Center for Quality and Applied Statistics
Kate Gleason College of Engineering
Rochester Institute of Technology

Technical Report 2005–3

May 12, 2005

A Study of the Bland-Altman Plot and its Associated Methodology

Joseph G. Voelkel
Center for Quality and Applied Statistics
Rochester Institute of Technology
joseph.voelkel@rit.edu

Bruce E. Siskowski
Reichert, Inc.
bsiskowski@reichert.com

A Study of the Bland-Altman Plot and its Associated Methodology

by

Joseph G. Voelkel and Bruce E. Siskowski

Abstract

Consider the situation in which two measurement devices are compared but no "true" value exists. For this problem, when a set of subjects are measured with both devices, Bland and Altman have strongly criticized several methods of analysis. They advocated a plot of the differences in each subject's readings versus the average of such readings and statistics associated with it. They recommended this plot both for checking assumptions, such as homogeneity of variances, and for assessing the agreement between the devices. They also advocated several other methods associated with such comparisons. We argue that a sound comparison of devices should do more than measure agreement—for example, every comparison of devices should also compare the repeatability of the devices, a measure that cannot be obtained by simply measuring agreement. We show how the Bland-Altman plot and associated methods can itself be misleading through several examples. We propose another method, structural equation modeling, as the mathematical framework for such studies. We critique another method suggested by Bland and Altman for measuring agreement when the devices are measuring the phenomenon on different scales that are thought to be linearly related, and suggest alternative methods. We continue to advocate the use of the Bland-Altman plot, when used cautiously, as one of several ways for checking model adequacy.

1 Introduction

An important consideration in many areas of medicine is the comparison of measurement devices, especially when no "true" value can be measured. As Bland and Altman (1986) noted, "Clinicians often wish to have data on, for example, cardiac stroke volume or blood pressure where direct measurement without adverse effects is difficult or impossible. The true values remain unknown. Instead indirect methods are used, and a new method has to be evaluated by comparison with an established technique rather than with the true quantity. *If the new method agrees sufficiently well with the old, the old may be replaced.* This is very different from calibration, where known quantities are measured by a new method and the result compared with the true value or with measurements made by a highly accurate method. When two methods are compared neither provides an unequivocally correct measurement, *so we try to assess the degree of agreement.* But how?" (Italics ours.) The idea of *agreement* plays a key role in Bland and Altman analyses of device comparisons, a point to which we will return later.

Earlier, Altman and Bland (1983) had addressed this question because of the statistically incorrect methods of analysis they had witnessed in the medical literature, including correlation analysis, regression analysis, and hypothesis tests to compare means. To correct these deficiencies, they proposed a simple graphical technique that has become widely used in the medical literature. A set of subjects is selected, preferably at random from the population of interest, but at least to cover the range of values over which the devices should be compared. Each subject's feature is measured on each of the two devices in such a way that it is reasonable to compare the measurements. A graph is then made of (a) the differences $Y - X$ between the two readings vs (b) the average $(X + Y)/2$ of the two readings. Note that this basic plot is based on exactly one measurement from each device for each subject—we refer to this as the *one-measurement case.* This plot, commonly called the Bland-Altman plot, has gained wide acceptance in the medical literature. Similarly, their recommendations for investigating the data using this plot, as well as related recommendations, may be called the Bland-Altman method.

The plot and the method underlying it does essentially *examine the agreement* between the two techniques—large differences (where "large" should be clinically determined) indicate the two device do not agree well. Altman and Bland (1983) continued their argument by stating "[t]he main emphasis in method comparison studies clearly rests on a direct comparison of the results obtained by the alternative methods. *The question to be answered is whether the methods are comparable to the extent that one might replace the other with sufficient accuracy for the intended purpose of measurement.*" (Italics ours.)

By using this plot instead of a plot of Y vs. X, Altman and Bland (1983) noted that it is "much easier to assess the magnitude of the disagreement (both error and bias)" as well as other features in the data such as outliers and to "see if there is any trend, for example an increase in $Y - X$ for high values." (Bland and Altman used A and B instead of Y and X, but we substitute our terminology in their quotes for consistency in this paper.) If such an increase (or decrease) takes place, they suggested attempting a transformation of the raw data. They continue, "[i]n the absence of a suitable transformation it may be reasonable to describe the differences between the method by regressing $Y - X$ on $(X + Y)/2$." The same point is stated in Bland and Altman (1995): "There may also be a trend in the bias, a tendency for the mean difference to rise or fall with increasing magnitude... In [the] figure ... for example, there is an increase in bias with magnitude, shown by the positive slope of the regression line." They again suggest that a transformation may eliminate this effect. The model they appear to have considered in their article will be shown in

Section 3.

The 1983 article, which included a number of excellent insights on measurement studies from the view of applied statistics, had a strong impact on the ways in which such studies were analyzed. The purposes of our article are to review this plot and the associated methodology proposed by Bland and Altman, to note some difficulties with this methodology, and to recommend alternative analyses.

2 On Comparing Measurement Devices

In our experience, a measurement-device study is usually conducted to address the following questions:

1. Are the devices *identical* in their ability? This is equivalent to the word "interchangeability" used in fields of testing and instrumentation.
2. If they are not identical as is, (a) can they be made so after *calibration*? Or (b) are they measuring the same features but with *different precision*?
3. If they are not identical even after calibration or allowing for different precision, are they measuring the *same features* on the subjects, after calibration and allowing for different precision?
4. If they are not measuring the same features, to what extent are they measuring *different features* on the subjects?

From the italicized section of the Bland-Altman quotes that we have cited, and from the extensive writings of Bland and Altman, their key emphasis is to try to *assess the extent to which the devices agree.* "*The question to be answered is whether the methods are comparable to the extent that one might replace the other with sufficient accuracy for the intended purpose of measurement.*"

Whether two measurements are close enough must be determined from clinical considerations, so suppose that "close" has been defined. If a study is conducted that can address the four questions we pose, then it can address the questions posed by Bland and Altman. However, if the study only addresses the question posed by Bland and Altman, it can not necessarily address the questions we have posed.

Consider the one-measurement case. The assumption that the current device provides consistent readings cannot be ascertained from the one-measurement case, so it is possible that the current device does not agree very well with (is not "close to") itself. If the current measurement device is not particularly precise, it is possible—using the BA method—that the new device is more precise than the old one, but cannot be used to replace it.

3 Mathematical Model

The model we use in this paper is the common and useful *structural model* (e.g. Fuller(1987)). This is a natural model for the comparison of two devices. See, e.g., Mandel (1984), who used this technique on data from the U.S. National Institute of Standards and Technology (NIST). This model is a simple example of structural equations with latent variables, e.g. Bollen (1989).

Consider the case of two measuring devices. Let x_i (y_i) represents the long-term average ("true") value of the measurement for the i^{th} subject when measured on the x-device (y-device) at a some fixed point in time. (The "fixed point in time" is needed in the medical field, because the underlying values often change over time for subjects.) The structural model assumes that these (x_i, y_i) values lie on a straight line—this means that the two devices are measuring the same feature on a given subject except that a possible linear calibration needs to be made. However, we cannot observe the values directly. Instead, we only observe readings X_{ji} and Y_{ji}, $j = 1, \ldots, J$, for each subject. The associated model that is used is the following, where $i = 1, ..., N$.

$$\begin{gathered} y_i = \beta_0 + \beta_1 x_i \qquad (1) \\ x_i \sim \text{ind } N\left(\mu_x,\, \sigma_x^2\right) \\ X_{ji} = x_i + \delta_{ji},\ \delta_{ji} \sim \text{ind } N\left(0,\, \sigma_\delta^2\right) \\ Y_{ji} = y_i + \varepsilon_{ji},\ \varepsilon_{ji} \sim \text{ind } N\left(0,\, \sigma_\varepsilon^2\right) \\ \text{all} \left\{x_i\right\},\ \left\{\delta_{ji}\right\},\ \left\{\varepsilon_{ji}\right\} \text{ are independent} \end{gathered}$$

When $J > 1$, we are considering the case in which the same observer is making repeat measurements over a short period of time. We will not consider more complex, yet important cases here, such as when the study is designed to include measurements by multiple observers (a topic avoided by Bland and Altman as well, e.g. (2003).)

Assuming we have a data set rich enough to estimate all parameters, this model allows us to address questions 1–3 that we had posed. For example, we would say the two devices are *identical* if $\beta_0 = 0, \beta_1 = 1$, and $\sigma_\varepsilon^2 = \sigma_\delta^2$.

The case in which the (x_i, y_i) values do not lie on a straight line is also important. In particular the larger model in which we replace the first line in (1) with

$$y_i = \beta_0 + \beta_1 x_i + \xi_i, \xi_i \sim \text{ind } N\left(0,\, \sigma_\xi^2\right) \qquad (2)$$

should usually be considered as well, as this addresses question 4 above. We will use this model only in Section 8. We will refer to this model as the *extended structural model* in this paper. This model is a general alternative to the first line in (1), so it will have some power in detecting smooth non-linear alternatives, random differences from the linear fit, and occasional outliers.

When we follow the examples typically presented by Bland and Altman, we need to consider the *one-measurement case* in which $J = 1$, and here we write X_i and Y_i instead of X_{ji} and Y_{ji}.

