Objective Statistics For The Measurement Of Agreement

Elizabeth Perkin Moen
Wayne State University,
OBJECTIVE STATISTICS FOR THE MEASUREMENT OF AGREEMENT

by

ELIZABETH PERKIN MOEN

DISSERTATION

Submitted to the Graduate School

of Wayne State University

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2016

MAJOR: EVALUATION AND RESEARCH

Approved By:
DEDICATION

This dissertation is dedicated to the memories of Dr. William Fendley and Mr. Thomas J. Wilhelm. Their teachings and support proved invaluable in my education and career, and my gratitude remains profound.
ACKNOWLEDGEMENTS

The encouragement, support, and teaching of Dr. Shlomo S. Sawilowsky made this journey possible. My experience with the doctorate began and has ended with Professor Shlomo; I am humbled and proud to have him as my major adviser. I wish to also acknowledge both the late Drs. Gail Fahoome and Donald Marcotte for establishing a firm foundation in and appreciation for my work and education.

Many thanks to my dissertation committee members (Drs. Timothy Butler, Michael Owens, Michael Lance, and Sarah A. Rose) for their support and input into this process. I am indebted as well to my family at home and my work family, led by Dean Laurie M. Lauzon Clabo in the College of Nursing at Wayne State for their unwavering enthusiasm and encouragement.
# TABLE OF CONTENTS

DEDICATION.................................................................................................................. ii

ACKNOWLEDGEMENTS................................................................................................. iii

LIST OF TABLES ............................................................................................................. vii

LIST OF FIGURES .......................................................................................................... viii

CHAPTER 1 INTRODUCTION ......................................................................................... 1

Agreement...................................................................................................................... 1

Bland-Altman Method ................................................................................................. 2

Potential Solutions ...................................................................................................... 3

Statement of the problem ............................................................................................ 4

Importance of the Study .............................................................................................. 4

Key terms: .................................................................................................................... 5

CHAPTER 2 LITERATURE REVIEW ............................................................................. 6

The Correlation Coefficient ......................................................................................... 8

t-Test .......................................................................................................................... 10

Regression ................................................................................................................. 11

Bland-Altman Method ............................................................................................... 12

Kolmogorov-Smirnov (KS test) (test of general differences) ................................. 16
APPENDIX A: CORRELATIONS AND SAMPLE SIZES FOR ALGORITHM .....50

APPENDIX B: BLAND-ALTMAN PLOTS.................................................................53

REFERENCES .................................................................................................69

ABSTRACT .....................................................................................................76

AUTOBIOGRAPHICAL STATEMENT ...............................................................77
LIST OF TABLES

Table 1. Sample Blood Pressure Readings ................................................................. 17
Table 2. Blood Sugar Reading CDFs........................................................................ 17
Table 3. Letter Sequence and CDF.......................................................................... 18
Table 4. Order of Experiments.................................................................................. 31
Table 5. Fit Index Results ....................................................................................... 31
Table 6. Correlations Starting at .9 ........................................................................ 33
Table 7. Bland-Altman Computations..................................................................... 41
Table 8. Fit Index Cut Scores ................................................................................... 46
Table 9. Correlation Results and Sample Sizes ....................................................... 50
Figure 22. Sample Size 30, Correlations .55 and .7 ................................................................. 56
Figure 23. Sample Size 30, Correlations .6 and .7 ................................................................. 57
Figure 24. Sample Size 30, Correlations .65 and .7 ................................................................. 57
Figure 25. Sample Size 30, Correlations .7 and .7 ................................................................. 58
Figure 26. Sample Size 40, Correlations .5 and .7 ................................................................. 58
Figure 27. Sample Size 40, Correlations .55 and .7 ................................................................. 59
Figure 28. Sample Size 40, Correlations .60 and .7 ................................................................. 59
Figure 29. Sample Size 40, Correlations .65 and .7 ................................................................. 60
Figure 30. Sample Size 40, Correlations .7 and .7 ................................................................. 60
Figure 31. Sample Size 50, Correlations .5 and .7 ................................................................. 61
Figure 32. Sample Size 50, Correlations .55 and .7 ................................................................. 61
Figure 33. Sample Size 50, Correlations .60 and .7 ................................................................. 62
Figure 34. Sample Size 50, Correlations .65 and .7 ................................................................. 62
Figure 35. Sample Size 50, Correlations .7 and .7 ................................................................. 63
Figure 36. Sample Size 100, Correlations .5 and .7 ................................................................. 63
Figure 37. Sample Size 100, Correlations .55 and .7 ................................................................. 64
Figure 38. Sample Size 100, Correlations .6 and .7 ................................................................. 64
Figure 39. Sample Size 100, Correlations .65 and .7 ................................................................. 65
Figure 40. Sample Size 100, Correlations .7 and .7 ................................................................. 65
Figure 41. Sample Size 200, Correlations .5 and .7 ................................................................. 66
Figure 42. Sample Size 200, Correlations .55 and .7 ................................................................. 66
Figure 43. Sample Size 200, Correlations .6 and .7 ................................................................. 67
Figure 44. Sample Size 200, Correlations .65 and .7.......................... 67

