,$\sigma_m \gg \sigma_o$). In this case, there is no loss in efficiency. Also (not shown here), there is no material difference between $A(4,2,32)$ and the best plan A for any value of γ and δ in our array.

Because these comparisons are based on an asymptotic criterion, we also checked the relative efficiency of $A(4,2,32)$ versus $SP(16,1)$ using simulation. We generated 10,000 samples for each plan using all pairs of values for $\gamma = 0.5$, 0.3, 0.1 and $\delta = 0.1$, 0.5, 0.9. We then calculated the maximum likelihood estimates of the parameters and their associated asymptotic standard deviations for each sample. We provide a summary of these estimates for both plans in Table 2.5. In this table we define the 'Bias' to be the absolute difference between the true value of γ and the average of the 10 000 estimates. We similarly define the 'Standard Deviation' to be standard deviation of the 10 000 estimates of γ. The ratio of these values for the two plans defines the simulated efficiency which is to be compared with the theoretical efficiency (that is based on asymptotic calculations).

		$A(4,2,32)$		$SP(16,1)$			
γ	δ	Bias	Standard Deviation	Bias	Standard Deviation	Simulated Efficiency	Theoretical Efficiency
0.5	0.1	0.011	0.052	0.025	0.072	1.38	1.37
0.5	0.5	0.007	0.070	0.022	0.080	1.14	1.12
0.5	0.9	0.009	0.086	0.022	0.086	1.00	0.99
0.3	0.1	0.009	0.038	0.021	0.057	1.50	1.39
0.3	0.5	0.007	0.048	0.020	0.060	1.25	1.17
0.3	0.9	0.005	0.055	0.012	0.065	1.18	1.09
0.1	0.1	0.003	0.014	0.008	0.022	1.57	1.39
0.1	0.5	0.002	0.017	0.008	0.023	1.35	1.20
0.1	0.9	0.001	0.018	0.007	0.024	1.33	1.14

Table 2.5: Simulated and Theoretical Comparison of $A(4,2,32)$ and $SP(16,1)$ for $m = 4$, $N = 64$

The plan $A(4,2,32)$ provides a less biased estimate of γ with smaller standard deviation over all of the parameter values except when $\gamma = 0.5$, $\delta = 0.9$, as predicted by the theoretical information calculations. In all cases, the simulated efficiency is higher than predicted by the asymptotic calculations.

We investigated and compared the best plans of each type for $m = 2, 3, 4$ and $N \approx 60, 90$ with γ and δ as described above. To save space, we do not present all of the results here. We do however draw the following general conclusions when there is no subject-by-observer interaction based on this empirical investigation:

- The best standard plans have $r = 1$ and $n = N/m$. That is, we maximize the number of subjects in the study. We can justify this conclusion by noting that maximizing the number of subjects maximizes the degrees of freedom for estimating the between-subject variation and, because we are using the subject-by-observer sum of squares to estimate σ_m^2, increasing the number of subjects also increases the degrees of freedom for estimating σ_m^2. And because each subject is measured by each observer, subjects act as blocks, so we also get good estimates of the observer means $\mu_1, \dots, \mu_m$ and hence σ_o^2 by increasing the number of subjects. Note that this conclusion is contrary to the AIAG [2010] recommended plans (see sample forms pp. 224-225) that suggest setting $n = 10$, and $r = 2$ or 3.
- In all cases, the best standard plan is superior to the best plan B.
- The best augmented plans have $r = 2$ and use a small number of subjects in the SP component.
- For two or three observers, augmentation provides little gain unless σ_m is relatively small compared with σ_o. If $\delta = 0.1$ and $m = 2, 3$, the best plan A is about 6% ($m = 2$) and 20% ($m = 3$) more efficient in estimating γ than the best SP. These results are independent of N in the range $60 < N < 100$.
- With four observers and $N = 64$, the plan $A(4,2,32)$ is (almost) uniformly better than the best SP and the gains in efficiency are relatively large when $\delta \leq 0.5$. For any value of N, we can scale this plan, according to Section 2.1.1, and see the same gains in efficiency. For example, if $N = 96$, the plan $A(6,2,48)$ has the same good properties.
- The simulated results show that the asymptotic calculations are conservative. The actual efficiency of the recommended augmented plans is better than predicted by these calculations

2.2.3 Plans with More than One Observer and Possible Subject-by-Observer Interaction

Now we consider a measurement system with two or more observers in which we allow for the possibility of subject-by-observer interaction. We proceed as in the case with no interaction with $m = 2$, 3, 4 and $N \approx 60$, 90. There are two added complications. First, we have an extra parameter, σ_{so}, and γ is now given by equation [2.2]. We set $\sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2 = 1$ without loss of generality so that $\gamma^2 = \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2$. The first two terms are due to observer-to-observer differences, so, for a given value of γ, we look at three cases of

$$\delta = \frac{\sigma_m^2}{\sigma_o^2 + \sigma_{so}^2 + \sigma_m^2} \qquad [2.15]$$

with $\delta = 0.1$, 0.5, 0.9, so the repeatability contribution to the measurement system variation (repeatability and reproducibility) is relatively small to large. Algebraically we have $\sigma_m^2 = \delta\gamma^2$. In this situation we also define

$$\beta = \frac{\sigma_o^2}{\sigma_o^2 + \sigma_{so}^2} \qquad [2.16]$$

to be the proportion of the overall observer contribution to the measurement system variation that is attributable to σ_o^2. Then, for given values of γ and δ, we consider three cases with $\beta =$ 0.1 0.5, 0.9, so the contribution of σ_o^2 is a relatively small to large proportion of $(\sigma_o^2 + \sigma_{so}^2)$. Algebraically, we have $\sigma_o^2 = \beta(1-\delta)\gamma^2$ and $\sigma_{so}^2 = (1-\beta)(1-\delta)\gamma^2$.

The second complication is that we must decide if we are going to entertain plans with $r = 1$. In this case, we cannot separately estimate σ_{so}^2 and σ_m^2, but we can estimate $\sigma_{so}^2 + \sigma_m^2$ and hence γ. Including $r = 1$ plans is equivalent to assuming that there is no interaction (or more accurately, that any interaction is subsumed by the repeatability σ_m^2), so that we should compare $r = 1$ plans with those alternatives considered in the previous subsection. Here we do not allow $r = 1$ plans, so that we can get separate estimates of σ_{so}^2 and σ_m^2. Obtaining a separate estimate of σ_{so}^2 allows the practitioner to clearly judge the magnitude of a subject-by-observer interaction and allows them to respond appropriately. This information is also useful in the planning of future MSA studies, where it may be beneficial to know beforehand whether such an interaction exists. As we will see, design recommendations differ depending on the existence and size of a subject-by-observer interaction.

For given values of N and m, we look at the best plans of each type as we consider the 27 combinations of γ, δ, and β. To illustrate, Table 2.6 gives the best plans of each type when $m = 2$, $N = 60$, $\gamma = 0.3$, $\delta = 0.5$ (the repeatability σ_m^2 is the same as the reproducibility $\sigma_o^2 + \sigma_{so}^2$), and $\beta = 0.5$ (observer and subject-by-observer effects are equal). For the sake of comparison, we also include the AIAG [2010] recommended $SP(10,3)$.

Plan	SE($\hat{\gamma}$)	SE($\hat{\sigma}_m$)	SE($\hat{\sigma}_{so}$)	SE($\hat{\sigma}_o$)	Relative Efficiency
$B(2,2,26)$	0.0494	0.0713	0.1097	0.0341	1.23
$A(11,2,16)$	0.0552	0.0320	0.0678	0.0445	1.10
$SP(15,2)$	0.0607	0.0274	0.0581	0.0387	1.00
$SP(10,3)$	0.0713	0.0237	0.0570	0.0433	0.85

Table 2.6: Comparison of plans with $m = 2$, $N = 60$, $\gamma = 0.3$, $\delta = 0.5$, and $\beta = 0.5$

Compared with the best standard plan, the asymptotic standard error for estimating γ is about 23% smaller for the best plan B and about 10% smaller for the best plan A. In the best plan B, we use only two subjects with two replicate measurements by each observer in the SP component. In the augmented component, we have a large number of subjects ($n_B = 26$) measured once by each observer. For this plan, the estimates of σ_{so}^2 and σ_m^2 are highly correlated because most of the information is about their sum. The plan $SP(10,3)$ is much less efficient than both the best plan B and best plan A. We found that a plan B with 2 subjects measured twice by each of the two observers in the SP component was uniformly the best plan. This result is not surprising because this plan is very close to the corresponding standard plan with $r = 1$ that is more efficient for estimating γ but cannot separately estimate σ_{so}^2 and σ_m^2.

We also include the results of a simulation with 10,000 samples to demonstrate how well the asymptotic calculations rank the plans. Table 2.7 displays the results of comparing $B(2,2,26)$ with the best standard plan $SP(15,2)$ when $N = 60$ and $m = 2$. We display the results only for $\beta = 0.5$ because both the simulated and theoretical calculations do not depend significantly on β.

We see that the augmented plan provides a less biased estimate of γ with smaller standard deviation, across all nine combinations of γ and δ. Like the no-interaction case (see Section 2.2.2) the actual efficiency is always larger than predicted by the asymptotic calculations.

		$B(2,2,26)$		$SP(15,2)$			
γ	δ	Bias	Standard Deviation	Bias	Standard Deviation	Simulated Efficiency	Theoretical Efficiency
0.5	0.1	0.011	0.069	0.022	0.095	1.37	1.33
0.5	0.5	0.012	0.077	0.024	0.093	1.21	1.20
0.5	0.9	0.013	0.082	0.030	0.093	1.13	1.11
0.3	0.1	0.009	0.049	0.020	0.068	1.39	1.33
0.3	0.5	0.011	0.051	0.022	0.070	1.37	1.23
0.3	0.9	0.009	0.056	0.027	0.068	1.21	1.16
0.1	0.1	0.005	0.018	0.007	0.025	1.39	1.34
0.1	0.5	0.004	0.019	0.006	0.025	1.32	1.24
0.1	0.9	0.003	0.019	0.008	0.026	1.37	1.17

Table 2.7: Simulated and Theoretical Comparison of $B(2,2,26)$ and $SP(15,2)$ for $m = 2$, $N = 60$, $\beta = 0.5$

For $m = 3$ and 4, we see a very different behavior. In this case, there are a number of type A plans that are (almost) uniformly better than the best standard plans and always better (over our grid of parameter values) than any type B plan. In Table 2.8, we show the relative efficiencies of a few type A plans compared with the best SP. There are significant improvements possible over the best standard plan. We checked these results (not shown here) using a simulation with 10,000 runs. For example, when $m = 3$, $N = 60$, we compared $A(6,2,24)$ to $SP(10,2)$ over the complete grid of values for γ, δ, and β. In all cases, there is less bias with the augmented plan and the estimated efficiencies are substantially higher than those predicted by the asymptotic calculations.

			$m = 3, N = 60$			$m = 4, N = 64$		
γ	δ	β	$A(5,2,30)$	$A(6,2,24)$	$A(7,2,18)$	$A(4,2,32)$	$A(5,2,24)$	$A(6,2,16)$
0.5	0.1	0.1	0.99	1.04	1.07	1.10	1.15	1.16
0.5	0.1	0.5	1.17	1.19	1.19	1.31	1.32	1.29
0.5	0.1	0.9	1.51	1.47	1.40	1.70	1.63	1.49
0.5	0.5	0.1	1.05	1.09	1.11	1.17	1.21	1.20
0.5	0.5	0.5	1.15	1.18	1.18	1.29	1.30	1.27
0.5	0.5	0.9	1.28	1.28	1.26	1.44	1.42	1.35
0.5	0.9	0.1	1.09	1.12	1.13	1.21	1.24	1.22
0.5	0.9	0.5	1.10	1.14	1.14	1.23	1.26	1.23
0.5	0.9	0.9	1.12	1.15	1.15	1.25	1.27	1.24
0.3	0.1	0.1	1.08	1.12	1.15	1.21	1.26	1.26
0.3	0.1	0.5	1.22	1.25	1.24	1.37	1.39	1.35
0.3	0.1	0.9	1.53	1.50	1.43	1.73	1.66	1.52
0.3	0.5	0.1	1.16	1.20	1.20	1.31	1.35	1.32
0.3	0.5	0.5	1.24	1.27	1.26	1.41	1.42	1.36
0.3	0.5	0.9	1.36	1.36	1.33	1.53	1.51	1.43
0.3	0.9	0.1	1.22	1.25	1.24	1.38	1.40	1.35
0.3	0.9	0.5	1.24	1.26	1.25	1.40	1.41	1.36
0.3	0.9	0.9	1.25	1.27	1.26	1.41	1.42	1.37
0.1	0.1	0.1	1.12	1.17	1.19	1.27	1.32	1.31
0.1	0.1	0.5	1.23	1.27	1.27	1.39	1.42	1.38
0.1	0.1	0.9	1.54	1.51	1.44	1.73	1.68	1.54
0.1	0.5	0.1	1.21	1.25	1.25	1.38	1.41	1.38
0.1	0.5	0.5	1.28	1.31	1.30	1.45	1.47	1.41
0.1	0.5	0.9	1.39	1.39	1.36	1.57	1.55	1.47
0.1	0.9	0.1	1.29	1.31	1.30	1.47	1.49	1.42
0.1	0.9	0.5	1.30	1.32	1.31	1.48	1.49	1.42
0.1	0.9	0.9	1.31	1.33	1.32	1.50	1.50	1.43

Table 2.8: Efficiencies of some good type A plans when $m = 3$, $N = 60$ and $m = 4$, $N = 64$ relative to $SP(10,2)$ and $SP(8,2)$, the best standard plans when $m = 3$, $N = 60$ and $m = 4$, $N = 64$, respectively

We summarize our findings when we allow for the possibility of subject-by-observer interaction as follows:

- The best plans have $r = 2$ in the SP component. That is, there are minimal replicate measurements by the same observer on the same subject.
- The best plans use few subjects in the SP component and a large number of subjects in the augmented components.
- With $m = 2$ observers, a good plan for estimating γ is $B(2,2,28)$ when $N = 64$ or a scaled version for other values of N. For example, if $N = 96$, the scaled version is $B(3,2,42)$. Note that the best plan in this case is close to $SP(30,1)$, so the estimates of σ_{so}^2 and σ_m^2 are highly correlated.
- With $m = 3$ or 4 observers, there are good type A plans, e.g., $A(6,2,24)$ for $m = 3$, $N = 60$ and $A(5,2,24)$ for $m = 4$, $N = 64$, with relatively few subjects in the SP component. We can realize substantial benefits in estimating γ using one of these plans, or scaled versions for other values of N.
- Simulation results suggest that the asymptotic results are conservative. Actual efficiencies of the good augmented plans are higher than predicted.

We provide software at http://www.bisrg.uwaterloo.ca/ that can be used to select and compare good plans using the asymptotic calculations. In any given situation, a practitioner can investigate a wide variety of potential plans and select one that meets his or her needs. At the same website, we also provide software that will calculate the maximum likelihood estimate and the standard error for γ (and other parameters) for either type of augmented plan, given the data.

2.3 Discussion and Conclusions

The idea of augmented plans raises several design and analysis issues.

In many situations, augmented assessment plans provide a means to estimate γ more efficiently than the best standard plan with the same number of observers and total measurements. One drawback of the type A augmented plans is that we cannot use the ANOVA method of variance component estimation because the design is unbalanced in the sense that some subjects are only measured by one of the observers. With type B augmented plans however, we can apply the results of Chapter 7 in Burdick et al. [2005] to get approximate confidence intervals. However, in such an unbalanced design, the properties of the ANOVA-based estimates of γ can be examined only by simulation and so are not useful in the planning stage of the study. The easily calculated Fisher information is a convenient basis for comparison of plans and, with maximum likelihood estimation, provides a method of analysis. To derive approximate confidence intervals for γ, we suggest using the asymptotic standard errors which are found by substituting the maximum likelihood estimates into the asymptotic standard deviations developed in Section 2.1. A practitioner may use the available software recommended in Section 2.2.3 to calculate this asymptotic standard error for a given set of data. We have not explored the properties of such approximate confidence intervals; this stands as a possible extension to be pursued.

Recall in Section 1.3 we discussed a series of papers in which Browne et al. [2009a, 2009b, 2010], in a manufacturing context, consider the use of leveraging to increase the efficiency of standard plans. In these plans, the order of the two components is important. In the first stage of a leveraged plan, each observer measures a separate set of subjects once. Then a standard plan is carried out using extreme subjects selected from those measured in stage 1. Browne et al. do not consider the possibility of a subject-by-observer interaction. Note that the leveraged plans use fewer subjects, so it is not clear how their performance compares with the augmented plans described here. This is another issue for future investigation.

Augmented plans, on the other hand, are not sequential. We can carry out the components in any order. Another possibility is to carry out the SP component first and then select an augmented component based on a preliminary analysis of the SP data. Such a design may have superior performance over the augmented plans recommended here.

The results of the simulations were somewhat surprising. Good plans (as ranked by asymptotic standard deviation of the estimator $\tilde{\gamma}$) were typically close to unbiased and the actual standard deviations in the simulations were larger for both the augmented and corresponding standard plans but the efficiencies of the augmented plans were larger than those predicted by the asymptotic calculations.

The idea of augmentation is to use more subjects, consistent with the recommendation of Burdick and Larsen [1997]. Typical standard plans with only 10 subjects do not provide sufficient information to adequately estimate σ_s and hence γ. The plans we recommend all increase the number of subjects relative to the standard plans recommended by the Automotive Industry

Action Group [2010], and reduce the number of replicate measurements on the same subject by each observer. That is, we recommend only plans with $r = 1$ or 2. Also note that the recommended augmented plans are almost uniformly more efficient than the best SP over a wide range of the parameter values. We summarize our recommended plans as follows:

1. For a system with a single observer or with no observer effects, use the standard plan with $N/2$ subjects each measured r =2 times.
2. If you are willing to assume no subject-by-observer interaction or if you are willing to confound estimation of the subject-by-observer interaction and the measurement system repeatability,
   a. For $m = 2$ or 3 observers, use the standard plan with $r =$ 1 to maximize the number of subjects in the study.
   b. For m = 4 observers, use a type A plan with r = 2 and a small number of subjects in the SP component. For example, use $A(4,2,32)$ if N = 64, and a scaled version of this plan for other values of N.
3. If you wish to include the possibility of subject-by-observer interaction (and wish to separately estimate σ_{so}^2 and σ_m^2),
   a. For m = 2 observers, use a type B plan with a small number of subjects in the SP component, e.g., $B(2,2,(N - 8)/2)$.
   b. For more observers, use a type A plan with r = 2. For m = 3, use a scaled version of $A(6,2,24)$, and for $m =$ 4, use a scaled version of $A(5,2,24)$ with scaling depending on the ratio $N/60$ and $N/64$, respectively.
4. If you have some knowledge of the possible parameter values γ, δ, β, the software provided at http://www.bisrg.uwaterloo.ca/ can be used to investigate a number of plans over the restricted range of parameter values, and that can be used to analyze data from an augmented MSA study.

It is clear that increasing the number of subjects in an MSA study increases the precision for estimating γ. For a fixed number of measurements N, augmented plans look for alternate allocations of those measurements, in terms of the number of subjects and number of replicates, that yield more precise estimates of γ. Simply put, the more subjects included in the study, the better. However, the constraint that at most N measurements can be made in the study, prevents us from investigating the effect of having information from a very large number of subjects.

However, in many contexts the measurement system being assessed is used routinely, and records of the single measurements from day-to-day use are kept. This information is, in a sense, free; it does not cost any extra money, time, or man-power to obtain, and so it should be incorporated into the assessment of the measurement system. By incorporating this information into the analysis of an assessment study, we effectively increase the number of subjects being studied, and more precise estimates of σ_s^2 and γ result [Stevens et al., 2013]. If we have a lot of this historical, “baseline” data, the number of subjects and measurements in the MSA study itself can be reduced. We investigate the effect of incorporating such baseline information into the planning and analysis of MSA studies, in Chapter 3.

Chapter 3

Incorporating Baseline Data into the Assessment of a Measurement System

In Chapter 1, we introduced the idea of a measurement system assessment (MSA) study to assess the variability of the measurement system. We refer to the typical design of an assessment study as a standard plan (SP) in which n randomly selected subjects are measured r times by each of m observers, for a total of $N = nmr$ measurements. For a fixed number of observers m, we denote a standard plan with n subjects and r replicate measurements by $SP(n, r)$. Recall that the Automotive Industry Action Group [2010] suggest $n = 10$; m =2,3; $r = 2{,}3$ so $40 \leq N \leq 90$, and when a measurement system is automated, or has one observer, $n = 10$; $r = 6$ so $N = 60$.

In Chapter 2 we considered altering the design of the study to more precisely estimate the gauge repeatability and reproducibility (GR&R) ratio γ given by [2.2]. We introduced augmented designs [Stevens et al., 2010] which modify the allocation of measurements in the study. For a fixed total number of measurements N, the augmented plans included more subjects with fewer replicate measurements, and in many scenarios provided more precise estimates of γ than the best SP with the same total number of measurements.

In this chapter, we consider assessing a measurement system that is used routinely and that has a record of single measurements on many subjects from regular use. We call these measurements the *baseline data*, which we can and should incorporate into both the planning and analysis of an MSA study [Stevens et al., 2013].

In Section 1.3 we briefly mentioned that the use of baseline information has been recommended in the literature, but it does not seem to be well studied. Danila et al. [2008, 2010] quantified the substantial advantage of using available information in the assessment of a binary measurement system. For continuous measurements, Steiner and MacKay [2005] suggested incorporating available baseline information into the MSA analysis when there are no observer effects. However, in the analysis they assumed that the total variation (due to both the measurement system and between-subject variation) is known rather than estimated, and did not quantify the effect of their proposal. In a manufacturing context, Browne et al. [2009] quantified the gains in power for testing a hypothesis about the intraclass correlation ρ in the special case when the process mean and standard deviation are known. Minitab [Minitab Inc., 2013] allows users to specify a "historical standard deviation" in an R&R analysis. The Automotive Industry Action Group [2010, p. 121] also has a short section on "Using Historical Variation Information" in measurement system assessment. However, both the Minitab and AIAG suggestions ignore observer information available in the baseline data and, as in Steiner and MacKay [2005], assume the total variation is known rather than estimated. Finally, all of these previous references other than Steiner and MacKay [2005] did not suggest altering the design of an MSA study when baseline data are available.

The two primary goals of this chapter are to quantify the effect of (properly) including baseline data in the MSA study analysis for a continuous measurement characteristic, and to consider the best standard plans when baseline data are available [Stevens et al., 2013]. Specifically, we investigate the effects of supplementing the information from the SP data with the single measurements from the baseline data.

This design should sound familiar; it is similar to the type A augmented designs discussed in Chapter 2. In fact, supplementing SP data with baseline data in an MSA study is statistically equivalent to type A augmentation. However, from a practical standpoint the two designs are very different. For the augmented plans the extra data were collected as part of the MSA study by including more subjects. However, including more subjects in the study can be costly. Sometimes the process of taking the measurement can be very time-consuming, and sometimes the cost of measuring subjects can be very expensive [Aguirre-Torres and Lopez-Alvarez, 2013]. In these instances measuring fewer subjects would be beneficial, and the augmented plans may not be practical. Here we assume that the baseline data are readily available from previous use of the measurement system, and no additional cost is incurred to obtain them. Thus, studies with fewer subjects may be performed because the baseline data increases the amount of information about subjects, effectively increasing the number of subjects in the study, without any additional cost.

The remainder of the chapter is organized as follows. In Section 3.1, we discuss the model and derive the likelihood function for the $SP(n,r)$ plan when it is augmented by baseline data and apply the likelihood analysis to an example. Then in Section 3.2 we explore the design of an MSA study when baseline information is available; we use the asymptotic standard deviation of the estimator $\tilde{\gamma}$ to compare various $SP(n,r)$ plans for different baseline sample sizes and to suggest the optimal plans. We also report the results of a simulation study that demonstrate that the likelihood-based asymptotic results can be safely used to rank plans. As in Chapter 2 we consider the one (no) observer case as well as situations involving multiple observers with and without subject-by-observer interaction. We see a large improvement in precision for estimating γ even when the baseline sample size is small. Also, in most situations the recommended plans are markedly superior to those used in practice. Section 3.3 ends the chapter with a discussion and a summary of the results.

3.1 Modeling and Likelihood Analysis of an MSA Study with Baseline Data

3.1.1 The Model

Here we adopt the same two-way mixed effects model [2.1] as in Chapter 2 to specify the attributes of the measurement system and describe the data collected according to a standard plan. For convenience we re-state this model:

$$Y_{ijk} = S_i + o_j + SO_{ij} + M_{ijk} \qquad [3.1]$$

Here Y_{ijk} is a random variable that represents the k^{th} measurement on subject i by observer j, where $i = 1,2,\dots,n,\ j = 1,2,\dots,m,\ k = 1,2,\dots,r$. As before $S_i \sim N(\mu, \sigma_s^2)$ represents the unknown true value of the measurand for subject i; o_j represents the fixed effect of observer j; $SO_{ij} \sim N(0, \sigma_{so}^2)$ allows for the observer effect to change from subject to subject (i.e. allows for an interaction between observers and subjects); and $M_{ijk} \sim N(0, \sigma_m^2)$ represents the measurement error when the same observer takes replicate measurements of the same subject (i.e. the repeatability). We further assume that S_i, SO_{ij}, and M_{ijk} are all mutually independent.

We also define $\sigma_o^2 = \sum_{j=1}^{m} \left(\mu_j - \mu\right)^2 / m$, with $\mu_j = \mu + o_j$ representing the expected measurement by observer j and μ representing the overall mean value of the measurand. Thus σ_o^2 quantifies the measurement variation due to the relative biases among observers (i.e. the reproducibility). We then define $\sigma_t^2 = \sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2$ to be the total variation in the observed measurements.

With the parameters associated with model [3.1] we define the GR&R ratio γ, which quantifies the measurement system variability relative to the total variability:

$$\gamma = \sqrt{\frac{\sigma_o^2 + \sigma_{so}^2 + \sigma_m^2}{\sigma_s^2 + \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2}} \qquad [3.2]$$

As in Chapter 2, we assume this is the metric of primary interest; we use the asymptotic standard deviation of the estimator $\tilde{\gamma}$ to rank plans. The plans that we recommend are optimal in terms of their ability to estimate γ precisely, but they may not be optimal if estimation of a different metric is important.

We consider, as before, three cases for the design and analysis of an MSA study based on the number of observers included in the study, and the possible existence of a subject-by-observer interaction. These three cases correspond to three different formulations of model [3.1] and hence three different versions of γ.

First, we consider the case when the measurement system is operated by just one observer ($m = 1$), or is automated, and hence has no observer effects. Here $\sigma_o^2 = 0$ and we cannot estimate σ_{so}^2 separately from σ_m^2 so we set $\sigma_{so}^2 = 0$. In this case model [3.1] reduces to the random effects model [1.1] and $\gamma = \sqrt{\frac{\sigma_m^2}{\sigma_s^2 + \sigma_m^2}}$. For $m \geq 2$ we also consider the case when there is no subject-by-observer interaction by setting $\sigma_{so}^2 = 0$ and allowing $\sigma_o^2 > 0$. In this case we drop the SO_{ij} term from model [3.1] and $\gamma = \sqrt{\frac{\sigma_o^2 + \sigma_m^2}{\sigma_s^2 + \sigma_o^2 + \sigma_m^2}}$. When we consider the multiple-observer case and allow for a subject-by-observer interaction, we adopt the full model [3.1] and the corresponding definition of γ [3.2].

We again define two parameters of secondary interest, δ and β, to partition the variation due to the measurement system, as

$$\delta = \frac{\sigma_m^2}{\sigma_o^2 + \sigma_{so}^2 + \sigma_m^2} \tag{3.3}$$

and

$$\beta = \frac{\sigma_o^2}{\sigma_o^2 + \sigma_{so}^2} \tag{3.4}$$

Since γ, δ, and β are defined as ratios, we assume, without loss of generality, that the total variation σ_t^2 equals one, so $\gamma^2 = \sigma_o^2 + \sigma_{so}^2 + \sigma_m^2$. To assess plans, we look at three cases where $\sigma_m^2 = \delta\gamma^2$ for $\delta = 0.1$, 0.5, 0.9 so the repeatability contribution to γ is relatively small to large, respectively. Then, for given values of γ and δ, we consider the three cases $\beta = 0.1$, 0.5, 0.9 where the contribution of σ_o^2 is a relatively small to large proportion of $\sigma_o^2 + \sigma_{so}^2$, the observer contribution to the measurement system variation (reproducibility). Algebraically, we have $\sigma_o^2 = \beta(1-\delta)\gamma^2$ and $\sigma_{so}^2 = (1-\beta)(1-\delta)\gamma^2$. Note that $\delta = 1$ corresponds to $\sigma_o^2 = 0$; i.e., the when there are no observer effects in the model. And note that $\beta = 1$ corresponds to $\sigma_{so}^2 = 0$; i.e., the case where there is no subject-by-observer interaction.

Model [3.1] describes the data collected according to the standard plan $SP(n,r)$ with m observers. We assume that the baseline data has a balanced form where there are single measurements on b different subjects, with each of the m observers making b/m measurements. We further assume that we can associate each baseline measurement with a specific observer. We represent the data as z_{jl}, $j = 1,2,\ldots,m$, $l = 1,2,\ldots,b_j = b/m$, and according to the model [3.1], we have the corresponding independent random variables $Z_{jl} \sim N(\mu_j, \sigma_s^2 + \sigma_{so}^2 + \sigma_m^2)$. Note that in addition to information about σ_s^2, the baseline data also gives information about the observer means μ_j and the within-observer variation.

When there are data from an $SP(n,r)$ plan with no baseline data, we can estimate γ using ANOVA methods [Burdick et al., 2005]. However, when we add baseline data, we apply the likelihood methods as outlined in Section 3.1.2, to produce the estimates as it is not clear how to adapt the ANOVA analysis.

3.1.2 The Likelihood

As stated, we rank plans using the asymptotic standard deviation of $\tilde{\gamma}$ which is found using the inverse of the Fisher information matrix. To derive the information matrix, we need the likelihood function, which we develop here.

In Section 2.1 we derived the log-likelihood function for a standard plan with n subjects, m observers, and r replicate measurements. This log-likelihood contribution is

$$\begin{aligned} &l_{SP}(\mu_1, \dots \mu_m, \sigma_s^2, \sigma_{so}^2, \sigma_m^2) = \\ &-nmrln(2\pi) - \frac{n}{2}\ln\left[(\sigma_m^2 + r\sigma_{so}^2 + mr\sigma_s^2)(\sigma_m^2 + r\sigma_{so}^2)^{m-1}(\sigma_m^2)^{m(r-1)}\right] \\ &-\frac{1}{2}\left\{ a_1 \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}\left(y_{ijk} - \mu_j\right)^2 \right. \\ &\qquad + a_2 \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\sum_{k=1}^{r}\left(y_{ijk} - \mu_j\right)^2\right] \\ &\qquad \left. + a_3 \sum_{i=1}^{n}\left[\sum_{j=1}^{m}\sum_{k=1}^{r}\left(y_{ijk} - \mu_j\right)\right]^2 \right\} \end{aligned} \tag{3.5}$$

where

$$a_1 = \frac{1}{\sigma_m^2}$$

$$a_2 = \frac{-\sigma_{so}^2}{\sigma_m^2(\sigma_m^2 + r\sigma_{so}^2)}$$

$$a_3 = \frac{-\sigma_s^2}{(\sigma_m^2 + r\sigma_{so}^2)(\sigma_m^2 + r\sigma_{so}^2 + mr\sigma_s^2)}$$

For the baseline data, each observer measures different subjects once and measurements on all subjects are independent. For a baseline with a sample of size b, this is equivalent to the augmented component of a type A augmented plan, with n_A subjects. Thus, the log-likelihood contribution from the baseline data is equivalent to the log-likelihood contribution given by equation [2.7] in Section 2.1. With a small change in notation, the log-likelihood function for the baseline data is

$$l_b(\mu_1, \dots \mu_m, \sigma_s^2, \sigma_{so}^2, \sigma_m^2) = \\ -bln(2\pi) - \frac{b}{2}\ln(\sigma_s^2 + \sigma_{so}^2 + \sigma_m^2) \\ - \frac{1}{2(\sigma_s^2 + \sigma_{so}^2 + \sigma_m^2)} \sum_{j=1}^{m} \sum_{l=1}^{\frac{b}{m}} \left(z_{ij} - \mu_j\right)^2 \qquad [3.6]$$

where b is the total baseline sample size, and each observer measures b/m subjects once. Recall that we assume that each baseline measurement can be associated with a particular observer. In Section 3.3 we briefly discuss the scenario when we cannot trace the baseline measurements back to specific observers.

Assuming that the subjects in the baseline and the $SP(n,r)$ study are different and hence independent, the overall log-likelihood is the sum of the two log-likelihood components given by equations [3.5] and [3.6], namely,

$$l_{SP}(\mu_1, \dots \mu_m, \sigma_s^2, \sigma_{so}^2, \sigma_m^2) + l_b(\mu_1, \dots \mu_m, \sigma_s^2, \sigma_{so}^2, \sigma_m^2) \qquad [3.7]$$

To estimate γ, δ, and β we must estimate σ_s^2, σ_{so}^2, σ_m^2, and σ_o^2, and hence μ_j, $j = 1,2, \dots, m$. To do so we numerically maximize the log-likelihood function given by equation [3.7] using Matlab [The MathWorks Inc., 2013]. We find the Fisher information matrix by first taking the appropriate second partial derivatives and then by substituting the expected sums of squares [2.8-2.11] for the observed versions. Inverting this matrix gives the asymptotic variances for $\tilde{\sigma}_s^2$, $\tilde{\sigma}_{so}^2$, $\tilde{\sigma}_m^2$, and $\tilde{\mu}_j$, $j = 1,2, \dots, m$. To obtain the asymptotic variance for $\tilde{\gamma}$, and other functions of these parameters, we apply the delta method and pre- and post-multiply the inverse of the information matrix by a change-of-variables matrix of suitable partial derivatives, as in Section 2.1. As before, we use Maple [Maplesoft, 2014] to symbolically calculate all partial derivatives to avoid errors.

The likelihood described above corresponds to the multiple observer scenario in which we allow for the possibility of a subject-by-observer interaction. When we consider the special cases without subject-by-observer interaction, and without observer effects, we set $\sigma_{so}^2 = 0$ and $\sigma_o^2 =$

0 as is appropriate, and we alter the information matrix and the change-of-variable matrix as described in Section 2.1.

3.1.3 Example

In this subsection we provide an example from the manufacturing industry to illustrate the likelihood analysis, and highlight the benefit of incorporating baseline information, when analyzing standard plan data from an MSA study.

In the production of aluminum pistons, an automated gauge is used for 100% inspection of the skirt diameter at a particular height as well as many other key characteristics of the piston. There is a routine assessment of the system every 6 months using an $SP(10,6)$ plan. Because the gauge is automated, there are no observer effects. The data from the most recent MSA study are shown in Table 3.1. The gauge reports the deviation from nominal in micrometers. From process monitoring, there were also single measurements from 96 pistons available from the previous 24 hours of production. The mean and standard deviation for these baseline measurements are 0.56 and 2.88, respectively. With no observer effects, these are sufficient statistics for the baseline data, which follow a $N(\mu, \sigma_s^2 + \sigma_m^2)$ distribution.

	Replicate Measurements					
Part	1	2	3	4	5	6
1	-2.8	-3.6	-1.8	-2.5	-5.5	-3.3
2	1.8	2.7	2.0	2.9	0.8	-0.2
3	-3.7	-3.6	-3.9	-3.2	-3.4	-3.9
4	0.1	1.6	1.1	0.2	0.9	-0.2
5	2.3	3.6	1.0	2.0	5.2	3.4
6	-2.7	-3.0	-3.2	-1.5	-1.9	-4.7
7	0.9	2.8	1.2	0.2	0.7	0.9
8	0.9	0.8	2.1	-0.2	0.8	2.1
9	1.3	1.6	1.5	0.3	-0.5	0.8
10	-0.3	-0.6	0.1	0.3	-0.3	0.2

Table 3.1: Piston diameter data from an $SP(10,6)$ MSA study

Using maximum likelihood estimation of equation [3.7] and the observed information, we have $\hat{\gamma} = 0.354$ with standard error 0.0424 and an approximate 95% confidence interval for γ of (0.271, 0.437) when we include the baseline data. If we ignore the baseline data, the ANOVA estimate is $\hat{\gamma} = 0.408$ with standard error 0.086, which gives the 95% confidence interval of (0.230, 0.576). Clearly we get a large improvement in the precision of the estimate of γ by incorporating the available baseline data into the analysis. Software and instructions for its use to calculate the maximum likelihood estimate and the associated standard error of $\hat{\gamma}$, in light of baseline data, are provided at the website http://www.bisrg.uwaterloo.ca/.

We note that to avoid bias we need to be careful that the baseline data reflect the current state of the measurement system, and typical subject-to-subject variation. The analysis described in Section 3.1.2 assumes that the statistical properties of the measurement system are the same for the time interval that covers both the collection of the baseline data and the MSA study. To ensure that this is true, we suggest checking for stability in the baseline data as recommended by the Automotive Industry Action Group [2010].

3.2 Planning an MSA Study when Baseline Data are Available

In this section, we look at the effect of baseline information for a variety of standard plans and we recommend good choices for a $SP(n,r)$ plan given a fixed number of observers m, a fixed number of measurements $N = nmr$, and observer-specific baseline information. We consider three cases based on the number of observers m and whether or not the interaction is included in model [3.1]:

- Case 1: no (or one) observer ($m = 1$) – here there is no subject-by-observer interaction;
- Case 2: multiple observers ($m > 1$) – assuming no subject-by-observer interaction exists;
- Case 3: multiple observers ($m > 1$) – assuming subject-by-observer interaction exists.

In Chapter 1 we described a measurement system as being linear if its bias and variability do not depend on the true value of the measurand for the subject being measured. To check this property, we must have at least two subjects in the assessment study. However, for a better understanding of the possible relationship between bias and variability and the true value of the measurand, we recommend having three or more subjects in the study. As a result, in the following comparisons, we consider only plans with $n \geq 3$. We also restrict the comparisons to plans that produce estimates for all of the parameters in the model [3.1]. For example, in Case 3, without replicate measurements (i.e., $r = 1$) we can estimate $\sigma_{so}^2 + \sigma_m^2$, and hence γ and δ, but we cannot separately estimate σ_{so}^2, meaning that we cannot estimate β. As such, we restrict attention to plans with $r \geq 2$ in this case to allow estimation of all parameters.