Let "$MVN\left(\boldsymbol{\mu}',\ \boldsymbol{\Sigma}'\right)$" represent a multivariate normal distribution with mean vector $\boldsymbol{\mu}'$ and variance-covariance matrix $\boldsymbol{\Sigma}'$. From the model, it is straightforward to show that (X, Y) has a bivariate normal distribution:

$$\begin{aligned}\begin{bmatrix} X \\ Y \end{bmatrix} &\sim MVN\left(\begin{bmatrix} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_x^2 + \sigma_\delta^2 & \beta_1 \sigma_x^2 \\ \text{Sym} & \beta_1^2 \sigma_x^2 + \sigma_\varepsilon^2 \end{bmatrix}\right) \qquad (3) \\ &\sim MVN\left(\begin{bmatrix} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{bmatrix}, \sigma_x^2 \begin{bmatrix} 1 + R_{\delta x}^2 & \beta_1 \\ \text{Sym} & \beta_1^2 + R_{\delta x}^2 R_{\varepsilon\delta}^2 \end{bmatrix}\right),\end{aligned}$$

where we write $R_{\delta x} = \sigma_\delta / \sigma_x$ and $R_{\varepsilon\delta} = \sigma_\varepsilon / \sigma_\delta$ to denote key ratios of standard deviations. Note that $R_{\delta x} = 0$ corresponds to no measurement error for the x device, and $R_{\varepsilon\delta} = 1$ corresponds to equally precise devices. (We will not consider the $R_{\delta x} = 0$ case, but doing so would require that $R_{\delta x} R_{\varepsilon\delta}$ (the only way in which $R_{\varepsilon\delta}$ appears) be rewritten as $\sigma_\varepsilon / \sigma_x$.)

It follows that the data used in the BA plot is a sample from a bivariate normal distribution:

$$\begin{bmatrix} Y - X \\ (Y + X)/2 \end{bmatrix} \sim MVN\left(\boldsymbol{\mu},\ \boldsymbol{\Sigma}\right),$$

where

$$\begin{aligned}\boldsymbol{\mu} &= \begin{bmatrix} \beta_0 + (\beta_1 - 1)\mu_x \\ (\beta_0 + (\beta_1 + 1)\mu_x)/2 \end{bmatrix} \qquad (4) \\ \boldsymbol{\Sigma} &= \begin{bmatrix} \sigma_x^2 (\beta_1 - 1)^2 + \sigma_\delta^2 + \sigma_\varepsilon^2 & \sigma_x^2\left(\beta_1^2 - 1\right)/2 + \left(\sigma_\varepsilon^2 - \sigma_\delta^2\right)/2 \\ \text{Sym} & \sigma_x^2 (\beta_1 + 1)^2/4 + \left(\sigma_\delta^2 + \sigma_\varepsilon^2\right)/4 \end{bmatrix} \qquad (5) \\ &= \sigma_x^2 \begin{bmatrix} (\beta_1 - 1)^2 + R_{\delta x}^2\left(R_{\varepsilon\delta}^2 + 1\right) & \left(\beta_1^2 - 1 + R_{\delta x}^2\left(R_{\varepsilon\delta}^2 - 1\right)\right)/2 \\ \text{Sym} & \left((\beta_1 + 1)^2 + R_{\delta x}^2\left(R_{\varepsilon\delta}^2 + 1\right)\right)/4 \end{bmatrix} \qquad (6)\end{aligned}$$

Using a sample of data, (X_i, Y_i), $i = 1, 2, ..., N$, we can calculate the sample mean and variance-covariance matrix of the original data, or equivalently of the data corresponding to the BA plot $(Y_i - X_i, (Y_i + X_i)/2)$, $i = 1, 2, ..., N$. Either set of the sample means and variance-covariance matrix provide minimal sufficient statistics for the mathematical model. Mathematically, this minimal sufficiency holds for the normal distribution—from a more general point of view, we could consider any case where we believe the sample mean and variance-covariance matrix provide a good summary of the data.

There is an immediate problem here. It is well known that the *parameter estimation in the model based on (1) is indeterminate (not identifiable) in the one-*

measurement case. The reason is clear—there are six parameters that need to be estimated in the model, but the sufficient statistics only provide five quantities. It is therefore also true that the sufficient statistics associated with (4)-(5), and therefore the Bland-Altman plot of the associated raw data, can not provide enough information to estimate the parameters. We refer to this as the *problem of indeterminacy.* Altman and Bland (e.g., 1983) discuss the usefulness of estimating repeatability (which can be done in the *multiple-measurement case* for which $J > 1$). However, their wording suggests that they view repeatability as somehow distinct from a comparison of methods. It is true that repeatability estimates for a device require only that device. However, a comparison of methods needs to include information on repeatability, not be isolated from it.

The normality assumption on the δ and ε error terms is common, theoretically defensible, and is reasonable for many data sets. The normality assumption on the x term is also reasonable, although situations may arise in which it is questionable. For example, the researcher may wish to screen subjects to try to make an x distribution that is more uniform over the range of values. In such a case, a theoretical difficulty immediately arises. If all the assumptions of (1) hold except that x is non-normal (but is still sampled from a continuous distribution), then the regression of Y on X is no longer linear (Lindley, 1947). This means that non-linear features observed in such a graph may reflect only how the x values were sampled, an awkward situation for a data analyst. We have not investigated to what extent this non-linearity is of practical importance for the problem at hand.

The normality assumption also plays a role in the non-identifiability problem. For example, Reiersøl (1950) showed that β_1 is identifiable if x is non-normal. However, this appears to be mostly of theoretical interest. See Spiegelman (1979) and the references contained therein for details on estimation of this parameter.

4 Bland and Altman Papers: A Review

We shall examine several publications to which one or both of these authors contributed, with an emphasis based on the modeling that underlies their work.

In the Altman and Bland (1983) paper, the model they refer to in their Appendix is restricted to $\beta_1 = 1$ and $\beta_0 = 0$. They essentially employ means, variances, and covariances (or correlations, or looking for trend in the BA plot) in their paper, although they do refer to the normal distribution on occasion. In their analysis, however, they clearly don't restrict themselves to $\beta_0 = 0$, and it appears that the $\beta_1 = 1$ restriction is assumed to be tested by the data through the BA plot. For example, when comparing the BA plot to the Y vs. X plot, they use the phrase "...it is much easier to...see whether there is any trend, for example an increase in $Y - X$ for high values [of $(X + Y)/2$]." However, they do not appear to say what such a trend means, except to note that a test of $\rho_{Y-X,(X+Y)/2} = 0$ is equivalent to a test of $\sigma_X^2 = \sigma_Y^2$. This statement, although true, does not directly aid the user in comparing the two measurement devices because the test "$\sigma_X^2 = \sigma_Y^2$" is equivalent to the test "$\sigma_x^2 + \sigma_\delta^2 = \beta_1^2\sigma_x^2 + \sigma_\varepsilon^2$"–see (3). Unless the analyst is simply willing to assume that $\beta_1 = 1$ or that $\sigma_\delta^2 = \sigma_\varepsilon^2$, the results of this test do not lend themselves to simple interpretation. For example, Maloney and Rastogi (1970), based on earlier work of Pitman (1939) and Morgan (1939), test $H_0 : \sigma_\delta^2 = \sigma_\varepsilon^2$, but assume $\beta_1 = 1$ in doing so.

The sample standard deviation of $Y - X$ is said to be "the estimate of *error*" (italics theirs), although "error" is not defined there, for example in terms of parameters. Based on the structural model, their error term estimates $\sqrt{\sigma_x^2\left(\beta_1 - 1\right)^2 + \sigma_\delta^2 + \sigma_\varepsilon^2}$, which also does not lend itself to simple interpretation unless the user wants to assume artificially that that $\beta_1 = 1$. It is reasonable to state that, if this error term is suitably small (clinically), the methods could likely be used interchangeably in a clinical setting—even here, though, estimates of all the parameters in the structural equation model would provide additional scientific insight. Of course, one doesn't have this information when the study is being designed.

It is also not clear to us Altman and Bland proposed that "[i]f there is an association between the differences and the size of the measurements, then as before, a transformation (of the raw data) may be successfully employed." Unless larger measurements are associated with larger repeatability values or unless the association is nonlinear, a nonlinear transformation appears unwarranted.

The authors noted, after a review of some incorrect methods of analysis that we mentioned earlier in this article, "[n]one of the previously discussed approaches tells us whether the methods can be considered equivalent." Unfortunately, neither does theirs. It is precisely the use of structural equation models (including the extended model as a starting point) along with checks of assumptions, that can answer the equivalence question directly.

We do not agree with Altman and Bland (1983), who rejected structural equation models as being too complex. Good researchers should use the best techniques available that can directly answer the questions we have posed for measurement studies—the fact that many researchers do not should not dissuade us. As those authors correctly note, "[i]t is difficult to produce a method that will be appropriate for all circumstances...clearly the various possible complexities that could arise might require a modified approach, involving additional or even alternate analyses." The authors' claim that "[t]he majority of medical method comparison studies seem to be carried out without the benefit of professional statistical expertise" suggests to us that such expertise should be obtained.

Bland and Altman (2003) cited more recent references in medical statistics journals to again point out the errors of the use of correlation and regression analysis. All the examples, unfortunately, seem to be the one-measurement case, for which no method of analysis can address the questions that should be answered in such a study. The authors provide a nice example, using fetal lung measurements, of the use of the BA plot in checking assumptions. In this example, a non-constant variance was evident. However, their statement, "when X and Y have the same SD, as they should if they are measurements of the same thing, $Y-X$ and $(Y+X)/2$ should not be correlated at all in the absence of a true relationship," is not quite accurate, as (5) indicates. (They stated this more carefully in Bland and Altman (1995).) They noted the design they mentioned in their 1986 paper in which replicate measurements were taken on each subject and "regret that this has not been more widely adopted by researchers."