Figure 45. Sample Size 200, Correlations .7 and .7............................ 68
CHAPTER 1 INTRODUCTION

In research, particularly in education, psychology, and medicine, new techniques or equipment for the testing of students, clients, and patients are frequently developed, and it becomes necessary to ensure that the new method is as reliable and accurate as the standing, gold-standard method (Zaki, Bulgiba, Ismail, & Ismail, 2012). For example, a new method of measuring one’s blood pressure may be designed that does away with the use of a cuff. Of critical importance is the ability to test that the two methods are “measuring the same thing in an equivalent manner” (Hanneman, 2008, p. 223). The validity of both methods is not questioned in these cases; it is assumed (having already been established in earlier tests). The processes are designed to determine if the two methods are interchangeable; if one may be used in place of the other and achieve a similar result (Fernandez & Fernandez, 2009). For example, in the case of a new blood pressure measurement device, “It would be unwise simply to assume that measures made on the same person using different methods of measurement will agree” (Bunce, 2009, p. 4). It is necessary to determine that the two methods are “sufficiently similar” (Bartlett & Frost, 2008, p. 469).

Agreement

Agreement is the degree to which one device’s measures resemble the results of the alternate device. Bland and Altman state that the differences should not be sufficient to raise concern in a clinical environment (Bland & Altman, 1986). Overall agreement can be measured by the mean of differences of paired measures, and the limits of agreements for
individual measures constructed by the Bland & Altman levels of agreement (LoA), or the variance of the paired measures (Bunce, 2009). Bartlett and Frost (2008) noted that agreement is to be “measured on the same scale as the measurements themselves” (p. 466). The population from which the two measures are drawn should not be an issue for determining agreement unless there is a problem with the measurement devices (bias or instrument imprecision) (Bartlett & Frost, 2008).

Various techniques have been used to determine if two measurement methods are comparable, particularly correlation procedures and hypothesis testing (e.g., t-Test, linear regression). A prevalent method since its inception in the early 1980s is the Bland Altman method. (Zaki et al., 2012).

**Bland-Altman Method**

Bland and Altman (1986) described issues presented in the other techniques for device comparison and dismissed them as being inappropriate for what would amount to a measurement of agreement between two instruments. Their procedure is a graphical display of agreement – a line of agreement combined with levels of agreement (LoA) which are typically two standard deviations above and below the line of agreement (Bland & Altman, 1986, 2003). Bias is the difference between the two measures (O’Connor, Mahar, Laughlin, & Jackson, 2011).

The researcher/clinician would determine prior to the construction of the Bland Altman plot the degree of variability between the two measures that would be acceptable. The measurements would then be taken, the results plotted, and the levels of agreement constructed. Inspection of the plot would determine if there is sufficient agreement for the
researcher/clinician to agree that the methods are indeed interchangeable (Bland & Altman, 1986), or not so far apart to cause a problem. Ninety-five percent of the subject measurements should fall between the LoA to make a determination that the devices can be used interchangeably (Bunce, 2009).