In what follows, we compare plans using the asymptotic standard error for the estimate of γ, which we denote $SE(\hat{\gamma})$, calculated from the Fisher information matrix as described in Section 3.1.2. To check that the asymptotic results will allow us to appropriately rank the possible $SP(n, r)$ plans for different baseline sizes we first conducted a simulation study. In the simulation we compared the simulated and asymptotic standard errors for a variety of plans and parameter values. We considered:

- Total number of measurements: $N =$ 60, 90 and 120;
- Number of observers: $m =$ 1, 2, 3, and 4;
- Number of subjects: $n =$ 3 to a maximum depending on N and m;
- Per-observer baseline sample sizes of $b_j =$ 0, 10, 30, and 100 (recall that $b = mb_j$);
- Parameter values $\gamma =$ 0.1, 0.3, $\delta =$ 0.1, 0.5, 0.9, and $\beta =$ 0.1, 0.5, 0.9.

For each plan and set of parameter values, we generated 10 000 samples from model [3.1] and for each sample determined the maximum likelihood estimate of γ. The results show that the asymptotic standard error for $\hat{\gamma}$ closely matches the simulated results for all plans when the baseline sample size is larger than 30. For simulations based on small baseline sample sizes, the asymptotic results underestimate the simulated results with increasing large differences for plans with fewer subjects, n. Where there was a large difference between the asymptotic and simulated results, the estimate also has substantial bias. For additional information regarding the results of this simulation study, see Section B.1 of Appendix B.

These differences are important if we wish to select the overall number of measurements N to meet a goal in terms of the standard error of the estimate for γ when the baseline sample size is small. However, for fixed values of N, m, b, and the parameter values, the asymptotic and simulated standard errors provided the same ranking of plans as n and r varied. As a result, we proceed to rank plans based on the asymptotic results. Note that we never recommend a plan whose asymptotic properties do not closely match simulation results. In addition, based on simulation results there is an even larger benefit from using available baseline data than suggested in Figures 3.1 to 3.3.

To address the issue of bias when there was no or little baseline data we tried using REstricted Maximum Likelihood (REML) estimation [Corbeil and Searle, 1976]. Using the transformation suggested by Corbeil and Searle [1976], we can split the log-likelihood given in [3.7] into two pieces, one that depends only on the variance components (σ_s^2, σ_{so}^2, and σ_m^2) and the other that depends on all of the variance components and the observer means ($\mu_j, j = 1,2, \dots, m$). We found the REML estimates for the variance components by maximizing the first piece of the log-likelihood. For balanced plans without baseline data the REML estimates match those obtained by the usual ANOVA estimation. We subsequently obtained estimates for the observer means by maximizing the overall log-likelihood with the variance components fixed at the REML estimates. To explore the usefulness of the REML approach, we conducted a factorial simulation study similar to that described earlier in this section that compared the REML and usual maximum likelihood estimates for a number of different plans, baseline sizes, and values for δ and β. The results suggest that the REML estimator of γ is indeed substantially less biased than the usual maximum likelihood estimator (though still not unbiased) when there are no baseline data, especially when the number of subjects in the SP is small. However, when we add even a small amount of baseline data, say 30 observations, the difference in bias between the two estimation approaches disappears and in some combinations of the parameter values the usual maximum likelihood estimators are less variable than the REML estimators. For this reason, and the additional complexity of the REML approach, we continue to use standard maximum likelihood estimation. For additional information regarding the results of this simulation study, see Section B.2 of Appendix B.

3.2.1 Plans with One or No Observer

When there are no observer effects or only one observer ($m = 1$) we have $\sigma_o = 0$, and $\sigma_{so} = 0$. As a result, there are only three unknown parameters μ, σ_s, and σ_m, in model [3.1]. Recall that with no observer effects the parameter of interest simplifies to $\gamma = \sqrt{\frac{\sigma_m^2}{\sigma_s^2+\sigma_m^2}}$, and we do not consider δ or β, which in this case both equal one. To illustrate the contribution of the baseline data, we use a total of $N = 60$ measurements in the $SP(n, r)$ study and consider three different plans. The plan $SP(10,6)$ matches the AIAG [2010] recommendation, $SP(30,2)$ is the plan with the maximum number of subjects as suggested by Shainin [1992], and $SP(3,20)$ is the plan with the minimum number of subjects, as used by Steiner and MacKay [2005] when baseline data are also available.

Figure 3.1 shows the asymptotic standard error of the estimate of γ by baseline size for the three plans when $\gamma = 0.2$. The pattern of the results is similar for other values of γ and N though the specific values of $SE(\hat{\gamma})$ change.

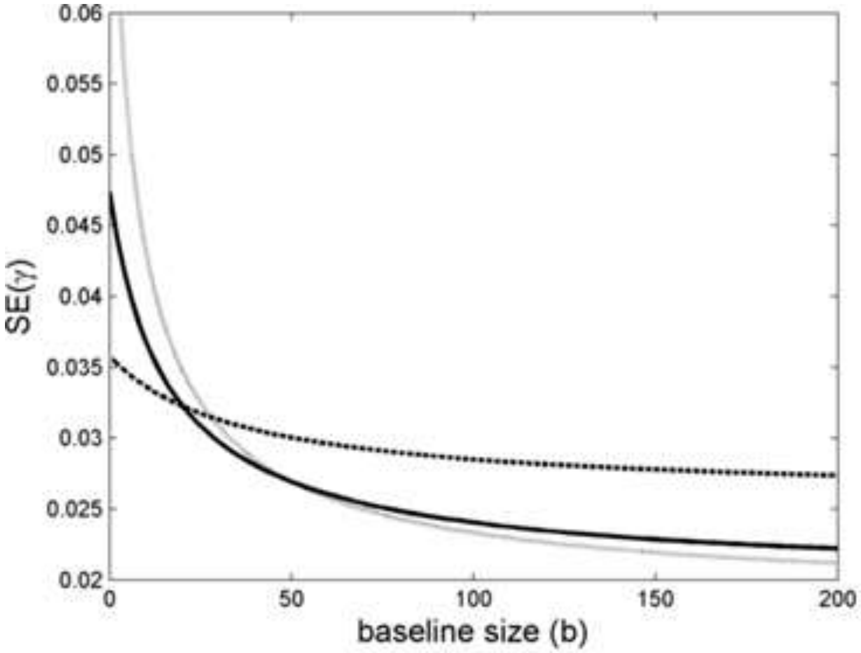

Figure 3.1: $SE(\hat{\gamma})$ as a function of the baseline size with $N = 60$, $m = 1$, $\gamma = 0.2$
Black line: $SP(10,6)$ – Grey line: $SP(3,20)$ – Dotted line: $SP(30,2)$

We draw two conclusions. First, the value of including the baseline data is substantial especially for standard plans with few subjects. This is not surprising because without the baseline data the standard plans with few subjects provide little information about σ_s.

Second, the best plan depends on whether or not baseline data are available. If there are no baseline data, $SP(30,2)$ is the best for estimating γ and it results in a substantial reduction in the standard error for $\hat{\gamma}$ compared with $SP(10,6)$, the default AIAG plan. With the $SP(30,2)$ plan and $b = 0$ (i.e. no baseline data), we match the degrees of freedom available to estimate the two unknown variance components σ_s^2 and σ_m^2. This finding corroborates what was found in Section

2.2.1; when there is no additional information from a baseline or augmentation, the Shainin proposal for an Isoplot^{TM} study [Shainin, 1992] is the optimal plan for estimating γ. However, even with a small number of baseline observations, say $b \geq 30$, the other two plans are better than the plan with the maximum number of subjects, and they are much better with larger baseline sample sizes.

Note that when b is large and there are only two variance components, the baseline sample gives a precise estimate of the total variation $\sigma_t^2 = \sigma_s^2 + \sigma_m^2$, and hence the best plan for estimating γ is to make replicate measurements on a single subject. However, we do not recommend this plan since with it we cannot check the linearity assumptions as discussed earlier.

As a general guideline, with a single observer, if $b > N/2$, we recommend the three-subject plan ($n = 3$, $r \approx N/3$). Otherwise, we suggest the plan with the maximum number of subjects ($n \approx N/2$, $r = 2$). For a more detailed analysis, software is available for use at the website http://www.bisrg.uwaterloo.ca/ to find optimal plans with the following inputs:

- the number of observers m (here $m = 1$);
- the baseline size b;
- a range of possible values for γ;
- the maximum number of measurements N used in the SP so that $nmr \leq N$.

For specified values of N, m and b, and a range for γ, the output includes the number of subjects, n, and the number of replicate measurements per subject, r, that minimize the asymptotic standard error of the maximum likelihood estimate $\hat{\gamma}$. As well, we show the ratio of the standard errors for the optimal plan compared with the three-subject plan and the plan with the maximum number of subjects. Note that the standard error associated with the optimal plan is in the numerator of these ratios.

For example, if we specify $m = 1$, $b = 60$, and $N = 60$ with $0.05 \leq \gamma \leq 0.40$, in increments of 0.05, we get the output as shown in Table 3.2. In this case, the best plan is somewhat sensitive to the unknown value of γ but the standard error of the recommended three-subject plan is virtually identical to that of the optimal plan over the entire range of values for γ.

Optimal Plan				Relative Efficiency of:	
				Three-Subject	Max-Subject
n	r	γ	$SE(\hat{\gamma})$	Plan	Plan
3	20	0.05	0.0065	1.0000	0.8674
3	20	0.10	0.0129	1.0000	0.8682
3	20	0.15	0.0194	1.0000	0.8695
3	20	0.20	0.0258	1.0000	0.8715
3	20	0.25	0.0322	1.0000	0.8740
3	12	0.30	0.0386	0.9996	0.8769
5	10	0.35	0.0448	0.9983	0.8798
6	10	0.40	0.0510	0.9961	0.8827

Table 3.2: Optimal choice of n and r with $m = 1$, $b = 60$, $N = 60$

3.2.2 Plans with More than One Observer and No Subject-by-Observer Interaction

Now we consider the case of two or more observers with no subject-by-observer interaction; i.e., we set $\sigma_{so} = 0$ in model [3.1]. Recall that with multiple observers we use δ to describe the proportion of the overall measurement system variability attributable to repeatability. In this case [3.3] simplifies to

$$\delta = \frac{\sigma_m^2}{\sigma_o^2 + \sigma_m^2}$$

Note that as δ gets closer to one, the repeatability contribution increases. Without subject-by-observer interaction, we do not consider the parameter β, which in this case equals one. And as before we assume that γ is the primary parameter of interest, as we base comparisons of plans on the asymptotic standard error of $\hat{\gamma}$. Here, $\gamma = \sqrt{\frac{\sigma_m^2+\sigma_o^2}{\sigma_s^2+\sigma_o^2+\sigma_m^2}}$.

With m observers we assume that each measures b/m subjects in the baseline so the total baseline sample size is b. In Figure 3.2, we compare the performance of plans with $n = 3$, 10, or 30 subjects for varying b when $\gamma = 0.2$, $m = 2$, $N = 120$, and $\delta = 0.1$, 0.5, 0.9. Two of the selected plans have the minimum, ($SP(3,20)$), and maximum, ($SP(30,2)$), number of subjects for the given values of m and N. The third plan, $SP(10,6)$, is close to that recommended by the Automotive Industry Action Group [2010].

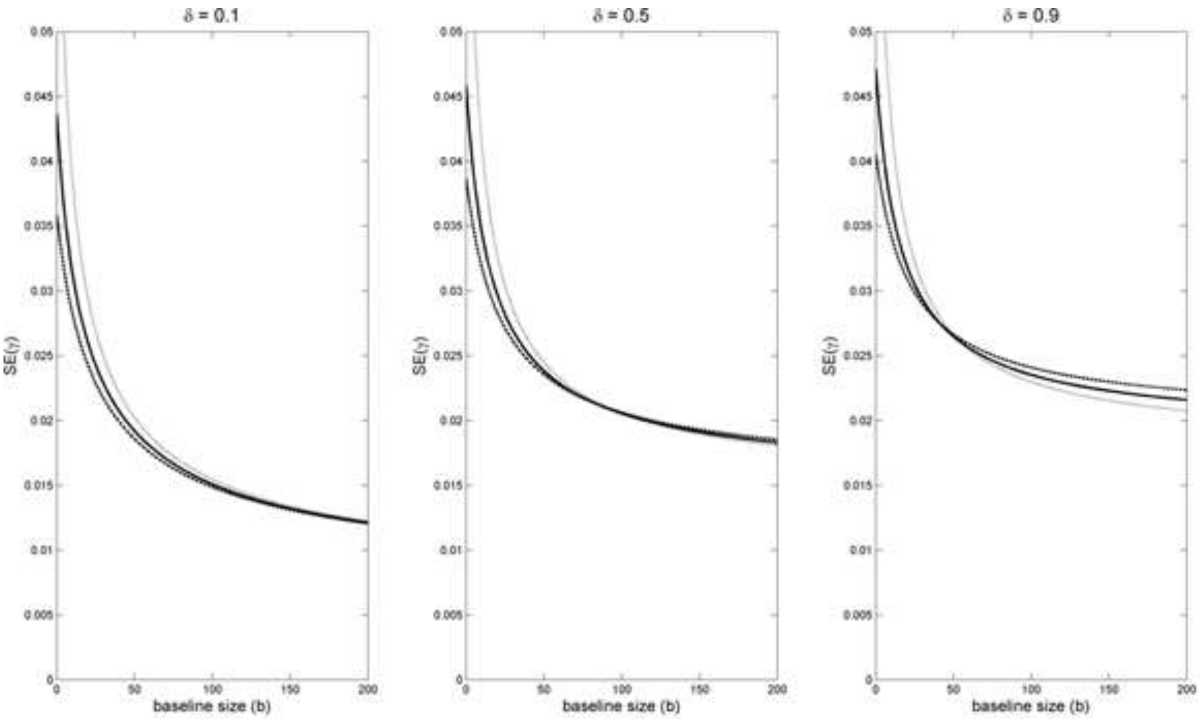

Figure 3.2: $SE(\hat{\gamma})$ as a function of the baseline size with $N = 120$, $m = 2$, $\gamma = 0.2$
Black line: $SP(10{,}6)$ – Grey line: $SP(3{,}20)$ – Dotted line: $SP(30{,}2)$

We see again the large benefit of the baseline information for the precision of the estimate of γ in all of the plans and that much of the benefit is obtained with b as small as 80. As in the single-observer case, plans with few subjects benefit the most. For small values of b, the best plan uses the maximum number of subjects and as b becomes large, the best plan uses just three subjects. The switching point depends on δ, the relative contribution of the repeatability to the overall measurement system variation. If δ is close to one, the three-subject plan becomes optimal for smaller values of b. We see the same general pattern when we look at similar plots for varying values of N, $m = 2$, 3, 4, and γ . For plans with m observers we suggest the three-subject plan for $b > 40 + 20m$ and the plan with a maximum number of subjects otherwise.

We can use the software described earlier for a more detailed analysis. Suppose we have $N = 60$, $m = 3$, $b = 60$, $\gamma = 0.1, 0.2, 0.3, 0.4$ and $\delta = 0.1, 0.5, 0.9$. According to the proposed planning guidelines, we should use the plan with the maximum number of subjects, $SP(20{,}1)$. From Table 3.3, we see that the optimal design depends on γ and δ. However, the recommended 20-subject plan has very high efficiency and we would consider using another plan only if we were confident that δ was close to one. If we set $\delta = 1$, there are no differences among the observers and we return to the case with no observer effects discussed in Section 3.2.1.

Optimal Plan					Relative Efficiency of:	
					Three-Subject Plan	Max-Subject Plan
n	r	γ	δ	$SE(\hat{\gamma})$		
20	1	0.1	0.1	0.0088	0.9009	1.0000
20	1	0.1	0.5	0.0116	0.9495	1.0000
4	5	0.1	0.9	0.0128	0.9711	0.9628
20	1	0.2	0.1	0.0171	0.9007	1.0000
20	1	0.2	0.5	0.0226	0.9489	1.0000
5	4	0.2	0.9	0.0255	0.9709	0.9651
20	1	0.3	0.1	0.0244	0.9002	1.0000
20	1	0.3	0.5	0.0328	0.9479	1.0000
5	4	0.3	0.9	0.0378	0.9699	0.9688
20	1	0.4	0.1	0.0302	0.8995	1.0000
20	1	0.4	0.5	0.0415	0.99457	1.0000
10	2	0.4	0.9	0.0494	0.9659	0.9729

Table 3.3: Optimal choice of n and r with $m = 3$, $b = 60$, $N = 60$

3.2.3 Plans with More than One Observer and Possible Subject-by-Observer Interaction

Lastly we consider the case where we include the subject-by-observer interaction term given in model [3.1]. We now consider all three parameters γ, δ, and β, as defined in [3.2-3.4], that describe the performance of the measurement system. As β gets closer to one, the relative contribution of the interaction to the reproducibility decreases. We note in passing that it would be surprising to have small values of β where the reproducibility is dominated by the interaction, as a large interaction effect would often coexist with a large observer effect.

Again we base our comparisons on the estimation of the primary parameter γ and include only plans that provide estimates of all the parameters. That is, we now require $r \geq 2$, as discussed earlier. In Figure 3.3, we look at the case of two observers ($m = 2$) with $\gamma = 0.2$ and compare the three plans $SP(10,3)$, $SP(15,2)$, and $SP(3,10)$, each with a total of 60 measurements over the usual range of values for δ and β. We see substantial improvement in the precision of $\hat{\gamma}$ by including the baseline data regardless of the plan. However, in comparing the three plans, over almost the whole range of values for δ, β, and all values of b, the plan with the maximum number of subjects, $SP(15,2)$, is optimal and substantially better than the other plans if δ and β are small. We see the same behavior for other values of N, $2 \leq m \leq 4$ and γ. When δ, β, and b are large, the three-subject plan becomes optimal but there is little loss in efficiency in using the plan with a maximum number of subjects.

For example, when $N = 120$, $m = 4$, $\delta = 0.99$, $\beta = 0.9$, and $b = 400$, the relative efficiency of $SP(15,2)$ compared to $SP(3,10)$ is at least 0.95 as γ varies between 0.1 and 0.3. We can use the software described previously to investigate other specific combinations of parameter values, but we make the following general recommendation: when there are multiple observers and we wish to estimate a possible subject-by-observer interaction effect we recommend the SP with the maximum number of subjects in all situations.

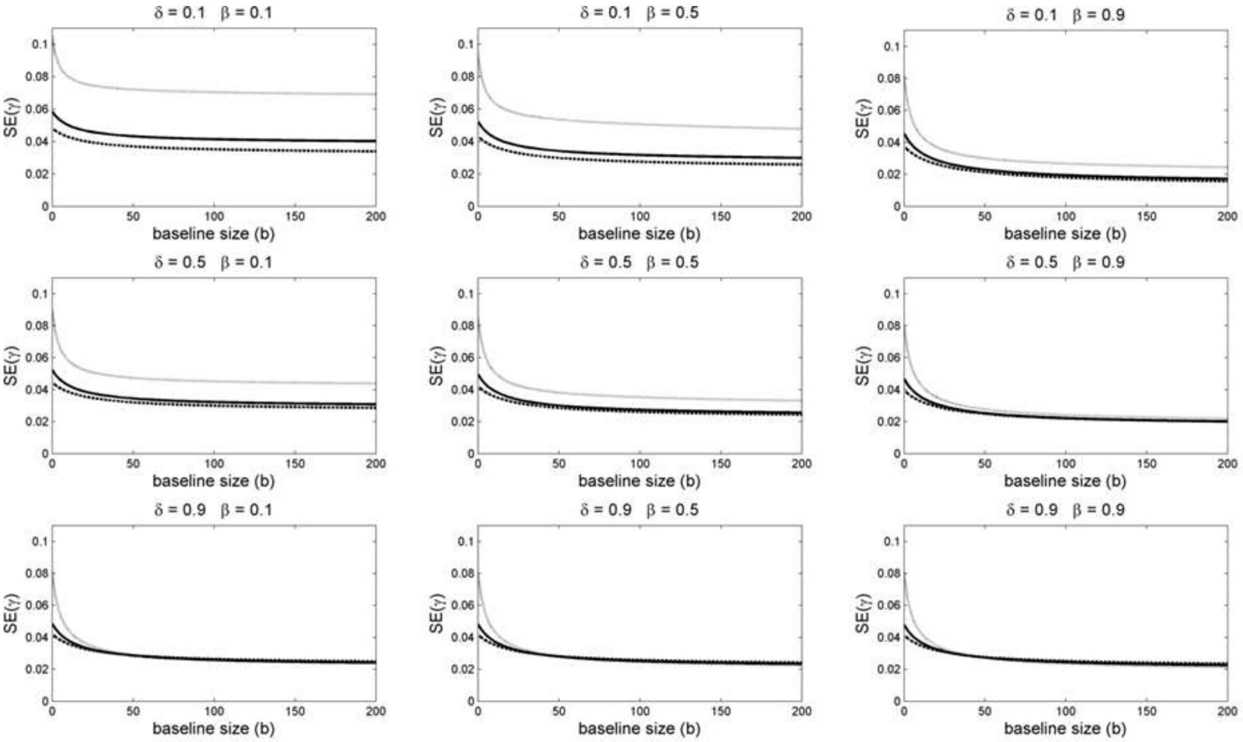

Figure 3.3: $SE(\hat{\gamma})$ as a function of the baseline size with $N = 60$, $m = 2$, $\gamma = 0.2$

Black line: $SP(10,3)$ – Grey line: $SP(3,10)$ – Dotted line: $SP(15,2)$

3.3 Discussion and Conclusions

In this chapter, we address the planning and analysis of measurement system assessment studies when baseline data are available. This is a common situation for measurement systems that are used routinely where the baseline data are available at no additional cost. We quantify the benefits of incorporating baseline data (subjects measured once) into the measurement system assessment study analysis and show that substantial improvements in precision are possible and attained even with small baseline sample sizes. We also recommend changes to the usual MSA study plan in terms of the number of subjects and replicate measurements taken. With a fixed baseline size, and a fixed total number of measurements $N = nmr$, we recommend standard plans that use either the minimum or maximum number of subjects, unlike the AIAG [2010] recommendation that suggests $n = 10$ subjects. To summarize, incorporating the baseline data into the analysis and selecting the standard plan with either the recommended maximum or minimum number of subjects dramatically increases the precision of the estimate of γ, the GR&R ratio.

We see most of the benefit from incorporating the baseline data in the analysis with total baseline sample sizes as small as 60. As mentioned, in order to avoid bias we need to be careful that the baseline data reflect the current state of the measurement system, and typical subject-to-subject variation. In the analysis, we assume that the measurement system is stable for the time interval that covers both the collection of the baseline data and the MSA study. This suggests a question of "how much baseline data should we use?" In particular, how far back in time should we go? If the measurement system changes at some point during the collection of the baseline data, then the estimates of $\mu_1, \mu_2, \dots., \mu_m$ and σ_t^2 from the baseline will be biased. To address this potential problem, we suggest checking for stability in the baseline data as recommended by the Automotive Industry Action Group [2010].

With a single observer and a large baseline sample (say, larger than 50 observations), we saw in Figure 3.1 that the three-subject plan is the best. With multiple observers, if δ (the proportion of measurement variability due to repeatability) approaches one and b (the baseline sample size) is large, then the three-subject plan is also optimal, but the gain in precision for estimating γ is smaller (Figure 3.2). This occurs because we are still estimating the other parameters using the full model [3.1]. If we first collapse the model by setting $\delta = 1$, then we get the full benefit of the three-subject plan. Of course, there is no need to use multiple observers in the MSA study in this case.

If the measurement system has a single observer or no observer effects and we use $SP(3,r)$, the likelihood analysis is somewhat unnecessary because the three subjects provide relatively little extra information about σ_s and we can estimate σ_m using ANOVA with little loss of precision. This idea was proposed by Steiner and MacKay [2005], who, in the situation where there is a single (or no) observer, recommend an MSA study with three subjects with small, medium, and large values (as defined by the baseline data), rather than subjects selected at random. In this case, the estimation approach suggested here is not applicable. Instead, we use the MSA data only to estimate σ_m. A more efficient (but complicated) maximum likelihood estimation procedure for this case that takes into account the selection process is given by Browne et al. [2009] and extended in Browne et al. [2010] to situations involving multiple observers. This series of papers shows the major benefit of selecting subjects with extreme values for the MSA study in terms of the precision of the estimate of γ.

In this chapter we have assumed that when multiple observers are concerned, the baseline measurements can be traced back to specific observers. However, this may not always be realistic; there may be no record of the observers in the baseline data. In Stevens et al. [2013] we consider the effect of this missing information. Through simulation we show that for $\gamma \leq 0.3$ and $\delta = 0.1$, 0.5, 0.9 and $\beta = 0.1$, 0.5, 0.9, the loss of precision for estimating γ from not knowing the baseline observers is surprisingly small. And in all cases, there is little loss as the baseline size changes. We propose that, given the small differences in precision when we have baseline data without knowing the associated observers, we continue to use the planning guidance provided for the case of known observers. This suggests that much of the impact of baseline data is in improved estimates of variance components, and less so on estimates of observer specific biases. For more details on the simulation and results, see Stevens et al. [2013].

We have concentrated on the estimation of the primary parameter γ. If interest centers on measurement system characteristics such as the precision-to-tolerance ratio (PTR) [1.4], that do not involve the between-subject variation σ_s, there is little value in the baseline data.

With this work we have again modeled observers as fixed effects. In other situations, it may be more reasonable to assume that observers are random effects. As noted previously, MSA studies with random observers, but without baseline data, were considered by Steiner et al. [2011]. To model the observers as random effects and incorporate baseline data we could consider a similar analysis to that proposed in this Chapter. If we further assume that the observers used in the baseline and MSA study are different (and randomly selected) we can write down the likelihood and determine the Fisher information matrix. With the same observers used in the MSA and the baseline studies, the likelihood expression is complicated and finding the Fisher information matrix is more difficult.

Another possible extension that we do not explore further here is the possibility of replacing the standard plans with plans that are not balanced in terms of number of measurements made by each observer. An unbalanced plan may be better, for instance, when observers are not balanced in the baseline data.

Chapter 4

Introduction: Measurement System Comparison

The first part of this thesis (Chapters 1, 2 and 3) was concerned with assessing the adequacy of a single measurement system. It is clear that a measurement system which is both accurate and precise, is ideal. However, accuracy and precision may come at a cost; an accurate and precise measurement system, defined again to be the devices, people, and protocol used to make a measurement, may be costly in terms of time, money or man-power, or may be invasive. In this case, new measurement systems that are less expensive, less time-consuming, less labour-intensive or less invasive may be developed. In order to decide whether a new measurement system can be used in place of an existing one, a Measurement System Comparison (MSC) study should be undertaken.

A common purpose of an MSC study is to determine whether the measurements by the new system sufficiently agree with those from the existing system, and hence determine whether the two systems can be used interchangeably. Another important goal, although not the focus of this work, is to decide whether the new measurement system is better (in terms of accuracy and precision) than the existing. However, given our focus, we assume "comparing measurement systems" is synonymous with "assessing interchangeability".

In a typical MSC study we measure some characteristic, the measurand, of a number of randomly chosen subjects, one or more times by each measurement system [Barnett and Youden, 1965; 1970; Westgard and Hunt, 1973]. Notice that this corresponds to what we have referred to as the single-observer standard plan (SP) in the previous chapters. We will herein similarly refer to this design as the standard plan.

As in the previous chapters we assume both measurement systems are non-destructive, meaning the act of measuring does not alter the true value of the measurand, and so multiple measurements on each subject are possible. As well, we only consider the interchangeability of two measurement systems for a single continuous measurand.

To describe the data collected during such a study, we can use the following mixed effects model:

$$\begin{aligned} Y_{i1k} &= S_i + M_{i1k} \\ Y_{i2k} &= \alpha + \beta S_i + M_{i2k} \end{aligned} \qquad [4.1]$$

Thus Y_{ijk} is a random variable which represents the value observed on system j's k^{th} measurement of subject i, where $i = 1,2, \dots, n$ indexes the subjects, $j = 1$ indexes the reference measurement system, $j = 2$ indexes the new measurement system and $k = 1,2, \dots, r$ indexes the replicate measurements. In [4.1], S_i is a random variable that represents the unknown true value of the measurand for subject i, with the distributional assumption $S_i \sim N(\mu, \sigma_S^2)$. Here μ is a parameter which represents the overall mean true value of the measurand, and σ_S^2 is the variance component which quantifies the variability in true values about the mean μ. M_{ijk} is a random variable which represents the measurement error when system j makes multiple measurements on subject i. We further assume that the M_{ijk} are independent of each other, independent of S_i, and that they are distributed $N\left(0, \sigma_j^2\right)$ where σ_j quantifies the measurement variation, or repeatability, of system j. Note that for a given measurement system, we denoted this by σ_m in the previous chapters.

Model [4.1] assumes that σ_j is constant across true values and hence the variability of each measurement system is linear [The Automotive Industry Action Group, 2010]. In the context of measurement system comparison, we will refer to such a measurement system as homoscedastic. In other situations σ_j may depend in some way on the true value, in which case we call measurement system j heteroscedastic, and a different model must be used. This is discussed in Chapter 6.

We mention in passing that model [4.1] does not include observer effects. If the measurement systems used in the MSC study have multiple observers, their effects are not separately considered. Instead they are subsumed in the error term and we treat this situation as if there were only a single observer operating each measurement system. As such, the repeatability σ_j parsimoniously summarizes the overall measurement system variation (i.e. no reproducibility).

Recall that in the context of measurement system assessment, the main focus was to assess the variability of a measurement system, while bias was de-emphasized because in manufacturing settings, it is typically thought that bias can be eliminated through calibration [Burdick et al., 2005]. In the context of measurement system comparison, however, accounting for bias is as important as accounting for variability. Because the true values of the measurand are unknown, we cannot estimate the absolute bias of the measurement systems, but we can estimate the bias of the two systems relative to one another. Model [4.1] accounts for this.

The parameters $-\infty < \alpha < \infty$ and $\beta > 0$ quantify the bias of the second (new) measurement system relative to the reference system. Here we assume that the reference measurement system is unbiased, and inferences regarding bias are made relative to it. We refer to α as the fixed bias since it increases or decreases the average measurement of the second system by a fixed amount. We call β the proportional bias because it biases the second system's measurements by an amount that is proportional to the true value [Ludbrook, 2010].

Based on [4.1], we say that the two measurement systems are identical if $\alpha = 0, \beta = 1$ and $\sigma_1 = \sigma_2$. However, the two systems do not need to be identical to be used interchangeably. Informally we say that two systems can be used interchangeably if, most of the time, their measurements on the same subject are similar. In other words, two measurement systems agree and could be used interchangeably, if $Y_{i1k} \approx Y_{i2k}$, most of the time. Typically this happens when $\alpha \approx 0$, $\beta \approx 1$, and when both σ_1 and σ_2 are small, relative to σ_s. We will further develop the notion of interchangeability below.

A variety of techniques exist for analyzing MSC data and hence judging interchangeability, and we use the remainder of this chapter to review a number of them. Specifically, we discuss comparison of means (Section 4.1), comparison of repeatabilities (Section 4.2), correlation (Section 4.3), regression (Section 4.4), and the limits of agreement approach (Section 4.5). Throughout, we highlight challenges associated with these techniques in an effort to demonstrate the need for a new method which overcomes these challenges, and that accurately assesses interchangeability in a wider set of situations.

Before beginning this discussion, we note that many of the techniques we address do not explicitly assume model [4.1], and some do not assume a model at all, but we use [4.1] to succinctly describe the techniques and illustrate their drawbacks. In [4.1], we have assumed that each system measures each subject $r \geq 2$ times. Sometimes replicate measurements are not made in an MSC study, in which case $r = 1$, and we drop the subscript k, reducing model [4.1] to

$$\begin{aligned} Y_{i1} &= S_i + M_{i1} \\ Y_{i2} &= \alpha + \beta S_i + M_{i2} \end{aligned} \qquad [4.2]$$

When appropriate, we discuss how each technique compares measurement systems in the context of [4.2] with $r = 1$ replicate measurements, and we make comments about how each deals with $r \geq 2$.

4.1 Comparing Means

Perhaps the simplest and most straight-forward method of assessing the agreement between two measurement systems is to compare the means of their measurements. Here it is assumed that the study design is the standard one in which each of n subjects are measured once ($r = 1$) by each system [Barnett and Youden, 1965; 1970]. Using the notation of [4.2] we define the difference between two measurements on a given subject i as

$$D_i = Y_{i2} - Y_{i1} \qquad [4.3]$$

As a consequence of the normal distribution of the Y_{ij}'s and the independence of S_i and M_{ij}, the distribution of the differences is $D_i \sim N(\mu_d, \sigma_d^2)$ where

$$\mu_d = \alpha + (\beta - 1)\mu$$

and

$$\sigma_d^2 = (\beta - 1)^2 \sigma_s^2 + (\sigma_1^2 + \sigma_2^2)$$

These population parameters are respectively estimated by:

$$\hat{\mu}_d = \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad [4.4]$$

and

$$\hat{\sigma}_d^2 = s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2 \qquad [4.5]$$

where $d_i = y_{i2} - y_{i1}$ is the observed value of D_i and y_{ij} is the observed value of Y_{ij}. The normality assumption $D_i \sim N(\mu_d, \sigma_d^2)$ can be assessed by constructing a histogram of the differences d_i [Bland and Altman, 1983; 1986] or by a QQ-plot since sample sizes are often small [Barnett and Youden, 1965; 1970].

The means of each system's measurements are compared by comparing the mean difference, μ_d, to 0 by way of a simple *t*-test [Barnett and Youden, 1965; 1970; Westgard and Hunt, 1973]. Specifically, we test the hypothesis

$$H_0: \mu_d = 0 \text{ versus } H_A: \mu_d \neq 0 \qquad [4.6]$$

Such a test looks for evidence of a non-zero average difference between the measurements made by each system. If no significant difference is found (i.e., not enough evidence to reject H_0), then the two measurement systems are thought to give agreeable results, and so they could be used interchangeably.

This *t*-test however, does not compare the precision of each system. In fact, the amount of variability in each system can render the test misleading. Bland and Altman [1983] point out that if σ_1 and σ_2 are large, then the two systems are unlikely to agree, but the test statistic will be small, leading to the acceptance of H_0 in [4.6]. They facetiously state that in using this criterion to judge agreement, "the greater the measurement error, and hence the less chance of a significant difference, the better" [Altman and Bland, 1983, p. 308]. While statistically speaking this is true, this issue could be overcome by framing [4.6] as an equivalence test where the null hypothesis assumes inequivalence until evidence suggests anything to the contrary [Wellek, 2010]. The real issue, however, is that testing the equality of means (by [4.6] or an equivalence test) ignores a direct comparison of repeatabilities. As such, a comparison of relative precisions between the two measurement systems is also necessary.

4.2 Comparing Repeatabilities

In industrial contexts, it is common to compare measurement systems by comparing metrics computed in MSA studies [Burdick et al., 2002; Majeske, 2012]. For instance, if an MSC study is conducted in which each system measures n subjects $r \geq 2$ times each, we can use the resulting data to estimate the metrics discussed in Chapter 1 for each system: we may wish to compare discrimination ratios (D), GR&R ratios (γ), intraclass correlation coefficients (ρ), or precision-to-tolerance ratios (PTR).

We could do so by calculating the desired metric for both measurement systems, and then examine their ratio. Using the subscript j to distinguish measurement systems ($j = 1$ corresonds to the existing system, and $j = 2$ corresponds to the new system), we define the following ratios:

$$\frac{D_1}{D_2} = \frac{\sigma_S/\sigma_1}{\sigma_S/\sigma_2} \tag{4.7}$$

$$\frac{\gamma_1}{\gamma_2} = \frac{\sigma_1/\sqrt{\sigma_S^2 + \sigma_1^2}}{\sigma_2/\sqrt{\sigma_S^2 + \sigma_2^2}} \tag{4.8}$$

$$\frac{\rho_1}{\rho_2} = \frac{\sigma_S^2/(\sigma_S^2 + \sigma_1^2)}{\sigma_S^2/(\sigma_S^2 + \sigma_2^2)} \tag{4.9}$$

$$\frac{PTR_1}{PTR_2} = \frac{k\sigma_1/(USL - LSL)}{k\sigma_2/(USL - LSL)} = \frac{\sigma_1}{\sigma_2} \tag{4.10}$$

where USL, LSL, and k are as in [1.4].

When interest lies in comparing two measurement systems using these ratios, hypothesis tests may be constructed in which the specified ratio is compared to 1. If it is found that the ratio differs significantly from 1, then the measurement variability for the two systems is significantly different. If, however, a significant difference is not found, then the measurement variation in the two systems is similar, indicating that they could be used interchangeably. Majeske [2012] develops a variety of two-sample tests to address hypotheses of this nature for [4.7-4.10].

However, due to the typical design of an MSC study, these multiple tests are redundant. If each system measures the same subjects, or if the subjects measured by each system come from the same population, then σ_s in the definitions of D_j, γ_j and ρ_j will be the same for $j = 1{,}2$. When this is the case, the ratios [4.7-4.9] simplify considerably, reducing to a direct comparison of σ_1 and σ_2, like in [4.10]. Thus, any hypothesis that compares the ratios [4.7-4.11] to 1, can simply be stated as

$$H_0\colon \sigma_1 = \sigma_2 \text{ versus } H_A\colon \sigma_1 \neq \sigma_2 \qquad [4.11]$$

Burdick et al. [2002] test this hypothesis by developing modified large sample (MLS) techniques for calculating confidence intervals for σ_1/σ_2, and interchangeability is determined by whether or not the value one is contained within the interval.

Similar to [4.6], the null hypothesis in [4.11] assumes equality of repeatabilities and is only rejected with sufficient evidence. This hypothesis could be more appropriately framed as an equivalence test in which the null hypothesis assumes the repeatabilities are not equivalent [Wellek, 2010]. In doing this, the test is protected against the phenomenon that a large enough sample will always provide evidence against equality.