In that paper, they also suggested an "appropriate use of regression" for the case in which a new method has different units than the old method (but, presumably, is linearly related in some sense). We examine that recommendation later in this article.

5 Example 1: PEFR

The mathematical indeterminacy in the one-measurement case is well known. Bland and Altman often avoid an explicit model in their work so it is not clear if they are aware of this problem. In fact, they state (1983) that the "considerable extra complexity of such analysis will not be justified if a simple comparison is all that is required" and that this is especially true when the "results must be conveyed to and used by non-experts, e.g. clinicians." We disagree with this comment, and show in our Example 4 that summaries of results need not be complex. We are also preparing a paper for publication that we hope will make such modeling easier for practitioners to understand, perform, and summarize.

Using the structural equation model, what effect does this problem of indeterminacy have on making practical decisions when comparing two devices? It can have a very important effect, as we illustrate with an example. Keep in mind that we are only considering the one-measurement case—if multiple measurements are made on each subject this indeterminacy disappears, although the BA plot alone (now of averages for each subject) still does not capture all the information needed.

Our example focuses on parameter estimation, not parameter uncertainty. The results we show hold true whether the data sets are based on $N=10$ or $N=10,000$—parameter uncertainty is not the issue in parameter indeterminacy.

Here is our approach. To get estimates of the parameters, we will perform a sensitivity analysis. We do this by pretending that one of the six quantities is known, and then varying this quantity over a range of values. Each fixed value of this quantity allows us to find estimates of the remaining quantities.

An examination of (4)-(5) reveals that the values of β_0 and μ_x only appear in (4). The μ_x parameter is not useful when comparing two methods. The β_0 parameter, while important (measuring one aspect of bias), is fairly straightforward to correct. For these reasons, we will simplify our analysis by ignoring this information. This leaves us with four unknown parameters for which we have three estimates, as shown in (5) or (6). We will use the latter form of the variance-covariance matrix, so

$(\sigma_x, \beta_1, R_{\delta x}, R_{\varepsilon\delta})$ are the unknown parameters.

One of the strong disagreements we have with the BA method is that *the BA method cannot indicate which device is more precise.* But we believe that this precision question, and *not* a measure of agreement, is often a crucial measure by which a device should be judged. For this reason we select a variety of $R_{\varepsilon\delta}$ values for our sensitivity analysis. The $R_{\varepsilon\delta}$ parameter also has a range of possible values that is independent of the data, making it a natural choice for sensitivity work.

The relationship of $(\sigma_x, \beta_1, R_{\delta x})$ to $(\Sigma, R_{\varepsilon\delta})$ is presented in Appendix A. For a set of data, we can obtain the sample variance-covariance matrix S from the differences and averages. We replace Σ with S in (6) and equate the terms to find estimates of $(\sigma_x, \beta_1, R_{\delta x})$ as a function of $R_{\varepsilon\delta}$ for this S.

The example we use is from Altman (1995), p. 270, and details are provided in Bland and Altman (1986). We selected this example because Bland and Altman have used it repeatedly to explain their methods. The data set is based on $N = 17$ subjects. PEFR (peak expiratory flow rate) measurements were obtained for each subject, both on a Wright meter (X) and a Mini meter (Y). (In fact, several measurements were made on each device for each subject, but Altman used only the first measurement from each device as an example of the BA plot.) We will usually suppress the units, liters/min, for brevity. The sample variance-covariance matrix of this set of data is (variance of $Y - X$) $S_{11} = 1502.7353$, (variance of $(Y + X)/2$) $S_{22} = 12786.1324$, and (covariance) $S_{12} = 366.8015$.

Based on this matrix, we generated the results in Figure 1. We have chosen to display this graph over a wide range of $R_{\varepsilon\delta} = \sigma_\varepsilon/\sigma_\delta$ values to emphasize that the *data themselves provide no evidence of this value.* Without additional information, such as that provided by making repeat measurements on each subject, *any* of the values shown on this graph are equally likely (equal likelihoods) to have given rise to the data observed.

For example, consider three $R_{\varepsilon\delta}$ scenarios:

1. Mini meter much less precise than the Wright meter, by a factor of 10. With $\sigma_\varepsilon/\sigma_\delta = 10$, we find $\hat{\beta}_1 = 0.97$, $\hat{\sigma}_\delta = 3.84$, $\hat{\sigma}_\varepsilon = 38.4$.

2. Mini meter as precise as the Wright meter. With $\sigma_\varepsilon/\sigma_\delta = 1$, we find $\hat{\beta}_1 = 1.03$, $\hat{\sigma}_\delta = \hat{\sigma}_\varepsilon = 27.3$.

3. Mini meter much more precise than the Wright meter, by a factor of 10. With $\sigma_\varepsilon/\sigma_\delta = 0.1$, we find $\hat{\beta}_1 = 1.09$, $\hat{\sigma}_\delta = 37.4$, $\hat{\sigma}_\varepsilon = 3.74$.

Bland (1995, p. 272) noted "the standard deviation of the differences is estimated to be 38.8 litres/min ..." He further adds that (italics ours) "*[o]n the basis of these data we would not conclude that the two devices are comparable* or that *the mini-meter could reliably replace the Wright peak flow meter.*" (He added that the Mini meter had received considerable wear, but this is outside the scope of the data themselves.)

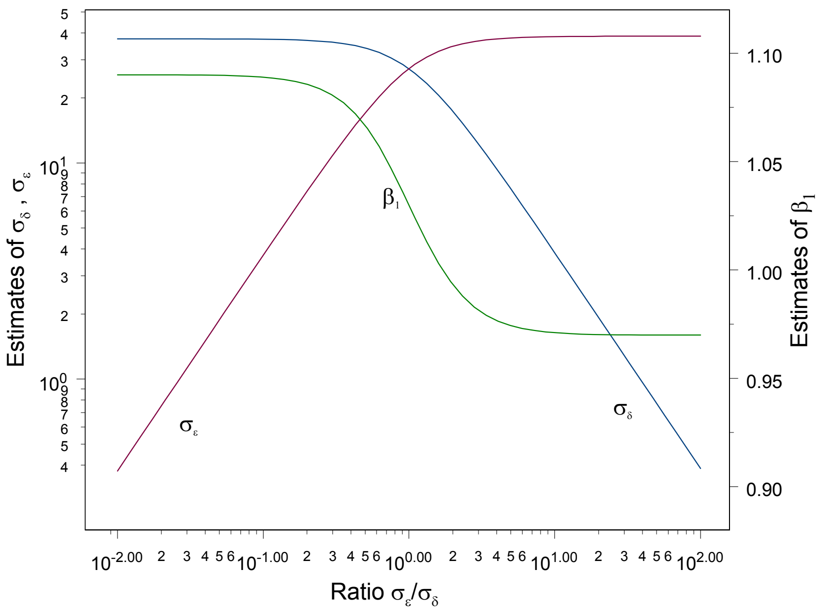

Figure 1: PEFR: Estimates of σ_ε, σ_δ , and β_1 as a function of $R_{\varepsilon\delta} = \sigma_\varepsilon/\sigma_\delta$.

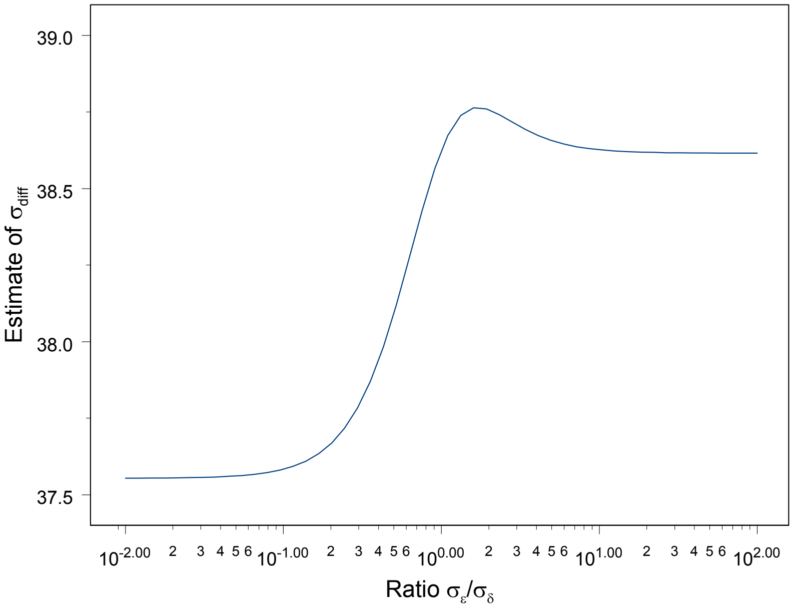

Figure 2: PEFR: Estimate of $\sigma_{diff} = \sqrt{\sigma_\delta^2 + \sigma_\varepsilon^2}$ as a function of $\sigma_\varepsilon/\sigma_\delta$.

We would instead conclude *on the basis of these data* that there is absolutely *no evidence* for whether the two devices are comparable or whether the Mini meter or the Wright meter is a more precise measuring device—such estimates can not be obtained from these data.