Bland & Altman’s limits of agreement are estimates that would shift with each new sample (Bland & Altman, 1986). However, there is no test statistic to aid making a judgement of the outcome of a Bland Altman plot. Hence, it can be argued that implications derived from the Bland Altman plot technique is subjective. This could lead to error in the conclusion that interchangeably exists where a statistical test may demonstrate that it does not.

**Potential Solutions**

A potential solution would be the employment of the Kolmogorov-Smirnov test, a non-parametric technique that assesses the maximum difference between two measurements. The results are still displayed graphically, however, a test statistic is calculated that would determine if the two sets of measurements differ in a statistically significant fashion (Neave & Worthington, 1988). This supplements the clinician’s judgement for a more objective result. Similarly, there are a plethora of fit indices, such as those used in Structural Equation modeling that might be usefully employed in this context.

Fit indices demonstrate either how well the data fit the model (goodness-of-fit) or conversely, how poorly (badness-of-fit) (Kline, 2011). There are also model fit test statistics in structural equations, and approximate fit indices. The model fit statistic is an examination of how closely the model covariance matrix resembles the sample covariance
matrix. Kline explained that most model fit statistics are badness-of-fit: the higher values are assigned to the worst fits (2011). A statistically significant result would indicate “problematic model-data correspondence” (Kline, 2011, p. 193).

Approximate fit indices are largely “goodness-of-fit” – higher values for better fit, and a continuous demonstration of sample fit to the model (Kline, 2011). Such fit indices can be absolute (the proportion of sample covariance explained by the model), incremental (an indication of the improvement of fit over a model with zero covariances), parsimony adjusted (preference for a simpler model), and predictive (model fit based on hypothetical replication samples) (Kline, 2011)

Statement of the problem

Beyond validating devices used in bio-medicine, it is often necessary to discern if one device can be used interchangeably for another (Bland & Altman, 1986). So far there have been insufficient techniques to make these kind of determinations accurately, reliably, and objectively (Hopkins, 2004; Zaki et al., 2012). Hence, the purpose of this study is to observe the differences between the Bland Altman method and techniques employing goodness of fit of the data to the model. This will enable making a determination if the Kolmogorov-Smirnov test, or various fit indices such as those based on the Chi-squared test, give a result that is more objective than the Bland Altman plot without losing the graphical aspect.

Importance of the Study

The Bland-Altman method is a popular technique used in bio-medicine device comparison (Zaki et al., 2012). However, it is problematic in several respects, such as being
subjective (Zaki et al., 2012) and exhibiting false biases in the data (Hopkins, 2004). Bias is present if the first method over or underestimates "measurements relative to the measurements of the second method (Bartlett & Frost, 2008). Bunce (2009) noted that the Bland & Altman limits of agreement are a "clinical, not statistical, decision" (Bunce, 2009, p. 5).

The consequences of using a technique understood to be subjective or otherwise problematic could be broad: "This has the potential to affect the management of patients, quality of care given to the patient, and worse still could cost lives" (Zaki et al., 2012, p. 5). It therefore becomes imperative to discern other methods known to be objective that will produce more accurate results over time.

**Key terms:**

**Accuracy.** How closely an instrument measures a true value; determined by comparing against a device already established as highly accurate (Hanneman, 2008).

**Limits of Agreement.** The confidence intervals for agreement: with a normal distribution of differences, the calculation for the limits of agreement is the mean of the differences ± 1.96 × the standard deviation of the difference (Bunce, 2009).

**Precision.** The consistency of the measures taken by the same instrument over time (Hanneman, 2008).
CHAPTER 2 LITERATURE REVIEW

From a classical measurement theory perspective, after reliability and validity have been established the next question when comparing two instruments intended to measure the same variable is whether they can be used interchangeably (or if the least expensive or perhaps more efficient method could replace the more standard method). This is to be distinguished from equating, by either score transformation or linear equating.