Even still, the comparison of repeatabilities (by [4.11 or an equivalence test) ignores the relative bias between the two measurement systems; two systems may be similarly precise, but if a large relative bias exists, it would be unwise to use the two systems interchangeably. As such, it is important to assess both the relative bias, and relative repeatability sizes when evaluating interchangeability.

4.3 Correlation

The correlation coefficient is a simple measure of linear association between Y_{i1} and Y_{i2}, measurements on the same subject by different systems. Note that we drop the subscript k because the correlation is based on single measurements ($r = 1$). The rationale for its use is that a correlation coefficient close to 1 may signify agreement between the two measurement systems. Using [4.2], the correlation between Y_{i1} and Y_{i2} (with $r = 1$) is:

$$Corr(Y_{i1}, Y_{i2}) = \frac{\beta\sigma_s^2}{\sqrt{(\sigma_s^2 + \sigma_1^2)(\beta^2\sigma_s^2 + \sigma_2^2)}} \qquad [4.12]$$

There are two common criticisms of this method; the first is that the correlation coefficient can be arbitrarily inflated by increasing the estimate of the between-subject variability, σ_s [Bland and Altman, 1983; 1986]. The definition of [4.12] assumes that subjects are randomly sampled from some population, but in some MSC studies, investigators non-randomly select individuals so that the sample intentionally covers a wide range of true values. When this is the case, we would expect $\hat{\sigma}_s$, and hence an estimate of [4.12] to be large, regardless of the level of agreement between the two measurement systems [Barnett and Youden, 1970; Bland and Altman, 1983; 1986]. Conversely, if σ_1 and σ_2 are large relative to σ_s then [4.12] will be small regardless of the agreement between the two measurements systems.

The second criticism is that the correlation coefficient is a measure of linear association, not agreement. We see that [4.12] does not depend on α, the fixed bias, and so two measurement systems can exhibit a strong linear relationship and hence be highly correlated even if the fixed bias is large [Bland & Altman 1983, 1986; Ludbrook 1997; 2002]. In this way, the correlation coefficient is "insensitive to inaccuracy" [Bookbinder and Panosian, 1987, p. 1170].

It is also possible that if β, the proportional bias, is very far from 1 the correlation between Y_{i1} and Y_{i2} can still be large. But if $\alpha \neq 0$ and $\beta \neq 1$, the measurements Y_{i1} and Y_{i2} are not likely to agree and so the two measurement systems should not be used interchangeably, despite a large correlation coefficient. For these reasons, the correlation coefficient is not an adequate measure of agreement, and so it should not be used to judge interchangeability.

When each measurement system makes multiple measurements ($r \geq 2$) on each subject, we may take the average of these replicate measurements and calculate the correlation for these averages. Based on [4.1], this correlation is given by

$$Corr(\bar{Y}_{i1\cdot},\bar{Y}_{i2\cdot}) = \frac{r\beta\sigma_s^2}{\sqrt{(r\sigma_s^2+\sigma_1^2)(r\beta^2\sigma_s^2+\sigma_2^2)}}$$

where $\bar{Y}_{ij\cdot} = \sum_{k=1}^{r} Y_{ijk}/r$. However, we see that this correlation can be arbitrarily inflated by increasing the number of replicate measurements r. As well, each of the problems discussed in the $r = 1$ case also exist for this case, and so we do not recommend use of the correlation coefficient to judge interchangeability when $r \geq 2$ either.

4.4 Regression

A closely related alternative to correlation is linear regression. The linear relationship between the single measurements ($r = 1$) made by each system on a particular subject i is given by:

$$Y_{i2} = \alpha + \beta Y_{i1} + M_i \qquad [4.13]$$

where $M_i = M_{i2} - \beta M_{i1}$, and M_{i1} and M_{i2} are as defined in [4.2].

The core of this analysis technique lies in the comparison of the fitted regression line $\hat{y}_2 = \hat{\alpha} + \hat{\beta}y_1$, to the line of equality $y_2 = y_1$, where y_j represents the observed measurements taken by system j on the same subject [Linnet, 1993; Ludbrook, 1997]. The idea is that the farther the fitted regression line is from $y_2 = y_1$, and hence $\hat{\alpha}$ from 0 and $\hat{\beta}$ from 1, the more evidence there is that the two measurement systems do not agree. Formally we can write the null and alternative hypotheses as follows [Linnet, 1993]:

$$H_0: (\alpha;\beta) = (0,1) \text{ versus } H_A:\ (\alpha;\beta) \neq (0,1) \qquad [4.14]$$

A variety of regression techniques can be used to test this hypothesis; the difference among them lies in the estimation procedure and its underlying assumptions. It is common to carry out the regression analysis based on ordinary least squares (OLS) [Linnet, 1993]. For OLS to be valid, we must assume that (i) the measurements made by the reference system are error free, i.e., $\sigma_1 = 0$, and (ii) the errors in the measurements made by the second system are constant, i.e., homoscedastic, and have mean 0 [Abraham and Ledolter, 2006]. Within the OLS framework, the regression coefficients α and β are estimated by $\hat{\alpha}$ and $\hat{\beta}$ which are calculated by solving the normal equations resulting from minimizing the sum of the squared vertical distance between each point and the fitted line.

However, in the context of MSC studies assumptions (i) and (ii) are often violated, in which case the OLS estimation approach is invalid [Linnet, 1993]. First, the independent variable (the measurements made by the reference system) is typically not measured exactly, i.e., $\sigma_1 \neq 0$. The exception to this is when the reference system is known to make measurements without error. Such a system is often referred to as a 'gold standard'. However, such a system is rare, and so the first assumption of OLS is usually broken.

Supposing we perform an OLS regression analysis even though the reference system does not produce error-free measurements, the slope estimate obtained by doing so would be biased toward zero [Berkson; 1950] and hence our conclusions could be misleading.

An alternative to OLS regression is ordinary Deming regression, which allows for error in the measurements of both systems. Estimation within this framework is achieved by minimizing the sum of squared deviations from the fitted line at an angle that is determined by the ratio of the repeatabilities (σ_1 and σ_2), which is assumed known [Deming, 1943; Linnet, 1993]. But unless σ_1 and σ_2 are known from a large prior study, this assumption is unreasonable. When $\sigma_1 \neq 0$ and $\sigma_2 \neq 0$ this method has been found to give unbiased and more efficient slope estimates than OLS [Linnet, 1993].

The necessity for weighted regression techniques arises when the second OLS assumption is violated: when the measurement variability is heteroscedastic, that is, not constant across the distribution of true values. This means that rather than σ_j being a constant value, it is instead some function of the true values. It is common in MSC studies for the variance of the error term to be proportional to the true values, or for the standard deviation to be proportional to the true values [Linnet, 1993].

In the weighted regression techniques, the variability of the random error need not be constant. Weighted least squares (WLS) regression is carried out by minimizing the sum of the weighted squared deviations from the fitted line in the vertical direction, where the weights are inversely proportional to the variability of the second measurement system at a given level [Linnet, 1993]. Although the WLS regression technique allows for heteroscedasticity in the second measurement system, it still assumes that the reference measurement system provides measurements without error. Thus the applicability of WLS is still minimal since a gold standard reference system is rarely available.

A more widely applicable, yet more complicated, approach is weighted Deming regression. Not only does it allow for measurement error to exist in both measurement systems, it allows for non-constant variability in both as well. This approach is similar to ordinary Deming regression except that here the sum of the weighted squared deviations from the fitted line are minimized. As before, the angle at which the deviations are minimized is determined by the ratio of repeatabilities σ_1 and σ_2. It is still assumed that this ratio is constant, which implies that although non-constant variability is allowed, its structure must be the same in both measurement systems. If the ratio is not constant then small biases may arise, yet the bias is at most of the same order of magnitude as that of OLS [Linnet, 1993]. Martin [2000] suggests that when the ratio of the standard deviations is not constant, a revised Deming approach called iteratively reweighted general Deming regression be used to obtain precise and unbiased estimates of the regression coefficients. We further discuss the comparison of heteroscedastic measurement systems in Chapter 6.

While different estimation techniques may be appropriate for different situations, the regression approach, regardless of estimation technique, is still flawed. As with [4.6] and [4.11], the hypothesis [4.14] could be more appropriately framed as an equivalence test [Wellek, 2010]. However, the chief weakness is that [4.14] focuses only on bias (α and β); it does not compare the repeatabilities of each system. It is possible that α may be close to 0, and β may be close to 1, but if one or both of the repeatabilities is large then the measurements Y_{i1} and Y_{i2} are unlikely to be similar, in which case using the two systems interchangeably is ill-advised. For example, if we assume H_0 in [4.14] is true, we still may not want to use the new system in place of the reference system if the new one much less precise.

Another weakness is that each of the estimation techniques discussed is predicated on the design in which subjects are measured just once by each system. When replicate measurements ($r \geq 2$) are available these estimation techniques are typically carried out on the averages of the replicate measurements, $\bar{Y}_{i1\cdot}$ and $\bar{Y}_{i2\cdot}$ [Linnet, 1993]. Because there is less variation between averages than between single measurements, this results in better estimates of α and β. But regardless of the number of replicate measurements, the hypothesis [4.14] still ignores σ_1 and σ_2, and so no comparison of repeatabilities is made. Thus the problematic scenario described in the previous paragraph still applies when $r \geq 2$.

4.5 The Limits of Agreement Approach

The usefulness of the previous four approaches in assessing interchangeability is limited, in part, because they do not simultaneously assess the relative bias and the relative repeatability of the two systems. One approach that attempts to do this is the 'limits of agreement' approach, due to Bland and Altman [1983; 1986]. This approach appears to be the most widely used technique for assessing interchangeability of measurement systems in clinical contexts. It was first introduced by Altman and Bland in 1983, but the wide uptake did not begin until the publication of their second paper on the topic which appeared in the *Lancet* in 1986 [Bland and Altman, 1986]. This article has been nearly 30 000 times and is one of the ten most frequently cited statistical articles ever [Ryan and Woodall, 2005].

4.5.1 The Standard Approach

To describe this technique, suppose we have only one measurement by each system on each subject, and hence model [4.2] applies. The limits of agreement approach characterizes the agreement between two measurement systems by evaluating the difference between measurements made on the same subject by each system: $D_i = Y_{i2} - Y_{i1}$.

Using a scatter plot, the observed differences, $d_i = y_{i2} - y_{i1}$, are compared to the observed average of the two measurements made on a given subject by each system: $a_i = (y_{i1} + y_{i2})/2$. This is known as the "difference plot" where, for a particular subject i, the cartesian points have the form (a_i, d_i), $i = 1,2, \dots, n$.

One purpose of the plot is to evaluate whether the differences are related to the averages, surrogates for the unknown true values. If no relationship appears to exist, the distribution of the differences is summarized by the limits of agreement, defined as:

$$\hat{\mu}_d \pm 1.96\hat{\sigma}_d \qquad [4.15]$$

where $\hat{\mu}_d = \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ and $\hat{\sigma}_d = s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}$ (as in [4.4] and [4.5]) are sample estimates of μ_d and σ_d. Assuming the differences roughly follow a normal distribution ($D_i \sim N(\mu_d, \sigma_d^2)$), these limits represent the interval within which we expect 95% of the differences to lie. Horizontal reference lines corresponding to the upper and lower limits of agreement and the average difference $\bar{d}$ are added to the plot.

To decide whether two measurement systems agree sufficiently to be used interchangeably, we compare the limits of agreement to the clinically acceptable difference (CAD) [Bland and Altman, 1983; 1986]. The CAD is the maximum allowable difference between two measurements that would not adversely affect clinical decisions. How far apart two measurements can be before it causes difficulties is not a statistical question, instead the answer must be based on clinical judgement and should be made prior to executing the MSC study [Bland and Altman, 1983; 1986]. In many situations the CAD is defined as a symmetric interval around zero: $(-c, c)$. Note that the term “clinically” is relevant in medical contexts, but we could similarly define a practically acceptable difference for use in other settings.

By incorporating a clinically, or practically, acceptable difference in the analysis, the limits of agreement approach provides context for the comparison, which the other methods of comparison fail to do. With this approach, interchangeability is not simply judged by a statistical test, practical considerations must also be made.

To describe the decision process, we refer to the upper and lower limits of agreement as ULA and LLA, respectively. If the limits of agreement are contained within the CAD, i.e. $-c \leq LLA < 0 < ULA \leq c$, we conclude that the differences will be clinically acceptable at least 95% of the time, and the measurement systems are interchangeable. Otherwise, if the limits of agreement fall outside the CAD, it is likely that measurements by the two systems will differ by more than the allowable amount. In this situation we conclude that the two measurement systems do not agree sufficiently and should not be used interchangeably.

Checking whether zero is contained within the CAD ensures that there is no significant relative bias, and comparing the limits of agreement to the CAD ensures that the variability of the differences is not excessive. In this way, the limits of agreement approach simultaneously addresses bias and variability.

While the difference plot addresses repeatability to a degree, Bland and Altman suggest that the explicit comparison of repeatabilities is also necessary for an informed comparison [1986; 1999].

> "Lack of repeatability can interfere with the comparison of two methods because if one method has poor repeatability, in the sense that there is considerable variation in repeated measurements on the same subject, the agreement between the two methods is bound to be poor. Even if the measurements by the two methods agreed very closely on average, poor repeatability of one method would lead to poor agreement between the methods for individuals." [Bland and Altman, 1999, p. 149]

However, in order to estimate the repeatability of *each* system, the MSC study must include $r \geq 2$ replicate measurements by each system on each subject, which is seldom done. Because the standard limits of agreement approach does not account for replicate measurements, Bland and Altman have suggested amendments to the approach that incorporate this extra information.

In fact Bland and Altman have authored many articles since the introduction of the limits of agreement technique which clarify the method and guide its use in non-standard situations. Specifically, they suggest alternate methods of calculating limits of agreement if the differences appear to depend in some way on the true value [Bland and Altman, 1999], or if each system makes replicate measurements on each subject [Bland and Altman, 2007]. We briefly describe these modifications in the following two subsections.

4.5.2 The Limits of Agreement when Differences are Related to True Values

When the differences depend in some way on the true values, the difference plot will display a non-random scatter of points (as the averages are thought of as a surrogate for the unknown true values). Two common non-random patterns that the difference plot may exhibit are a linear trend, which arises in the presence of proportional bias ($\beta \neq 1$), and a funnel-shaped trend, which arises when one or both systems is heteroscedastic (σ_j is non-constant). In both of these cases the limits of agreement as defined by [4.15] are no longer applicable; they would be too wide or too narrow for certain true values. As such, an alternate formulation must be used in these cases.

If the difference plot displays evidence of proportional bias or heteroscedasticity, Bland and Altman suggest that we deal with “such deviations from assumptions” by making a suitable transformation of the raw data [Bland and Altman, 1995, p. 1085]. The transformation that they suggest is the natural logarithm: “Under these circumstances, logarithmic (log) transformation of both measurements before analysis will enable the standard approach to be used. The limits of agreements derived from log transformed data can be back-transformed to give limits for the ratio of actual measurements” (Bland & Altman 1999, p. 143-144).

By following this prescription, the back transformed limits of agreement are no longer in terms of a simple difference, they are now in terms of a ratio of the original measurements. This may not be a large problem, but when deciding whether the two measurement systems are interchangeable, the pre-defined CAD needs to be in terms of a percent difference and not an absolute difference. For instance, a 3% difference in measurements may be tolerated.

It seems beneficial that if one is committed to using the limits of agreement approach to analyze the MSC data, the clinically acceptable difference should be defined in terms of both absolute and percent-differences. By doing this, the conclusions and interpretations are safeguarded against the possibility of having to log-transform the data. However, this might not be clinically feasible.

Bland and Altman also acknowledge however, that sometimes the relationship between the differences and averages is more complicated and log transformation may not solve the problem [Bland and Altman, 1999]. In this situation, they propose modeling the relationship between differences and averages, and performing formal tests for proportional bias and heteroscedasticity. They propose a simple linear regression of the differences, $D_i = Y_{i2} - Y_{i1}$, on the averages, $A_i = (Y_{i1} + Y_{i2})/2$ [Bland and Altman, 1999]. Based on model [4.2] we have

$$D_i = \vartheta_0 + \vartheta_1 A_i + M_i \qquad [4.16]$$

where $M_i = 2(M_{i2} - \beta M_{i1})/(\beta + 1)$, $\vartheta_0 = 2\alpha/(\beta + 1)$, and $\vartheta_1 = 2(\beta - 1)/(\beta + 1)$.

If the slope ϑ_1 is not significantly different from zero, then $\beta \approx 1$ and we return to the standard case discussed in Section 4.5.1. If, however, ϑ_1 is significant then the estimated difference between any two measurements depends on the average of those measurements, and is given by $\hat{\vartheta}_0 + \hat{\vartheta}_1 a_i$. The limits of agreement in this case are given by

$$\hat{\vartheta}_0 + \hat{\vartheta}_1 a_i \pm 1.96\hat{\sigma}_d^* \qquad [4.17]$$

where

$$\hat{\sigma}_d^* = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} r_i^2}$$

and r_i is the residual for subject i associated with the regression in [4.16]. Note that the '*' here is used to differentiate this estimate from $\hat{\sigma}_d$ given in [4.5].

The limits of agreement in [4.17] correspond to an interval around the fitted regression line associated with [4.16], and visually they are straight lines equidistant and parallel to this regression line.

To account for the possibility of heteroscedasticity Bland and Altman [1999] suggest regressing the absolute value of the residuals r_i, on the averages a_i, $i = 1,2 \dots., n$. If this regression is not significant then the variability in differences does not depend on the averages, and so we calculate the limits of agreement as in [4.17]. However, if the regression is significant, then heteroscedasticity is present and the limits of agreement must account for this.

When this is the case Bland and Altman [1999] suggest that the limits of agreement be defined as in [4.17] but where, for a particular subject i, $\hat{\sigma}_d^*$ is substituted for the fitted value associated with regression of $|r_i|$ on a_i. These limits of agreement have been dubbed "V-shaped limits" [Ludbrook, 2010] as they look like a rotated letter 'V' with the points scattered within.

As a test for heteroscedasticity, this works well if the structure is such that the standard deviation of the differences is a linear function of the true values, but if the relationship between differences and true values is something more complicated, this test may not accurately identify it. We further discuss the comparison of heteroscedastic measurement systems in Chapter 6.

The ability to calculate these more accurate and well-behaved limits of agreement might be mathematically attractive, but their applicability is still limited. These alternate formulations do not account for replicate measurements by each system on each subject, and as we will see in Section 5.1.2, there is not a similar adaptation when replicate measurements are available.

4.5.3 The Limits of Agreement with Replicate Measurements

When an MSC study includes $r \geq 2$ replicate measurement by each system on each subject, it is desirable to use all of the data to compare measurement systems. As such, Bland and Altman developed a modification to the limits of agreement approach that incorporates replicate measurements [1999; 2007].

In particular, they recommend averaging the replicate measurements on a single subject by a particular measurement system, and constructing the difference plot using the differences and averages of the averaged measurements on each subject.

If we denote the difference between the k^{th} replicate measurements by each system on subject i by $d_{ik} = y_{i2k} - y_{i1k}$, the difference in the averages of replicate measurements on subject i is given by

$$\bar{d}_{i\cdot} = \bar{y}_{i2\cdot} - \bar{y}_{i1\cdot}$$

where $\bar{y}_{ij\cdot} = \sum_{k=1}^{r} y_{ijk}/r$ and y_{ijk} is the observed value of Y_{ijk} as defined in [4.1]. The average of the averages of replicate measurements on subject i is similarly given by

$$\bar{a}_{i\cdot} = \frac{\bar{y}_{i1\cdot} + \bar{y}_{i2\cdot}}{2}$$

and the Cartesian points on the difference plot are now $\left(\bar{a}_{i\cdot}, \bar{d}_{i\cdot}\right)$ for $i = 1,2,\dots,n$.

When constructing limits of agreement for this plot, we can no longer use those defined in [4.15], as they will be too narrow; working with the average of replicate measurements instead of the individual measurements results in a reduction of measurement variation. Because the limits of agreement are meant to depict the typical range of differences between single measurements, they must be constructed with an estimate of σ_d that is based on the standard deviation of differences between single measurements, not between averages of several replicates. Accordingly, Bland and Altman adjust the calculation of $\hat{\sigma}_d$ to account for this.

When both measurement systems make r replicate measurements on each subject, σ_d is estimated by

$$\hat{\sigma}_d^{**} = \sqrt{\hat{\sigma}_{\bar{d}}^2 + \left(1 - \frac{1}{r}\right)(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)}$$

where $\hat{\sigma}_{\bar{d}}^2$ is the observed variance of the differences between the within subject means $\bar{d}_{i\cdot}$, $i = 1,2, \dots, n$, and where

$$\hat{\sigma}_j^2 = \frac{1}{n(r-1)} \sum_{i=1}^{n} \sum_{k=1}^{r} \left(y_{ijk} - \bar{y}_{ij\cdot}\right)^2$$

is the estimate of the within-subject variability (i.e. the repeatability), of system j. Note that we use '**' to distinguish $\hat{\sigma}_d^{**}$ from $\hat{\sigma}_d^{*}$ discussed in the previous subsection.

Thus, when each system makes $r \geq 2$ replicate measurements on each of the n subjects in the MSC study, and the difference plot is constructed using the differences and averages of the within-subject means, the limits of agreement are defined as

$$\hat{\mu}_d \pm 1.96\hat{\sigma}_d^{**} \qquad [4.20]$$

With the difference plot and limits of agreement constructed in the manner just described, the assessment of interchangeability proceeds as in the $r = 1$ case.

4.6 Looking Ahead

In this chapter, we have introduced the notion of comparing measurement systems. We have described a measurement system comparison (MSC) study, the goal of which is to determine if the two systems being compared agree well enough to be used interchangeably. In Sections 4.1-4.5 we described a number of methods for analyzing MSC study data, and for assessing the agreement between two measurement systems.

The comparison of means, comparison of repeatabilities, correlation and regression have all been criticized as methods for judging interchangeability because none of them simultaneously assess the bias and repeatability of the measurement systems, and because interchangeability is based on statistical significance, not practical importance. Bland and Altman proposed the limits of agreement approach as an alternative analysis method which aimed to overcome this challenge.

In Section 4.5 we described the limits of agreement approach in its simplest form, and in other more complicated situations. In Chapter 5 we identify a variety of problems associated with the limits of agreement approach that can result in misinterpretation of the relationship between the two measurement systems being compared, and that can ultimately lead to incorrect conclusions regarding their interchangeability.

The problems we discuss demonstrate the need for a method for quantifying the agreement between two measurement systems that is more transparent and more informative. In Chapter 5 we also introduce the *probability of agreement*, an intuitive and instructive metric, that more effectively and transparently quantifies the agreement between two measurement systems [Stevens et al., 2014 (*under revision*)]. The analysis method we propose consists of plotting the probability of agreement across the distribution of true values, thus summarizing the agreement between systems for any true value of interest. We also consider the design of an MSC study, in light of the probability of agreement, which allows for optimal estimation.

In Chapter 6 we relax various model assumptions made in Chapter 5 and consider the application of the probability of agreement when comparing two measurement systems in these more general settings. In particular, we consider the situation in which the true values of the measurand do not follow a normal distribution, and when the measurement variation of one or both systems depends on this unknown true value. One of the key advantages of the probability of agreement analysis is its intuitive interpretation; it can be easily understood by non-statisticians and this interpretation does not change if we change assumptions and adjust the model. Application of the probability of agreement in the aforementioned settings serves to demonstrate the versatility of the suggested analysis method.

Chapter 5

The Probability of Agreement: An Alternative to the Limits of Agreement Analysis

In Chapter 4 we introduced the idea of comparing measurement systems, where the goal is to determine whether a new measurement system agrees well enough with an existing one, for the two to be used interchangeably. This goal is achieved by a measurement system comparison (MSC) study in which a number of subjects are measured one or more times by each system. The data from this study are then analyzed and the agreement between the two system's measurements is assessed.

A variety of statistical techniques can be used to analyze the data from an MSC study and to help determine whether the two measurement systems can be used interchangeably. In Chapter 4 we reviewed five such techniques: comparison of means, comparison of repeatabilities, correlation, regression, and the limits of agreement approach. Four of these methods, correlation, regression and the comparison of means and repeatabilities, were deemed inappropriate measures of agreement because individually they do not simultaneously address the relative bias, and the relative repeatability, of the two systems. As well, in each of these methods the interchangeability of two systems is determined by the statistical significance of a hypothesis test; the practical significance, and hence the context of the comparison, is not considered.

In Section 4.5, we described the limits of agreement approach which was proposed by Bland and Altman [1983; 1986] as an alternative analysis method to overcome these challenges. However, this method faces its own challenges, which we critically examine in Section 5.1.

Given the problems with existing methods, we propose a new method for assessing interchangeability which is based on the probability of agreement [Stevens et al., 2014 (*under revision*)]. This method accurately and transparently quantifies the agreement between two measurement systems in an intuitive manner.

In Section 5.2 we introduce the analysis method by defining the probability of agreement and associated plot, we describe the maximum likelihood-based estimation procedure, and discuss the assessment of model assumptions. We also illustrate the analysis method with an example from the literature [Bland and Altman, 1999]. Then in Section 5.3 we discuss the design of an MSC study, and make recommendations for a design which facilitates precise estimation of the probability of agreement. We end the chapter with a summary and discussion in Section 5.4.

Before proceeding however, we restate model [4.1] as it will help to describe the problems associated with the limits of agreement approach, and it is necessary to describe the probability of agreement analysis.

To describe the data collected during an MSC study, we propose the following mixed effects model:

$$\begin{aligned} Y_{i1k} &= S_i + M_{i1k} \\ Y_{i2k} &= \alpha + \beta S_i + M_{i2k} \end{aligned} \qquad [5.1]$$

where Y_{ijk} is a random variable which represents the value observed on system j's k^{th} measurement of subject i; $i = 1,2,\dots,n$; $j = 1,2$; $k = 1,2,\dots,r$. Note that $j = 1$ indexes the reference measurement system and $j = 2$ indexes the new measurement system. Recall S_i is a random variable that represents the unknown true value of the measurand for subject i, with the distributional assumption $S_i \sim N(\mu, \sigma_S^2)$, and M_{ijk} is a random variable which represents the measurement error when system j makes multiple measurements on subject i. We further assume that the M_{ijk} are independent of each other and independent of S_i, and that $M_{ijk} \sim N\left(0, \sigma_j^2\right)$ where σ_j quantifies the measurement variation, or repeatability, of system j. The parameters $-\infty < \alpha < \infty$ and $\beta > 0$ quantify the bias of the second (new) measurement system relative to the existing system. We refer to α as the fixed bias and we call β the proportional bias [Ludbrook, 2010].

As in Chapter 4, model [5.1] does not include observer effects and so the repeatability σ_j quantifies the overall measurement system variation of system j (i.e. no reproducibility). As well, we assume that the measurement systems being compared are both homoscedastic; that is, σ_j is constant across the true values of the measurand for $j = 1,2$. We consider the heteroscedastic case in Chapter 6.

Based on [5.1], we say that the two measurement systems are identical if $\alpha = 0, \beta = 1$ and $\sigma_1 = \sigma_2$. In Chapter 4 we noted that two systems do not need to be identical to be used interchangeably, and we informally stated that two measurement systems could be used interchangeably if, most of the time, their measurements on the same subject are sufficiently similar, i.e., $Y_{i1k} \approx Y_{i2k}$.

We can reformulate this criteria in terms of the difference between single measurements by each system on a given subject, $D_i = Y_{i2k} - Y_{i1k}$: two measurement systems can be used interchangeably if, most of the time, the differences D_i are small relative to a clinically (or practically) acceptable difference. It is by this criterion that two systems will be deemed interchangeable by the limits of agreement analysis, and also by the probability of agreement analysis.

5.1 Problems with the Limits of Agreement Approach

As discussed in Section 4.5, the limits of agreement approach was proposed by Bland and Altman [1983; 1986] as an alternative to other methods of assessing interchangeability. By evaluating the differences between single measurements on a given subject by the two systems, this technique is meant to determine whether the measurements are similar enough for the two systems to be used interchangeably. This analysis method is based on the difference plot: a scatter plot of the observed differences versus the observed averages of the measurements made by both systems on each subject. Although more informative than correlation, regression, and a comparison of means and repeatabilities, this approach has challenges of its own that can effect judgements of interchangeability.

5.1.1 Problems when there are no replicate measurements ($r = 1$)

One of the most consequential problems is that, although Bland and Altman [1995; 1999] recommend measuring each subject two or more times by each measurement system, this is seldom done in practice. This could, in part, be because the example presented in their landmark *Lancet* paper ignores the fact that each system made two measurements on each subject, and uses only the first measurement on each subject to compare the two systems.

The difference plot, as described in Section 4.5.1, does not provide adequate information about the relationship between the two systems, and without the additional information gained by replicate measurements, the difference plot can be misleading. To fully understand the relationship between the two measurement systems, and hence decide if the two systems are interchangeable, it is important to estimate all of the parameters in [5.1].

However, by not making replicate measurements we cannot separately estimate these parameters, a limitation Voelkel and Siskowski [2005] referred to as the *problem of indeterminacy*. Specifically, we can only estimate the following five quantities: $E(Y_{i1}) = \mu$, $E(Y_{i2}) = \alpha + \beta\mu$, $Var(Y_{i1}) = \sigma_s^2 + \sigma_1^2$, $Var(Y_{i2}) = \beta^2\sigma_s^2 + \sigma_1^2$ and $Cov(Y_{i1}, Y_{i2}) = \beta\sigma_s^2$. Recall that in the absence of replicate measurements ($r = 1$), we eliminate the subscript k.

One consequence of no replicate measurements is that without separate estimates of α and β, we cannot distinguish between fixed and proportional bias, and so the biases become confounded. As well, without separate estimates of the two repeatabilities, σ_1 and σ_2, we cannot determine which system is more precise, and we risk rejecting interchangeability with a new measurement system which is more precise than the existing one. Unfortunately the difference plot cannot disentangle confounding biases, and it does not indicate which system is more precise, so replicate measurements are necessary to understand how the measurements by the two systems are related.

To illustrate the effect of not explicitly estimating and comparing σ_1 and σ_2, consider the comparison of two measurement systems where both systems are unbiased ($\alpha = 0$ and $\beta = 1$). In this situation, the standard deviation of the differences, $D_i = Y_{i2} - Y_{i1}$ is $\sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2}$, which is estimated by s_d, defined in [4.5]. When using the limits of agreement approach to decide whether a new measurement system is ‘as good as’ the reference one, the decision depends on how ‘good’ the reference system is. For example, when σ_1 is large but σ_2 is small, σ_d and hence s_d might still be large enough to push the limits of agreement outside the CAD, leading one to reject interchangeability. When this happens we reject interchangeability with a new measurement system that is more precise than the existing one, even though both are unbiased. Thus practitioners who uncritically apply the limits of agreement technique can be misled. Bland and Altman [1986; 1999] acknowledge that in using their technique this problem is a possibility. We feel that this is a potentially serious problem that practioners want to avoid.

The absence of replicate measurements and hence estimates of repeatability can result in another problem which we call *false correlation* [Stevens et al., 2014 (*under revision*)]. Recall that one purpose of the difference plot is to detect whether there is a relationship between the differences and the averages (which are a surrogate for the unknown true values of the measurand). By using the averages on the horizontal axis, we are supposedly protected against the appearance of a pattern when no real relationship between differences and true values exists. Bland and Altman suggest that the correlation between the differences and averages should be zero because the variability of each measurement system should be the same: "as they should if they are measurements of the same thing" [2003, p. 91]. However, just because both systems are measuring the same thing, does not imply that the repeatabilities should be the same. As such, assuming $\sigma_1 = \sigma_2$ may not be valid. The results below demonstrate that in situations where this assumption does not hold, the correlation between differences and averages is not zero, and in fact can be quite large.

Because we wish to determine whether a correlation exists in the absence of any true relationship, we use the components of model [5.1] with $\beta = 1$ to calculate the correlation between $D = Y_2 - Y_1$ and $A = (Y_1 + Y_2)/2$. Note that we exclude the subscripts i and k on Y_{ijk} because we are assuming replicate measurements have not been made, and the distribution of D and A is the same for all subjects. With this amendment in notation, the correlation between D and A is given by

$$Corr(D, A) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_s^2 + \sigma_1^2 + \sigma_2^2)}} \qquad [5.2]$$

Clearly the correlation is zero and hence D and A are uncorrelated if $\sigma_1 = \sigma_2$. But if they are not equal, which is a more realistic assumption, the differences and averages are correlated.

We have investigated the magnitude of this correlation for varying relative sizes of σ_1, σ_2 and σ_s. Figure 5.1 provides a contour plot which shows the correlation between differences and averages for various values of σ_s/σ_1 and σ_2/σ_1. As we can see, the resulting correlations can be different from zero, and in some cases very different from zero. It is clear that if σ_1 and σ_2 are small compared to the overall subject variation σ_s (i.e., σ_s/σ_1 is large and σ_2/σ_1 is not), then the correlation between differences and averages is not likely to differ materially from zero. But as the variation due to the measurement systems increases, the correlation increases. For example, if σ_s/σ_1 and σ_2/σ_1 are both 3, which might be common in practice, we get a correlation of -0.37.

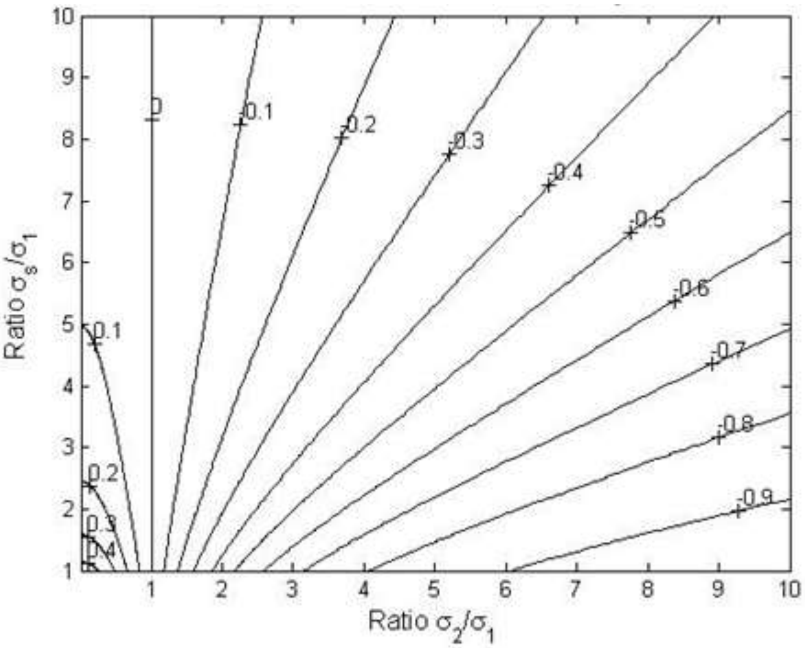

Figure 5.1: Contours of the correlation between differences and averages for various relative sizes of σ_1, σ_2 and σ_s

A serious issue arises here. In the absence of any true relationship between differences and true values, the difference plot can suggest a significant (positive or negative) relationship exists. This may lead an unsuspecting practitioner to log-transform their data when it is not warranted, or miscalculate the limits of agreement (see Section 4.5.2). As well, the presence of a false negative correlation could mask the existence of a true positive relationship, and vice versa. Thus the existence of a false correlation can confuse the relationship between two measurement systems and may lead to misinformed judgments of interchangeability. However, by taking replicate measurements on each subject with each system, we can obtain the estimates $\hat{\sigma}_1$, $\hat{\sigma}_2$ and $\hat{\sigma}_s$, and hence estimate the theoretical correlation in [5.2] to help determine whether any relationship seen on the difference plot is due to β being different from 1, or just a consequence of the relative sizes of the variance components.

It should be clear that without performing replicate measurements on each subject by each system, the limits of agreement approach can be misleading; without replicated measurements we cannot adequately assess bias or repeatability, potentially causing a researcher to make the wrong decision regarding interchangeability. As such we do not recommend the use of the limits of agreement technique when replicate measurements are not available. In fact, we do not recommend the comparison of measurement systems at all, if replicate measurements are not available.

5.1.2 Problems when there are replicate measurements ($r > 1$)

As noted in Section 4.5, Bland and Altman [1999; 2007] suggest extensions to the limits of agreement technique when replicate measurements are available ($r > 1$). In this case they recommend averaging the replicate measurements on a single subject by a particular measurement system, and constructing the difference plot using the differences and averages of the averaged measurements on each subject. By doing this, the typical limits of agreement are too narrow and so they adjust the estimate of the variability in differences to account for the reduction in measurement variation that results from working with the average of replicate measurements instead of individual measurements. Although this results in limits that more accurately reflect the distribution of differences in single measurements, the approach is not without difficulties.

First, by plotting averages of the replicate measurements, a transparent display of the raw data is unavailable. A plot of the averages can mask large differences in the replicate measurements on the same subject by each system, and can make the level of agreement between the two measurement systems appear stronger than it truly is. A second issue is that Bland and Altman's method of calculating the limits of agreement in this situation assumes that "the difference between the two methods is reasonably stable across the range of measurements" [2007, p. 572]. In other words, this technique assumes there is no proportional bias ($\beta = 1$), and so its applicability is limited. A third problem is that although replicate measurements are made, there is no explicit comparison of repeatabilities, i.e. σ_1 and σ_2 in [5.1], and so it is still possible to reject interchangeability with a more precise measurement system if the measurement variation in the reference system is large.

5.1.3 Misuse of the Technique

Another issue that exists – that is not a direct function of the limits of agreement method – is that, in general, the technique is widely misused. In fact, Bland and Altman acknowledge the misuse of their technique when they say "the 95% limits of agreement method has been widely cited and widely used, though many who cite it do not appear to have read the paper" [2003, p. 91]. To investigate this, Mantha et al. [2000] and Dewitte et al. [2002] undertook large-scale literature reviews of MSC studies analyzed by the limits of agreement technique, and found a variety of problems. For example, common errors were that repeatability was often not assessed, the limits of agreement were incorrectly calculated, the axes on the difference plots were mis-specified, and that relationships between the differences and averages were often ignored.