In the $R_{\varepsilon\delta}$ limits, the parameter β_1 (or its estimate, for a sample of data) is the regression parameter of Y on X (if $R_{\varepsilon\delta} \to \infty$) or the the inverse of the regression parameter of X on Y (if $R_{\varepsilon\delta} \to 0$). In the case where $R_{\varepsilon\delta} = 1$ $\left(\sigma_\varepsilon^2 = \sigma_\delta^2\right)$, β_1 is the orthogonal (principal component) regression parameter. For example, the slope estimate found by regressing the Wright data on the Mini data is 0.97, which is the limit of $\hat{\beta}_1$ as $R_{\varepsilon\delta} \to \infty$, and the slope estimate found by regressing the Mini data on the Wright data is 0.92, and $1/0.92 = 1.09$, the limit of $\hat{\beta}_1$ as $R_{\varepsilon\delta} \to 0$.

Bland's estimated variance of the differences is $(38.8)^2$, which is simply S_{11}. From (5), this estimate is biased upwards unless $\beta_1 = 1$. However, for this data set, the bias is slight—see Figure 2. The maximum occurs at 38.8 for the value of $R_{\varepsilon\delta}$ that corresponds to $\hat{\beta}_1 = 1$. The relatively stability shown in this example corresponds to comments made in Bland and Altman (1995).

Finally, we examine how $\hat{\sigma}_x$ varies as a function of $R_{\varepsilon\delta}$—see Figure 3. This feature

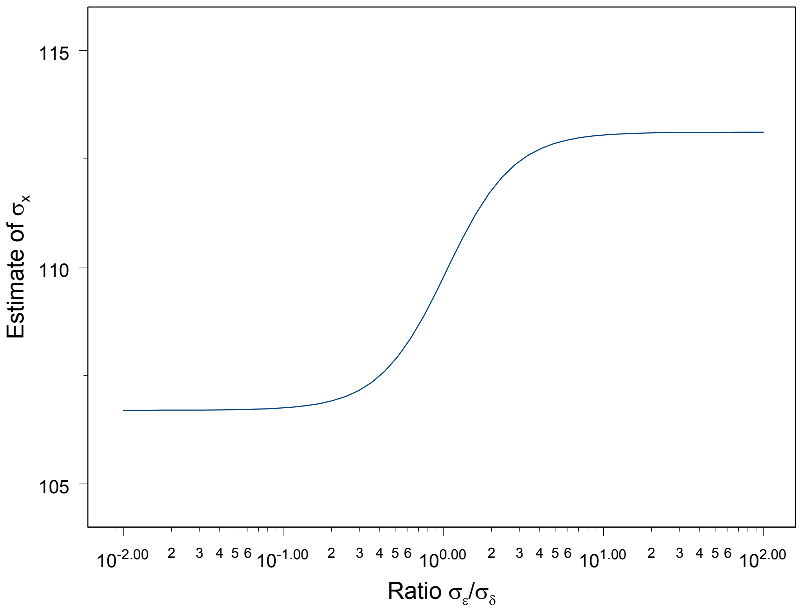

Figure 3: PEFR: Estimate of σ_x as a function of $R_{\varepsilon\delta} = \sigma_\varepsilon/\sigma_\delta$.

of the data exhibits relatively slight variation. Because this is a parameter associated with the population being measured rather than the measurement devices themselves, we will not examine this parameter any further.

To emphasize that *any* of the values associated with these figures are equally likely to have given rise the observed data, we generated 100 pairs of simulated data from each of the three $R_{\varepsilon\delta}$ scenarios listed above. We also estimated the corresponding μ_x and β_0 to center the results appropriately. We then graphed the PEFR data and also overlaid the PEFR data for each scenario. We selected $N = 100$ for each scenario to smooth out some of the random variation.

See Figure 4. The three scenarios have very different values of the underlying parameters, but result in bivariate data that are being sampled from the same distribution. The BA plots, which we do not graph here, are simply a rotation and possible re-scaling of these results.

6 Information Contained in the Bland-Altman Plot

A statistical analysis frequently consists of addressing two broad questions. First, using a tentative mathematical model and *assuming* the model provides a reasonable

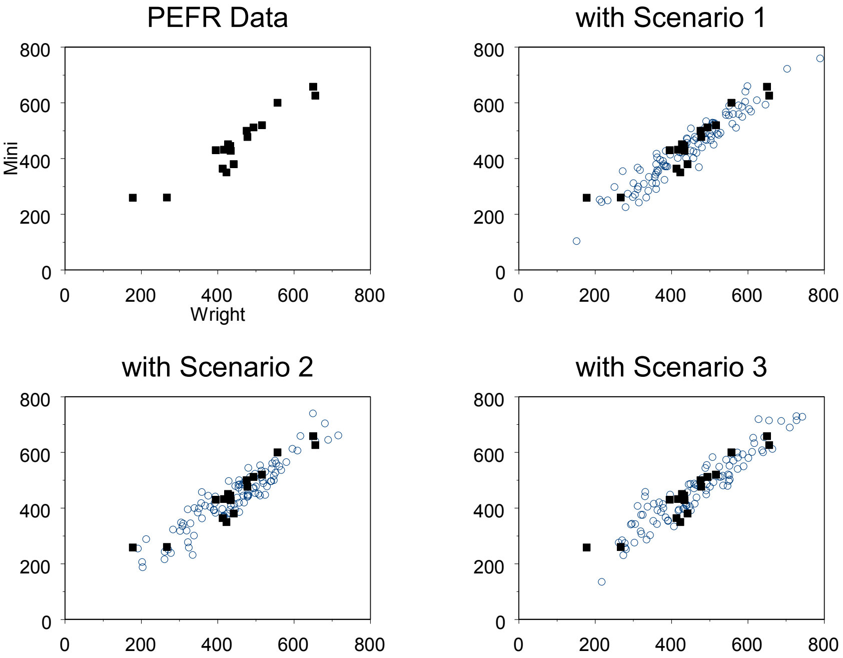

Figure 4: PEFR Data (solid squares) and Simulations (open circles) of Three Scenarios.

representation of the data, what estimates (and uncertainties) are associated with the parameters of the model? Such estimates, by providing a summary of the data *if* the model is reasonable, can then be used to make summaries and help aid in decisions. Second, *does* the tentative model in fact provide a reasonable representation of the data? If so, good; if not, decisions need to be made whether to use the model anyway, to use another model, or to abandon the idea of modeling. A simple example is in linear regression, in which a tentative model is used and the data are fit to the model assuming the model is reasonable; and then a residual analysis is performed to see if the model is reasonable.

This article has emphasized the first aspect of this analysis, for which the BA plot does not provide enough information. However, *the BA plot is valuable in addressing whether the model is reasonable.* For example, the plot can be useful in detecting outliers, in detecting higher variability at higher readings, and in detecting non-linear relationships. Bland and Altman use such checks as part of their method. Note however, that checks of a constant difference ($\beta_1 = 1$) can be misleading, as we illustrate in Example 2.

We have emphasized the one-measurement case, because this appears to be how the BA method is most frequently used. We strongly recommend that no measurement-comparison study ever be conducted unless it includes multiple measurements per subject on each machine (unless this is physically impossible). If so, then all six parameters may be estimated under the structural model. In fact, because the minimal sufficient statistics are of dimension seven, this structural model can be tested against the extended structural model. See Example 4.

In such a case, the BA plot should be based on the average readings for each subject on each machine. Without explicitly incorporating information on the estimates of σ_δ and σ_ε, the plot still suffers the same indeterminacy and its use should be restricted to graphical examination of the assumptions behind the model.

7 Example 2. Blood Pressure

Bland and Altman are aware that problems can exist in interpreting graphs in which measurement variation exists. For example, in their 1995 paper, they criticized the method that has sometimes been advocated of plotting the difference against one of the measurements, say X, often the “gold standard”—see their paper for references. They correctly point out how this can be misleading, and instead recommended the Bland-Altman approach.

Plotting the difference against the X measurements is misleading, because

$$\begin{bmatrix} Y - X \\ X \end{bmatrix} \sim MVN\left(\boldsymbol{\mu}'',\ \boldsymbol{\Sigma}''\right),$$

where (using the structural model again)

$$\begin{aligned}
\boldsymbol{\mu}'' &= \begin{bmatrix} \beta_0 + (\beta_1 - 1)\,\mu_x \\ \mu_x \end{bmatrix} \qquad (7)\\
\boldsymbol{\Sigma}'' &= \sigma_x^2 \begin{bmatrix} (\beta_1 - 1)^2 + R_{\delta x}^2 \left(R_{\varepsilon\delta}^2 + 1\right) & \beta_1 - 1 - R_{\delta x}^2 \\ \text{Sym} & 1 + R_{\delta x}^2 \end{bmatrix}
\end{aligned}$$

and the resulting correlation is

$$\rho_{Y-X,X} = \frac{\rho_{XY}\sigma_Y - \sigma_X}{\sqrt{\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y}} = \frac{\beta_1 - 1 - R_{\delta x}^2}{\sqrt{\left[(\beta_1 - 1)^2 + R_{\delta x}^2\left(R_{\varepsilon\delta}^2 + 1\right)\right]\left(1 + R_{\delta x}^2\right)}}.$$

The middle term, involving the observable (X, Y) data, appears in the Bland and Altman (1995) paper. Even if two measurement devices are identical ($\beta_0 = 0$, $\beta_1 = 1$, $R_{\varepsilon\delta} = 1$), a spurious correlation appears in the data because

$$\rho_{Y-X,X} = \frac{-R_{\delta x}}{\sqrt{2\left(1 + R_{\delta x}^2\right)}}$$

Bland and Altman (1995) also noted the correlation from their plot. It is (first line appears in their paper)

$$\begin{aligned}
\rho_{Y-X,(X+Y)/2} &= \frac{\sigma_Y^2 - \sigma_X^2}{\sqrt{\left(\sigma_X^2 + \sigma_Y^2\right)^2 - 4\rho_{XY}^2\sigma_X^2\sigma_Y^2}} \qquad (8)\\
&= \frac{\beta_1^2 - 1 + R_{\delta x}^2\left(R_{\varepsilon\delta}^2 - 1\right)}{\sqrt{\left[(\beta_1 - 1)^2 + R_{\delta x}^2\left(R_{\varepsilon\delta}^2 + 1\right)\right]\left[(\beta_1 + 1)^2 + R_{\delta x}^2\left(R_{\varepsilon\delta}^2 + 1\right)\right]}} \qquad (9)
\end{aligned}$$

They state "[t]his is zero if the variances are equal, and will be small unless there is a marked difference in the variability between subjects for the two methods." They precede this by stating "[i]f the study includes a wide range of measurements, and unless the two methods of measurement have very poor agreement, we expect σ_X^2 and σ_Y^2 to be similar and ρ to be fairly large, at least 0.7."

The set of data they examined was a random sub-sample of 200 blood pressure readings from a larger study—the sample of size 200 was used to avoid clutter in their graphs. For those data, the estimate of $\rho_{Y-X,(X+Y)/2}$ is 0.17, with a 95% confidence interval that excludes 0. They note, in line with (9) that this correlation may be non-zero if either there is a trend in the relationship (our $\beta_1 > 1$, for example) or if "one method has considerably more measurement error than the other..." and go on to note that this can be estimated only by making repeated measurements.

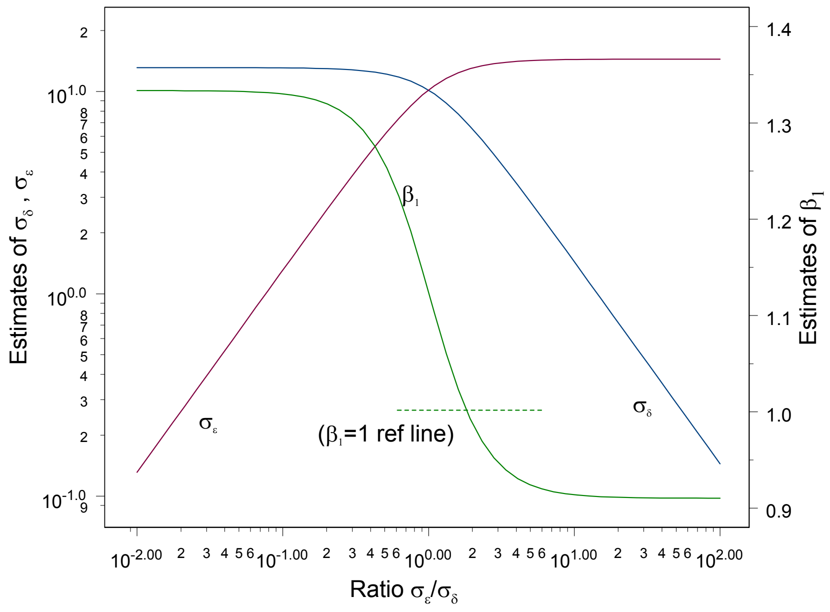

Figure 5:

From the information supplied in that article, a reasonable approximation to the sample variance-covariance matrix of the differences and averages may be retrieved, yielding $S_{11} = 213.16$, $S_{12} = 58.00$, $S_{22} = 546.71$. This was used to generate Figure 5. These figures are generally in line with their comments, but serve again to indicate the indeterminacy in the one-measurement case. For example, the changing bias in the sample could range as high as 33% per unit $\left(\hat{\beta}_1 = 1.33\right)$ to -9% per unit $\left(\hat{\beta}_1 = 0.91\right)$. We believe that the sample correlation of 0.17 and its statistical significance would likely be taken by most users to indicate that $\beta_1 > 1$. However, the entire range of values is once again equally likely to have generated these data, including $\hat{\beta}_1 = 1$, $\hat{\sigma}_\delta = 7.0$, $\hat{\sigma}_\varepsilon = 12.8$.

8 Example 3: PEFR Revisited

We will use this example to outline the approach we recommend in the multiple-measurement case. We plan to explain this approach elsewhere in a style designed for the non-statistician. We again use the PEFR data from Bland and Altman (1986). There were $N = 17$ subjects, each of whom was measured twice on each of the Wright

Wright Meter	Mini Meter
0 : xxxx	0 : xxxxxxxxxxxx
1 : xxxx	1 : x
2 : xxxx	2 : xxxx
3 : xxxx	3 : x
4 : xx	4 : xx
5 : xxxx	5 : xxxxx
6 : xxx	6 : x
7 : xxxxx	7 : xxx
8 : xxx	8 : xxxx
9 : x	9 : x

Figure 6: PEFR: Stem-and-Leaf Plots of Last Digits of Wright and Mini Meters.

and Mini meters. As those authors noted, those data were collected to illustrate a statistical method, and we will use it with this in mind as well—the small sample size would normally be too small for decision making.

Like Bland and Altman, we begin by a graphical examination of the data. However, those authors used only the first measurement of each method to illustrate the comparison of methods, and used the second measurement only for studying repeatability. Although it is clear to us why they may have done so, it is an inefficient use of data and, we believe, does not set a good example for data analysis. We will examine all the data simultaneously.

Before any means or differences are calculated, it is important to look at the raw data themselves. This can provide unexpected insights. Here, we discovered a difference between the two measuring devices that, to our knowledge, has not been noted by Bland and Altman in their publications. The existence of round-off error of a measuring device can often be considered by examining only the last digit—e.g., 516 is reduced to a 6. See Figure 6.

The actual readings had a range of over 400 for each meter, so the last digit is expected to be randomly distributed. This is observed for the Wright Meter. However, the Mini Meter readings suffer from rounding—this is most obvious by the large number of 0's, but also reveals itself in the peaks at what we shall call 2.5, 5, and 7.5. The smaller three peaks are of little consequence here, but the high peak at 0 suggests that some readings may have been misread by as much as 5 units. We do not examine this issue further in the analysis.

Our next plots continue to examine the features of each machine. For each machine, we recommend a plot of standard deviation vs mean for each subject. However, for the two-measurement case, a plot of difference vs mean can be used instead—Bland and Altman (1986) made these same, natural, suggestions. There are two main purposes for such a plot: to see if the standard deviation increases with the mean, and if so perform an alternate analysis (such as a log transformation of the raw data);

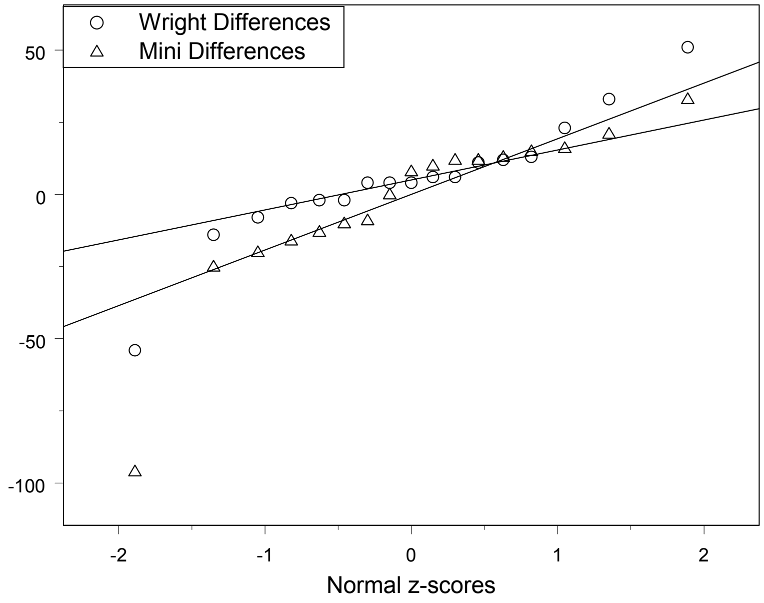

Figure 7: PEFR: Normal Quantile-Quantile Plot of Wright and Mini Differences.

and to look for outliers. Such graphs, not shown here, reveal a very large outlier for the Mini meter, which Bland and Altman (1986) noted. However, there appear to be problems with the Wright meter as well, which the authors did not note. Because the standard deviation does not increase with the mean, we show both sets of differences with normal quantile-quantile plots. See Figure 7.