However, the most pervasive and alarming was that in more than 90% of the articles examined the authors did not define a clinically acceptable difference, let alone compare their limits of agreement to one. Clearly these authors forgot, or were unaware that the crux of the limits of agreement approach, and the basis upon which interchangeability is determined, is the comparison of limits of agreement to the clinically acceptable difference. Without this comparison, the judgement of interchangeability is difficult and can be misinformed.

In the absence of a clinically acceptable difference, authors would conclude that agreement between systems is indicated by the fact that 'most' of the differences fall within the upper and lower limits of agreement. To this point Bland and Altman have replied:

> "The very wide uptake of the limits of agreement approach has naturally been very pleasing. We have been aware, however, that sometimes the method has not been adopted with full understanding. For example, we have seen it suggested that two methods agree well because most of the observations lie within the 95% limits of agreement. The limits are calculated so that this will always be the case." [2002, p. 802]

Because these two previous literature reviews were published over ten years ago, we have conducted an updated review. In line with these previous reviews, we consulted articles in the field of clinical chemistry. Specifically we reviewed articles in *Clinical Chemistry* and the

Annals of Clinical Biochemistry between 2002 and 2012, inclusive. In these journals there were respectively 85 and 33 articles which cited Bland and Altman's 1986 *Lancet* paper. Of these, 73 and 29 articles documented actual MSC studies. As in Mantha et al. [2000], we evaluated each of these 102 articles on four main criteria:

- Was a clinically acceptable difference defined and used to assess agreement?
- Was the repeatability of both systems assessed?
- Was the horizontal-axis correctly specified as the average of measurements by each system?
- Was the relationship between differences and averages considered?

	Total	*Clinical Chemistry*	*Annals of Clinical Biochemistry*
CAD	16/102 (15.69%)	11/73 (15.07%)	5/29 (17.24%)
Repeatability	25/102 (24.5%)	16/73 (21.92%)	9/29 (31.03%)
X-axis[a]	77/88 (87.50%)	56/63 (88.89%)	21/25 (84.00%)
Relationship[b]	25/73 (34.25%)	19/53 (35.85%)	6/20 (30.00%)

Table 5.1: Results of MSC literature review. Values are ratio (%).

[a] The denominators were 88, 63 and 25, respectively, corresponding to the number of articles which displayed a difference plot.

[b] The denominators were 73, 53 and 20, respectively, corresponding to the number of articles with evidence of proportional bias or heteroscedasticity.

Table 5.1 summarizes the results of this analysis. We highlight key findings below:

- Only 16% of studies defined a clinically acceptable difference and compared this with the limits of agreement to determine whether or not the systems were interchangeable. Although this percentage is higher than in the previous literature reviews, it is still very low.
- Roughly 58% of the studies assessed the repeatability of one or both of the systems being compared. However, in only 24.5% of the articles was the repeatability of both systems determined.
- Roughly 87% of the studies correctly specified the horizontal axis as the average of the measurements made by each method. Although this percentage seems high, given the

volume of literature Bland and Altman have devoted to this topic, this is a problem that should rarely be made.

- Proportional bias and/or heteroscedasticity was visually identified in 73 of the 88 articles in which a difference plot was presented. Among these 73 articles, 53 acknowledged that such a relationship was present, but in only 25 of them did the authors deal with this issue. Common methods for eliminating proportional bias and heteroscedasticity were to log-transform the data, plot the differences as a percent of the averages, or to plot the ratio of the two measurements vs. the averages.

The results of this literature review demonstrate that the application of the limits of agreement technique has improved somewhat in recent years, but, particularly with regards to defining clinically acceptable differences and assessing repeatability, there is still a long way to go.

The success of the limits of agreement approach for assessing interchangeability has largely been due to its perceived simplicity. In this section, we have demonstrated some of the challenges associated with the approach, many of which arise because the approach over-simplifies the relationship between two measurement systems. It is clear that there is need for an analysis method that more accurately quantifies the agreement between two measurement systems and that is safeguarded against misuse.

5.2 The Probability of Agreement Analysis

In this section we propose a new method of analysis for comparing measurement systems as an alternative to the limits of agreement approach. We propose a simple metric, the probability of agreement, and an associated plot to clearly quantify the agreement between two measurement systems and hence help to decide whether the two systems can be used interchangeably. This approach seeks to overcome the deficiencies of the limits of agreement technique described in the previous section, and the deficiencies of the approaches discussed in Sections 4.1 – 4.4.

5.2.1 The Probability of Agreement

As with the limits of agreement technique, this approach examines the difference between single measurements made by two systems, and compares this difference to a clinically acceptable difference. Note that the term “clinically acceptable difference” was coined in the medical context, but we could equivalently refer to the concept as a range of "practically acceptable differences" in other contexts.

Here we assume the clinically acceptable difference has the form $CAD = (-c, c)$. Using this and the notation associated with model [5.1] we define $\theta(s)$, the *probability of agreement* as

$$\theta(s) = P(|Y_2 - Y_1| \leq c | S = s) \qquad [5.3]$$

The probability of agreement is the probability that the difference between single measurements on the same subject by the two systems falls within the range that is deemed to be acceptable, conditional on the value of the measurand. Note that we exclude the subscripts i and k that are used in [5.1] because we are interested in the difference between single measurements and the distribution of S_i and $Y_{i2} - Y_{i1}$, conditional on the true dimension S_i, is the same for all subjects. Based on [5.1], $\theta(s)$ can be written as:

$$\theta(s) = \Phi\left(\frac{c - \alpha - (\beta - 1)s}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) - \Phi\left(\frac{-c - \alpha - (\beta - 1)s}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \qquad [5.4]$$

where $\Phi(x)$ is the standard normal cumulative distribution function evaluated at x.

Using probabilities of this form, we construct the *probability of agreement plot* which graphically displays the estimated probability of agreement across a range of plausible values for s, the true value of the measurand. On this plot we include approximate pointwise confidence intervals for each value of $\theta(s)$ which reflect the uncertainty associated with its estimation. In Section 5.2.2 we describe how to obtain the standard errors necessary for calculating such confidence intervals.

This plot serves as a simple tool for displaying the results when comparing two measurement systems; it summarizes agreement transparently and directly while accounting for possibly complicated bias and variability structures. While the modelling and estimation of $\theta(s)$ is somewhat complicated, its interpretation is extremely simple and one that most non-statisticians can understand.

Another benefit is that even if a more complicated model than [5.1] is assumed, the interpretation of the probability and the plot is unchanged. For example we may wish to relax the assumption that S_i is normally distributed, or perhaps we wish to model measurement variation heteroscedasticity. In both cases we might alter model [5.1], but our interpretation of the probability of agreement and of the probability of agreement plot remains the same. We address both of these scenarios in Chapter 6.

Note that for simplicity we have assumed $CAD = (-c, c)$; this technique, however, easily generalizes to the situation where the clinically acceptable difference is not symmetric: $CAD = (c_1, c_2)$ and when the endpoints depend on the true value of the measurand: $CAD = \left(c_1(s), c_2(s)\right)$. As an example, the clinically acceptable difference may increase proportionally with s, i.e. for non-zero s we may have $CAD = (-0.1s, 0.1s)$.

The probability of agreement, $\theta(s)$, is a conditional quantity that depends on the value $S = s$. If $\theta(s)$ is largely unchanged across the possible values for $S = s$, or if we simply wish to focus on the most likely values of $S = s$, we may summarize the probability of agreement, with a single number.

To do so, we define an *unconditional probability of agreement*, denoted θ, which is, in a sense, the average value of $\theta(s)$ across the distribution of values of S. Using model [5.1], the unconditional probability of agreement is:

$$\begin{aligned}\theta &= P(|Y_2 - Y_1| \leq c) \\ &= \Phi\left(\frac{c - \alpha - (\beta - 1)\mu}{\sqrt{(\beta - 1)^2\sigma_s^2 + \sigma_1^2 + \sigma_2^2}}\right) - \Phi\left(\frac{-c - \alpha - (\beta - 1)\mu}{\sqrt{(\beta - 1)^2\sigma_s^2 + \sigma_1^2 + \sigma_2^2}}\right)\end{aligned} \qquad [5.5]$$

Use of this single-number summary is valid when $\theta(s)$ is similar for all values of s, or when the range of measurand values of interest is close to the mean, μ.

Thus we first recommend constructing the probability of agreement plot which is based on the (conditional) probability of agreement $\theta(s)$, and then we recommend using the plot to determine whether $\theta(s)$ depends substantially on s, and hence whether the use of the unconditional probability of agreement θ is appropriate.

Whether using the probability of agreement analysis, the probability that is deemed to indicate acceptable agreement and hence interchangeability is context-specific and is not a statistical decision. Accordingly, we demonstrate how to estimate and interpret $\theta(s)$ and θ, but how large they need to be to indicate interchangeability must be decided by the user. One reasonable choice might be to require $\theta(s), \theta \geq 0.95$, similar to the limits of agreement approach.

If the probability of agreement plot does not indicate acceptable agreement, (i.e. $\theta(s)$ is too low in the range of interest for s), then we recommend looking at the separate estimates of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$ to determine the source of disagreement. Although the probability of agreement plot is informative and simple to interpret, examining the individual parameter estimates is the most informative description of the relationship between the two system's measurements. We describe how to obtain these estimates in Section 5.2.2.

Note that the individual measurements made in the MSC study are used in the estimation procedure to construct the probability of agreement plot, but they are not explicitly displayed. For this reason we also propose the use of diagnostic plots that display the individual data points and that allow us to check various assumptions made by model [5.1]. We discuss this further in Section 5.2.3.

5.2.2 Maximum Likelihood Estimation

In order to estimate $\theta(s)$ and construct the probability of agreement plot, we use maximum likelihood estimation to obtain estimates of the parameters in model [5.1]. Using Matlab [The MathWorks Inc., 2013], we numerically maximize the log-likelihood function to obtain the maximum likelihood estimates $\left(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_s\right)$ which are substituted into [5.4] and [5.5] to obtain $\hat{\theta}(s)$ and $\hat{\theta}$. We sketch the derivation of the log-likelihood function and the estimation procedure here.

Because we assume subjects are independent from one another, we can write the log-likelihood contribution for a single subject and obtain the full log-likelihood function by summing these components. Thus we begin by deriving the log-likelihood function for a single subject.

For a particular subject i, we order the random vector corresponding to its measurements by system and write $\vec{Y}_i = \left(\vec{Y}_{i1}^T, \vec{Y}_{i2}^T\right)^T$, where $\vec{Y}_{ij} = \left(Y_{ij1}, Y_{ij2}, \dots, Y_{ijr}\right)^T$ corresponds to the r measurements by system j on subject i. In what follows we let J_a be a column vector of a 1's, $J_{a\times b}$ be an $a \times b$ matrix of 1's, and I_a be the $a \times a$ identity matrix. From [5.1] we have for subject i, $\vec{Y}_i \sim MVN(\vec{\mu}, \Sigma)$ with

$$\vec{\mu} = (\mu, \alpha + \beta\mu)^T \otimes J_r$$

and

$$\Sigma = \sigma_s^2 \begin{bmatrix} 1 & \beta \\ \beta & \beta^2 \end{bmatrix} \otimes J_{r\times r} + \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \otimes I_r$$

where $\otimes$ denotes the Kronecker product.

In order to explicitly write down the log-likelihood function for subject i we must first obtain the inverse and determinant of the covariance matrix. The form of Σ allows us to write down Σ^{-1} and $|\Sigma|$ explicitly as

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \otimes I_r - \frac{\sigma_s^2}{1 + r\sigma_s^2\left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2}\right)} \begin{bmatrix} \frac{1}{\sigma_1^4} & \frac{\beta}{\sigma_1^2\sigma_2^2} \\ \frac{\beta}{\sigma_1^2\sigma_2^2} & \frac{\beta^2}{\sigma_2^4} \end{bmatrix} \otimes J_{r\times r}$$

and

$$|\Sigma| = (\sigma_1^2\sigma_2^2)^r \left\{1 + r\sigma_s^2\left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2}\right)\right\}$$

See Appendix C for more details on calculating the inverse and determinant of the variance covariance matrix Σ.

Denoting the observed data by y_{ijk} $i = 1,2, \dots, n$, $j = 1,2$ and $k = 1,2, \dots, r$ (we distinguish the random variable Y_{ijk} by using a lower-case y_{ijk} to denote the observed data), the log-likelihood contribution from subject i with r replicate measurements by both systems is

$$-rln(2\pi) - \frac{1}{2}ln|\Sigma| - \frac{1}{2}(\vec{y}_i - \vec{\mu})^T\Sigma^{-1}(\vec{y}_i - \vec{\mu})$$

since by model [5.1], $\vec{Y_i} \sim MVN(\vec{\mu}, \sum)$. We can explicitly write this as:

$$
\begin{aligned}
l_i(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s) = -rln(2\pi) - \frac{1}{2}\ln\left[1 + r\sigma_s^2\left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2}\right)\right] \\
-\frac{1}{2}\left\{\frac{1}{\sigma_1^2}\sum_{k=1}^{r}(y_{i1k} - \mu)^2 + \frac{a}{\sigma_1^4}\left[\sum_{k=1}^{r}(y_{i1k} - \mu)\right]^2 - \frac{1}{\sigma_2^2}\sum_{k=1}^{r}(y_{i2k} - \alpha - \beta\mu)^2\right. \qquad [5.6] \\
\left. + \frac{a\beta^2}{\sigma_2^4}\left[\sum_{k=1}^{r}(y_{i2k} - \alpha - \beta\mu)\right]^2 - \frac{2a\beta}{\sigma_1^2\sigma_2^2}\sum_{k=1}^{r}(y_{i1k} - \mu)(y_{i2k} - \alpha - \beta\mu)\right\}
\end{aligned}
$$

where

$$
a = \frac{\sigma_s^2}{1 + r\sigma_s^2\left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2}\right)}
$$

As noted, we assume measurements made on different subjects are independent and so we obtain the full log-likelihood function by summing the log-likelihood contribution [5.6] for each subject. That is

$$
l(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s) = \sum_{i=1}^{n} l_i \qquad [5.7]
$$

In order to calculate approximate confidence intervals for $\theta(s)$ we must obtain asymptotic standard deviations for $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$ which are found using the expected Fisher information matrix. The expected Fisher information matrix is found by taking second partial derivatives of [5.7], which are performed symbolically by Maple [Maplesoft, 2014] to avoid errors, and by calculating the expected values of the necessary sums of squares. We do not give all of the formulas here, but note that we use the following results

$$
E\left[\sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{i1k} - \mu)^2\right] = nr(\sigma_s^2 + \sigma_1^2)
$$

$$E\left[\sum_{i=1}^{n}\left\{\sum_{k=1}^{r}(Y_{i1k}-\mu)\right\}^2\right] = nr(r\sigma_s^2+\sigma_1^2)$$

$$E\left[\sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{i2k}-\alpha-\beta\mu)^2\right] = nr(\beta^2\sigma_s^2+\sigma_2^2)$$

$$E\left[\sum_{i=1}^{n}\left\{\sum_{k=1}^{r}(Y_{i2k}-\alpha-\beta\mu)\right\}^2\right] = nr(r\beta^2\sigma_s^2+\sigma_2^2)$$

$$E\left[\sum_{i=1}^{n}\left\{\sum_{k=1}^{r}(Y_{i1k}-\mu)\right\}\left\{\sum_{k=1}^{r}(Y_{i2k}-\alpha-\beta\mu)\right\}\right] = nr^2\beta\sigma_s^2$$

$$E\left[\sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{i1k}-\mu)\right] = 0$$

$$E\left[\sum_{i=1}^{n}\sum_{k=1}^{r}(Y_{i2k}-\alpha-\beta\mu)\right] = 0$$

We then invert the Fisher Information matrix numerically using Matlab [The MathWorks Inc., 2013]. This gives the asymptotic variances of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$. But because we are interested in $\theta(s)$ and θ, we find their asymptotic variances by applying the Delta Method [Lehmann and Casella, 1998]; we pre- and post-multiply the inverse of the Fisher information matrix by a suitable vector of partial derivatives: D_s for the asymptotic variance of $\theta(s)$, and D for θ.

$$D_s = \frac{\partial\theta(s)}{\partial(\mu,\alpha,\beta,\sigma_1,\sigma_2,\sigma_s)} \qquad D = \frac{\partial\theta}{\partial(\mu,\alpha,\beta,\sigma_1,\sigma_2,\sigma_s)}$$

Where $\theta(s)$ and θ are given by [5.4] and [5.5], respectively.

Again, we use Maple [Maplesoft, 2014] to calculate these partial derivatives and Matlab [The MathWorks Inc., 2013] to find their numerical values for any selected parameter values.

With these asymptotic results we calculate approximate confidence intervals for $\theta(s)$ and θ. In Section 5.3.1 we use simulation to demonstrate that for typical sample sizes ($40 \leq n \leq 120$, and

$2 \leq r \leq 5$) these asymptotic results match simulated results, suggesting that asymptotic confidence intervals are reliable.

5.2.3 Model Checking

The first step of any data analysis should be to look at the data and decide whether the intended analysis is appropriate. In this context we suggest that the assumptions of model [5.1] be checked. The two main assumptions are that (i) the unknown true values of the measurand are normally distributed, and (ii) the measurement variation is constant across the range of true values. We can assess each of these assumptions respectively with a *modified QQ-plot* and a *repeatability plot*.

To evaluate whether $S_i \sim N(\mu, \sigma_S^2)$, for each measurement system we average the replicate measurements on a particular subject and create a QQ-plot of these n averages. By working with the averages we reduce the effect of the measurement variation, allowing us to better examine the between-subject variation and the distribution of S_i. If the normality assumption holds the QQ-plots for both systems should yield a relatively straight line. To aid in their interpretation, we suggest overlaying the quantiles of 50 simulated normal datasets with mean and variance equal to the sample mean and sample variance of the n averages [Oldford, 2014]. Doing so allows us to judge more clearly if the sample data can be reasonably modeled with a normal distribution. If this modified QQ-plot suggests that the normal distribution is a reasonable assumption for S_i then model [5.1] is applicable. However, if it does not, then an alternative to the maximum likelihood approach should be used. In Chapter 6 we discuss a moment-based estimation procedure which does not require this normality assumption.

To decide whether the measurement variation for each system is constant across true values of the measurand, we suggest constructing what we call a repeatability plot for each measurement system, which is also useful for identifying outliers. The plot is an individual values plot of the residuals of the replicate measurements on each subject versus the average of those replicate measurements, ordered by size. If the residuals seem unrelated to the averages this suggests that the measurement variation is homoscedastic. If, however, there appears to be a dependency between the residuals and averages, for example if variability in the residuals increases as the average increases, we conclude the measurement variation is heteroscedastic. The exact structure of heteroscedasticity will depend on the nature of the relationship between the residuals and averages. If the repeatability plots suggest heteroscedasticity of any kind in one or both measurement systems then model [5.1] is no longer appropriate and another approach must be taken. We propose an alternate model, and the corresponding estimation procedure in Chapter 6.

5.2.4 Example

To illustrate how to determine whether two measurement systems are interchangeable using the probability of agreement and the associated plot, we use systolic blood pressure (in mmHg) data from an example published by Bland and Altman [1999]. In this example, 85 subjects are measured three times by each of two observers, labelled “J” and “R”, both using a sphygmomanometer. While this is technically a comparison of two observers using the same measurement system, it is statistically equivalent to the comparison of two measurement systems; we can think of observer J using the sphygmomanometer as measurement system 1 (MS1), and observer R using the sphygmomanometer as measurement system 2 (MS2). The data can be found in Table C.1 of Appendix C.

To justify the model assumptions underlying the probability of agreement approach, we present the modified QQ-plots and the repeatability plots for this data in Figure 5.2. These plots suggest that it is reasonable to assume S_i is normally distributed, and that the repeatability of each measurement system is homoscedastic. Together the diagnostic plots in Figure 5.2 suggest that the model and the analysis approach is valid, and that there are no large outliers.

Having verified that model [5.1] is appropriate, we use this data we obtain maximum likelihood estimates and asymptotic standard deviations for each of the parameters in this model. These results are presented in Table 5.2.

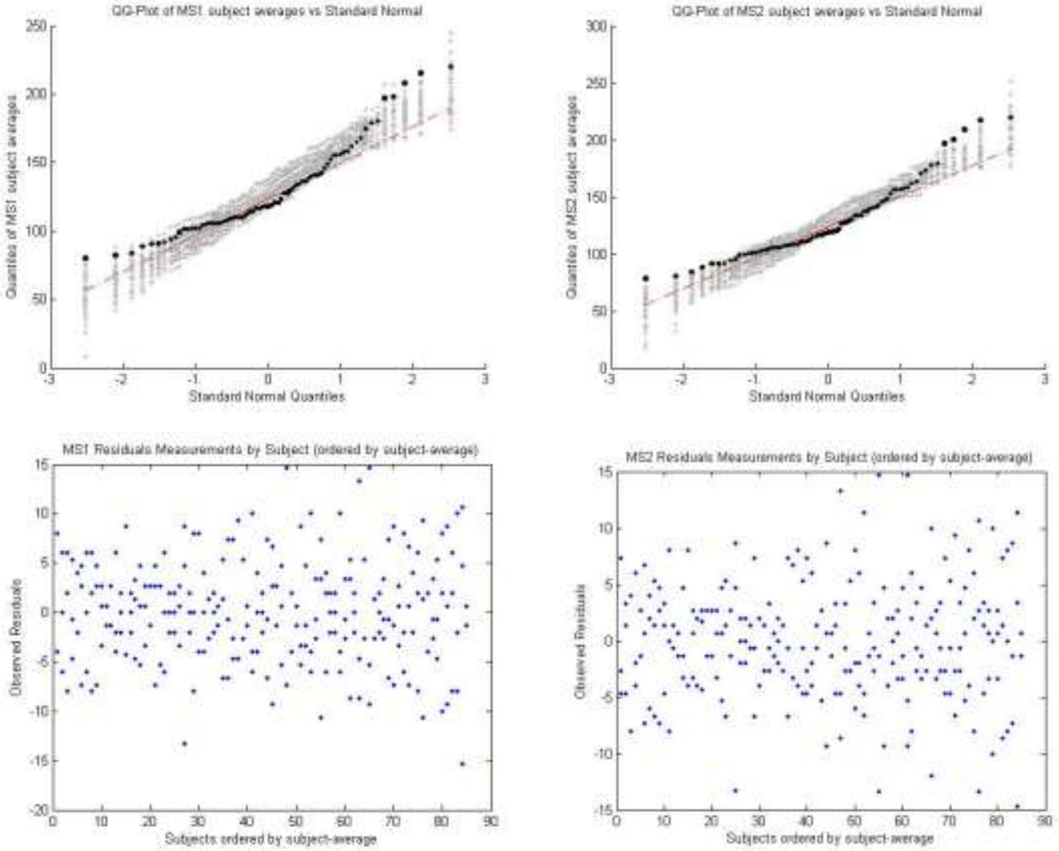

Figure 5.2: Modified QQ-Plot and Repeatability Plot for observers "J" (MS1) and "R" (MS2) from the example data. Left panels correspond to observer "J" and right panels correspond to observer "R"

	Estimate	Asy. Standard Error
μ	127.3612	3.2937
α	-1.3623	2.1432
β	1.0108	0.016377
σ_s	30.1959	2.3421
σ_1	5.5655	0.28559
σ_2	5.4955	0.28347
θ	0.7985	0.09511

Table 5.2: ML estimates and asymptotic standard deviations associated with the systolic blood pressure data

Using these maximum likelihood estimates, we calculate $\hat{\theta}(s)$ for s in the range $(\hat{\mu} - 3\hat{\sigma}_s, \hat{\mu} + 3\hat{\sigma}_s)$ and construct the probability of agreement plot given in Figure 5.3. Note that the calculation of the probabilities in this plot assumes a clinically acceptable difference with $c = 10$. This is somewhat arbitrarily chosen since surprisingly Bland and Altman [1999] do not explicitly provide a clinically acceptable difference for these data.

To justify our assumed CAD we note that when assessing systolic blood pressure measuring devices, O'Brien et al. [1993] provide criteria for grading such measurement systems. A blood pressure measurement device can be graded as A, B, C, or D depending on the proportion of differences that lie within ±5, ±10, and ±15 mmHg. These criteria are based on the difference between measurements by a new system and a sphygmomanometer, and are intended for assessing the adequacy of a new system relative to this standard. Our goal (assessing interchangeability) is different, but we assume these cut-off values are still relevant and use $c = 10$ for illustration. Note that the probably of agreement will increase for larger values of c and decrease for smaller values.

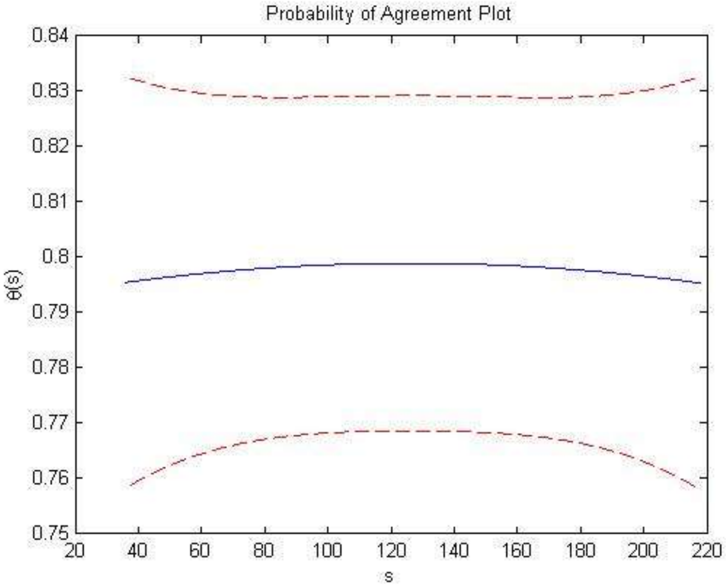

Figure 5.3: Likelihood-Based Probability of Agreement Plot comparing "J" and "R" from the example data with $c = 10$

In Figure 5.3 we see that the probability of agreement is relatively constant (roughly 0.8) across the range of common systolic blood pressures. It is not surprising then to find that the estimate of the unconditional probability of agreement θ is 0.799 with an approximate confidence interval given by (0.61, 0.98). For this example, use of the unconditional probability seems reasonable.

Whether these results indicate good enough agreement for the two measurement systems to be used interchangeably depends on whether the investigators deem $\theta \approx 0.8$ to be sufficiently large. Such a judgement will also depend on the assumptions that the chosen CAD of $c = 10$ is appropriate. If an acceptable value of c cannot be agreed upon, multiple probability of agreement plots could be constructed to investigate the relationship between the level of agreement and c.

Suppose that the choice of $c = 10$ is appropriate, but that $\theta \approx 0.8$ is not sufficiently large (perhaps $\theta \geq 0.95$ is necessary), leading us to conclude that the two measurement systems do not agree well enough to be used interchangeably. To identify the source of this disagreement we should examine the individual parameter estimates and their asymptotic standard errors, which are shown in Table 5.2. Note that we define the asymptotic standard error for a parameter as the corresponding asymptotic standard deviation (as determined by the Fisher information) evaluated at the maximum likelihood estimates of all of the relevant parameters.

In light of this apparent disagreement, it is perhaps surprising to find that the fixed and proportional biases are negligible ($\alpha \approx 0, \beta \approx 1$) and the repeatabilities are very similar ($\sigma_1 \approx \sigma_2$), indicating that the distribution of the measurements made by each system are similar. The issue here is that although $\sigma_1 \approx \sigma_2$, both σ_1 and σ_2 are large relative to σ_s leading to large differences between individual measurements made by each system, causing the probability of agreement to be small. This is an example in which the reference system is highly variable, and so even replicate measurements by that system will not often closely agree.

For this example, and others like it, whether or not we use the new system interchangeably may not be based solely on the probability of agreement; the probability of agreement may be low, but if the reference system is used routinely, perhaps a justification can be made for using a new system that is equally imprecise, if it is, say, cheaper to operate. Such a decision cannot be made by looking at the probability of agreement plot alone; although it accounts for complicated bias and repeatability structures, the probability of agreement masks the individual values of these parameters. Accordingly, we recommend that if the plot suggests disagreement between two measurement systems, the individual parameter estimates be examined for guidance on a final decision.

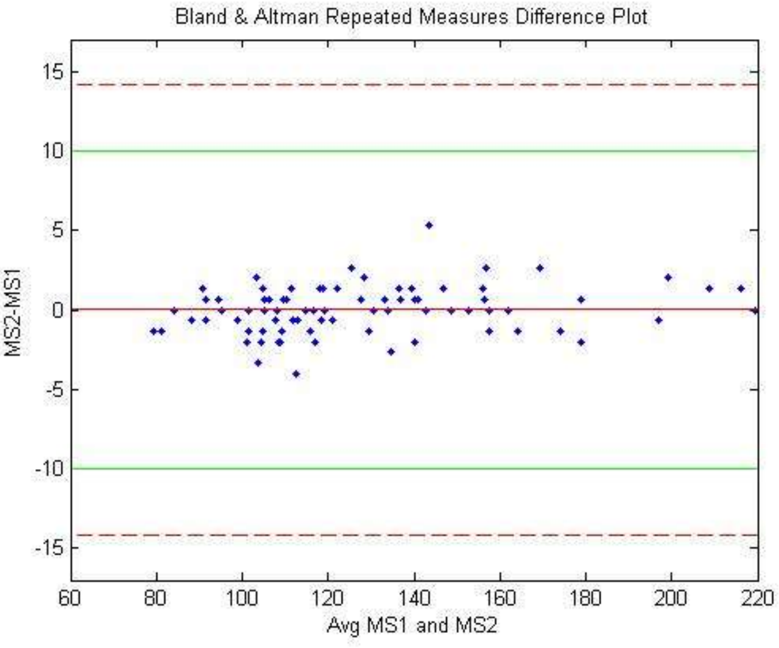

Figure 5.4: Repeated measures difference plot comparing “J” (MS1) and “R” (MS2) from the example data with $c = 10$

We also present the repeated measures difference plot for these data in Figure 5.4. This plot also indicates disagreement, as the limits of agreement lie outside $CAD = (-10,10)$. However, the difference plot does not quantify the disagreement as concisely as does the probability of agreement plot, nor does it offer any indication of the source of this disagreement.

The probability of agreement analysis technique and plot have been automated, and Matlab [The MathWorks Inc., 2013] software is available to practitioners who wish to use it to determine whether two measurement systems are interchangeable.

5.3 Planning an MSC Study

When using the probability of agreement to decide whether two measurement systems are interchangeable, it is important to consider the design of the MSC study. The typical recommendation, although not always followed, is for each measurement system to measure n subjects $r > 1$ times for a total of $N = nr$ measurements each. As we have stated, replicate measurements are necessary to ensure that the parameters in model [5.1], and hence the probability of agreement, can be estimated.

For a fixed total number of measurements N, we investigate the effect of the number of subjects n and the number of replicate measurements r on the precision with which θ can be estimated. Note that we base these comparisons on the unconditional probability of agreement, θ, instead of $\theta(s)$ because it is difficult to determine in general which values of s are relevant. As such, we investigate the effect of n and r on θ, the unconditional probability of agreement. We compare designs using the asymptotic standard deviations of $\tilde{\theta}$, calculated from the expected Fisher information matrix, as described in Section 5.2.2.

To ensure that the asymptotic results allow us to appropriately rank the possible designs, we first conduct a simulation study to compare the asymptotic and simulated standard errors of $\hat{\theta}$ which we describe in Section 5.3.1. This simulation confirmed that the simulated and asymptotic results agree, justifying the use of asymptotic results to investigate possible (n, r) combinations for a given value of N. In Section 5.3.2 we make design recommendations for optimal estimation of the probability of agreement.

5.3.1 Simulation Study

In the simulation study we compared the simulated and asymptotic standard errors of $\hat{\theta}$ for a variety of (n, r) combinations and parameter values. To cover a wide range of sample sizes, replicate measurements and parameter values, we considered:

- $n =$ 40 to 120 in steps of 10
- $r =$ 2 to 5 in steps of 1
- $\mu =$ 1, 10, 100
- $\sigma_s = \mu/10$, $\mu/4$
- $\sigma_1 = \sigma_s/10$, $\sigma_s/4$
- $\sigma_2 = 3\sigma_1/4$, σ_1, $5\sigma_1/4$
- $\alpha =$ 0, 0.05μ
- β = 1, 1.1

For each of the 5,184 combinations of n, r and the parameters, we generated 10,000 samples according to model [5.1] and for each sample determined the maximum likelihood estimate of θ and the asymptotic standard error associated with that estimate. Note that $SE(\hat{\theta})$ is defined as the asymptotic standard deviation of $\tilde{\theta}$, evaluated at the maximum likelihood estimates the other parameters.

We compare the simulated and asymptotic results by dividing the standard deviation of the 10,000 estimates of $\hat{\theta}$ by the average of the 10,000 asymptotic standard errors. Across all combinations of n, r and the parameters, the average of this ratio was 0.9915 and it ranges between 0.89 and 1.11 with the middle 50% lying between 0.97 and 1.02. A graphical analysis of the simulation results suggest that this ratio does not depend materially on n or r, and only very minimally on the other parameters. For more information regarding the results of this simulation study, see Section C.3 of Appendix C.

Overall the results suggest that the asymptotic standard deviation of $\tilde{\theta}$ closely matches the simulated results for all designs. Accordingly we proceed to rank designs based on the asymptotic results.

5.3.2 Recommendations for MSC study design

For a particular combination of the parameter values and $N = 40$, 60, 100, 120, 200, we iterate through $2 \leq r \leq 10$ and take $n = N/r$. In the case that N/r is not an integer, we round this quantity down to the nearest integer to determine n, in which case $nr < N$. We then rank the designs according to the asymptotic standard deviation of $\tilde{\theta}$, and consider the design associated with the smallest asymptotic standard deviation to be the 'best'. In doing this it became clear that the design in which each subject is measured twice, corresponding to $(n, r) = (N/2,2)$, always has the smallest, or nearly the smallest, asymptotic standard deviation.

To investigate this further we compare the asymptotic standard deviations of $\tilde{\theta}$ associated with the 'best' design and the design with two replicate measurements, i.e. $(n, r) = (N/2,2)$. Specifically we divide the standard deviation corresponding to the $(n, r) = (N/2,2)$ design by that of the best design. For $N = 40$, 60, 100, 120, 200, $2 \leq r \leq 10$, and the 729 combinations of $(\mu, \sigma_s, \alpha, \beta, \sigma_1, \sigma_2)$ outlined in C.4 of Appendix C , we found the average of this ratio to be 1.01. Thus the asymptotic standard deviation associated with the $(n, r) = (N/2,2)$ design is on average only 1% larger than the best design. We found the maximum of this ratio to be 1.065, indicating that the $(n, r) = (N/2,2)$ design is at most 6.5% worse than the best design. Such large values occur when α is different from 0, β is different from 1 and when σ_1 and σ_2 are large relative to σ_S. See Section C.4 of Appendix C for more details.

Matlab [The MathWorks Inc., 2013] software is available to practitioners which provides the best design for a particular combination of parameter values and maximum number of measurements N. However, because the best design depends on the values of the unknown parameters, and the $(n, r) = (N/2,2)$ design is good across all parameter values we considered, we recommend its use.

5.4 Discussion and Conclusions

When a new measurement system is available, a potential user must decide whether the measurements made by this new system agree suitably with those made by the existing system, and hence decide whether the two can be used interchangeably. To do so a measurement system comparison (MSC) study must be undertaken. Arguably the most widely used statistical technique for analyzing MSC data, and judging interchangeability, is the limits of agreement technique due to Bland and Altman. We have discussed this approach, and have shown that there are problems associated with it and that misjudgments regarding interchangeability are possible.

In this chapter, we proposed an alternative analysis technique: the probability of agreement [Stevens et al., 2014 (*under revision*)]. The probability of agreement is, for a particular value of the measurand, the probability that the difference between two measurements made by different systems falls within a user specified interval that is deemed to be practically acceptable. This quantity can be translated into an informative plot which depicts the probability of agreement across a range of possible true values for the measurand. The result is a simple and intuitive summary of the agreement between two measurement systems. The benefit of this approach is that while the statistical modelling and estimation may be complicated for a non-statistician, the interpretation is straight forward, intuitive, and easy to understand.

Many of the problems with the limits of agreement approach stem from the over-simplicity of the difference plot, and the misuse of the technique in general. For example, the difference plot can be uninformative and misleading, particularly when it is not supplemented by information from replicate measurements, and in the absence of a clinically acceptable difference. Another benefit of the probability of agreement analysis is that it cannot be performed without replicate measurements on each subject by each system, and without the definition of a clinically acceptable difference. As such the analysis implicitly ensures that it will be performed correctly, since otherwise it cannot be performed at all.

Throughout this chapter we have stressed the importance of replicate measurements in an MSC study; in order to estimate $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$ and the probability of agreement, two or more measurements must be made by each system on each subject. Although the probability of agreement is the metric of primary interest, the separate estimation of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$ is also important. Should the probability of agreement plot suggest disagreement between the two measurement systems, estimates of the bias and precision parameters may help to identify the source of this disagreement. As well, these estimates allow for the direct comparison of repeatabilities, which helps to avoid the possibility of rejecting interchangeability for a new system when $\sigma_2 \ll \sigma_1$.

The estimation of each parameter in [5.1] has other benefits as well: for example, we can calculate other metrics such as $P(|Y_j - s| \leq c|S = s)$, which quantifies how closely the measurements by system j agree with the true value of the measurand. This quantity could be used when assessing system j on its own, or when system j is being compared to a ‘gold standard’ for which the true value of the measurand is known. We may also use the parameter estimates, in particular $\hat{\alpha}$ and $\hat{\beta}$, to adjust the measurements by the new system to eliminate bias. We could then calculate an adjusted probability of agreement that accounts for the estimation error of $\hat{\alpha}$ and $\hat{\beta}$, and examine the agreement between the two measurement systems after calibrating the new one.

Given that replicate measurements are necessary for this analysis method, we have evaluated MSC study designs based on their ability to precisely estimate θ, the probability of agreement. We suggest that if $N = nr$ measurements can be made by each system in the study, the study should consist of $n = N/2$ subjects that are measured $r = 2$ times by each system. This design provides optimal or near-optimal estimation of θ in all situations.

Here we have considered the case where each measurement system measures each subject r times. A straight forward extension of this work would be to adapt the model and consider the case when the two systems make a different number of replicate measurements per subject, i.e. $r_1 \neq r_2$, or a different number of measurements on each subject.