The outlying value of -96 stands out for the Mini Meter, but there appears to be at least one outlier for the Wright data as well. Using the method of Grubbs (1969) and an algorithm supplied by NIST (http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm) based on $G = \max\left(D_i - \bar{D}\right)/s_D$, where the $\{D_i, i = 1, ..., 17\}$ are the differences using one of the devices and s_D is the sample standard deviation of the the $\{D_i\}$, we find $G = 3.23$ for Mini, which is significant at $P = 0.0006$. This is the only such outlier in the Mini data, and we will set it aside (subject #7). However, the Wright data has $G = 2.71$ and $P = 0.03$. If this point is set aside and we re-run the test, we have $G = 2.64$ and $P = 0.04$, so there is some evidence of two outliers. However, this evidence of outliers had $P > 0.01$, and so we decided to keep these points. This is in line with Grubbs' (1969), who wrote "it is generally recommended that a low significance level, such as 1%, be used...". (The NIST site uses G above, but the Grubbs statistic used only the "one-sided" version of this test.

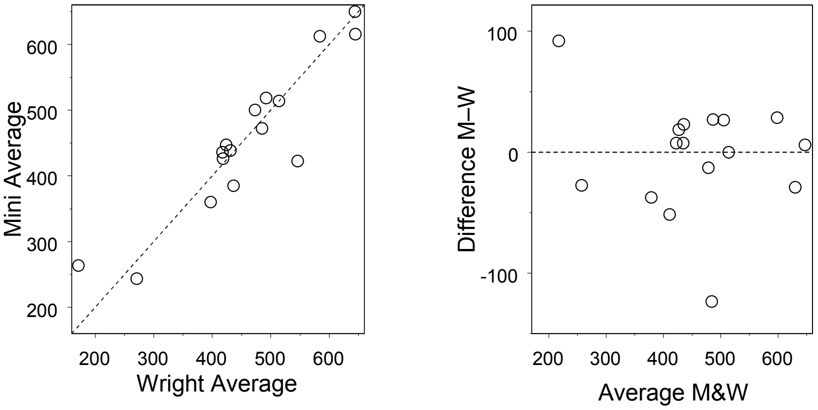

Figure 8: PEFR: Average Plot, and Difference–Average Plot.

Model #	Model
4	Same as 3 except equation (2) is used
3	See equation (1): $\sigma_\xi^2 = 0$
2	Same as 3 but $\sigma_\varepsilon^2 = \sigma_\delta^2$: same measurement variance
$2'$	Same as 3 but $\beta_0 = 0,\ \beta_1 = 1$: same readings on average
1	Same as 3 but $\sigma_\varepsilon^2 = \sigma_\delta^2$ and $\beta_0 = 0,\ \beta_1 = 1$: identical devices

Table 1: Hierarchy of Models

For this reason, the α levels and P-values cited there need to be multiplied by 2, and we have done so here.)

Next, we graphically examine how the two devices compare, using the means for each subject and device. The main reasons for this plot are twofold: to look for outliers (a different type than before) and to look for a non-linear relationship. (Increases in variability with the means would normally be detected in the earlier plots.) See Figure 8, where the outlier from subject 7 has been removed.

Neither plot suggests any smooth departures from linearity, but the plot on the right suggests that two additional subjects (numbers 2 and 15) were measured differently on the two machines.

The standard statistical approach to formally examine such data is based on a series of structural models. The models we examine are shown in Table 1 and correspond to the numbered listing at the beginning of Section 2.

The models form a hierarchy with the exception of 2 and $2'$. To examine the models, we start with the largest and search for the smallest model consistent with the data. We use maximum likelihood to fit the models. The deviance (-2log-likelihood) for Model 4 is 519.30, while for Model 3 it is 532.64. The difference of 13.34 is highly significant ($P = 0.0003$) when using the standard χ^2_1 reference distribution, although the sample sizes are too small here to justify using this reference distribution with high precision. The outliers in the right-hand plot of Figure 8 and the dramatic increase in the estimate of measurement standard deviations in moving from Model 4 ($\hat{\sigma}_\xi = 44, \hat{\sigma}_\delta = 16, \hat{\sigma}_\varepsilon = 11$) to Model 3 ($\hat{\sigma}_\xi = 0, \hat{\sigma}_\delta = 25, \hat{\sigma}_\varepsilon = 19$) also indicate that Model 3 does not fit these data well.

One can only theorize why these two subjects were measured so differently for these data. One would normally want to re-measure them to see if there are consistent differences in the devices or if these were anomalies. For purposes of this example, we will assume the latter and set these two subjects aside as well. The deviances for Model 4 (442.99) and Model 3 (447.50) differ by 4.51, which yields $P = 0.033$ with the χ^2_1 reference distribution. Because likelihood methods generally reject hypotheses at a higher than nominal rate for small sample sizes, there is not strong evidence to reject Model 3 for Model 4. The increase in the estimate of measurement standard deviations in moving from Model 4 ($\hat{\sigma}_\xi = 21, \hat{\sigma}_\delta = 17, \hat{\sigma}_\varepsilon = 12$) to Model 3 ($\hat{\sigma}_\xi = 0, \hat{\sigma}_\delta = 22, \hat{\sigma}_\varepsilon = 14$) was also less, and there was also no evidence of other outliers in the right-hand plot of Figure 8.

The deviances for Model $2'$ (449.6), Model 2 (448.52) , and Model 1 (451.17, for which $\hat{\sigma}_\delta = \hat{\sigma}_\varepsilon = 19$) show that Model 1 is the simplest model consistent with the data for the $N = 14$ subjects. Based on these results, it is reasonable to summarize the analysis as follows:

1. The Mini meter measurement on one subject was a clear outlier. (Subject's data removed from analysis.)

2. Measurements for two subjects were inconsistent across devices. (May indicate fundamental differences in devices for certain subjects or at certain times. In this analysis, we set these subjects' data aside.)

3. For the remaining $N = 14$ subjects, the data suggest the devices may function identically, with a estimated measurement standard deviation of 19 liters/min. This means that the repeatability is estimated to be 54 liters/min.

The last number is based on the British Standard Institute's (BSI's) definition of repeatability (1979), the value in which 95% of differences in reading will lie.

Our analysis treats the two devices in a symmetric way and finds no evidence of a difference. This contradicts the analysis of Bland (1995, p. 272) who performs an analysis only on the first reading, and estimates a repeatability value of 78 liters/min. He states that "on the basis of these data we would not conclude that the two methods are comparable or that the mini-meter could reliably replace the Wright peak flow meter." In fact, after excluding the subject with the obvious outlier from the analysis, the estimates of within-device measurement error based solely on the duplicate set of readings were 16 for the Wright meter and 12 for the Mini meter, less (but not statistically so) than the Wright meter.

9 Example 4. Use of Regression

Bland and Altman (2003) stated that an appropriate use of regression exists for the case in which a new method has different units from the old (but are, presumably, linearly related). A regression of Y (old) on X (new), when sampling from a bivariate normal distribution, estimates the regression line $E[Y|X] = E[Y] + \rho_{XY}(\sigma_Y/\sigma_X)(X - E[X])$, with $Var(Y|X) = \sigma_Y^2(1 - \rho_{XY}^2)$. Using the structural equation model, this becomes

$$\begin{aligned}
\rho_{XY} &= \frac{\beta_1\sigma_x^2}{\sqrt{(\sigma_x^2+\sigma_\delta^2)(\beta_1^2\sigma_x^2+\sigma_\varepsilon^2)}} \qquad (10)\\
&= \frac{\beta_1}{\sqrt{(1+R_{\delta x}^2)(\beta_1^2+R_{\delta x}^2R_{\varepsilon\delta}^2)}}\\
E[Y|X] &= \beta_0+\beta_1\mu_x+\frac{\beta_1\sigma_x^2}{\sigma_x^2+\sigma_\delta^2}(X-\mu_x)\\
&= \beta_0+\beta_1\mu_x+\frac{\beta_1}{1+R_{\delta x}^2}(X-\mu_x)\\
Var(Y|X) &= \sigma_Y^2\left(1-\frac{\beta_1^2\sigma_x^4}{(\sigma_x^2+\sigma_\delta^2)(\beta_1^2\sigma_x^2+\sigma_\varepsilon^2)}\right)\\
&= \sigma_Y^2\left(1-\frac{\beta_1^2}{(1+R_{\delta x}^2)(\beta_1^2+R_{\delta x}^2R_{\varepsilon\delta}^2)}\right)
\end{aligned}$$

As a special case, first consider when the two devices are in fact equivalent, which is the same as comparing the device to itself. In this case

$$\begin{aligned}
E[Y|X] &= \mu_x+\frac{\sigma_x^2}{\sigma_x^2+\sigma_\delta^2}(X-\mu_x)\\
&= \mu_x+\frac{1}{1+R_{\delta x}^2}(X-\mu_x).
\end{aligned}$$

Unless there is no measurement error in X, this produces a slope less than 1, as is well known, and thus gives the regression to the mean phenomenon that Bland and Altman have repeatedly criticized, including in the (2003) paper itself. (It is true that this estimate of the next measurement made on the same device, for example, is a better estimate than the reasonable (but naïve) estimate X itself, under the random-x model that is used here. See Fuller (1987, p. 75) for additional comments. However, we believe most researchers would be more interested in what we shall call the unadjusted corrected value, which here would be X , and which in general would be $\hat{\beta}_0+\hat{\beta}_1 X$, where the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ would be obtained through the structural-equation model approach, based on the data from the multiple-measurement case.)

In the general case, it is clear from (10) that the unadjusted corrected value for Y will be correct only when $X = \mu_X$ (in practice, if the observed X is very close to the observed mean of the X's), but otherwise suffers from the same regression-to-the-mean phenomenon. As before, in the one-measurement-case, the estimate of β_1 cannot be obtained.