We assume that the primary goal of the MSC study is to decide whether two measurement systems agree well enough to be used interchangeably, and we answer this question with the probability of agreement. The likelihood framework however, also allows us to decide whether the two measurement systems are identical. Specifically, we can use a likelihood ratio test to simultaneously test $H_0: \alpha = 0$ and $\beta = 1$ and $\sigma_1 = \sigma_2$. However, this test suffers from the phenomenon whereby the null hypothesis is always rejected if the number of measurements in the study is suitably large, even if the differences between the two measurement systems are very small. Alternatively we could frame the hypothesis in terms of an equivalence test to avoid this issue [Wellek, 2010].

We note that although the emphasis of this work is on the comparison of two different measurement systems, the methodology also applies to the comparison of two observers using the same measurement system, as we saw in Section 5.2.4. Another possible extension is the simultaneous comparison of three or more measurement systems (or observers).

Model [5.1] does not include observer effects; as in Bland and Altman's limits of agreement analysis, we implicitly assume that the measurement systems being compared are each operated by a single observer, or if operated by multiple observers, we assume that their effects are the same. Another possible extension to this work is to incorporate observer effects into the probability of agreement analysis.

Another notable advantage of the proposed method is that we can alter the model assumptions and account for generalizations like observer effects and unbalanced replicate measurements, and still calculate the probability of agreement and interpret it in the same way. The modeling and estimation procedure may change, but the simple interpretation remains.

To help decide which model to use, and which assumptions are valid we suggest that model diagnostics be performed. We propose the use of a modified QQ-plot to decide whether the true values of the measurand follow a normal distribution and the use of a repeatability plot to decide whether or not the measurement system repeatabilities are homoscedastic, and to check for outliers.

In this chapter we have assumed that the unknown true values follow a normal distribution, but in some cases this assumption may be unreasonable. When we relax this assumption we must alter the method by which we estimate the probability of agreement $\theta(s)$, but its interpretation does not change. In Chapter 6 we propose moment-based estimates of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$, and hence $\theta(s)$ that do not require any distributional assumptions for the true values.

We have also assumed that the repeatability of each measurement system is homoscedastic. That is, the measurement variability does not depend on the unknown true value of the measurand. Although this assumption is often valid, situations may arise in which it is not. When this is the case we suggest using a model different from [5.1] that accounts for a dependence between the measurement variation and the unknown true value of the measurand. In Chapter 6 we propose such a model, and illustrate the probability of agreement analysis in light of this change. Application of the probability of agreement in these two settings serves to demonstrate the versatility of the suggested analysis method.

Chapter 6

Generalizing the Probability of Agreement Analysis

In Chapter 5 we introduced the probability of agreement and an associated plot as a tool for summarizing the agreement between two measurement systems. Recall the probability of agreement is defined to be the probability that the difference between single measurements on the same subject by the two systems falls within the range that is deemed to be acceptable, conditional on the value of the measurand. Statistically the probability of agreement is:

$$\theta(s) = P(|Y_2 - Y_1| \leq c | S = s) \quad [6.1]$$

where Y_j represents the single measurements made by measurement system $j = 1,2$ on a particular subject, and absolute differences less than or equal to c are considered clinically (or practically) acceptable.

Thus the probability of agreement is the probability that the difference between single measurements on the same subject by the two systems falls within the range that is deemed to be acceptable, conditional on the true values of the measurand. With estimates of $\theta(s)$, we construct the probability of agreement plot which graphically displays the estimated probability of agreement across a range of plausible values for s. Whether two systems are deemed to be interchangeable depends on whether $\theta(s)$ (for s in the range of interest) is suitably large. The determination of how large $\theta(s)$ must be to indicate acceptable agreement and hence interchangeability is context-specific and, like c, must be decided by the user.

This technique is proposed as an alternative to the limits of agreement method (due to Bland and Altman [1983; 1986]), for determining whether two measurement systems can be used interchangeably. The technique quantifies the agreement between two measurement systems in an intuitive and informative manner. With the estimation and plot-construction automated, the probability of agreement analysis can be used and understood by non-statisticians.

In addition to the ease of interpretation, in Chapter 5 we noted that another important advantage of the analysis is its versatility. If we generalize various model assumptions, we can still estimate $\theta(s)$, construct the probability of agreement plot, and interpret the results in exactly the same manner as described in the previous chapter. Note that the when model assumptions are changed, the method of estimating $\theta(s)$ may also change, but the interpretation remains the same.

In Chapter 5 we proposed the following model to describe the data that arise as a result of a measurement system comparison (MSC) study:

$$\begin{aligned} Y_{i1k} &= S_i + M_{i1k} \\ Y_{i2k} &= \alpha + \beta S_i + M_{i2k} \end{aligned} \qquad [6.2]$$

Thus Y_{ijk} is a random variable which represents the value observed on system j's k^{th} measurement of subject i, where $i = 1,2, \dots, n$ indexes the subjects, $j = 1$ indexes the reference measurement system, $j = 2$ indexes the new measurement system and $k = 1,2, \dots, r$ indexes the replicate measurements. Recall $S_i \sim N(\mu, \sigma_s^2)$ is a random variable that represents the unknown true value of the measurand for subject i, and $M_{ijk} \sim N\left(0, \sigma_j^2\right)$ is a random variable which represents the measurement error of system j. Here σ_j quantifies the measurement variation, or repeatability, of system j, and in [6.2] we assume it is a constant. Model [6.2] also assumes that the reference (existing) measurement system is unbiased and the parameters $-\infty < \alpha < \infty$ and $\beta > 0$ quantify the bias of the second (new) measurement system relative to the existing one. We refer to α and β as the fixed and proportional bias, respectively [Ludbrook, 2010].

In this chapter we discuss relaxing some of these model assumptions, and illustrate the application of the probability of agreement analysis in these more general settings. In Section 6.1 we discuss performing the analysis when we no longer assume that S_i, the true values of the measurand, are normally distributed; in this case we propose a moment-based estimate of $\theta(s)$ which does not require this normality assumption [Stevens et al., 2014 (*under revision*)]. In

Section 6.2 we consider the case when at least one of the measurement systems is heteroscedastic. That is, when the repeatability (σ_j), depends on the true value of the measurand. We conclude with a summary and discussion in Section 6.3.

6.1 The Probability of Agreement Analysis with a Non-Normal Measurand

In some cases, it may be unreasonable to model the unknown true values of the measurand with a normal distribution. This may be known prior to analyzing the MSC study data, but in case it is not, in Section 5.2.3 we proposed the use of a modified QQ-plot to assess this assumption. When modeling the true values of the measurand with a normal distribution is unjustified, we should relax this assumption and alter the method by which we estimate the probability of agreement.

In this section we propose a moment-based approach to the probability of agreement analysis [Stevens et al., 2014 (*under revision*)], and we illustrate the procedure on the systolic blood pressure example introduced in Section 5.2.4.

6.1.1 Moment-Based Estimation Procedure

We begin this subsection by deriving moment-based estimates of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$ that do not require any distributional assumptions for S_i. We then use these estimates to obtain a moment-based estimate of $\theta(s)$.

To begin, we note that given $S_i = s$, model [6.2] becomes

$$\begin{aligned} Y_{i1k} &= s + M_{i1k} \\ Y_{i2k} &= \alpha + \beta s + M_{i2k} \end{aligned}$$

where randomness enters only through the measurement error terms $M_{ijk}, j = 1,2$. As such the difference $Y_{i2k} - Y_{i1k}$ conditional on $S_i = s$ is given by

$$\alpha + (\beta - 1)s + (M_{i2k} - M_{i1k})$$

and so

$$Y_{i2k} - Y_{i1k} | S_i = s \sim N(\alpha + (\beta - 1)s, \sigma_1^2 + \sigma_2^2)$$

even though S_i is non-normal.

Thus, because $\theta(s)$ is defined as a probability of the difference $Y_{i2k} - Y_{i1k}$ conditional on the value of $S_i = s$, we can still write the probability of agreement as

$$\theta(s) = \Phi\left(\frac{c - \alpha - (\beta - 1)s}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) - \Phi\left(\frac{-c - \alpha - (\beta - 1)s}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \qquad [6.3]$$

where $\Phi(x)$ is the standard normal cumulative distribution function evaluated at x. Thus we substitute $(\check{\mu}, \check{\alpha}, \check{\beta}, \check{\sigma}_1, \check{\sigma}_2, \check{\sigma}_s)$, the moment-based estimates that we derive later in this subsection, into [6.3] to obtain $\check{\theta}(s)$. Note that we overscore the Greek letters with an inverted-circumflex to distinguish the moment-based estimates from the maximum likelihood estimates.

For the derivation of the moment-based estimate $\check{\theta}(s)$ we adopt model [6.2] and all of the assumptions associated with it, except that we do not specify a distribution for S_i. Instead we simply assume that the mean and variance of the unknown true values are respectively given by $E(S_i) = \mu$ and $Var(S_i) = \sigma_s^2$, for $i = 1,2,\dots,n$. As in Section 5.2.2 we order the random data vector for subject i by measurement system, and denote it by $\vec{Y}_i = \left(\vec{Y}_{i1}^T, \vec{Y}_{i2}^T\right)^T$, where $\vec{Y}_{ij} = \left(Y_{ij1}, Y_{ij2}, \dots, Y_{ijr}\right)^T$ corresponds to the r measurements by system j on subject i. The $2r \times 1$ random vector $\vec{Y}_i$ has mean and variance given by

$$\vec{\mu} = (\mu, \alpha + \beta\mu)^T \otimes J_r$$

and

$$\Sigma = \sigma_s^2 \begin{bmatrix} 1 & \beta \\ \beta & \beta^2 \end{bmatrix} \otimes J_{r\times r} + \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \otimes I_r$$

where J_a is a column vector of a 1's, $J_{a\times b}$ is an $a \times b$ matrix of 1's, I_a is the $a \times a$ identity matrix, and $\otimes$ denotes the Kronecker product.

We can estimate $\vec{\mu}$ and $\sum$ non-parametrically with the sample mean (denoted $\check{\vec{\mu}}$), and sample variance-covariance matrix (denoted $\check{\sum}$) associated with the $\vec{Y}_i$'s, $i = 1,2,\dots,n$. These two sample estimators are respectively given by

$$\check{\vec{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \vec{Y}_i = (\bar{Y}_{\cdot 11}, \dots, \bar{Y}_{\cdot 1r}, \bar{Y}_{\cdot 21}, \dots, \bar{Y}_{\cdot 2r})^T$$

and

$$\check{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} \left(\vec{Y}_i - \check{\vec{\mu}}\right)\left(\vec{Y}_i - \check{\vec{\mu}}\right)^T$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\begin{bmatrix} (Y_{i11}-\bar{Y}_{\cdot 11})^2 & \cdots & (Y_{i11}-\bar{Y}_{\cdot 11})(Y_{i1r}-\bar{Y}_{\cdot 1r}) & (Y_{i11}-\bar{Y}_{\cdot 11})(Y_{i21}-\bar{Y}_{\cdot 21}) & \cdots & (Y_{i11}-\bar{Y}_{\cdot 11})(Y_{i2r}-\bar{Y}_{\cdot 2r}) \\ & \ddots & \vdots & \vdots & \ddots & \vdots \\ & & (Y_{i1r}-\bar{Y}_{\cdot 1r})^2 & (Y_{i11}-\bar{Y}_{\cdot 11})(Y_{i21}-\bar{Y}_{\cdot 21}) & \cdots & (Y_{i11}-\bar{Y}_{\cdot 11})(Y_{i2r}-\bar{Y}_{\cdot 2r}) \\ & & & (Y_{i21}-\bar{Y}_{\cdot 21})^2 & \cdots & (Y_{i21}-\bar{Y}_{\cdot 21})(Y_{i2r}-\bar{Y}_{\cdot 2r}) \\ & symmetric & & & \ddots & \vdots \\ & & & & & (Y_{i2r}-\bar{Y}_{\cdot 2r})^2 \end{bmatrix}$$

where $\bar{Y}_{\cdot jk} = \sum_{i=1}^{n} Y_{ijk}/n$.

Note that many of the components of $\check{\vec{\mu}}$ and $\check{\Sigma}$ are estimates of the same quantities. For example $\bar{Y}_{\cdot 1k}$, $k = 1,2,\dots,r$ all estimate $E(Y_{i1k}) = \mu$, and $\sum_{i=1}^{n}(Y_{i2k} - \bar{Y}_{\cdot 2k})^2/(n-1)$, $k = 1,2,\dots,r$ all estimate $Var(Y_{i2k}) = \beta^2\sigma_s^2 + \sigma_2^2$. Thus, averaging the components that estimate the same quantities (i.e. averaging over replicate measurements) yields the following estimators:

$$\tilde{\psi}_1 = \frac{1}{nr}\sum_{i=1}^{n}\sum_{k=1}^{r} Y_{i1k}$$

$$\tilde{\psi}_2 = \frac{1}{nr}\sum_{i=1}^{n}\sum_{k=1}^{r} Y_{i2k}$$

$$\tilde{\psi}_3 = \frac{1}{(n-1)r}\sum_{i=1}^{n}\sum_{k=1}^{r} (Y_{i1k} - \bar{Y}_{\cdot 1k})^2$$

$$\tilde{\psi}_4 = \frac{1}{(n-1)r}\sum_{i=1}^{n}\sum_{k=1}^{r} (Y_{i2k} - \bar{Y}_{\cdot 2k})^2$$

$$\tilde{\psi}_5 = \frac{1}{(n-1)r^2}\sum_{i=1}^{n}\sum_{k=1}^{r}\sum_{l=1}^{r} (Y_{i1k} - \bar{Y}_{\cdot 1k})(Y_{i2l} - \bar{Y}_{\cdot 2l})$$

$$\tilde{\psi}_6 = \frac{1}{(n-1)r(r-1)}\sum_{i=1}^{n}\sum_{k=1}^{r}\sum_{\substack{l=1\\ l\neq k}}^{r}(Y_{i1k}-\bar{Y}_{\cdot 1k})(Y_{i1l}-\bar{Y}_{\cdot 1l})$$

$$\tilde{\psi}_7 = \frac{1}{(n-1)r(r-1)}\sum_{i=1}^{n}\sum_{k=1}^{r}\sum_{\substack{l=1\\ l\neq k}}^{r}(Y_{i2k}-\bar{Y}_{\cdot 2k})(Y_{i2l}-\bar{Y}_{\cdot 2l})$$

Taking their expected values yields:

$$E(\tilde{\psi}_1) = \mu$$

$$E(\tilde{\psi}_2) = \alpha + \beta\mu$$

$$E(\tilde{\psi}_3) = \sigma_s^2 + \sigma_1^2$$

$$E(\tilde{\psi}_4) = \beta^2\sigma_s^2 + \sigma_2^2$$

$$E(\tilde{\psi}_5) = \beta\sigma_s^2$$

$$E(\tilde{\psi}_6) = \sigma_s^2$$

$$E(\tilde{\psi}_7) = \beta^2\sigma_s^2$$

Let us denote the observed version of $\tilde{\psi}_l$ by $\breve{\psi}_l$ $(l = 1,2,\ldots,7)$ which is found by substituting the observed data, y_{ijk}, for the random variables, Y_{ijk}. Solving the previous equations for the parameters of interest, and replacing $E(\tilde{\psi})$ with $\breve{\psi}_l$ yields what we refer to as the moment-based estimates of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$:

$$\breve{\mu} = \breve{\psi}_1$$

$$\breve{\alpha} = \breve{\psi}_2 - \breve{\psi}_1\breve{\psi}_5/\breve{\psi}_6$$

$$\breve{\beta} = \breve{\psi}_5/\breve{\psi}_6$$

$$\breve{\sigma}_1^2 = \breve{\psi}_3 - \breve{\psi}_6$$

$$\breve{\sigma}_2^2 = \breve{\psi}_4 - \breve{\psi}_7$$

$$\breve{\sigma}_s^2 = \breve{\psi}_6$$

As mentioned, we then substitute the moment-based estimates $(\breve{\mu}, \breve{\alpha}, \breve{\beta}, \breve{\sigma}_1, \breve{\sigma}_2, \breve{\sigma}_s)$ into [6.3] to obtain $\breve{\theta}(s)$. For values of s in the range $(\breve{\mu} - 3\breve{\sigma}_s, \breve{\mu} + 3\breve{\sigma}_s)$ we calculate $\breve{\theta}(s)$ and construct the probability of agreement plot. To complete the plot, we must also calculate pointwise confidence intervals for $\theta(s)$, which requires estimation of the standard error for $\breve{\theta}(s)$.

Because we do not wish to assume a distribution for S_i, we estimate the standard error for $\check{\theta}(s)$ by bootstrapping [Efron, 1979]. We construct $B = 10\ 000$ bootstrap samples in the usual way: we randomly sample n subjects with replacement from the original sample of n subjects. If a subject is included in a bootstrap sample, all of their replicate measurements are also included. For each sample we then derive the moment-based estimates of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$, and the corresponding bootstrap estimates are calculated as the average of these 10 000 estimates. To then calculate $SE\left(\check{\theta}(s)\right)$ we must estimate the covariance matrix for $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$. The estimated covariance matrix is calculated as the sample covariance of the 10 000 bootstrap estimates of $(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)$. We obtain $SE\left(\check{\theta}(s)\right)$ in accordance with the delta method [Lehmann and Casella, 1998] by pre- and post-multiplying this matrix by a change-of-variables vector of suitable partial derivatives:

$$D_s = \frac{\partial \theta(s)}{\partial(\mu, \alpha, \beta, \sigma_1, \sigma_2, \sigma_s)}$$

Thus with the moment-based estimate $\check{\theta}(s)$ and its corresponding standard error, which do not assume a distribution for S, we construct and interpret the probability of agreement plot as before.

Recall in Chapter 5 we discussed the unconditional probability of agreement θ given in [5.4]. To calculate this unconditional probability of agreement, we must assume S_i is normally distributed, and so we do not use it in this scenario.

6.1.2 Example

In this subsection we illustrate the moment-based analysis on the systolic blood pressure data used in Section 5.2.4. Recall that this data come from an MSC study in which the systolic blood pressure of 85 subjects is measured three times by each of two observers labelled “J” and “R”. The data can be found in Table C.1 of Appendix C.

In Figure 5.2 we displayed the modified QQ-plot for these data, which we interpreted as evidence that did not contradict the $S_i \sim N(\mu, \sigma_s^2)$ assumption, since the plots depicted relatively straight lines. However, judging straightness is somewhat subjective, and some might argue that the plot suggests that the normality assumption is invalid. Those who are skeptical of the normality assumption may perform the moment-based probability of agreement analysis instead.

	Estimate	Standard Error
μ	127.3608	3.2786
α	-3.0337	0.75564
β	1.0239	0.00604
σ_s	30.0772	2.8509
σ_1	6.0731	0.35814
σ_2	5.9637	0.38181

Table 6.1: Moment-based estimates and bootstrapped standard errors associated with the blood pressure data

The moment-based estimates of the parameters in model [6.2] and their associated standard errors (determined using the bootstrap) are presented in Table 6.1 and the probability of agreement plot is given in Figure 6.1. Note that as in Figure 5.3, these probabilities are calculated assuming a clinically acceptable difference with $c = 10$.

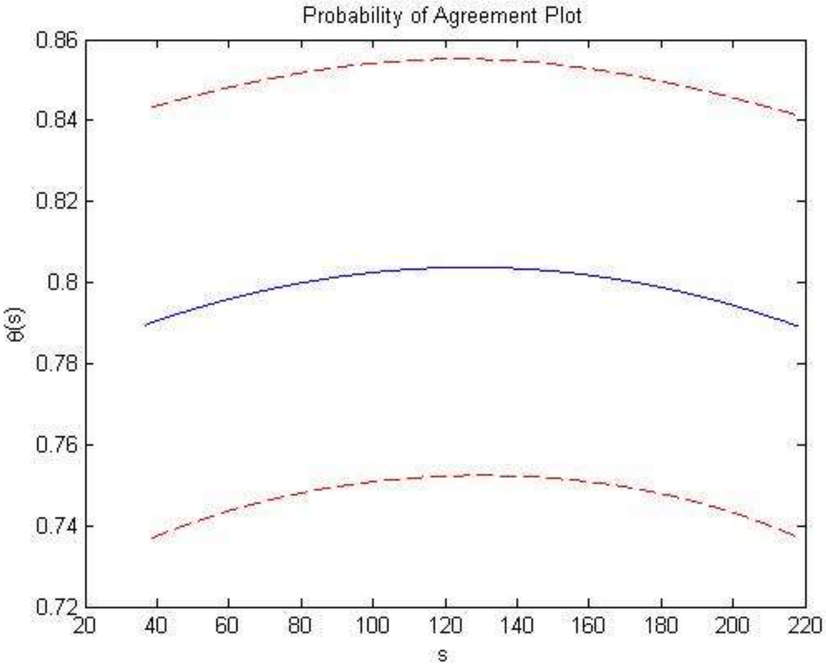

Figure 6.1: Moment-Based Probability of Agreement Plot comparing “J” and “R” from the blood pressure data with $c = 10$

In examining these results, we see small differences between the moment-based and maximum-likelihood-based estimates (see Table 5.2) of some of the model parameters, α in particular, but the estimated probability of agreement is similar (roughly 0.8) across the plausible values of S. When we compare the standard errors associated with the two methods of estimation (see Table

5.2) we see that the standard errors for μ, σ_s, σ_1, and σ_2 are similar (but slightly larger with the moment-based method), and the standard errors for α and β associated with the moment-based method are much smaller. These differences result in slightly larger standard errors for $\theta(s)$ under the moment-based method, which are evident when we compare Figures 5.2 and 6.1. Despite this small difference, the results indicate that for these data, the conclusions drawn about the agreement between "J" and "R" are not affected by the distributional assumption of the true values.

As with the maximum-likelihood approach, Matlab [The MathWorks Inc., 2013] software for the moment-based probability of agreement analysis is available to practitioners.

6.2 The Probability of Agreement Analysis with Heteroscedastic Measurement Error

In some situations, the variability of a measurement system may depend in some way on the true value of the measurand. This is particularly common in the field of clinical chemistry where it is typical for measurement variation to increase as the true value increases [Pollack et al., 1992]. In Chapter 4 we discussed a few analysis methods that have been developed to assess agreement in this situation. For example, we discussed the weighted Deming and weighted least squares regression techniques that, unlike ordinary least squares, allow for non-constant variability. We also discussed Bland and Altman's [1999] recommendation to log-transform the data and carry out the limits of agreement analysis on the log scale. And we described the V-shaped limits of agreement Bland and Altman [1999] recommend if log-transformation does not eliminate non-random patterns on the difference plot.

When it comes to assessing the interchangeability of two measurement systems when one or both is heteroscedastic, these techniques are a step in the right direction, but they do have deficiencies. First, the weighted least squares technique allows for non-constant variability in the new measurement system, but it still assumes that the reference system is error free (i.e., $\sigma_1 = 0$). We noted in Chapter 4 that this assumption is rarely valid because gold standard measurement systems are uncommon. Weighted Deming regression is more general in the sense that it allows for heteroscedastic measurement variation, but it assumes that the ratio of repeatabilities is constant. That is, it assumes that both systems are heteroscedastic, and the structure of heteroscedasticity is the same for both.

Within the limits of agreement framework, the log-transformation is problematic because it also assumes that both systems are heteroscedastic, which may not be the case. As well, the transformation implicitly assumes that the measurement variability is log-linear, which also may not be the case. Furthermore, when working on the transformed scale, one must be careful when interpreting results. Bland and Altman suggest back-transforming the limits of agreement so that they are in terms of the original units. However, in doing so the results are now based on percent-differences and not absolute differences, and this needs to be accounted for when comparing the limits of agreement to the clinically acceptable difference.

Working with the raw data on the original scale is clearly desirable, and Bland and Altman meet this desire with the V-shaped limits of agreement that more accurately reflect the distribution of differences when heteroscedasticity is present. However, this approach is not ideal because the problems that plague the difference plot (see Section 5.1) still exist in this situation: this approach assumes each system measures each subject once ($r = 1$) and without the extra information gained by replicate measurements the difference plot cannot disentangle confounding biases, and it does not provide insight into the magnitude of each system's repeatability. As well, the difference plot does not indicate if heteroscedasticity is present in both, or just one system.

Therefore, as with the homoscedastic case, we propose that the relationship between two measurements systems be modelled directly. This allows us to specify the structure of heteroscedasticity that seems plausible, it allows us to work with the raw data on the original scale, and when accompanied by replicate measurements ($r > 1$) it allows us to separately estimate the parameters of interest, allowing us to construct the probability of agreement plot.

In Section 6.2.1 we propose a model which can be used to describe the relationship between two possibly heteroscedastic measurement systems, and we extend the probability of agreement analysis to this scenario. In Section 6.2.2 we provide details of the maximum-likelihood estimation procedure, and in Section 6.2.3 we demonstrate the analysis with an example from the literature. We close this section with a brief discussion of further extensions and other considerations in Section 6.2.4.

6.2.1 The Model

In this subsection we propose an extension to model [6.2] in which the measurement variation is a function of the true value of the measurand. Because the variance of each system depends on the unknown true value, we use a latent variable model which describes the measurements by each system on a given subject, conditional on the true value of the measurand for that subject. Given $S_i = s$ we have

$$\begin{aligned} Y_{i1k} &= s + M_{i1k} \\ Y_{i2k} &= \alpha + \beta s + M_{i2k} \end{aligned} \qquad [6.4]$$

As in model [6.2] Y_{ijk} represents the k^{th} measurement by system j on subject i, where $i = 1,2,\ldots,n$; $j = 1,2$; $k = 1,2,\ldots,r$. We represent the latent variable, the true value of the measurand for subject i, by s. The estimation procedure described in Section 6.2.2 assumes $S_i \sim N(\mu, \sigma_s^2)$, but we discuss other choices for this distribution in Section 6.2.4. As in [6.2] we assume that the existing reference system is unbiased, and that inferences regarding bias are made from the new system relative to it. The fixed and proportional bias of the new system relative to the existing are respectively quantified by $-\infty < \alpha < \infty$ and $\beta > 0$.

The random variable M_{ijk} represents the measurement error that arises when measurement system j makes multiple measurements of subject i, and we assume $M_{ijk} \sim N\left(0, \sigma_j^2(s)\right)$. Thus we continue to assume that $E\left(M_{ijk}\right) = 0$, but now we assume that the measurement variability of system j, $Var\left(M_{ijk}\right) = \sigma_j^2(s)$, is a function of the true value s. This model allows for some flexibility with regards to the form of the variance function $\sigma_j^2(s)$; for illustration we assume that the standard deviation of the measurement error is a linear function of the true value s [Rocke and Lorenzato, 1995]. Specifically we assume:

$$\sigma_j^2(s) = \left(\omega_j + \tau_j s\right)^2 \qquad [6.5]$$

where $\tau_j \geq 0$ is the proportionality constant, and $\omega_j > 0$ is a constant term which represents the repeatability of measurement system j if it is not heteroscedastic (i.e. if $\tau_j = 0$). Note that restricting τ_j to be non-negative corresponds to an increase in variability with an increase in true values, which is common in the field of clinical chemistry [Pollack et al., 1992]. We also assume that the true values of the measurand are positive to ensure that $\sigma_j(s) = \omega_j + \tau_j s > 0$, and $\sigma_j(s) = \omega_j$ only when $\tau_j = 0$. If the true values of the measurand lie close to zero, but remain positive, then the normality assumption for S_i may be invalid, and assuming a skewed distribution for the true values may be reasonable. We discuss this further in Section 6.2.4.

Note that when neither measurement system is heteroscedastic, i.e. $\tau_1 = \tau_2 = 0$, model [6.4] reduces to the homoscedastic model [6.2]. In Section 5.2.3 we proposed the use of a repeatability plot to determine whether the measurement systems are heteroscedastic or homoscedastic. This graphical method is effective, but informal; we can formally assess whether heteroscedasticity (of the form specified by [6.5]) is present in both systems, by testing the hypothesis $H_0\colon \tau_1 = \tau_2 = 0$. Within the maximum likelihood framework described in Section 6.2.2 we can use a likelihood ratio test to test this hypothesis. We can also use a similar test to decide if just one of the two systems is heteroscedastic.

If one wishes to allow for a different type of heteroscedasticity, we can change the structure of the variance function $\sigma_j^2(s)$ to allow for this. For example, to model heteroscedasticity such that the measurement variability increases as the true values decrease we could restrict $\tau_j \leq 0$, but we would need to ensure $\sigma_j(s) \geq 0$, for all s. Or if the heteroscedasticity is more complicated than a linear function of s, we could specify a polynomial or exponential function instead.

Having introduced this extended model we can now define the probability of agreement, $\theta(s)$, in this scenario. The general definition in [6.1] still holds, but based on the assumptions of model [6.4] we have $Y_{i2k} - Y_{i1k} | S_i = s \sim N(\alpha + (\beta - 1)s, (\omega_1 + \tau_1 s)^2 + (\omega_2 + \tau_2 s)^2)$ for all i and k. As such $\theta(s)$ can be written as

$$\theta(s) = \Phi\left(\frac{c - \alpha - (\beta - 1)s}{\sqrt{(\omega_1 + \tau_1 s)^2 + (\omega_2 + \tau_2 s)^2}}\right) - \Phi\left(\frac{-c - \alpha - (\beta - 1)s}{\sqrt{(\omega_1 + \tau_1 s)^2 + (\omega_2 + \tau_2 s)^2}}\right) \quad [6.6]$$

where $\Phi(x)$ is the standard normal cumulative distribution function evaluated at x.

With estimates of $\theta(s)$ across a range of values for s, and associated standard errors, we construct the probability of agreement plot as we would in the homoscedastic case. However, as we will see in Section 6.2.2, the estimation procedure for model [6.4] is much more complicated. That being said, one of the key advantages of the probability of agreement analysis is that the results are interpreted in exactly the same way, even when the underlying model is more complicated.

6.2.2 Maximum Likelihood Estimation

In this subsection we describe the maximum likelihood estimation procedure associated with model [6.4]. In order to obtain estimates of $(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$ and hence $\theta(s)$, we must obtain the marginal likelihood of the data Y_{ijk}, $i = 1,2, \dots, n$; $j = 1,2$; $k = 1,2, \dots, r$. However, model [6.4] is written as a latent variable model, which describes the measurements by each system on a given subject, conditional on the unknown true value of the measurand for that subject. As such, we obtain the marginal likelihood by integrating the joint density function of S_i and the data, over the support of S_i, thus eliminating the latent variable.

As usual, for a particular subject i, we order the random vector corresponding to its measurements by system, and write $\vec{Y}_i = \left(\vec{Y}_{i1}^T, \vec{Y}_{i2}^T\right)^T$ where $\vec{Y}_{ij} = \left(Y_{ij1}, Y_{ij2}, \dots, Y_{ijr}\right)^T$ corresponds to the r measurements by system j on subject i.

From [6.4] we know $Y_{i1k}|S_i = s \sim N(s, (\omega_1 + \tau_1 s)^2)$ and $Y_{i2k}|S_i = s \sim N(\alpha + \beta s, (\omega_2 + \tau_2 s)^2)$ and so

$$\vec{Y}_i|S_i = s \sim MVN\left(\vec{\mu}(s), \Sigma(s)\right)$$

where

$$\vec{\mu}(s) = (s, \alpha + \beta s)^T \otimes J_r$$

and

$$\Sigma(s) = \begin{bmatrix} (\omega_1 + \tau_1 s)^2 & 0 \\ 0 & (\omega_2 + \tau_2 s)^2 \end{bmatrix} \otimes I_r$$

where J_r is an $r \times 1$ vector of ones, I_r is the $r \times r$ identity matrix, and $\otimes$ denotes the Kronecker product.

Denoting the observed data by y_{ijk} $i = 1,2,\dots,n$, $j = 1,2$ and $k = 1,2,\dots,r$, and using a lower-case y to denote the observed data vectors, the conditional probability density function of $\vec{Y}_i|S_i$ is given by

$$g(\vec{y}_i|S_i = s) = (2\pi)^{-r}|\sum(s)|^{-1/2}\exp\left\{-\frac{1}{2}\left(\vec{y}_i - \vec{\mu}(s)\right)^T\sum(s)^{-1}\left(\vec{y}_i - \vec{\mu}(s)\right)\right\}$$

To explicitly write $g(\vec{y}_i|S_i = s)$ we need the inverse and determinant of the covariance matrix $\sum(s)$. These are calculated as

$$\sum(s)^{-1} = \begin{bmatrix} 1/(\omega_1 + \tau_1 s)^2 & 0 \\ 0 & 1/(\omega_2 + \tau_2 s)^2 \end{bmatrix} \otimes I_r$$

and

$$|\sum(s)| = [(\omega_1 + \tau_1 s)(\omega_2 + \tau_2 s)]^{2r}$$

giving

$$g(\vec{y}_i|S_i = s) = [2\pi(\omega_1 + \tau_1 s)(\omega_2 + \tau_2 s)]^{-r} \times \\ \exp\left\{\frac{-1}{2(\omega_1 + \tau_1 s)^2}\sum_{k=1}^{r}(y_{i1k} - s)^2 - \frac{1}{2(\omega_2 + \tau_2 s)^2}\sum_{k=1}^{r}(y_{i2k} - \alpha - \beta s)^2\right\} \quad [6.7]$$

As mentioned, the marginal likelihood function for the measurements on subject i is found by integrating the joint density function of the data and S_i across the support of S_i. This joint density function is given by

$$h(\vec{y}_i, s; \mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2) = g(\vec{y}_i|S_i = s; \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)f(s; \mu, \sigma_s) \quad [6.8]$$

where $g(\vec{y}_i|S_i = s; \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$ is as given in [6.7] and $f(s; \mu, \sigma_s)$ is the probability density function for S_i which has a normal distribution with mean μ and variance σ_s^2:

$$f(s; \mu, \sigma_s) = \frac{1}{\sqrt{2\pi\sigma_s^2}}\exp\left\{\frac{-(s-\mu)^2}{2\sigma_s^2}\right\} \quad [6.9]$$

Note that we will most often suppress notation and denote $h(\vec{y}_i, s; \mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$, $g(\vec{y}_i | S_i = s; \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$ and $f(s; \mu, \sigma_s)$ by $h(\vec{y}_i, s)$, $g(\vec{y}_i | S_i = s)$ and $f(s)$, respectively.

Thus the likelihood function for the measurements on subject i by both systems is given by:

$$L_i(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2) = \int_{-\infty}^{\infty} h(\vec{y}_i, s) ds \qquad [6.10]$$

where $h(\vec{y}_i, s)$ is as in [6.8]. Because the integral in [6.10] cannot be solved analytically we must approximate it. We do so using a Riemann sum with p partitions: $\{[s_1, s_2], [s_2, s_3], \ldots, [s_p, s_{p+1}]\}$. Specifically we approximate [6.10] with the following 'middle' Riemann sum:

$$L_i(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2) \approx \sum_{l=1}^{p} h(\vec{y}_i, s_l^*)\,(s_{l+1} - s_l) \qquad [6.11]$$

where $s_l^* = (s_l + s_{l+1})/2$ is the midpoint of the partition $[s_l, s_{l+1}]$, $l = 1,2, \ldots, p$. The corresponding log-likelihood function is approximated by

$$l_i(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2) \approx \log\left(\sum_{l=1}^{p} h(\vec{y}_i, s_l^*)\,(s_{l+1} - s_l)\right) \qquad [6.12]$$

The larger the number of partitions p, the better these approximations become. In fact, as $p \to \infty$ the approximations in [6.11] and [6.12] become exact [Stewart, 2003]. The choice of p is discussed in Appendix D.

Because we assume that measurements on different subjects are independent, the overall log-likelihood function for all subjects is found by summing [6.11] over $i = 1,2, \ldots, n$:

$$l(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2) \approx \sum_{i=1}^{n} \log\left(\sum_{l=1}^{p} h(\vec{y}_i, s_l^*)\,(s_{l+1} - s_l)\right) \qquad [6.13]$$

Estimates of $(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$ are found by numerically maximizing [6.13], which we achieve with the constrained optimization routine "fmincon" in Matlab [The MathWorks Inc., 2013] that accounts for $\omega_j > 0$ and $\tau_j \geq 0$, $j = 1,2$. These estimates are then substituted into [6.6] to obtain the maximum likelihood estimate, $\hat{\theta}(s)$.

In order to correctly build the probability of agreement plot, we must also estimate the standard error of $\hat{\theta}(s)$ in order to display approximate confidence intervals for $\theta(s)$. To find the standard error of $\hat{\theta}(s)$, we calculate the observed information matrix for $(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$.

To obtain this observed information matrix we calculate negative second partial derivatives of [6.13], and substitute the observed data into the appropriate sums of squares. We obtain a particular derivative of [6.13] by finding the corresponding derivative of [6.12] and then summing over $i = 1,2, \dots, n$. The derivatives of [6.12] have the following form:

$$\frac{\partial^2 l_i}{\partial \lambda^* \partial \lambda} = \frac{\partial}{\partial \lambda^*}\left(\frac{\partial l_i}{\partial \lambda}\right) = \frac{\sum_{l=1}^{p} \frac{\partial^2 h(\vec{y}_i, s_l^*)}{\partial \lambda^* \partial \lambda}}{\sum_{l=1}^{p} h(\vec{y}_i, s_l^*)} - \left(\frac{\partial l_i}{\partial \lambda}\right)\left(\frac{\partial l_i}{\partial \lambda^*}\right)$$

where $\lambda, \lambda^* \in \{\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2\}$ and

$$\frac{\partial l_i}{\partial \lambda} = \frac{\sum_{l=1}^{p} \frac{\partial h(\vec{y}_i, s_l^*)}{\partial \lambda}}{\sum_{l=1}^{p} h(\vec{y}_i, s_l^*)}$$

Note that to avoid errors, we use Maple [Maplesoft, 2014] to symbolically take the derivatives of $h(\vec{y}_i, s_l^*)$. Also note that we use the observed information matrix, and not the expected information matrix (as has been done in previous chapters), because the expected values of the derivatives of [6.13] are difficult to obtain.

Once the observed information matrix for $(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$ is calculated, we obtain the asymptotic standard deviation of $\tilde{\theta}(s)$ by pre- and post-multiplying this matrix by a change-of-variables vector in accordance with the Delta method [Lehmann and Casella, 1998]. We then substitute the maximum likelihood estimates of $(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$ into this asymptotic standard deviation to obtain the asymptotic standard error of $\hat{\theta}(s)$, which is used in the construction of confidence intervals for $\theta(s)$.

6.2.3 Example

Here we illustrate the application of the probability of agreement analysis when accounting for heteroscedasticity. To do so, we use systolic blood pressure data from the same example used earlier, published by Bland and Altman [1999]. Previously we have discussed comparing observers “J” and “R”, but in this section we focus on the comparison of the measurements by observer “J” using a sphygmomanometer, and the measurements made by a semi-automatic blood pressure monitor, denoted “S”. It is this comparison that is presented in the cited article, where observer “J” is treated as the reference system (MS1), and the semi-automatic monitor “S” is treated as the new system (MS2). As before, the systolic blood pressure (in mmHg) of 85 subjects is measured three times by both systems (in this case “S” and “J”). These data are given in Table C.1 of Appendix C. Note that we use the same data for “J” as in Sections 5.2.4 and 6.1.2.

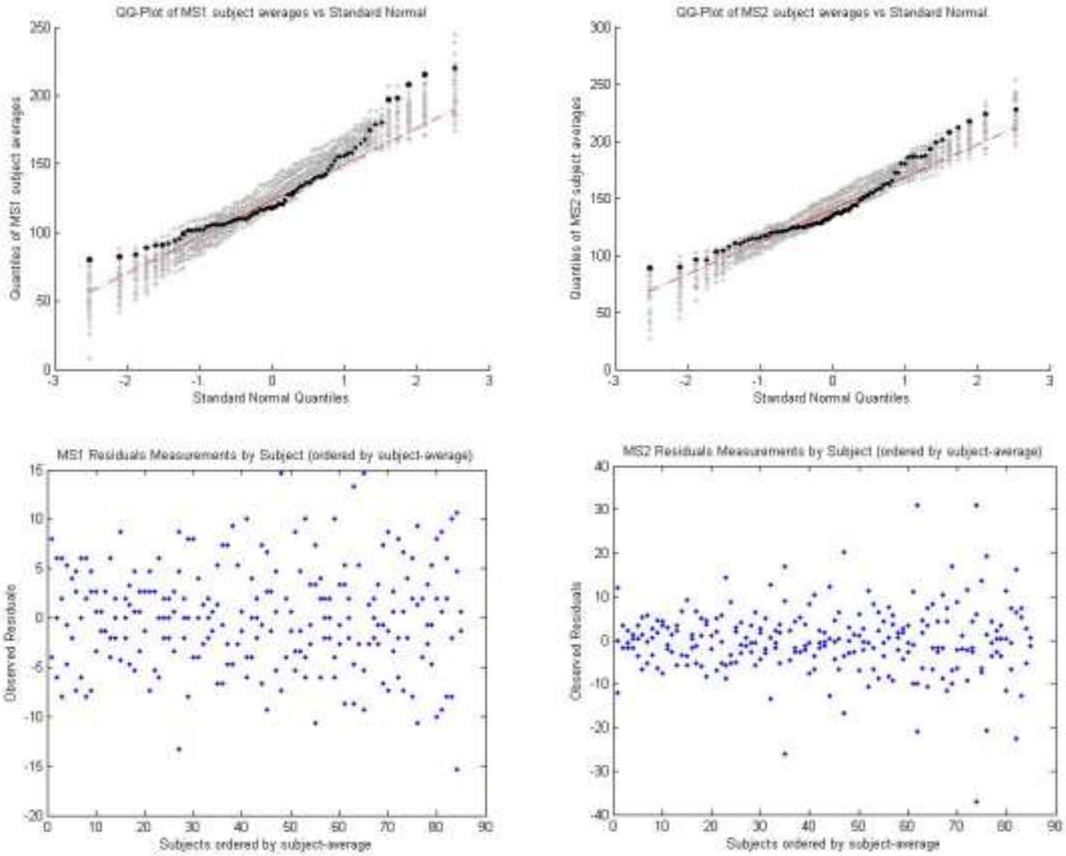

Figure 6.2: Modified QQ-Plot and Repeatability Plot for observers “J” (MS1) and “S” (MS2) from the example data.

Left panels correspond to observer “J” and right panels correspond to observer “S”

Before fitting model [6.4] and estimating the associated parameters, we should first justify its use. In Figure 6.2 we present the modified QQ-plots and the repeatability plots for these data. The modified QQ-plots suggest that it is reasonable to model the true values of the measurand with a normal distribution. When we examine the repeatability plots for evidence of heteroscedasticity, a funnel pattern on the plot for system “S” is apparent. Specifically, the plot depicts an increase in residuals as the subject-average increases suggesting that the measurement variability for system “S” increases as the true value of the measurand increases. As such, the homoscedastic model [6.2] should not be used, and instead we proceed with the use of model [6.4]. Note that, as before, the repeatability plot for observer “J” does not suggest much of a dependence between true values and the variability of observer “J”.

With the use of model [6.4] warranted, we use the data in Table C.1 to obtain the maximum likelihood estimates of $(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$, and their corresponding asymptotic standard errors. These results are presented in Table 6.2. To obtain these estimates we used $p = 150$ partitions to approximate the likelihood function [6.11]. See Appendix D for details on this choice of p.

	Estimate	Standard Error
μ	127.5222	3.1496
σ_s	27.9784	2.3278
α	3.4501	5.1584
β	1.0943	0.0429
ω_1	0.0000	6.2649
ω_2	0.0000	4.0452
τ_1	0.0995	0.0454
τ_2	0.0779	0.0339

Table 6.2: Maximum likelihood estimates and standard errors associated with the J vs. S example

Using these estimates we calculate $\hat{\theta}(s)$ for s in the range $(\hat{\mu} - 3\hat{\sigma}_s, \hat{\mu} + 3\hat{\sigma}_s)$ and construct the probability of agreement plot shown in Figure 6.3. Note that the calculation of the probabilities in this plot assumes a clinically acceptable difference with $c = 10$ as has been done in the previous examples. Given this clinically acceptable difference, the plot suggests that system “S” and observer “J” do not agree very well, with the probability of agreement decreasing from approximately 0.6 to 0.2 across a typical range of systolic blood pressure values. This probability will increase for larger values of c and decrease for smaller values, but it is unlikely that a practitioner would recommend that system “S” be used interchangeably with observer “J”.

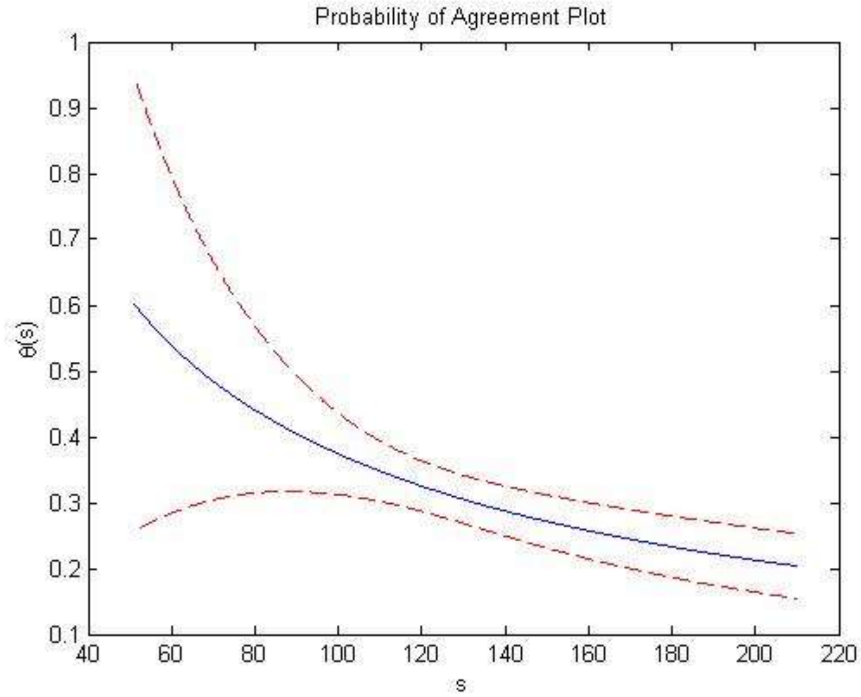

Figure 6.3: Probability of Agreement Plot comparing “J” and “S” from the example data with $c = 10$

The source of disagreement is clear when we examine the estimates in Table 6.2. The average blood pressure measurement by observer J is estimated to be $\hat{\mu} \approx 127$ mmHg, while the average measurement by system S is estimated to be $\hat{\alpha} + \hat{\beta}\hat{\mu} \approx 143$ mmHg. Thus observer J and system S differ on average by roughly 16 mmHg. This difference arises partially because the fixed bias is non-zero, but it is largely due to the fact that the proportional bias is significantly different from 1 ($\hat{\beta} = 1.09$). The proportional bias also accounts, to some extent, for the decrease in agreement across the range of S_i. This dependence of agreement on the true systolic blood pressure is also due to the presence of heteroscedasticity. Because $\hat{\tau}_1$and $\hat{\tau}_2$ are both non-zero, this indicates that both systems are somewhat heteroscedastic. And because $\hat{\tau}_1 \neq \hat{\tau}_2$, the degree of heteroscedasticity is not the same in the two systems, which is another source of disagreement. We note that $\widehat{\omega}_1$ and $\widehat{\omega}_2$ are both approximately zero, which suggests that the structure of heteroscedasticity in each system is strictly proportional to the true values. The large standard errors associated with these estimates reflects the difficulty of estimating ω_j near the boundary.

The estimates in Table 6.2 demonstrate that the underlying distribution for the two measurement system's measurements are different. In addition to using the probability of agreement plot to assess the level of agreement, we could also use a likelihood ratio test to formally test for equality between the two measurement systems with a hypothesis such as $H_0\colon \alpha = 0, \beta = 1, \omega_1 = \omega_2, \tau_1 = \tau_2$. In general, we would expect that if a formal test rejects equality, the probability of agreement would be low.

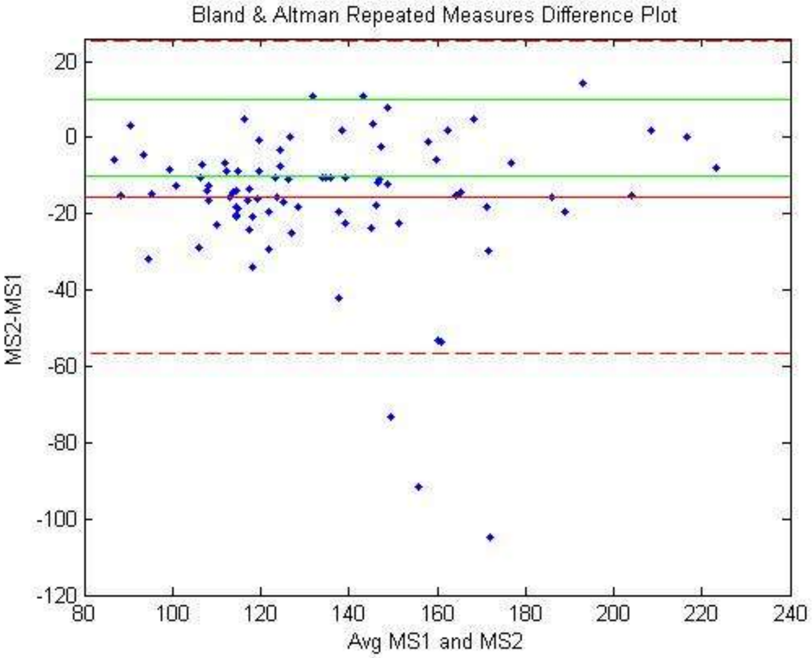

Figure 6.4: Repeated measures difference plot comparing "J" and "S" from the example data with $c = 10$

For comparison, we present the repeated measures difference plot for these data in Figure 6.4. From this plot Bland and Altman [1999] draw a similar conclusion, citing a "lack of agreement" between "S" and "J", but they do not quantify this lack of agreement. They also do not acknowledge the obvious relationship between the differences and averages. By using the repeatability plots in Figure 6.2, we were able to diagnose the heteroscedasticity of system "S", and use the appropriate model. In doing so we estimated the parameters $(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$ giving an accurate summary of the relationship between "S" and "J". As in the homoscedastic case, the probability of agreement analysis provides a much more informative assessment of agreement than the limits of agreement analysis, when heteroscedasticity is present.

The probability of agreement analysis technique has been automated for this heteroscedastic scenario as well, and Matlab [The MathWorks Inc., 2013] software is available to practitioners who wish to use it to determine whether two (possibly heteroscedastic) measurement systems are interchangeable.

6.2.4 Other Considerations

In this subsection we comment on a few additional factors to consider when using the probability of agreement analysis to determine whether two, possibly heteroscedastic, measurement systems are interchangeable. We first point out that unlike the homoscedastic case, we do not consider an unconditional version of the probability of agreement here. Recall that we defined θ in [5.4] which could be used when the probability of agreement $\theta(s)$ was relatively unchanged across the range of true values. This unconditional version is based on the marginal distribution of the Y_{ijk}'s and does not depend on the true value of the measurand, s. Because the model we propose in the heteroscedastic case ([6.4]) explicitly depends on s, we cannot generalize θ to this scenario.

Model [6.4] also assumes that the unknown true values of the measurand follow a normal distribution, which may not be realistic. For example, when measuring a characteristic whose values are positive but clustered near zero, a right-skewed distribution may more accurately describe the behaviour of the true values. A skewed distribution may also be reasonable if measurements are being made on a mixture of healthy and diseased individuals. In cases like these we could relax the normality assumption, and specify a different distribution: perhaps a two-parameter gamma distribution.

Fortunately the maximum likelihood procedure discussed in Section 6.2.2 could still be applied in this case. For example if we assume S_i has a gamma distribution we need only substitute the gamma density function for the normal density function in [6.8], the joint density function for S_i and the data. In doing this, the parameters μ and σ_s would no longer be relevant, and instead the two parameters of the gamma distribution are pertinent. Note that because $\theta(s)$ does not explicitly depend on μ and σ_s, they are just nuisance parameters, and not an integral part of the current analysis method; as long as some distribution is assumed for S_i, $\theta(s)$ can be estimated, and interchangeability can be assessed. Thus the normality assumption for the true values is not overly restrictive- it can easily be changed.

If, however, it is reasonable to assume that the true values are normally distributed, we could use Gaussian Quadrature [Lindsey, 2001] to approximate the likelihood function [6.10], instead of the Riemann sum approximation. With Gaussian Quadrature integrals of the form

$$\int_{-\infty}^{\infty} e^{-x^2} q(x) dx$$

are approximated by sums of the form

$$\sum_{l=1}^{p} w_l q(x_l)$$

where the points x_l are called nodes and the w_l are weights. By making the substitution $x = \frac{s-\mu}{\sqrt{2}\sigma_s}$ in [6.10] we get:

$$L_i(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-x^2} g\left(\vec{y}_i \middle| S_i = \sqrt{2}\sigma_s x + \mu\right) dx$$

According to Gauss-Hermite integration, this can be approximated by

$$L_i(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2) \approx \frac{1}{\sqrt{\pi}} \sum_{l=1}^{p} w_l g\left(\vec{y}_i \middle| S_i = \sqrt{2}\sigma_s x_l + \mu\right)$$

Where the weights w_l are functions of the roots of the Hermite polynomial [Abramowitz and Stegun, 1965].

Liu and Pierce [1994] show that the error associated with this approximation has order $O\left(n^{-\left[\frac{p}{3}+1\right]}\right)$ where p is the number of nodes used in the approximation. However, through simulation we found that the asymptotic standard errors associated with the estimates were more accurate with the Riemann sum approximation than with the Gauss-Hermite approximation, even for large p. Both methods provide accurate estimates of the parameters themselves, but we use the Riemann sum as the preferred approximation because it also provides accurate asymptotic standard errors. And as mentioned, the Riemann sum approximation has the added benefit of being flexible with respect to the distribution of S_i. For a comparison of the two approximation methods on a simulated example, see Appendix D.

The last point we discuss in this subsection is the design of an MSC study when heteroscedasticity is present. Until now Section 6.2 has focused largely on analyzing MSC data in this case. In Section 5.3 we recommended an optimal design for precisely estimating the probability of agreement, in the homoscedastic case. In particular, when each system can make a total of N measurements in an MSC study, we proposed that $N/2$ subjects be measured twice by each system. We have not extensively studied the effect of the study design in this case, but based on preliminary investigation it appears that more than $r = 2$ replicate measurements by each system are necessary to provide sufficiently precise estimates of the variability parameters ω_j and τ_j, and hence $\theta(s)$. We remark that this is an important topic, and one that we plan to pursue in future work.

6.3 Discussion and Conclusions

One of the major benefits of the probability of agreement analysis is that it can be easily adapted to more general, and potentially complicated, settings. But no matter the how complicated the generalization might be, the simplicity and the explanatory nature of the probability of agreement does not change.

In this chapter we have developed two generalizations of the probability of agreement analysis: we have suggested modifications to the analysis that can be applied when the true values of the measurand do not follow a normal distribution, and when one or both measurement systems is heteroscedastic. In each of these modifications, we alter the method by which the probability of agreement $\theta(s)$ is estimated, but in each case we show that the analysis still informatively quantifies agreement, and the intuitive interpretation remains the same. For added accessibility, the probability of agreement analysis in these two situations has been automated, and Matlab [The MathWorks Inc., 2013] software is available to practitioners.

In future work we plan to continue adapting the probability of agreement analysis for use in other settings. For example, we may consider the case when the number of replicate measurements

(r_{ij}) by each system and on each subject is different. We may also consider the case when both measurement systems are biased. In this case we might account for fixed and proportional bias, α_j and β_j, in each system. And lastly, it is of interest to adapt the probability of agreement analysis to account for observer effects when each system is operated by more than one individual.

Chapter 7

Conclusion and Extensions

7.1 Summary

The objective of this thesis was to expand upon the statistical methodologies associated with the assessment and comparison of continuous measurement systems. Measurement systems are an important part of many industrial and medical processes, where understanding a complex system or process cannot be done without high quality measurements. As such, evaluating the adequacy of a single measurement system, or comparing the performance of two measurement systems, is essential.

In the first part of this thesis (Chapters 1-3), we focused on assessing measurement systems where a measurement system's adequacy is determined by the results of a measurement system assessment (MSA) study. The standard design of such a study, which we refer to as the standard plan (SP), involves multiple measurements on each of a random sample of subjects. The adequacy of the measurement system is quantified by the proportion of overall variation due to the measurement system (gauge R&R ratio, γ). Using the data from the MSA study, γ is estimated and the measurement system's adequacy is evaluated.

The goal of the work in Chapters 2 and 3 was to consider alternative designs to the standard plan, which provide a more precise estimate of the measurement system's performance. In Chapter 2, we propose the use of unbalanced 'augmented' designs which augment the standard plan with single measurements on additional subjects [Stevens et al., 2010]. The goal of these designs is to supplement the information gained from the standard plan, so as to more precisely estimate the gauge R&R ratio γ, and hence the measurement system's performance. When a measurement system is operated by multiple observers we show that use of an appropriate augmented plan can produce substantial gains in precision for estimating γ compared to the best standard plan with the same total number of measurements.

In Chapter 3 we consider the assessment of a measurement system that is used routinely, and for which a record of historical (baseline) measurements are kept. In this case we propose incorporating these baseline measurements in both the planning and analysis of an MSA study [Stevens et al., 2013]. We demonstrated the substantial benefits of incorporating baseline data into the analysis, where most of the gains in precision can be obtained with relatively small baseline sample sizes. We also recommended good standard plans with a fixed total number of measurements that take into account available baseline data.

The second part of this thesis (Chapters 4-6) is concerned with the comparison of two measurement systems, where the goal is to decide whether two systems agree well enough to be used interchangeably. To make this determination, a measurement system comparison (MSC) study is undertaken in which a random sample of subjects are each measured multiple times by both systems. Chapters 5 and 6 develop a novel method for analyzing MSC study data that adequately quantifies and summarizes the agreement between two measurement systems.

In Chapter 5 we propose the probability of agreement analysis as a statistical tool for assessing the interchangeability of two measurement systems [Stevens et al., 2014 (*under revision*)]. This analysis is to be used as an alternative to various other methods which fail to evaluate agreement concisely and accurately. The proposed method is based on the probability of agreement which, for a particular value of the measurand, is the probability that the absolute difference between measurements by each system on the same subject is small enough to be considered clinically (or practically) acceptable. We then use estimates of this probability to construct the probability of agreement plot which quantifies the agreement between two measurement systems across a range of plausible values for the measurand.

The probability of agreement analysis has three main benefits. The first is that the results are intuitive, and can be interpreted by non-statisticians. The second is that the estimation procedure can be generalized to account for alterations to model assumptions, and the probability of agreement plot can still be constructed, and the results are interpreted in exactly the same way. We illustrated this versatility in Chapter 6 where we considered comparing measurement systems when the true values of the measurand do not follow a normal distribution, or when the measurement variation of one or both systems depends on this unknown true value. The final benefit is that the analysis cannot be performed without the definition of a clinically acceptable difference, and replicate measurements on each subject by each system must be taken. As such, the analysis is safeguarded against misuse, unlike the limits of agreement approach due to Bland and Altman.

7.2 Extensions

In this section we outline several extensions of the thesis work that can be pursued. Like the thesis itself, we consider these extensions separately in terms of measurement system assessment (subsection 7.2.1) and measurement system comparison (subsection 7.2.2). We then discuss the dissemination of results in subsection 7.2.3.

7.2.1 Measurement System Assessment Extensions

In Chapter 2 we chose to use augmented plan A or B and hence chose values of n, m, r and n_A or n_B based on the ability of the plan to precisely estimate γ, the gauge R&R ratio. An alternative method of calculating sample sizes would be to choose n, m, r and n_A or n_B based on their ability to achieve a desired power associated with a hypothesis test such as [1.13]:

$$H_0: \gamma \geq \gamma_0 \text{ versus } H_A: \gamma < \gamma_0$$

where reasonable choices for γ_0 might be the acceptability/unacceptability criteria 0.1 and 0.3 suggested by the Automotive Industry Action Group [2010]. It would be interesting to compare the optimal allocation of subjects, observers and replicate measurements found with this method to that which has already been considered.

In Chapter 1 we discussed leveraged plans [Browne et al., 2009a; 2009b; 2010] as an alternative to the use of standard plans in MSA studies. This series of papers provides one of the only other recommendations for an alternate study design, and so it is of interest to compare the performance of augmented plans with the performance of these leveraged plans. Note that the leveraged plans use fewer subjects, so it is not immediately clear how their performance will compare with the augmented plans.

Two final measurement system assessment extensions have arisen from work on measurement system comparison. The first is to explore is the possibility of modeling and testing for the presence of heteroscedasticity. MSA studies in manufacturing contexts typically assume that the measurement system is linear, and hence homoscedastic. However, in Chapter 6, we saw that heteroscedasticity can be present in clinical contexts and so MSA studies could benefit from the added generality of being able to assess the adequacy of a measurement system whose variability depends on the true value of the measurand. In Chapter 6 we also discussed the comparison of two measurement systems when the true values of the measurand are non-normally distributed. A final extension in the assessment context would be to consider assessing a single measurement system when the true values are non-normally distributed.

7.2.2 Measurement System Comparison Extensions

In future work we plan to extend the work in Chapter 6 and continue to adapt the probability of agreement analysis for use in other settings. Specifically, we suggest various model assumptions that could be altered, thus expanding the applicability of the probability of agreement analysis.

One primary extension to consider is the inclusion of observer effects in the analysis. Often a measurement system is operated by multiple individuals, but models [5.1] and [6.4] do not include observer effects. As such [5.1] and [6.4] implicitly assume that the measurement systems being compared are each operated by a single observer, or if operated by multiple observers, we assume that their effects are the same. Any variation that is attributable to observers is then confounded with the measurement variation. We may relax this assumption and include a fixed or random observer effect in these models, as is often done in the analysis of MSA studies.

Another assumption that we have made in models [5.1] and [6.4] is that when each measurement system makes r replicate measurements on each subject, we assume that r is the same for each system and each subject. We may modify the modeling and estimation to accommodate the scenario in which the two systems make a different number of replicate measurements per subject, i.e. $r_1 \neq r_2$, or a different number of measurements on each subject, i.e. $r_{ij} \neq r_{lj}$, where $i, l = 1,2, \dots, n$ index subjects and $j = 1,2$ indexes measurement systems.

As well, we have assumed that the existing (reference) system is unbiased, and all inferences regarding bias are made relative to it through α and β. However, in some situations it might not be reasonable to assume the existing system is unbiased. In this case we may model both systems as being biased and account for fixed and proportional bias, α_j and β_j, in each system. By including bias parameters for both systems, the models become over-parameterized and separate estimation of each individual parameter becomes infeasible. However, even in such an over-parameterized model, the probability of agreement is still estimable.

Another area of extension concerns the design of MSC studies when using the probability of agreement analysis. In Chapter 5 we recommended an optimal design for precisely estimating the probability of agreement, in the homoscedastic case. We have not extensively studied the effect of the study design in the heteroscedastic case, but this is an important topic, and one that we plan to pursue in future work.

With regard to MSC study design (in the homoscedastic or heteroscedastic case), we might also consider the effect of incorporating baseline information into the planning and analysis. If the measurement system being compared to the existing one is new, baseline data may not be available for it. But if baseline data is available for the existing system it could help to accurately describe the distribution of true values, which may help to estimate the probability of agreement.

One other possible extension is the simultaneous comparison of three or more measurement systems. Current analysis methods, the probability of agreement included, compare measurement systems in a pairwise fashion. However, in some situations it is of interest to compare multiple measurement systems. For example, Ungerer et al. [2012] compare four methods of measuring cardiac troponin, and Manley et al. [2007] compare eleven assays for measuring insulin concentration. In these cases it is of interest to compare each measurement system to each other, and not just to the reference system. Using a series of pairwise comparisons to achieve this goal is inefficient (particularly when the number of systems is large). As such, we wish to modify the probability of agreement analysis to simultaneously compare more than two measurement systems.

7.2.3. Dissemination of Results

The last extension we discuss concerns the dissemination of results. The results in this thesis are largely based on the content of two published papers [Stevens et al., 2010; Stevens et al., 2013], and two papers which are currently in progress. The papers that are in progress propose the probability of agreement analysis, and contain the content of Chapters 5 and 6.

The journals in which these articles are published, or intended to be published, are statistical in nature and the readership consists primarily of statisticians. However, because the content is very applied, the practitioners that will potentially use these methods are not likely to be statisticians, and not likely to read these statistical journals. As such, the ideas developed in this thesis are not likely to be used, unless we disseminate them to the non-statisticians that are likely to apply them.

In order to increase the chance that practitioners will learn about (and use) these ideas, we plan to publish non-technical articles such as case studies, commentaries, and instructive examples in more accessible journals such as *Quality Progress*, *Quality Engineering*, or *Clinical Chemistry*. In fact, the benefit and impact of incorporating baseline information in MSA studies (as discussed in Chapter 3) has been presented in a relatable and understandable way in a *Quality Progress* article which is to appear in December 2014 [Stevens et al., 2014].

Another way to ensure the use of these ideas is to automate the methods with software that is freely available, and easy to use. Such software is available for planning and analyzing MSA studies at www.bisrg.uwaterloo.ca, but we also intend to make software associated with MSC studies and the probability of agreement analysis available.

Appendix A

A.1 Inverse and Determinant of Covariance Matrix: MSA

In this appendix we derive the inverse and determinant of the covariance matrix $\sum$ associated with model [2.1]. In Section 2.1 we saw that this covariance matrix is given by:

$$\sum = \sigma_s^2 J_{mr\times mr} + \sigma_{so}^2 I_m \otimes J_{r\times r} + \sigma_m^2 I_{mr}$$

where $J_{a\times b}$ is an $a \times b$ matrix of 1's, I_a is the $a \times a$ identity matrix and $\otimes$ denotes the Kronecker product.

Based on the Shermin-Morrison formula [1950] and the Matrix Determinant Lemma [Harville, 2008] we state the following theorem and corollary. We apply this theorem and corollary to find the inverse and determinant of $\sum$.

Theorem 1: *If* $W = A + vv^T$ *where* A *is non-singular then*

1. $W^{-1} = A^{-1} - \frac{A^{-1}vv^TA^{-1}}{1+v^TA^{-1}v}$
2. $|W| = |A|(1 + v^TA^{-1}v)$

Corollary 1: *If* $W = aI_n + bJ_{n\times n}$ *then*

1. $W^{-1} = \frac{1}{a}I_n - \frac{b}{a(a+bn)}J_{n\times n}$
2. $|W| = a^{n-1}(a + bn)$

We begin by letting $A = \sigma_{so}^2 I_m \otimes J_{r\times r} + \sigma_m^2 I_{mr}$ and $v = \sigma_s J_{mr}$, where J_{mr} is a column vector of mr 1's. Since $vv^T = \sigma_s^2 J_{mr\times mr}$ we see that the covariance matrix $\sum$ can be written in the rank-one update form $\sum = A + vv^T$, allowing us to apply the results of Theorem 1. We will compute

A^{-1}, $A^{-1}vv^TA^{-1}$, $v^TA^{-1}v$ and finally $|A|$ to obtain all of the elements of the formulae necessary to calculate $\sum^{-1}$ and $|\sum|$.

Before calculating A^{-1}, note that $A = \sigma_{so}^2 I_m \otimes J_{r\times r} + \sigma_m^2 I_{mr}$ can be written as

$$A = I_m \otimes [\sigma_{so}^2 J_{r\times r} + \sigma_m^2 I_r]$$

Because $(P \otimes Q)^{-1} = P^{-1} \otimes Q^{-1}$ [Jemderson et al., 1983], we have

$$A^{-1} = I_m \otimes [\sigma_{so}^2 J_{r\times r} + \sigma_m^2 I_r]^{-1}$$

To find the inverse of $\sigma_{so}^2 J_{r\times r} + \sigma_m^2 I_r$ we note that it has the correct form to apply the results of Corollary 1. Based on the first result of Corollary 1, we have

$$\begin{aligned} A^{-1} &= I_m \otimes \left[\frac{1}{\sigma_m^2} I_r - \frac{\sigma_{so}^2}{\sigma_m^2(\sigma_m^2 + r\sigma_{so}^2)} J_{r\times r}\right]^{-1} \\ &= \frac{1}{\sigma_m^2} I_{mr} - \frac{\sigma_{so}^2}{\sigma_m^2(\sigma_m^2 + r\sigma_{so}^2)} I_m \otimes J_{r\times r} \\ &= a_1 I_{mr} + a_2 I_m \otimes J_{r\times r} \end{aligned} \qquad [A.1]$$

where we let $a_1 = \frac{1}{\sigma_m^2}$ and $a_2 = \frac{-\sigma_{so}^2}{\sigma_m^2(\sigma_m^2 + r\sigma_{so}^2)}$.

Next we calculate $A^{-1}vv^TA^{-1}$:

$$\begin{aligned} A^{-1}vv^TA^{-1} &= [a_1 I_{mr} + a_2 I_m \otimes J_{r\times r}][\sigma_s^2 J_{mr\times mr}][a_1 I_{mr} + a_2 I_m \otimes J_{r\times r}] \\ &= \sigma_s^2[a_1 I_{mr} + a_2 I_m \otimes J_{r\times r}][a_1 J_{mr\times mr} + ra_2 J_{mr\times mr}] \\ &= \sigma_s^2(a_1 + ra_2)[a_1 I_{mr} + a_2 I_m \otimes J_{r\times r}][J_{mr\times mr}] \\ &= \sigma_s^2(a_1 + ra_2)[a_1 J_{mr\times mr} + ra_2 J_{mr\times mr}] \\ &= \sigma_s^2(a_1 + ra_2)^2 J_{mr\times mr} \end{aligned} \qquad [A.2]$$

Next we calculate $v^TA^{-1}v$:

$$\begin{aligned} v^TA^{-1}v &= \sigma_s^2 J_{mr}^T[a_1 I_{mr} + a_2 I_m \otimes J_{r\times r}]J_{mr} \\ &= \sigma_s^2[a_1 J_{mr}^T + ra_2 J_{mr}^T]J_{mr} \end{aligned}$$

$$= \sigma_s^2(a_1 + ra_2)J_{mr}^T J_{mr}$$

$$= mr\sigma_s^2(a_1 + ra_2) \qquad [A.3]$$

To calculate $|A|$ we note that if P is a $p \times p$ matrix and Q is a $q \times q$ matrix, then $|P \otimes Q| = |P|^q|Q|^p$ [Jemderson et al., 1983]. Applying this result we find

$$|A| = |I_m|^r|\sigma_{so}^2 J_{r\times r} + \sigma_m^2 I_r|^m$$

Noting again that $\sigma_{so}^2 J_{r\times r} + \sigma_m^2 I_r$ has the correct form to apply Corollary 1, we apply the second result giving

$$|A| = \left((\sigma_m^2)^{r-1}(\sigma_m^2 + r\sigma_{so}^2)\right)^m \qquad [A.4]$$

Using Equations [A.1-A.3], and applying the first result of Theorem 1, we have

$$\sum^{-1} = a_1 I_{mr} + a_2 I_m \otimes J_{r\times r} - \frac{\sigma_s^2(a_1 + ra_2)^2}{1 + mr\sigma_s^2(a_1 + ra_2)} J_{mr\times mr}$$

$$= a_1 I_{mr} + a_2 I_m \otimes J_{r\times r} - \frac{\sigma_s^2}{(a_1 + ra_2)^{-2} + mr\sigma_s^2(a_1 + ra_2)^{-1}} J_{mr\times mr}$$

Noting that $(a_1 + ra_2) = (\sigma_m^2 + r\sigma_{so}^2)^{-1}$, we have

$$\sum^{-1} = a_1 I_{mr} + a_2 I_m \otimes J_{r\times r} + a_3 J_{mr\times mr} \qquad [A.5]$$

where

$$a_1 = \frac{1}{\sigma_m^2}$$

$$a_2 = \frac{-\sigma_{so}^2}{\sigma_m^2(\sigma_m^2 + r\sigma_{so}^2)}$$

$$a_3 = \frac{-\sigma_s^2}{(\sigma_m^2 + r\sigma_{so}^2)(\sigma_m^2 + r\sigma_{so}^2 + mr\sigma_s^2)}$$

Using equations [A.3] and [A.4], and applying the second result of Theorem 1, we have

$$|\sum| = \left((\sigma_m^2)^{r-1}(\sigma_m^2 + r\sigma_{so}^2)\right)^m \left(1 + mr\sigma_s^2(a_1 + ra_2)\right)$$

Substituting $(a_1 + ra_2) = (\sigma_m^2 + r\sigma_{so}^2)^{-1}$ yields

$$|\Sigma| = (\sigma_m^2)^{m(r-1)}(\sigma_m^2 + r\sigma_{so}^2)^{m-1}(\sigma_m^2 + r\sigma_{so}^2 + mr\sigma_s^2) \qquad [A.6]$$

The inverse and determinant given respectively by [A.5] and [A.6] are the ones used in Section 2.1 to construct the log-likelihood function associated with the standard plan measurements for a single subject. Note that when we consider the special case where we assume there is no subject-by-observer interaction we set $\sigma_{so}^2 = 0$ in the above calculations.

Appendix B

B.1 Comparing Asymptotic and Simulated Standard Errors

In this section we expand on the results of the simulation study described in Section 3.2. Because we compare plans using the asymptotic standard error for γ, we must check that the asymptotic results match simulated results. This will ensure that we can use the asymptotic results to appropriately rank the possible $SP(n, r)$ plans for different baseline sizes. In the simulation we compare the simulated and asymptotic standard errors for a variety of plans and parameter values. We consider:

- Total number of measurements: $N =$ 60, 90 and 120;
- Number of observers: $m =$ 1, 2, 3, and 4;
- Number of subjects: $n =$ 3 to a maximum depending on N and m;
- Per-observer baseline sample sizes: $b_j =$ 0, 10, 30, and 100 (recall that $b = mb_j$);
- Parameter values: $\gamma = 0.1, 0.3$, $\delta = 0.1, 0.5, 0.9$, and $\beta = 0.1, 0.5, 0.9$.

For each plan and set of parameter values, we generate 10 000 samples from model [3.1] and for each sample determine the maximum likelihood estimate of γ. We then define the standard deviation of the 10 000 estimates of γ to be the simulated standard error, which we compare to the asymptotic standard error as calculated by the expected Fisher Information matrix (see Section 3.1.2).

We illustrate the results in Figures B.1-B.4. In these figures, “ratio” represents the asymptotic standard error divided by the simulated standard error corresponding to each combination of n, b, and γ. In these figures we see that the asymptotic standard error for γ closely matches the simulated results for all plans when the baseline sample size (b) is large. For simulations based on small baseline sample sizes, the asymptotic results underestimate the simulated results with increasing large differences for plans with fewer subjects, n. We also see that the results are similar for one, two, three, or four observers. In the multiple observer cases, we do not stratify by δ and β. In Figures B.2-B.4, each cluster of points corresponds to the different values of δ and β for a specific combination of n, b, and γ. We see that the conclusions we have drawn do not materially depend on the values of δ and β.