Bland and Altman (2003) also recommend that 95% prediction-interval limits be used as a substitute for their standard 95% limits of agreement. It is difficult for us to understand why they would make such a recommendation, because the two limits have little in common, as we now show. The 95% prediction-interval limit is meant to capture one future value with a 95% chance, where the 95% takes into account both the estimate of variability of a future value Y, given X, and the estimated variability of the regression coefficients involved. These estimated limits are a strongly affected by the sample size of the data used in the regression. The 95% limits of agreement take into account the the estimate of the variability of $Y - X$, but do not take into account any estimated variability of the estimates. Such estimated limits are not very dependent on the sample size. Suppose we have N pairs of values, and call the i^{th} pair (X_i, Y_i), with resulting least square equation of $\hat{Y}_X = \hat{\gamma}_0 + \hat{\gamma}_1 \left(X - \bar{X}\right)$, where $\bar{X}$ is the average of the X_i's in the sample, and $(\hat{\gamma}_0, \hat{\gamma}_1)$ are the usual least square estimates. Let the estimated residual standard be denoted by $s_{Y|X}$ Then the 95% P.I. at a new value X is

$$\begin{aligned} & \hat{Y}_X \pm t_{N-2,0.975} \sqrt{\widehat{Var\left(\hat{\gamma}_0\right)} + \left(X - \bar{X}\right)^2 \widehat{Var\left(\hat{\gamma}_1\right)} + s_{Y|X}^2} \\ = \quad & \hat{Y}_X \pm t_{N-2,0.975} s_{Y|X} \sqrt{\frac{1}{N} + \left(X - \bar{X}\right)^2 \frac{1}{(N-1)\, s_X^2} + 1}, \end{aligned}$$

where $s_X^2 = \sum \left(X_i - \bar{X}\right)^2 / (N-1)$ is the sample variance of the X_i's, and the dependence of the width on the sample size is clear. Now, these actual values are based on a particular sample. However, we can examine the properties of the width of this interval by first replacing these estimates with the parameters they are estimating. This produces a width of

$$2t_{N-2,0.975}\sigma_x\sqrt{\beta_1^2+R_{\delta x}^2R_{\varepsilon\delta}^2}\sqrt{1-\frac{\beta_1^2}{\left(1+R_{\delta x}^2\right)\left(\beta_1^2+R_{\delta x}^2R_{\varepsilon\delta}^2\right)}}\times$$
$$\sqrt{\frac{1}{N}+\left(X-\mu_x\right)^2\frac{1}{\left(N-1\right)\sigma_x^2\left(1+R_{\delta x}^2\right)}+1},$$

a very complex function of the parameters. Consider only the simple case in which $\beta_1 = 1$, as Bland and Altman did. The width reduces to

$$2t_{N-2,0.975}\sigma_x\sqrt{1+R_{\delta x}^2R_{\varepsilon\delta}^2}\sqrt{1-\frac{1}{\left(1+R_{\delta x}^2\right)\left(1+R_{\delta x}^2R_{\varepsilon\delta}^2\right)}}\times$$
$$\sqrt{\frac{1}{N}+\left(X-\mu_x\right)^2\frac{1}{\left(N-1\right)\sigma_x^2\left(1+R_{\delta x}^2\right)}+1}.$$

For a sample of size N, the largest value of X in the data, the N^{th} order statistic, can be approximated by $\mu_x + z_{N/(N+1)}\sigma_X$. From this, the smallest and largest estimated widths are

$$2t_{N-2,0.975}\sigma_x\sqrt{1+R_{\delta x}^2R_{\varepsilon\delta}^2}\sqrt{1-\frac{1}{\left(1+R_{\delta x}^2\right)\left(1+R_{\delta x}^2R_{\varepsilon\delta}^2\right)}}\times \tag{11}$$
$$\sqrt{\frac{1}{N}+z_{N/(N+1)}^2\frac{I_{\max}}{\left(N-1\right)}+1},$$

where $I_{\max} = 1$ for the maximum width and $I_{\max} = 0$ for the minimum width. The 95% width for limits of agreement when $\beta_1 = 1$, on the other hand, is simply an estimate of

$$2z_{0.975}\sqrt{\sigma_\delta^2+\sigma_\varepsilon^2}.$$

We now consider three special cases of (11). First assume that $\sigma_\delta = 0$, but that σ_x and σ_ε are fixed. Then (11) reduces to

$$2t_{N-2,0.975}\sigma_\varepsilon\sqrt{\frac{1}{N}+z_{N/(N+1)}^2\frac{I_{\max}}{\left(N-1\right)}+1},$$

a standard regression result. Second, assume that $\sigma_\varepsilon = 0$, but that σ_x and σ_δ are fixed. Then (11) reduces to

$$2t_{N-2,0.975}\sigma_\delta\sqrt{\frac{1}{1+R^2_{\delta x}}}\times \tag{12}$$
$$\sqrt{\frac{1}{N}+z^2_{N/(N+1)}\frac{I_{\max}}{(N-1)}+1}.$$

Third, assume that $\sigma_\varepsilon = \sigma_\delta$. We get

$$2t_{N-2,0.975}\sigma_\varepsilon\sqrt{1+\frac{1}{1+R^2_{\delta x}}}\times \tag{13}$$
$$\sqrt{\frac{1}{N}+z^2_{N/(N+1)}\frac{I_{\max}}{(N-1)}+1}.$$

To examine these, consider a variety of ways in which graphs of Y vs. X might appear. The unitless measure of this association is ρ_{XY}, and we consider ρ_{XY} values of $(0.5, 0.7, 0.9, 0.95)$ for our comparisons. We select $(\sigma_\varepsilon,\ \sigma_\delta)$ values for three cases to be $(10,\ 0)$, $(0,\ 10)$, and $(10,\ 10)$, respectively, to create the graphs—this choice of scale has no effect on the relative comparisons. The ρ_{XY} values for each scenario let us determine the corresponding σ_x value based on (10) with $\beta_1 = 1$.

Rather than use these unusual intervals, the user could consider $\hat{Y} \pm z_{0.975}s_{Y|X}$. This leads to the following estimated interval widths, which we call *regression widths.* For the $\sigma_\delta = 0$ case, the width estimates $2z_{0.975}\sigma_\varepsilon$, which is identical to the agreement width, and independent of ρ_{XY}. For the $\sigma_\varepsilon = 0$ case, the width estimate $2z_{0.975}\sigma_\delta\sqrt{1/\left(1+R^2_{\delta x}\right)} = 2z_{0.975}\sigma_\delta\rho_{XY}$, which will be slightly smaller than the $2z_{0.975}\sigma_\delta$ agreement width, and especially so for smaller values of ρ_{XY}. (The P.I. approach makes this slightly larger for smaller samples sizes, but not because of any good theoretical reasons.) For the $\sigma_\varepsilon = \sigma_\delta$ case, the width estimate $2\sqrt{2}z_{0.975}\sigma_\varepsilon\sqrt{1-R^2_{\delta x}/2\left(1+R^2_{\delta x}\right)} = 2\sqrt{2}z_{0.975}\sigma_\varepsilon\sqrt{\left(1+\rho^2_{XY}\right)/2\rho^2_{XY}}$, which is again slightly less than the $2\sqrt{2}z_{0.975}\sigma_\varepsilon$ difference width.

In fact, we can do better than this—from the data, we can estimate ρ_{XY} and this in turn lets us find an upper bound for the width. Because $\rho_{XY} < \sqrt{\left(1+\rho^2_{XY}\right)/2\rho^2_{XY}}$ for $0 < \rho_{XY} < 1$, this upper bound is $\hat{Y} \pm z_{0.975}\left(s_{Y|X}/\rho_{XY}\right)$. (Note: it appears that this upper bound would always based on the $\sigma_\varepsilon = 0$ case, but but we have not proven this.) Call this width the *regression max width.* From this, we compare the BA difference widths (which Bland and Altman are trying to estimate) to the BA prediction interval widths (minimum and maximum), the regression width, and the regression max width. See Figures 9 to 11. From these, it is clear that the regression

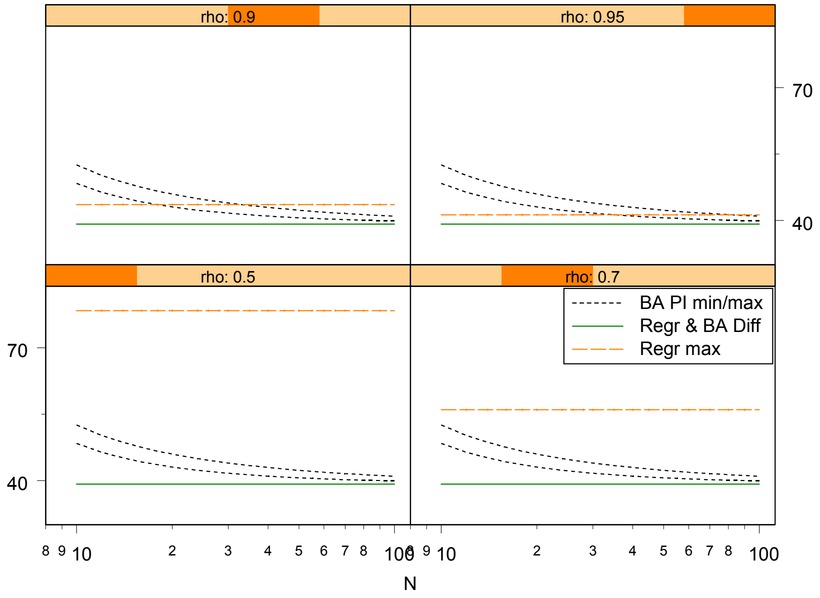

Figure 9: Interval Widths, Four Methods, vs ρ and N, when $\sigma_\delta = 0$.

and regression max widths estimate reasonable ranges for the BA difference widths, while the BA prediction interval widths do not.