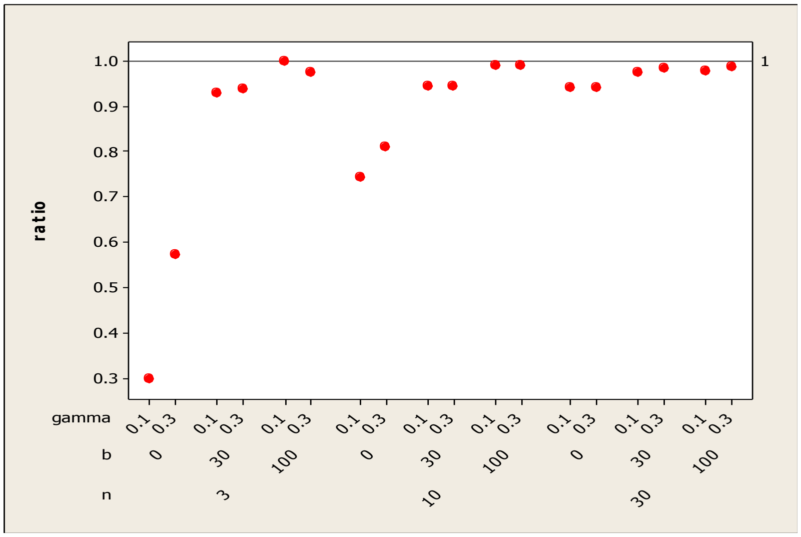

Figure B.1: Individual Value Plot of “ratio” versus n, b, and γ when $m = 1$ and $N = 60$

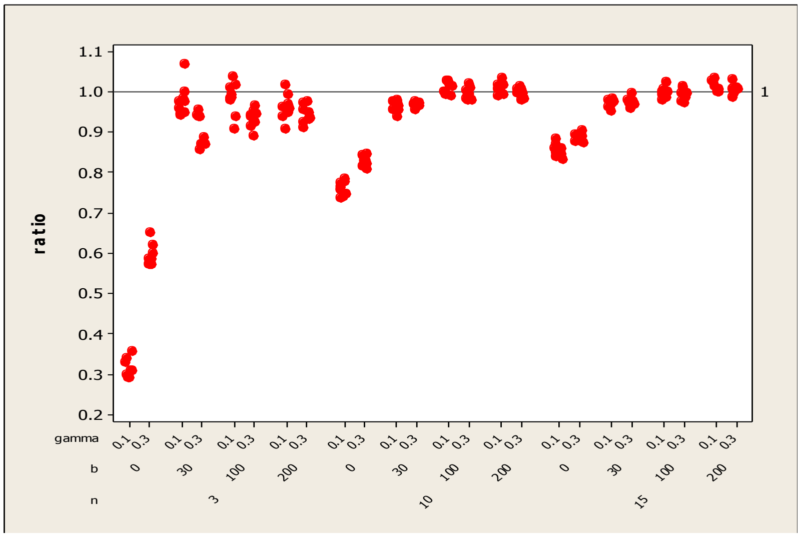

Figure B.2: Individual Value Plot of “ratio” versus n, b, and γ when $m = 2$ and $N = 60$

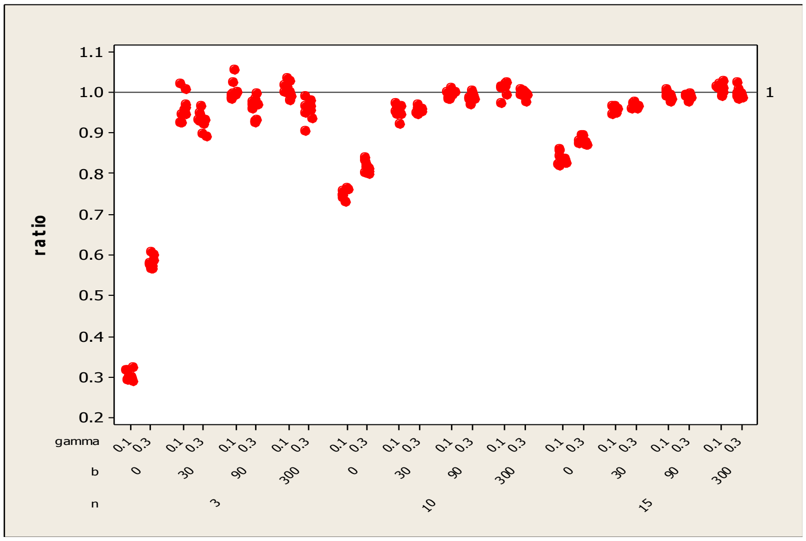

Figure B.3: Individual Value Plot of "ratio" versus n, b, and γ when $m = 3$ and $N = 90$

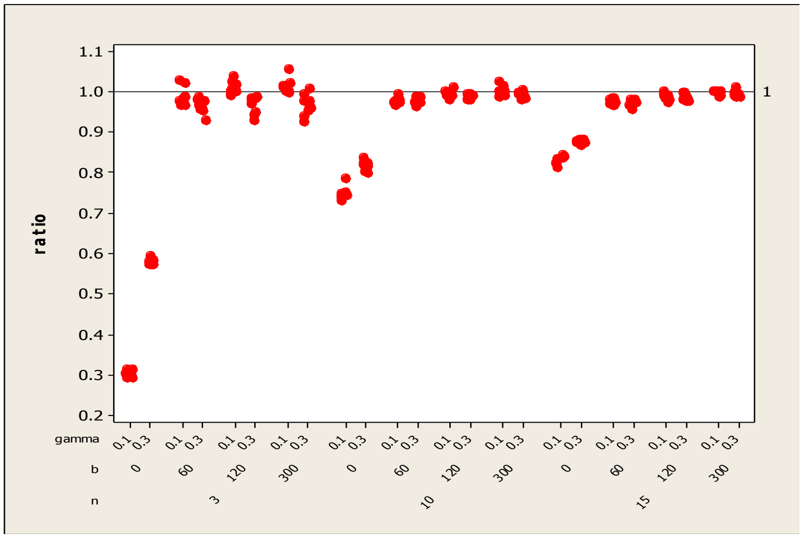

Figure B.4: Individual Value Plot of "ratio" versus n, b, and γ when $m = 4$ and $N = 120$

These results suggest that were are warranted in using the asymptotic results to rank plans when baseline sizes are large. When the baseline sample size is small, and the asymptotic and simulated results to not closely agree, we have found that, the asymptotic and simulated standard errors still provide the same ranking of plans as n and r vary. As such, we proceed to rank plans based on the asymptotic results in all scenarios.

B.2 REML Simulation Results

To compare the restricted maximum likelihood (REML) approach and the usual maximum likelihood (ML) approach, we conduct a factorial simulation study where we compared the estimates of γ for each approach for a number of different plans, baseline sizes, and values for γ, δ, and β. Specifically, we consider:

- Total number of measurements: $N =$ 60, 90, 120;
- Number of observers: $m =$ 1, 2, 3, and 4;
- Number of subjects: $n =$ 3 and 10;
- Baseline sample sizes: $b =$ 0, 25, 50, 75, 100 when $m =$ 1; $b =$ 0, 30, 60, 90 when $m >$ 1;
- Parameter values: $\gamma =$ 0.1, 0.3; $\delta =$ 0.1, 0.5, 0.9; and $\beta =$ 0.1, 0.5, 0.9.

For each plan and set of parameter values, we generated 10 000 samples from model [3.1] and for each sample determined both the REML and usual ML estimates of γ. We then take the mean of these 10 000 estimates and subtract the true value of γ, giving the bias of the estimator. We define the standard deviation of the 10 000 estimates of γ to be the standard error of the estimator.

We illustrate the results of the one-observer case in Figures B.5 and B.6, and the results of the two-observer case in Figures B.7 and B.8. The conclusions we draw from these plots are independent of the number of observers m, which is why the $m =$ 3 and 4 cases are not presented here. Figures B.5 and B.7 display the bias associated with estimating γ with each technique, and Figures B.6 and B.8 display the associated standard errors. In each plot the clusters of points represent the results of each parameter combination at the specified values of b and n.

Figures B.5 and B.7 suggest that the REML estimator of γ is indeed substantially less biased than the usual maximum likelihood estimator (though still not unbiased) when there are no baseline data, especially when the number of subjects (n) in the SP is small. However, when we add even a small amount of baseline data, say $b =$ 30 observations, the difference in bias between the two estimation techniques disappears.

Figures B.6 and B.8 suggest that that variability of the usual maximum likelihood estimators and REML estimators are similar, but for some combinations of the parameter values the usual maximum likelihood estimators are in fact less variable. Because the ML and REML estimators are similarly biased for non-zero baseline sample sizes, and because the variability of ML estimators are less than or equal to that of the REML estimators, we continue to use standard maximum likelihood estimation.

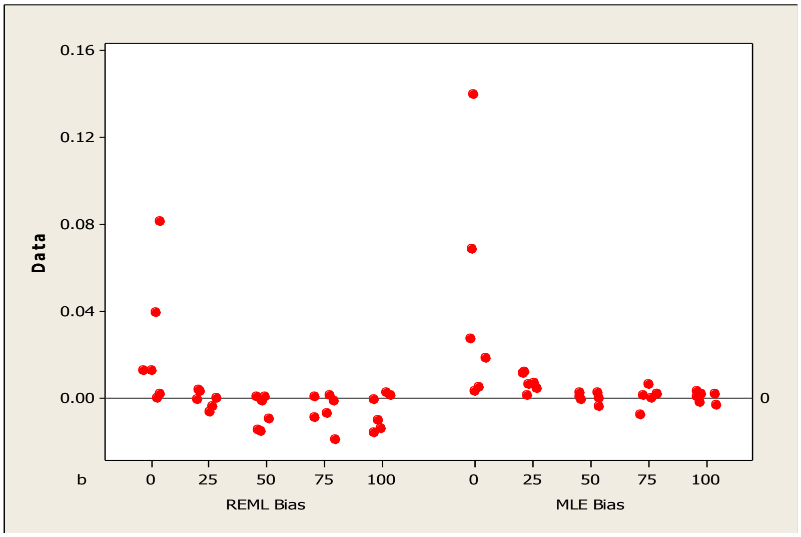

Figure B.5: Individual Value Plot of the bias associated with REML and ML estimates of γ by b when $N = 60$ and $m = 1$

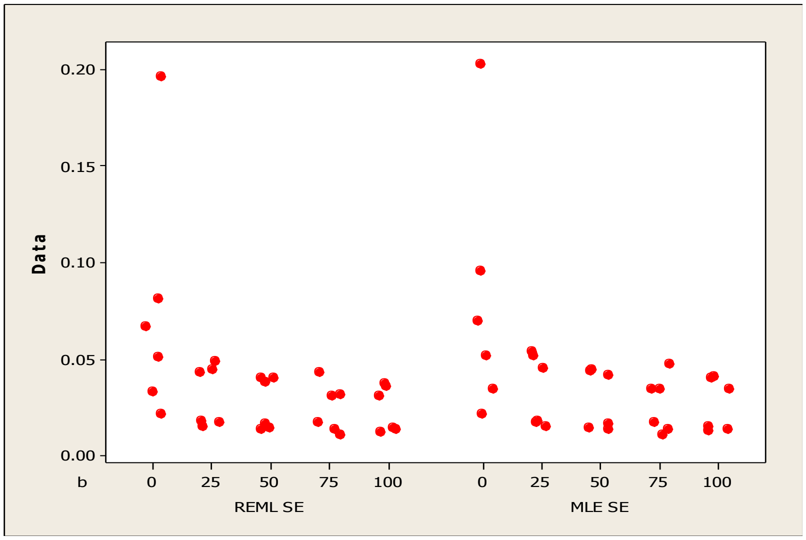

Figure B.6: Individual Value Plot of the standard error associated with REML and ML estimates of γ by b when $N = 60$ and $m = 1$

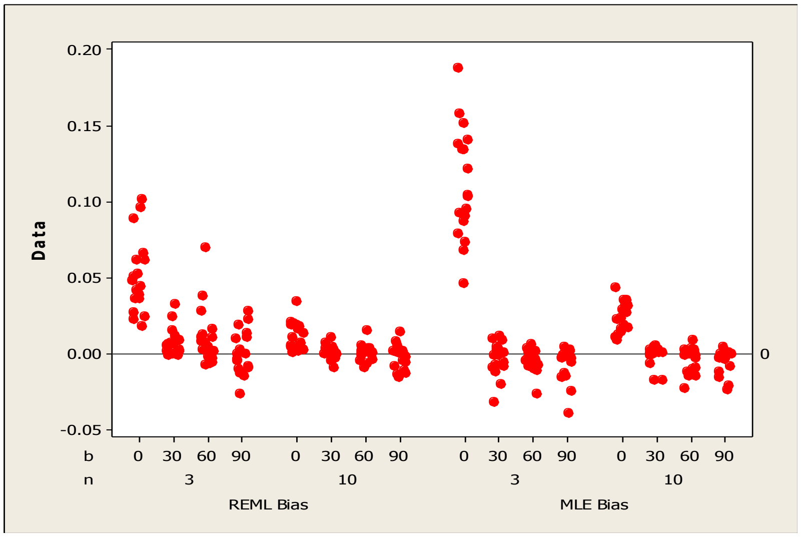

Figure B.7: Individual Value Plot of the bias associated with REML and ML estimates of γ by b and n when $N = 60$ and $m = 2$

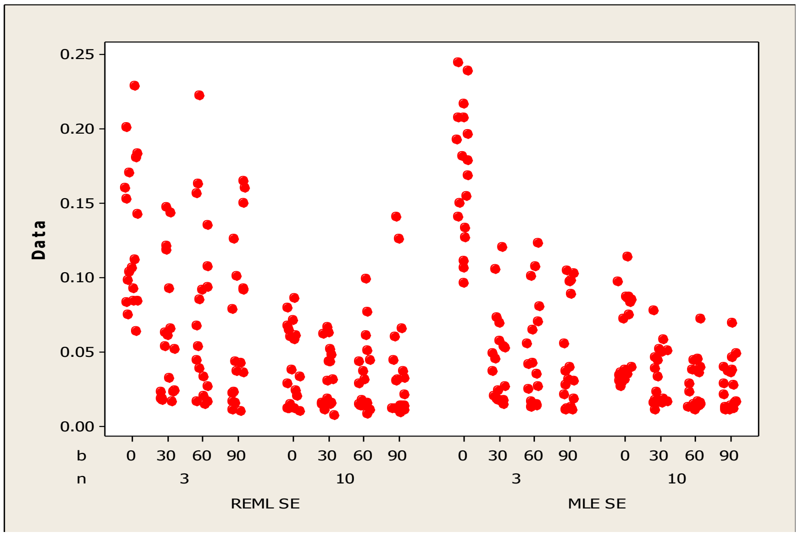

Figure B.8: Individual Value Plot of the standard error associated with REML and ML estimates of γ by b and n when $N = 60$ and $m = 2$

Appendix C

C.1 Inverse and Determinant of Covariance Matrix: MSC

In this section we derive the inverse and determinant of the covariance matrix $\sum$ associated with model [5.1]. In Section 5.2.2 we saw that this covariance matrix is given by:

$$\sum = \sigma_s^2 \begin{bmatrix} 1 & \beta \\ \beta & \beta^2 \end{bmatrix} \otimes J_{r\times r} + \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \otimes I_r$$

where $J_{a\times b}$ is an $a \times b$ matrix of 1's, I_a is the $a \times a$ identity matrix and $\otimes$ denotes the Kronecker product.

As in Appendix A we will compute the inverse, $\sum^{-1}$, and determinant, $|\sum|$, using the Sherman-Morrison formula [1950] and the Matrix Determinant Lemma [Harville, 2008] in accordance with Theorem 1. For convenience, we restate this theorem here.

Theorem 1: *If* $W = A + vv^T$ *where* A *is non-singular then*

1. $W^{-1} = A^{-1} - \frac{A^{-1}vv^TA^{-1}}{1+v^TA^{-1}v}$
2. $|W| = |A|(1 + v^TA^{-1}v)$

To apply the results of this theorem we let

$$A = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \otimes I_r$$

and

$$v = (\sigma_s, \dots, \sigma_s, \beta\sigma_s, \dots, \beta\sigma_s)^T$$

Note that we choose v to have this form because

$$vv^T = \begin{bmatrix} \sigma_s^2 & \cdots & \sigma_s^2 & \beta\sigma_s^2 & \cdots & \beta\sigma_s^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \sigma_s^2 & \cdots & \sigma_s^2 & \beta\sigma_s^2 & \cdots & \beta\sigma_s^2 \\ \beta\sigma_s^2 & \cdots & \beta\sigma_s^2 & \beta^2\sigma_s^2 & \cdots & \beta^2\sigma_s^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \beta\sigma_s^2 & \cdots & \beta\sigma_s^2 & \beta^2\sigma_s^2 & \cdots & \beta^2\sigma_s^2 \end{bmatrix} = \sigma_s^2 \begin{bmatrix} 1 & \beta \\ \beta & \beta^2 \end{bmatrix} \otimes J_{r\times r}$$

Thus the covariance matrix $\sum$ can be written in the rank-one update form $\sum = A + vv^T$, allowing us to apply the results of Theorem 1. We will compute A^{-1}, $A^{-1}vv^TA^{-1}$, $v^TA^{-1}v$ and $|A|$ to obtain all of the elements of the formulae necessary to calculate $\sum^{-1}$ and $|\sum|$.

Because $(P \otimes Q)^{-1} = P^{-1} \otimes Q^{-1}$ [Jemderson et al., 1983], we have

$$A^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \otimes I_r \tag{C.1}$$

Next we calculate $A^{-1}vv^TA^{-1}$:

$$A^{-1}vv^TA^{-1} = \left(\begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \otimes I_r\right)\left(\sigma_s^2 \begin{bmatrix} 1 & \beta \\ \beta & \beta^2 \end{bmatrix} \otimes J_{r\times r}\right)\left(\begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \otimes I_r\right)$$

$$= \left(\begin{bmatrix} \frac{\sigma_s^2}{\sigma_1^2} & \frac{\beta\sigma_s^2}{\sigma_1^2} \\ \frac{\beta\sigma_s^2}{\sigma_2^2} & \frac{\beta^2\sigma_s^2}{\sigma_2^2} \end{bmatrix} \otimes J_{r\times r}\right)\left(\begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \otimes I_r\right)$$

$$= \begin{bmatrix} \frac{\sigma_s^2}{\sigma_1^4} & \frac{\beta\sigma_s^2}{\sigma_1^2\sigma_2^2} \\ \frac{\beta\sigma_s^2}{\sigma_1^2\sigma_2^2} & \frac{\beta^2\sigma_s^2}{\sigma_2^4} \end{bmatrix} \otimes J_{r\times r} \tag{C.2}$$

In the preceding derivation we applied the mixed-product property $(P \otimes Q)(R \otimes S) = PR \otimes QS$ [Jemderson et al., 1983].

Next we calculate $v^TA^{-1}v$:

$$v^T A^{-1} v = [\sigma_s \quad \cdots \quad \sigma_s \quad \beta\sigma_s \quad \cdots \quad \beta\sigma_s] \begin{bmatrix} 1/\sigma_1^2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1/\sigma_2^2 \end{bmatrix} \begin{bmatrix} \sigma_s \\ \vdots \\ \sigma_s \\ \beta\sigma_s \\ \vdots \\ \beta\sigma_s \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sigma_s}{\sigma_1^2} & \cdots & \frac{\sigma_s}{\sigma_1^2} & \frac{\beta\sigma_s}{\sigma_2^2} & \cdots & \frac{\beta\sigma_s}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} \sigma_s \\ \vdots \\ \sigma_s \\ \beta\sigma_s \\ \vdots \\ \beta\sigma_s \end{bmatrix}$$

$$= \frac{\sigma_s^2}{\sigma_1^2} + \cdots + \frac{\sigma_s^2}{\sigma_1^2} + \frac{\beta^2\sigma_s^2}{\sigma_2^2} + \cdots + \frac{\beta^2\sigma_s^2}{\sigma_2^2}$$

$$= r\sigma_s^2 \left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2} \right) \qquad [C.3]$$

And lastly,

$$|A| = (\sigma_1^2 \sigma_2^2)^r \qquad [C.4]$$

Using Equations [C.1-C.3], and applying the first result of Theorem 1, we have

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{bmatrix} \otimes I_r - \frac{\sigma_s^2}{1 + r\sigma_s^2 \left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2} \right)} \begin{bmatrix} \frac{1}{\sigma_1^4} & \frac{\beta}{\sigma_1^2\sigma_2^2} \\ \frac{\beta}{\sigma_1^2\sigma_2^2} & \frac{\beta^2}{\sigma_2^4} \end{bmatrix} \otimes J_{r\times r} \qquad [C.5]$$

Using equations [C.3] and [C.4], and applying the second result of Theorem 1, we have

$$|\Sigma| = (\sigma_1^2\sigma_2^2)^r \left\{ 1 + r\sigma_s^2 \left(\frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_2^2} \right) \right\} \qquad [C.6]$$

The inverse and determinant given respectively by [C.5] and [C.6] are the ones used in Section 5.2.2 to construct the log-likelihood function associated with the r replicate measurements by each system on a single subject.

C.2 Systolic Blood Pressure Example Data

Subject	J1	J2	J3	R1	R2	R3	S1	S2	S3
1	100	106	107	98	98	111	122	128	124
2	108	110	108	108	112	110	121	127	128
3	76	84	82	76	88	82	95	94	98
4	108	104	104	110	100	106	127	127	135
5	124	112	112	128	112	114	140	131	124
6	122	140	124	124	140	126	139	142	136
7	116	108	102	118	110	102	122	112	112
8	114	110	112	112	108	112	130	129	135
9	100	108	112	100	106	112	119	122	122
10	108	92	100	108	98	100	126	113	111
11	100	106	104	102	108	106	107	113	111
12	108	112	112	108	116	120	123	125	125
13	112	112	110	114	112	110	131	129	122
14	104	108	104	104	108	104	123	126	114
15	106	108	102	104	106	102	127	119	126
16	122	122	114	118	122	114	142	133	137
17	100	102	102	102	102	100	104	116	115
18	118	118	120	116	118	118	117	113	112
19	140	134	138	138	136	134	139	127	113
20	150	148	144	148	146	144	143	155	133
21	166	154	154	164	154	148	181	170	166
22	148	156	134	136	154	132	149	156	140
23	174	172	166	170	170	164	173	170	154
24	174	166	150	174	166	154	160	155	170
25	140	144	144	140	144	144	158	152	154
26	128	134	130	128	134	130	139	144	141
27	146	138	140	146	138	138	153	150	154
28	146	152	148	146	152	148	138	144	131
29	220	218	220	220	218	220	228	228	226
30	208	200	192	204	200	190	190	183	184
31	94	84	86	94	84	88	103	99	106
32	114	124	116	112	126	118	131	131	124
33	126	120	122	124	120	120	131	123	124
34	124	124	132	126	126	120	126	129	125
35	110	120	128	110	122	126	121	114	125
36	90	90	94	88	88	94	97	94	96
37	106	106	110	106	108	110	116	121	127
38	218	202	208	218	200	206	215	201	207
39	130	128	130	128	126	128	141	133	146
40	136	136	130	136	138	128	153	143	138
41	100	96	88	100	96	86	113	107	102
42	100	98	88	100	98	88	109	105	97
43	124	116	122	126	116	122	145	102	137
44	164	168	154	164	168	154	192	178	171
45	100	102	100	100	104	102	112	116	116
46	136	126	122	136	124	122	152	144	147
47	114	108	122	114	108	122	141	141	137

48	148	120	132	146	130	132	206	188	166
49	160	150	148	160	152	146	151	147	136
50	84	92	98	86	92	98	112	125	124
51	156	162	152	156	158	152	162	165	189
52	110	98	98	108	110	98	117	118	109
53	100	106	106	100	108	108	119	131	124
54	100	102	94	100	102	96	136	116	113
55	86	74	76	88	76	76	112	115	104
56	106	100	110	106	100	108	120	118	132
57	108	110	106	106	118	106	117	118	115
58	168	188	178	170	188	182	194	191	196
59	166	150	154	164	150	154	167	160	161
60	146	142	132	144	142	130	173	161	154
61	204	198	188	206	198	188	228	218	189
62	96	94	86	96	94	84	77	89	101
63	134	126	124	132	126	124	154	156	141
64	138	144	140	140	142	138	154	155	148
65	134	136	142	136	134	140	145	154	166
66	156	160	154	156	162	156	200	180	179
67	124	138	138	122	140	136	188	147	136
68	114	110	114	112	114	114	149	217	192
69	112	116	122	112	114	124	136	132	133
70	112	116	134	114	114	136	128	125	142
71	202	220	228	200	220	226	204	222	224
72	132	136	134	134	136	132	184	187	192
73	158	162	152	158	164	150	163	160	152
74	88	76	88	90	76	86	93	88	88
75	170	174	176	172	174	178	178	181	181
76	182	176	180	184	174	178	202	199	195
77	112	114	124	112	112	126	162	166	148
78	120	118	120	118	116	120	227	227	219
79	110	108	106	110	108	106	133	127	126
80	112	112	106	112	110	106	202	190	213
81	154	134	130	156	136	132	158	121	134
82	116	112	94	118	114	96	124	149	137
83	108	110	114	106	110	114	114	118	126
84	106	98	100	104	100	100	137	135	134
85	122	112	112	122	114	114	121	123	128

Table C.1: Systolic Blood Pressure Measurements made by two observers (J and R) and an automatic blood pressure measuring machine (S), each making three observations in quick succession on 85 subjects. This table is reproduced from Bland and Altman [1999].

C.3 Comparing Asymptotic and Simulated Standard Errors

In this section we describe a simulation study which was used to compare the asymptotic and simulated standard errors of $\hat{\theta}$ for a variety of (n, r) combinations and parameter values. To cover a wide range of sample sizes, replicate measurements and parameter values, we considered:

- $n = 40$ to 120 in steps of 10
- $r = 2$ to 5 in steps of 1
- $\mu = 1$, 10, 100
- $\sigma_s = \mu/10$, $\mu/4$
- $\sigma_1 = \sigma_s/10$, $\sigma_s/4$
- $\sigma_2 = 3\sigma_1/4$, σ_1, $5\sigma_1/4$
- $\alpha = 0$, 0.05μ
- β = 1, 1.1

For each of the 5,184 combinations of n, r and the parameters, we generated 10,000 samples according to model [5.1] and for each sample determined the maximum likelihood estimate of θ and the asymptotic standard error associated with that estimate. Note that $SE(\hat{\theta})$ is defined as the asymptotic standard deviation of $\tilde{\theta}$, evaluated at the maximum likelihood estimates the other parameters.

We compare the simulated and asymptotic results by dividing the standard deviation of the 10,000 estimates of $\hat{\theta}$ (which we refer to as the simulated standard error) by the average of the 10,000 asymptotic standard errors. Across all combinations of n, r and the parameters, the average of this ratio was 0.9915 and it ranges between 0.89 and 1.11 with the middle 50% lying between 0.97 and 1.02.

Figures C.1 and C.2 depict boxplots of this ratio (labelled ‘ratio’) by n and r. In these plots we see that the value of ‘ratio’ does not depend materially on the number of subjects, or the number of replicate measurements. Similarly, Figure C.3 depicts boxplots of ‘ratio’ by μ. As with n and r, ‘ratio’ does not depend materially on μ. Similar boxplots (not shown here) demonstrate that this ratio also does not depend on α and β. However, the ratio does seem to depend on the relative sizes of the variance components.

The boxplots in Figure C.4 suggest that when σ_s is large relative to μ than the ratio is slightly less than one (i.e. the simulated standard errors are slightly less than the asymptotic standard errors). The boxplots in Figure C.5 demonstrate that this reduction in 'ratio' primarily happens when σ_s is large *and* σ_1 is large relative to σ_s. Otherwise the ratio of simulated and asymptotic standard errors are close to 1. Lastly, the boxplots in Figure C.6 indicate that this ratio is relatively unaffected by the size of σ_2 relative to σ_1.

The conclusions in the previous paragraph, and the results depicted in Figures C.4-C.6 are substantiated by the boxplots in Figure C.7 which indicate that there is slight disagreement between the simulated and asymptotic standard errors when σ_s is large relative to μ, and especially when σ_1 is large relative to σ_s. However, this disagreement is not substantial. Overall the results of this simulation suggest that the asymptotic standard deviation of $\tilde{\theta}$ closely matches the simulated results for all designs. Accordingly we proceed to rank designs based on the asymptotic results.

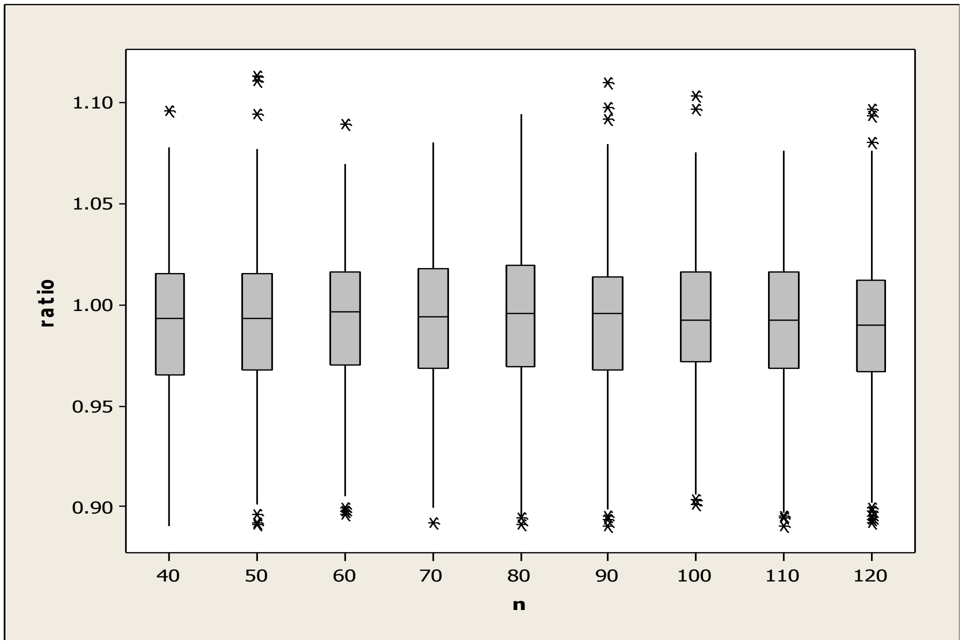

Figure C.1: Boxplots of 'ratio' by n

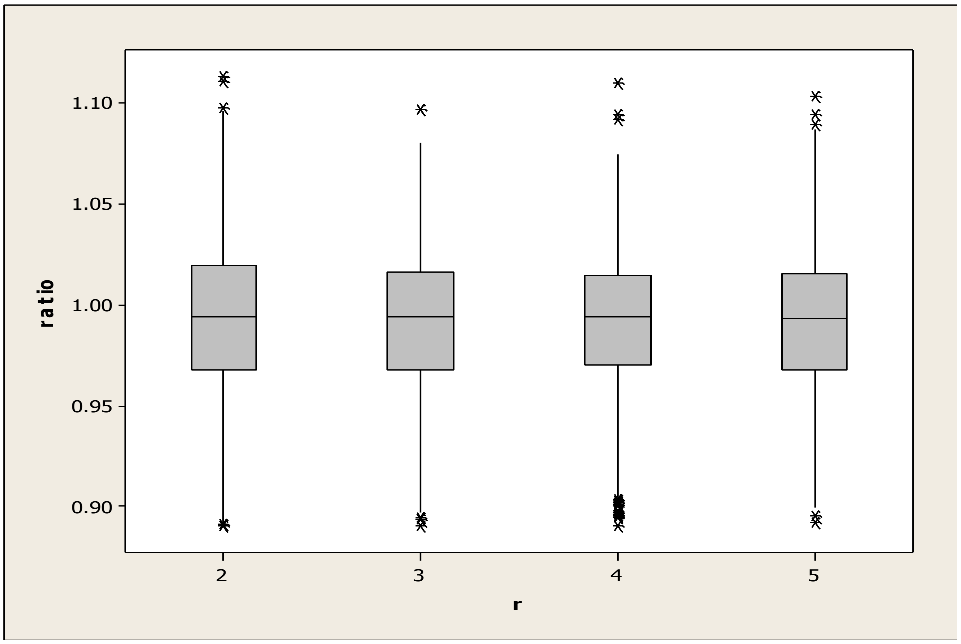

Figure C.2: Boxplots of ‘ratio’ by r

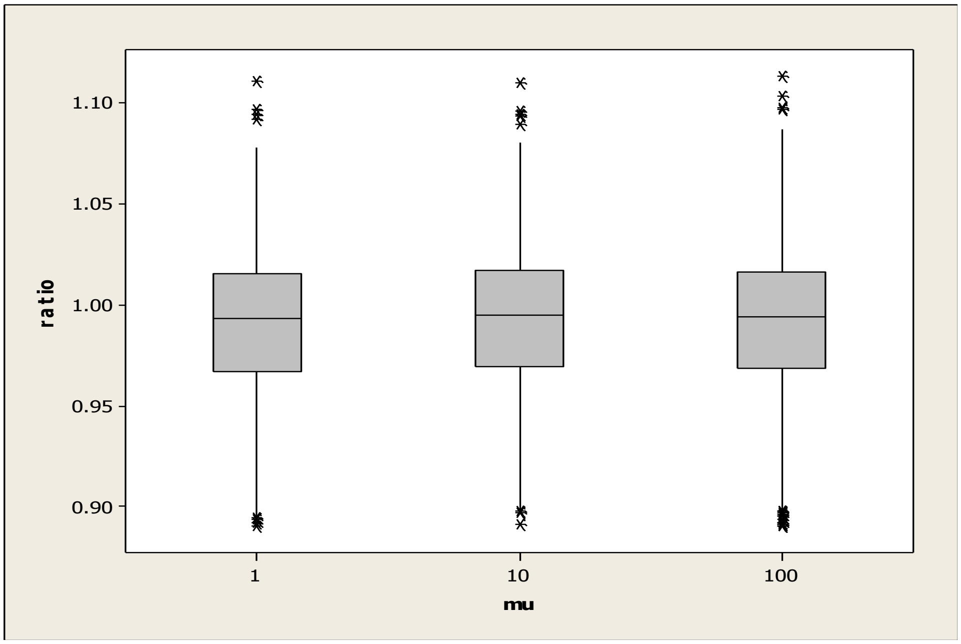

Figure C.3: Boxplots of ‘ratio’ by μ

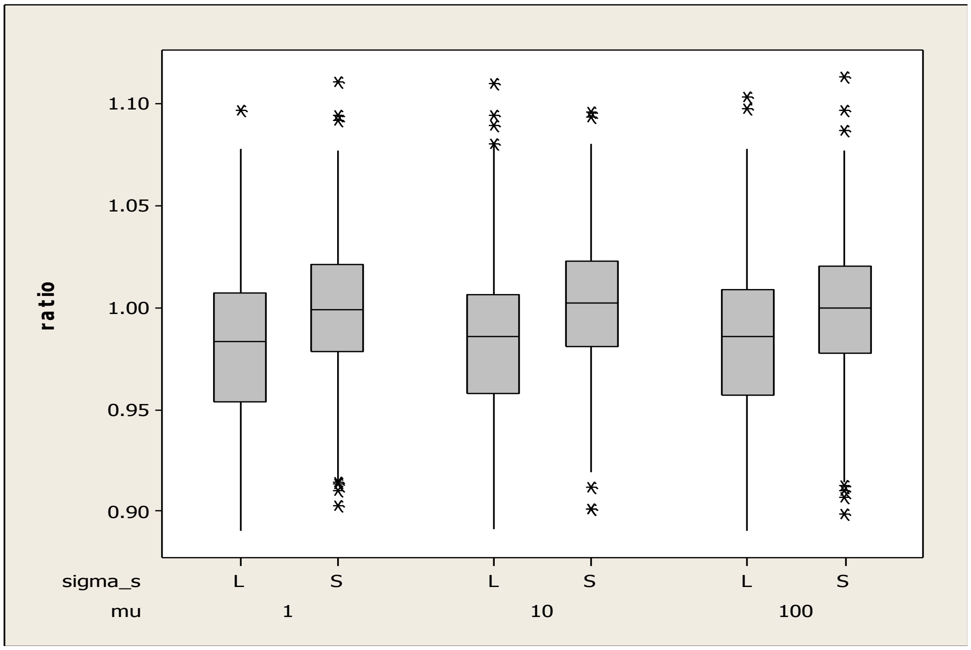

Figure C.4: Boxplots of ‘ratio’ by μ and σ_s, where σ_s is small (S) or large (L) relative to μ

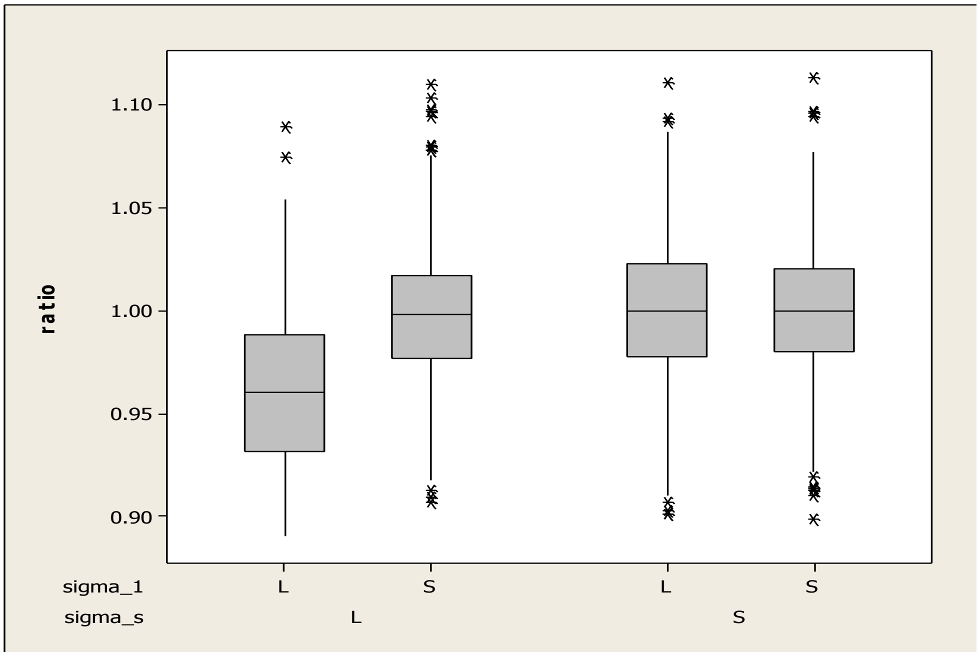

Figure C.5: Boxplots of ‘ratio’ by σ_s and σ_1, where σ_1 is small (S) or large (L) relative to σ_s

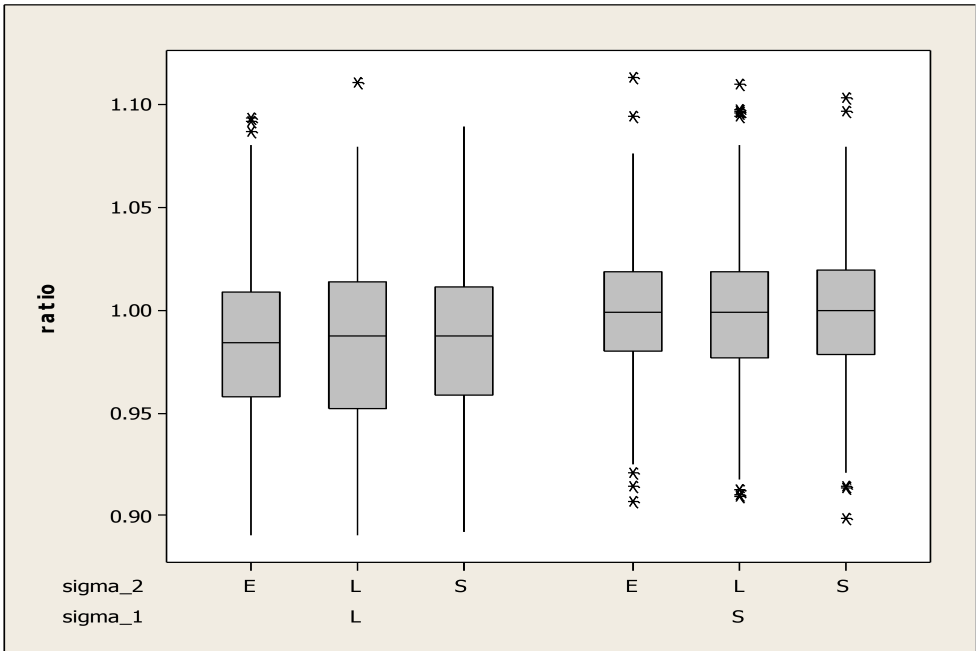

Figure C.6: Boxplots of ‘ratio’ by σ_1 and σ_2, where σ_2 is small (S) or large (L) relative to σ_1, or equal (E) to σ_1

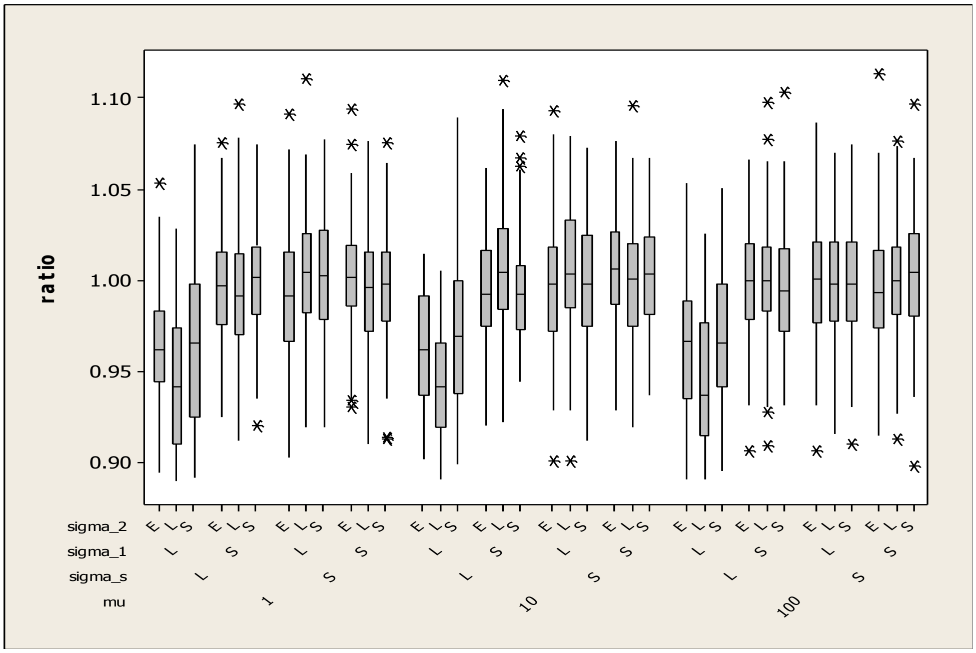

Figure C.7: Boxplots of ‘ratio’ by μ, σ_s, σ_1, and σ_2

C.4 Investigating the Recommended $r = 2$ MSC Study Design

Here we consider balanced designs for an MSC study in which both measurement system measures n subjects r times, for a total of $N = nr$ measurements each. For a fixed number of total measurements N, we refer to a particular allocation of n and r as the 'best' design if, at a particular combination of parameter values, the asymptotic standard deviation of $\tilde{\theta}$ is minimized. Not surprisingly, the best design depends on the values of the unknown parameters. However, through preliminary investigation we found that the design with $(n, r) = (N/2{,}2)$, always had the smallest, or nearly the smallest, asymptotic standard deviation.