10 Recommendations

The use of the BA methodology is certainly to be preferred to the other methods that Bland and Altman have criticized. However, we are very concerned on the emphasis they place on the one-measurement cases in their examples. This, we believe, inadvertently suggests to researchers that such a methodology is acceptable. In Altman's (1995) generally excellent textbook, for example, he stated "For simplicity, I shall use only one measurement by each method here. We could make use of the duplicate data by using the average of each pair first, but this introduces an extra stage in the calculation. Bland and Altman (1986) give details."

It is true that Bland and Altman (1986) give formulas for the multiple-measurement case, near the end of that article. Ironically though, even when they have two readings available per subject/device, they still only use first to illustrate the BA plot—this may be simpler to explain, perhaps, but it is not an appropriate way to analyze these data. We believe in simple analysis, but not at the expense of properly using the data

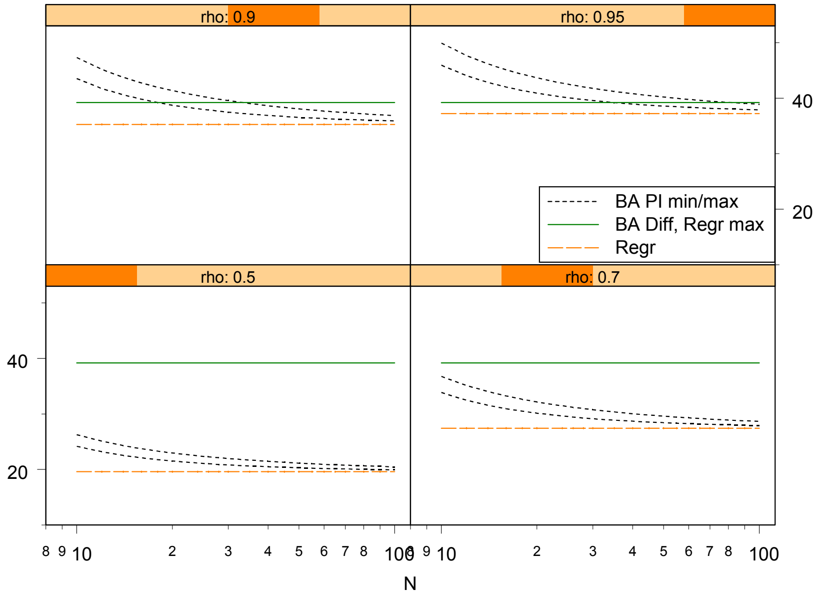

Figure 10: Interval Widths, Four Methods, vs ρ and N, when $\sigma_\varepsilon = 0$.

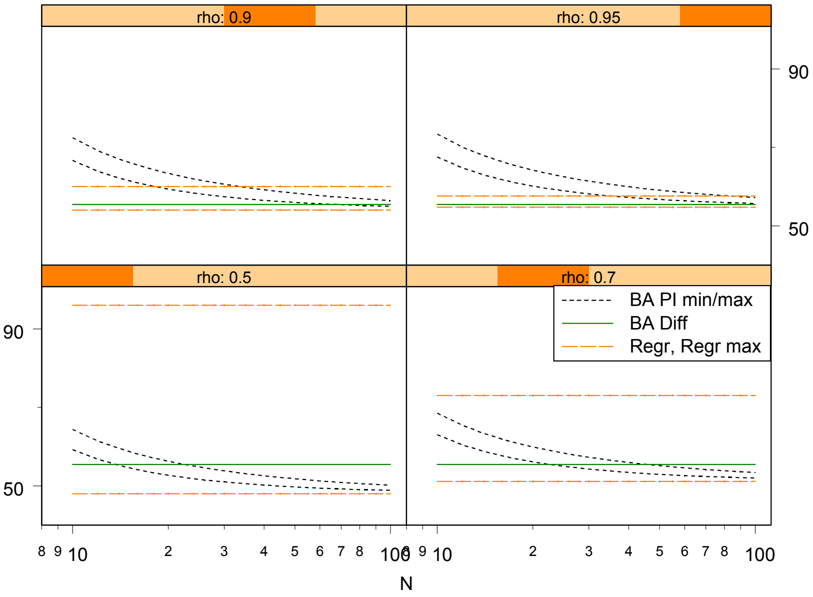

Figure 11: Interval Widths, Four Methods, vs ρ and N, when $\sigma_{\varepsilon} = \sigma_{\delta}$.

for modeling.

Our recommendations are simple.

1. If physically possible, repeated measurements on each device for each subject should be required.
2. The extended structural equation model should be used for analysis. Plots, including sample standard deviations versus means (or, in the case of two-measurement studies, second measurement versus first measurement) for each device, $\bar{Y}$ vs. $\bar{X}$, and the BA plot, should be used along with formal tests to check for model adequacy—this includes outliers, nonlinear relationships, and non-constancy of variance.
3. If the model is adequate, provide estimates and confidence intervals of the parameters and other related quantities. The latter may include estimates that are related to local standards, such as those of BSI.

We will illustrate the use of extended structural equation models, including model checking, in a separate article. It is our intent to do this is such a way to allow these methods to become easily accessible to researchers.

Appendix A. $(\sigma_x, \beta_1, R_{\delta x}) = f(\Sigma, R_{\varepsilon\delta})$

Based on (6), we want to find $(\sigma_x, \beta_1, R_{\delta x}) = f(\Sigma, R_{\varepsilon\delta})$. We will show it here in terms of population parameters. For a sample of data, simply replace Σ with its estimate S.

Define the ratios $R_1 = (2\Sigma_{12} + \Sigma_{11}) / (4\Sigma_{22} - \Sigma_{11})$, $R_2 = (2\Sigma_{12} - \Sigma_{11}) / (4\Sigma_{22} - \Sigma_{11})$, where Σ_{ij} is the $(i, j)^{th}$ element of Σ. Next, define $b = R^2_{\varepsilon\delta} - 1 - 2R_2 R^2_{\varepsilon\delta} - 2R_1$. Then, a lengthy but straightforward calculation shows that

$$\begin{aligned} \beta_1 &= \left(-b + \sqrt{b^2 + 4R^2_{\varepsilon\delta}}\right)/2 \\ R^2_{\delta x} &= -\beta_1 (\beta_1 - 1 - 2R_1) / R^2_{\varepsilon\delta} \\ \sigma^2_x &= (4\Sigma_{22} - \Sigma_{11})/(4\beta_1) \end{aligned}$$

From this, we can calculate related quantities of interest. For example, $\sigma^2_\delta = \sigma^2_x R^2_{\delta x}$ and $\sigma^2_\varepsilon = \sigma^2_x R^2_{\delta x} R^2_{\varepsilon\delta}$.

With the information on β_1, we can also calculate $\mu_x = 2(\mu_2 - \mu_1)/2$ and then $\beta_0 = \mu_1 - (\beta_1 - 1)\mu_x$.

References

Altman, D. G. and Bland, J. M. (1983), “Measurement in medicine: the analysis of method comparison studies,” *The Statistician*, 32, 307–317.

Bland, J. M. (1995), *An Introduction to Medical Statistics*, Oxford Medical Publications, New York.

Bland, J. M., and Altman, D. G. (1986), "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet* i, 307–310.

Bland, J. M., and Altman, D. G. (1995), "Comparing methods of measurement: why plotting difference against standard method is misleading," *Lancet*, 346, 1085–1087.

Bland J. M., and Altman D. G. (1999) "Measuring agreement in method comparison studies." *Statistical Methods in Medical Research* 8, 135–160.

Bland J. M., and Altman D. G. (2003), "Applying the right statistics: analysis of measurement studies," *Ultrasound Obstet Gynecol*, 22, 85–93.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, Wiley, New York.

British Standards Institution (1979), Precision of Test Methods, Part 1: Guide for the Determination of Repeatability and Reproducibility for a Standard Test Method. BS 5497, Part 1. London.

Fuller, W. (1987), *Measurement Error Models,* Wiley, New York.

Grubbs, F. (1969), "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, 11, 1–21.

Lindley, D. V. (1947), "Regression Lines and the Linear Functional Relationship," *Journal of the Royal Statistical Society Supplement*, 9, 218–244.

Mandel, J. (1984), "Fitting Straight Lines When Both Variables are Subject to Error," *Journal of Quality Technology*, 16, 1–14.

Maloney, C. J., and Rastogi, S. C. (1970), "Significance Test for Grubbs's Estimators," *Biometrics* 26, 671–676.

Morgan, W. A. (1939), "A test of significance of the differences between two variances in a sample from a normal bivariate population," *Biometrika*, 13–19.

Pitman, E. J. G. (1939), "A note on normal correlation," *Biometrika*, 31, 9–12.

Reiersøl, O. (1950), "Identifiability of a Linear Relation between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389.

Spiegelman, C. (1979), "On Estimating the Slope of a Straight Line when Both Variables are Subject to Error," *The Annals of Statistics*, 7, 201–206.