To investigate this we performed a simulation study in which we compared the asymptotic standard deviation of the $(n, r) = (N/2{,}2)$ design, with the 'best' design. To cover a wide range of sample sizes, replicate measurements and parameter values, we considered:

- $N = 40, 60, 100, 120, 200$
- $2 \leq r \leq 10$
- $\mu = 1$, 10, 100
- $\sigma_s = \mu/10$, $\mu/4$, $\mu/2$
- $\sigma_1 = \sigma_s/10$, $\sigma_s/4$, $\sigma_s/2$
- $\sigma_2 = 3\sigma_1/4$, σ_1, $5\sigma_1/4$
- $\alpha = -\mu/10$, 0, $\mu/10$
- β = 0.9, 1, 1.1

For a particular combination of the parameter values and $N = 40$, 60, 100, 120, 200, we iterate through $2 \leq r \leq 10$ and take $n = N/r$. In the case that N/r is not an integer, we round this quantity down to the nearest integer to determine n, in which case $nr < N$. We then rank the designs according to the asymptotic standard deviation of $\tilde{\theta}$ and we divide the standard deviation corresponding to the $(n, r) = (N/2{,}2)$ design by that of the best design.

For $N = 40$, 60, 100, 120, 200, $2 \leq r \leq 10$, and the 729 combinations of $(\mu, \sigma_s, \alpha, \beta, \sigma_1, \sigma_2)$ described above, we found the average of this ratio to be 1.01. Thus the asymptotic standard deviation associated with the $(n, r) = (N/2{,}2)$ design is on average only 1% larger than the best design. We also found that the maximum of this ratio is 1.065, indicating that the $(n, r) = (N/2,2)$ design is at most 6.5% worse than the best design.

Figure C.8 is a histogram of this ratio (labelled 'ratio') when $N = 100$. We see that most often this ratio is 1, or very close to 1, but sometimes it can get as large as 1.065. This supports the numerical summaries just mentioned. Histograms of 'ratio' for the other values of N are similar, so we do not present them here.

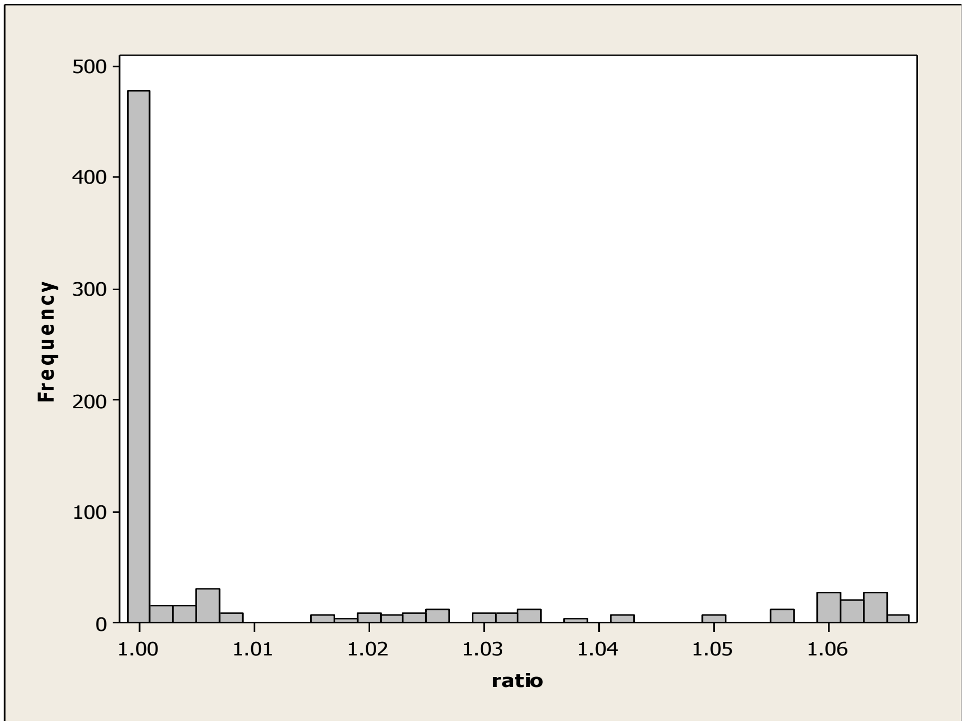

Figure C.8: Frequency Histogram of 'ratio' when $N = 100$

Figures C.9, C.10, and C.11 present boxplots of 'ratio' by α, β, and σ_1/σ_s, respectively (when $N = 100$). These plots demonstrate that larger values of 'ratio' occur when α is different from 0, β is different from 1 and when σ_1 and σ_2 are large relative to σ_s. Based on the boxplots depicted in Figure C.12, we deduce that 'ratio' attains its largest values when $\alpha \neq 0$ and $\beta \neq 1$, simultaneously. These results are the same for other values of N as well.

Thus the performance of the $(n, r) = (N/2,2)$ design is independent of N, and seems only to depend on α and β. Even still, the $(n, r) = (N/2,2)$ design is on average only 1% worse than the best design in terms of its ability to estimate θ precisely. Thus because the 'best' design depends on the values of all of the unknown parameters, and the $(n, r) = (N/2,2)$ design is good across all of the parameter values we considered, we recommend its use.

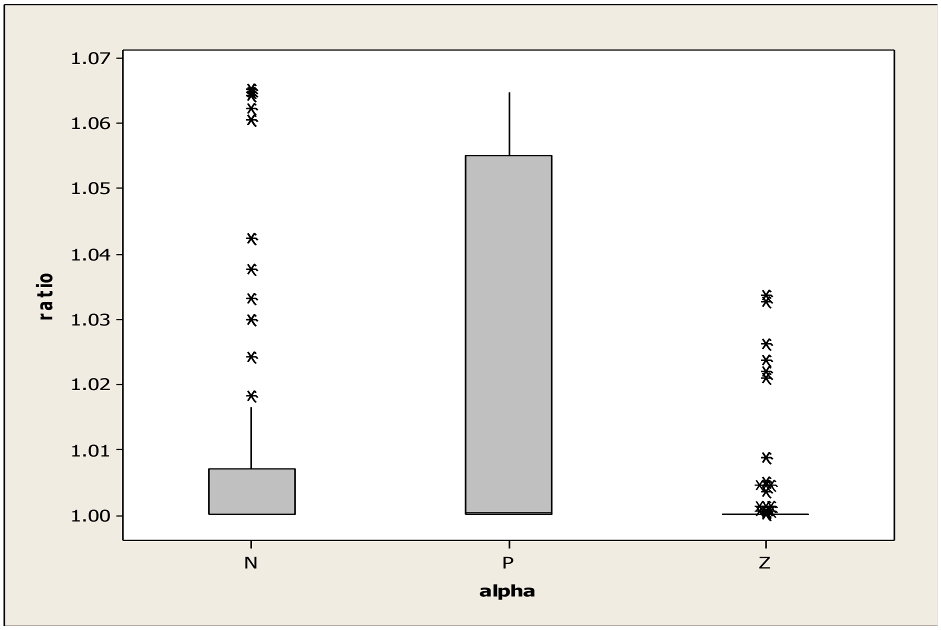

Figure C.9: Boxplots of ‘ratio’ by α for $N = 100$

N, P, Z correspond respectively to $\alpha < 0$, $\alpha > 0$, and $\alpha = 0$

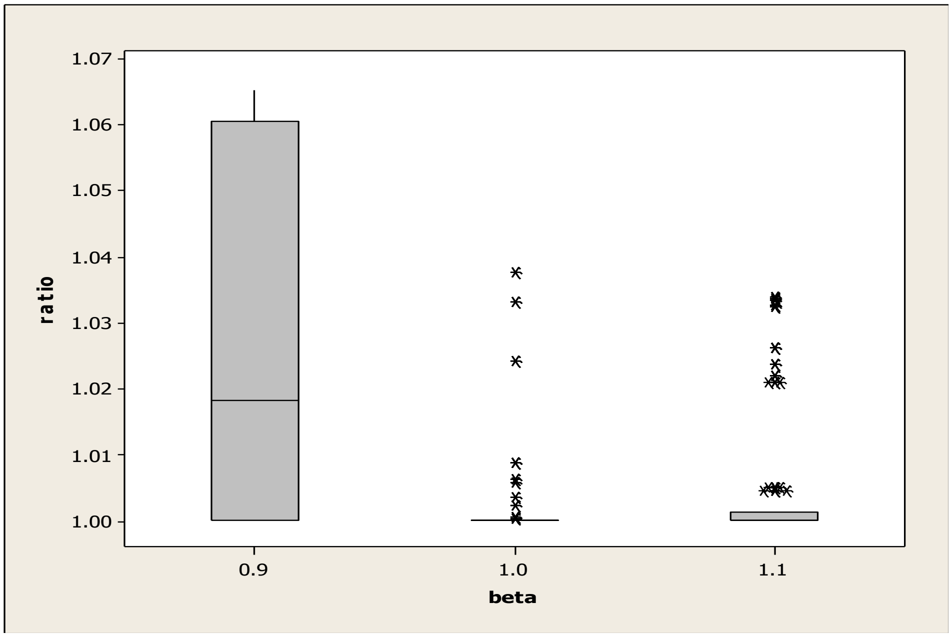

Figure C.10: Boxplots of ‘ratio’ by β for $N = 100$

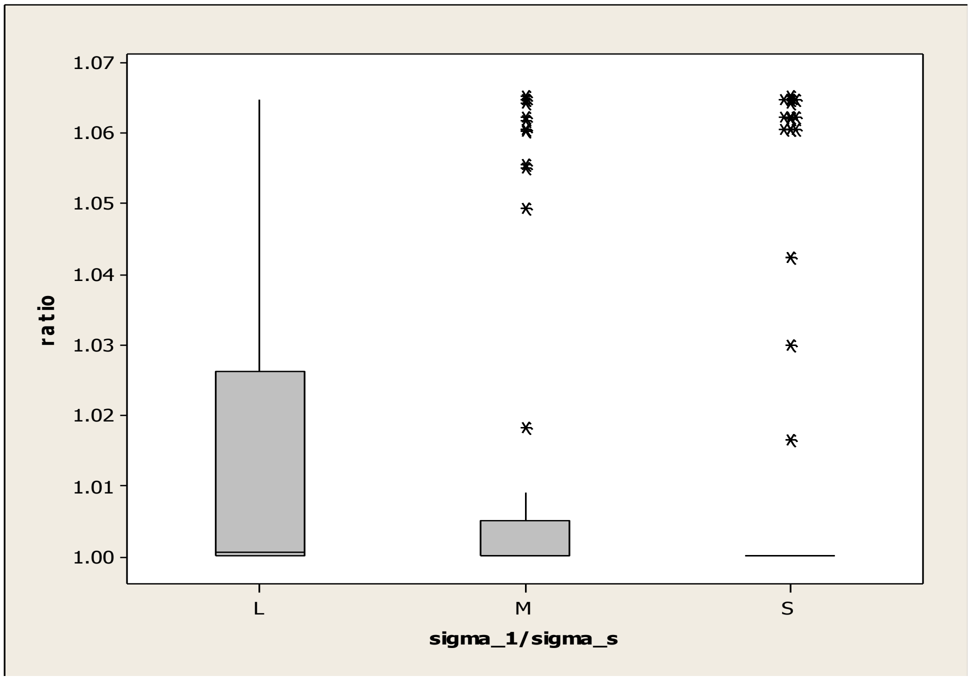

Figure C.11: Boxplots of ‘ratio’ by σ_1/σ_2 for $N = 100$
L, M, S correspond respectively to σ_1 being large, medium, or small relative to σ_s

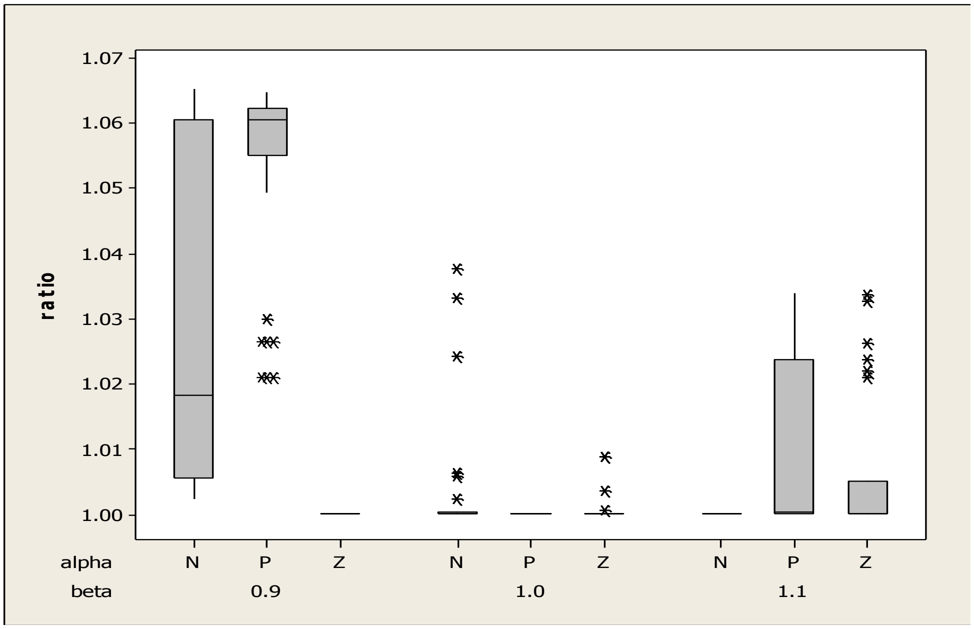

Figure C.12: Boxplots of ‘ratio’ by α and β for $N = 100$
N, P, Z correspond respectively to $\alpha < 0$, $\alpha > 0$, and $\alpha = 0$

Appendix D

D.1 Choosing the Number of Partitions p for the Riemann Sum Approximation

In this section we discuss choosing the number of partitions p for the Riemann sum approximation to the likelihood function in [6.11]. In Section 6.2.3 we discussed the systolic blood pressure data where we compared observer J using the sphygmomanometer and the semi-automatic blood pressure measuring device, S. In that section we noted that $p = 150$ was chosen to obtain the parameter estimates shown in Table D.1. We justify this choice of p here.

	Estimate	Standard Error
μ	127.5222	3.1496
σ_s	27.9784	2.3278
α	3.4501	5.1584
β	1.0943	0.0429
ω_1	0.0000	6.2649
ω_2	0.0000	4.0452
τ_1	0.0995	0.0454
τ_2	0.0779	0.0339

Table D.1: Maximum likelihood estimates and standard errors associated with the J vs. S example

The choice of p must balance accuracy of estimates, and computational efficiency. A small number of partitions is not computationally expensive, but inaccurate estimates may result. On the other hand, a large number of partitions will ensure accurate estimation, but the estimation procedure will take much longer.

Figure D.1 consists of 8 subplots which depict the estimates (blue lines) and associated standard errors (red lines) of the parameters associated with model [6.4] $(\mu, \sigma_s, \alpha, \beta, \omega_1, \omega_2, \tau_1, \tau_2)$, for varying values of p. In each of these subplots both the parameter estimates and standard errors asymptote (toward the values presented in Table D.1) as p becomes large.

We notice that these asymptotes are reached by $p \approx 100$. For the estimates in Table D.1, we chose $p = 150$ to ensure accurate estimation but also to keep computation times reasonable.

This choice of p is sensible for the current example, but it may not be reasonable for other data sets. We have not fully investigated the selection of p in general, so we recommend an exploratory approach similar to the one just described. Software to construct plots like those shown in Figure D.1 is available to aid in this selection process.

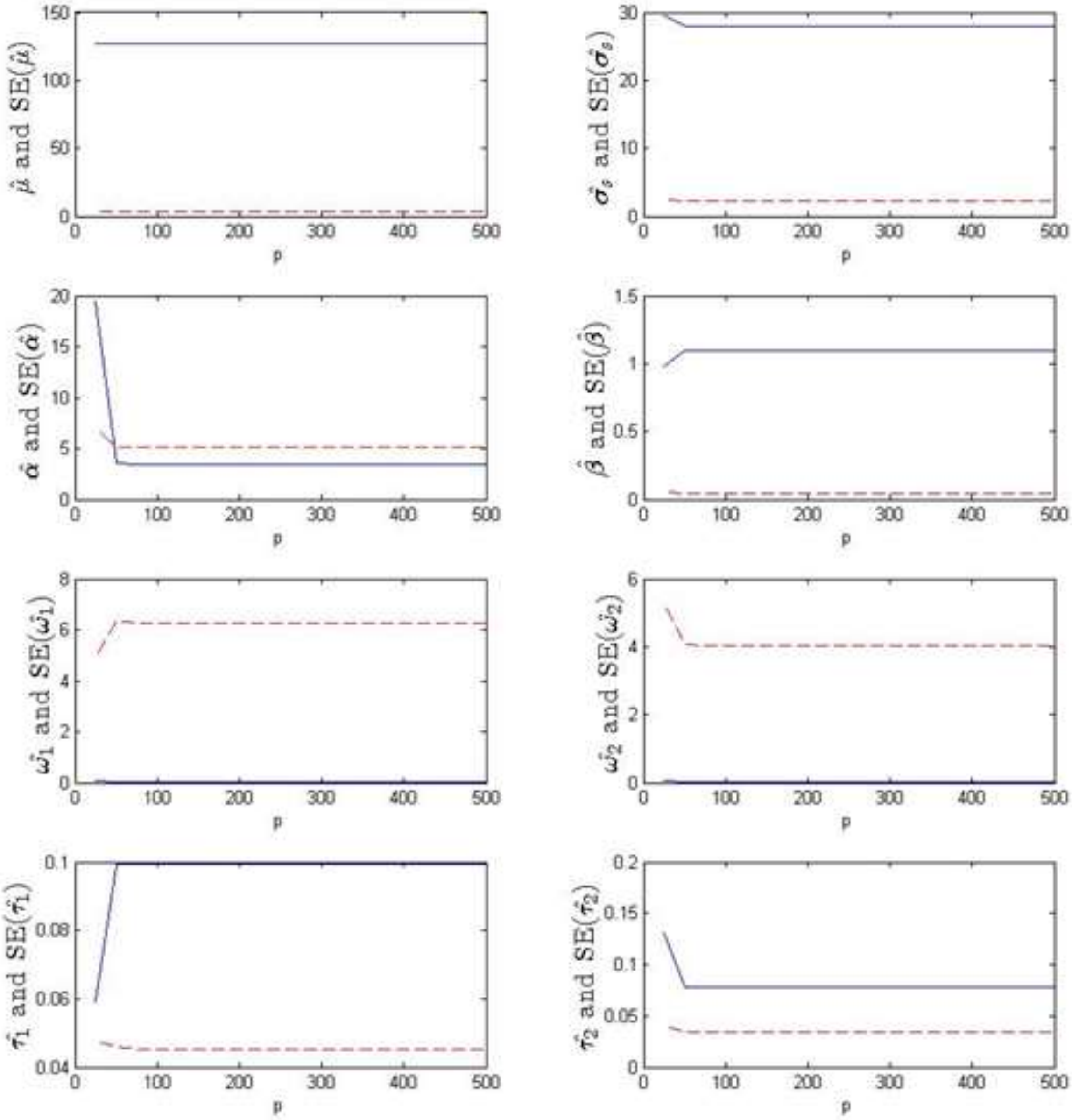

Figure D.1: Estimates (solid blue lines) and Standard Errors (dashed red lines) by Number of Partitions

D.2 Comparing the Riemann Sum and Gauss-Hermite Approximations

In this section we use simulation to highlight the difference between the Riemann sum and Gauss-Hermite methods of integral approximation. We do so by simulating 500 datasets according to model [6.4] with $\mu = 140$, $\sigma_s = 30$, $\alpha = -16$, $\beta = 1$, $\omega_1 = 3$, $\omega_1 = 2$, $\tau_1 = 0.04$ and $\tau_2 = 0.06$. For purposes of illustration we simulate datasets in which $n = 2000$ subjects are measured $r = 20$ times by each system, to ensure ample information is available for estimating each parameter.

For each dataset we calculate the maximum likelihood estimates of the eight parameters, and find their asymptotic standard errors. To assess the performance of each approximation method individually we compare the average of the 500 estimates for each parameter to its true value, and we compare the average of the 500 asymptotic standard errors to the standard deviation of the 500 parameter estimates. Checking that the asymptotic and simulated results match is important for the validity of inferences based on the asymptotic results. The results of this simulation are shown in Table D.2 for the Riemann sum approximation method, and in Table D.3 for the Gauss-Hermite approximation. Note that for each method we use $p = 200$ partitions to ensure optimal accuracy of the approximation.

We see that both methods provide accurate estimates of each of the parameters, but the asymptotic standard errors associated with the Riemann sum approximation match the simulated results more closely than do the asymptotic standard errors associated with the Gauss-Hermite approximation. In particular, the asymptotic and simulated standard errors associated with the Riemann sum approximation match across all parameters, but with the Guass-Hermite method, the asymptotic and simulated results only match for the variability parameters ω_j, τ_j, $j = 1,2$. As such, we use the Riemann sum method as the preferred approximation method. Note that even with this method, there is still small disagreement between the asymptotic and simulated results for the standard error of $\hat{\mu}$. However, this is inconsequential because μ is just a nuisance parameter, since it is not used in the calculation of the probability of agreement.

The results discussed here are for a particular combination of parameter values, but the superiority of the Riemann sum method is also observed for other parameters, and is exaggerated for smaller sample sizes (i.e., smaller n and r).

	μ	σ_s	α	β	ω_1	ω_2	τ_1	τ_2
True Value	140	30	-16	1	3	2	0.04	0.06
Average of 500 Estimates	140.0027	29.9996	-16.02	1.0002	3.0023	1.9819	0.04	0.0601
SD of 500 Estimates	0.7023	0.4750	0.2643	0.002	0.1343	0.1488	0.0010	0.0011
Average of 500 SEs	0.6722	0.4763	0.2780	0.0021	0.1324	0.1509	0.0010	0.0011

Table D.2: Simulated versus Asymptotic Performance of Riemann Sum Likelihood Approximation method

	μ	σ_s	α	β	ω_1	ω_2	τ_1	τ_2
True Value	140	30	-16	1	3	2	0.04	0.06
Average of 500 Estimates	139.152	30.06	16.04	1.0002	2.9962	1.9868	0.0400	0.0601
SD of 500 Estimates	2.8718	1.6647	0.9401	0.0069	0.4137	0.4972	0.0031	0.0037
Average of 500 SEs	2.1390	1.5252	0.8549	0.0065	0.4222	0.4852	0.0031	0.0037

Table D.3: Simulated versus Asymptotic Performance of Gauss-Hermite Likelihood Approximation method

References

Abraham, B., & Ledolter, J. (2006). *Introduction to Regression Modeling*. Belmont, CA: Thomson Brooks/Cole.

Abramawitz, M., & Stegun, I.A. (1965). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.

Aguirre-Torress, V.M., & Lopez-Alvarez, M.T. (2013). Parts of the Process. *Quality Progress*. June, 34-39.

Altman, D.G., & Bland, J.M. (1983). Measurement in Medicine: the Analysis of Method Comparison Studies. *The Statistician* 32, 307-317.

Automotive Industry Action Group. (2003). *Measurement Systems Analysis*, 3rd edition. Southfield, MI: AIAG.

Automotive Industry Action Group. (2009). *Quality Management System*, 3rd edition. Southfield, MI: AIAG.

Automotive Industry Action Group. (2010). *Measurement Systems Analysis*, 4th edition. Southfield, MI: AIAG.

Barnett, R.N. (1965). A scheme for the comparison of quantitative methods. *Amer. J. Clin. Pathol.* 43(6), 562-569.

Barnett, R.N., & Youden W.J. (1970). A revised scheme for the comparison of quantitative methods. *Amer. J. Clin. Pathol.* 54, 454-462.

Berkson, J. (1950). Are there two regressions? *J. Amer. Stat. Ass.* 45, 164-180.

Bland, J.M., & Altman D.G. (1986). Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet* i, 307-310.

Bland, J.M., & Altman D.G. (1995). Comparing Methods of Measurement: Why Plotting Differences Against Standard Method is Misleading. *Lancet* 346, 1085-1087.

Bland, J.M., & Altman D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8, 135-160.

Bland, J.M., & Altman D.G. (2003). Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet. Gynecol.* 22, 85-93.

Bland, J.M., & Altman, D.G. (2007). Agreement between methods of measurement with multiple observations per individual. *J. Biopharmaceut. Stat.* 17(4), 571-582.

Bonnett, D.G. (2002). Sample size requirements for estimating intraclass correlation with desired precision. *Statistics in Medicine*. 21, 1331-1335.

Bookbinder, M.J., & Panosian, K.J. (1987). Using the Coefficient of Correlation in Method-Comparison Studies. *Clinical Chemistry* 33(7), 1170-1176.

Breyfogle, F.W. (1999). *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*. New York, NY: John Wiley & Sons.

Browne, R.P., MacKay, R.J., & Steiner, S.H. (2009a). Improved Measurement-System Assessment for Processes with 100% Inspection. *Journal of Quality Technology*. 41(4), 376-388.

Browne, R.P., MacKay, R.J., & Steiner, S.H. (2009b). Two-Stage Leveraged Measurement System Assessment. *Technometrics*. 51(3), 239-249.

Browne, R.P., MacKay, R.J., & Steiner, S.H. (2010). Leveraged Gauge R&R Studies. *Technometrics*. 52(3), 294-302.

Burdick, R.K., & Larsen, G.A. (1997). Confidence intervals on measures of variability in R&R studies. *Journal of Quality Technology*. 29, 261-273.

Burdick, R.K., Borror, C.M., & Montgomery, D.C. (2005). *Design and Analysis of Gauge R&R Studies: Making Decisions with Confidence Intervals in Random and Mixed effects Models*. Philadelphia, PA: ASA-SIAM Series on Statistics in Applied Probability.

Casella, G., & Berger, R.L. (2002). *Statistical Inference,* 2nd edition. Andover, UK: Cengage Learning.

Corbeil, R.R. & Searle, S.R. (1976). Restricted maximum likelihood (REML) estimation of variance components in mixed models. *Technometrics*. 18, 31-38.

Danila, O., Steiner S.H., MacKay, & R.J. (2008). Assessing a Binary Measurement System. *Journal of Quality Technology*. 40(3), 312-320.

Danila, O., Steiner S.H., MacKay, & R.J. (2010). Assessment of a Binary Measurement System in Current Use. *Journal of Quality Technology*. 42(2), 152-164.

Deming, W.E. (1943). *Statistical Adjustment of Data*. New York, NY: Wiley.

Dewitte, K., Fierens, C., Stockl, D., & Thienpont, L.M. (2002). Application of the Bland-Altman Plot for Interpretation of Method-Comparison Studies: A Critical Investigation of its Practice. *Clinical Chemistry* 48, 799-801.

Donner A., & Eliasziw M. (1987). Sample Size Requirements for Reliability Studies. *Statistics in Medicine*. 6, 441-448.

Efron, B. (1979). Bootstrap Methods: Another look at the Jackknife. *Ann. Stat.* 7(1), 1-26.

Giraudeau, B., & Mary, J.Y. (2001). Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Statistics in Medicine*. 20, 3205-3214.

Hamada, M.S., & Borror, C.M. (2012). Analyzing Unreplicated Gauge R&R Studies. *Quality Engineering*. 24(4), 543-551.

Harville, D.A. (2008). *Matrix Algebra From a Statistician's Perspective*. New York, NY: Springer.

Jemderson, H.V., Pukelshiem, F., & Searle, S.R. (1983). On the history of the kronecker product. *Linear and Multilinear Algebra*. 14(2), 113-120.

Lehmann, E.L., & Casella, G. (1998). *Theory of Point Estimation*, 2nd Edition. New York, NY: Springer.

Lindsey, J.K. (2001). *Parametric Statistical Inference*. New York: Oxford.

Linnet, K. (1993). Evaluation of Regression Procedures for Methods Comparison Studies. *Clinical Chemistry*. 39(3), 424-432.

Liu, Q., & Pierce, D.A. (1994). A note on Guass-Hermite quadrature. *Biometrika*. 81(3), 624-629.

Ludbrook J. (1997). Comparing Methods of Measurement. *Clinical and Experimental Pharmacology and Physiology* 24, 193-203.

Ludbrook J. (2002). Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology* 29, 527-536.

Ludbrook J. (2010). Confidence in Altman-Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology and Physiology* 37, 143-149.

Mader, D.P., Prins, J., & Lampe, R.E. (1999). The Economic Impact of Measurement Error. *Quality Engineering*. 11(4), 563-574.

Majeske, K.D. (2012). Two-sample tests for comparing measurement systems. *Quality Engineering*. 24(4), 501-513.

Manley, S.E., Stratton, I.M., Clark, P.M., & Luzio, S.D. (2007). Comparison of 11 Human Insulin Assays: Implications for Clinical Investigation and Research. *Clinical Chemistry*. 53(5), 922-932.

Mantha, S., Roizen, M.F., Fleisher, L.A., Thisted, R., & Foss, J. (2000). Comparing Methods of Clinical Measurement: Reporting Standards for Bland and Altman Analysis. *Anesth. Analg.* 90, 593-602.

Maple 18. (2014). Maplesoft, Waterloo Maple Inc. Waterloo, ON, www.maplesoft.com

Martin, R.F. (2000). General Deming Regression for Estimating Systematic Bias and its Confidence Interval in Method-Comparison Studies. *Clinical Chemistry.* 46(1), 100-104.

Matlab 8.2.0. (2013). The MathWorks Inc. Natick, MA, www.mathworks.com

Mazu, M.J. (2006). Design and Analysis of Gauge R&R Studies- Book Review. *Technometrics*. 48(2), 305.

Montgomery, D.C., & Runger, G.C. (1993a). Gauge Capability and Designed Experiments Part I: Basic Methods. *Quality Engineering*. 6(1), 115-135.

Montgomery, D.C., & Runger, G.C. (1993b). Gauge Capability and Designed Experiments Part II: Experimental Design Models and Variance Component Estimation. *Quality Engineering*. 6(2), 289-305.

Montgomery, D.C. (2005). *Introduction to Statistical Quality Control*, 5th edition. Hoboken, NJ: John Wiley & Sons, Inc.

Myhrberg, E.V. (2009). *A Practical Field Guide to ISO 9000:2008*. Milwaukee, WI: ASQ Quality Press.

Pollok, M.A., Jefferson, S.G., Kane, J.W., Lomax, K., MacKinnon, G., & Winnard C.B. (1992). Method comparison- a different approach. *Ann. Clin. Biochem.* 29, 556-560.

Rocke, D.M., & Lorenzato, S. (1995). A two-component model for measurement error in analytical chemistry. *Technometrics*. 37(2), 176-184.

Ryan, T.P., & Woodall, W.H. (2005). The most-cited statistical papers. *J. Appl. Stat.* 32, 461-474.

Shainin, P.D. (1992). Managing SPC, A Critical Quality System Element. *The 46th Annual Quality Congress Proceedings, ASQC*, 251-257.

Sherman, J., & Morrison, W.J. (1950). Adjustment of an Inverse matrix Corresponding to a Change in One Element of a Given Matrix. *The Annals of Mathematical Statistics*. 21(1), 124-127.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*. 86(2), 420-428.

Steiner, S.H., & MacKay, R.J. (2005). *Statistical Engineering: An Algorithm for Reducing Variation in Manufacturing Processes*. Milwaukee, WI: ASQ Quality Press.

Steiner, S.H., Stevens, N.T., Browne R.P., & MacKay, R.J. (2011). Planning and Analysis of Measurement Reliability Studies. *Canadian Journal of Statistics*. 39(2), 344-355.

Stevens, N.T., Browne, R.P., Steiner, S.H., & MacKay, R.J. (2010). Augmented Measurement System Assessment. *Journal of Quality Technology*. 42(4), 388-399.

Stevens, N.T., Steiner, S.H., Browne, R.P., & MacKay, R.J. (2013). Gauge R&R studies that incorporate baseline information. *IIE Transactions*. 45, 1166-1175.

Stevens, N.T., Steiner, S.H., & MacKay, R.J. (2014). Assessing agreement between two measurement systems: an alternative to the limits of agreement approach. *Under Revision.*

Stevens, N.T., Steiner, S.H., & MacKay, R.J. (2014). 5 parts is too many. *Quality* Progress. To appear.

Stewart, J. (2003). *Calculus: Early Transcendentals Single Variable*, 5th edition. Belmont, CA: Brooks/Cole – Thomson Learning.

Swallow, W.H., & Monahan, J.F. (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML and ML estimators of variance components. *Technometrics*. 26(1), 47-57.

Traver, R.W. (1995). *Manufacturing Solutions for Consistent Quality and Reliability*. New York, NY: American Management Association.

Tsai, P. (1988). Variable Gauge Repeatability and Reproducibility Study Using the Analysis of Variance Method. *Quality Engineering*. 1(1), 107-115.

Ungerer, J.P.J., Marquart, L., O'Rourke, P.K., Wilgen, U., & Pretorius, C.J. (2012). Concordance, Variance, and Outliers in 4 Contemporary Cardiac Troponin Assays Implications for Harmonization. *Clinical Chemistry*. 58(1), 274-283.

Van den Heuvel, E.R., & Trip, A. (2002). Evaluation of Measurement Systems with a small Number of Operators. *Quality Engineering*. 15(2), 323-331.

Vardememan, S.B., & Van Valkenberg, E.S. (1999). Two-way Random Effects Analyses and Gauge R&R Studies. *Technometrics*. 41(3), 202-210.

Voelkel, J.G., & Siskowski, B.E. (2005). A study of the Bland-Altman plot and its associated methodology. Technical Report, Center for Quality and Applied Statistics, Rochester Institute of Technology.

Walter, S.D., Eliasziw M., & Donner A. (1998). Sample Size and Optimal Design for Reliability Studies. *Statistics in Medicine*. 17, 101-110.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd edition. Boca Raton, FL: CRC Press.

Westgard, J.O., & Hunt, M.R. (1973). Use and interpretation of common statistical tests in method-comparison studies. *Clinical Chemistry*. 19(1), 49-57.

Wheeler, D.J., & Lyday, R.W. (1989). *Evaluating the Measurement Process,* 2nd edition. Knoxville, TN: SPC Press, Inc.