In score transformation a “transformed score is equal to the difference between the original score and the mean of the original distribution, multiplied by the ratio of the standard deviation of the transformed distribution to the original distribution, plus the mean of the transformed distribution” (Sawilowsky & Fahoome, 2003a, p. 139). The formula is given below (Sawilowsky & Fahoome, 2003b)

\[
x^t = \frac{\sigma_t}{\sigma_o} (x_o - \mu_o) + \mu_t
\]  

(1)

Linear equating is the conversion of measurements from one scale to another; the mean and standard deviation are adjusted from one test form to make it comparable to another (Angoff, 1987). Angoff (1987) stated that “the ultimate aim of equating is to provide an equation that describes the nature of the conversion of units from one instrument to another, without regard to the nature of the particular groups to whom these instruments were administered in the equating study” (Angoff, 1987, p. 295).

However, in this case of determining the interchangeability of one bio-medical device to another, the goal is not to match a metric or convert scores. The goal is to determine the magnitude of the difference (if any) from one device to the other in terms of its measurements. It is therefore important to conduct method-comparison studies prior to
the implementation of the newer method (Hanneman, 2008). The variables being examined in a method-comparison study need to be on the same scale (Fernandez & Fernandez, 2009).

Terminology should also be clearly defined; Hanneman stated “‘Accuracy’ and ‘precision’ are often used when ‘bias’ and ‘repeatability’ are the properties being assessed.” (2008, p. 224). A necessary element is the assessment of agreement when comparing methods of measurement, (Luiz & Szklo, 2005). Twomey and Kroll (2008) stated agreement alone is insufficient, what is being assessed is “how well the two methods agree from the point of individual specimens across the whole data range found in clinical practice” (Twomey & Kroll, 2008, p. 529).

Among the procedures used in method-comparison studies have been correlation (both Pearson’s Product-Moment Correlation Coefficient, and Concordance Correlation Agreement), linear regression, comparing means from the data, and a graphical technique called the Bland-Altman analysis (Bland & Altman, 1986; Luiz, Costa, Kale, & Werneck, 2003). None of these methods have been determined to be wholly satisfactory in determining whether two separate measurement devices are measuring the same thing (and thus can be used interchangeably or one could replace the other, see, e.g.,(Bland & Altman, 1986; Zaki et al., 2012). A review was conducted (Zaki et al., 2012) to ascertain the most used methods in method-comparison studies; the third most popular technique was comparing the means of the data using means comparison (the t-Test), second was correlation coefficients, and the most popular was the Bland-Altman plot (Zaki et al.).
The Correlation Coefficient

Correlation is the measurement of association; it is used to determine the strength of a relationship between two variables. The relationship can be strong in either a positive (both variables increasing simultaneously) or negative (one variable increasing while the other decreases) direction. A common method for computing the correlation is calculating \( r \), the coefficient of correlation (McClave, Benson, & Sincich, 2005). Should \( r \) be zero, or close to zero, the implication is little or no relationship between the variables. An \( r \) value going to -1 or 1 indicates a strong negative or strong positive relationship, respectively ((McClave et al., 2005).

The Pearson product-moment correlation formula (McClave et al., 2005) is

\[
 r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\Sigma x^2 - (\Sigma x)^2][\Sigma y^2 - (\Sigma y)^2]}}
\]  

(2)

The relationship indicated by the Pearson r only an indication of a possible relationship. No causality should be inferred.

The problem with using correlation to establish the interchangeability of methods is that correlation measures only the linear relationship of the two methods, not the agreement between the two methods. Bland and Altman (1986) made the following arguments: agreement is affected by a change in measure scale where correlation is not; a stronger correlation will be found in a broader sample than a smaller one, and larger samples are typically used in method comparison studies. Also, “data which seem to be in poor agreement can produce quite high correlations,” (Bland & Altman, 1986, p. 308).

Another type of correlation that has been employed (by the evidence in the literature more prominently prior to the introduction of the Bland Altman plot) to assess
interchangeability between measurement devices as well as evaluating reproducibility is the Concordance Correlation. Lin (1989) stated that [when a new device is developed] it is “of interest to evaluate whether the new [device] can reproduce the results based on a traditional gold-standard [device]” (p. 255). Lin defined the Concordance Correlation Coefficient as the degree of concordance (agreement) between two pairs of samples and the expected value of the squared difference between them.

\[
\sum (Y_1 - Y_2)^2 = (\mu_1 - \mu_2)^2 + (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) = (\mu_1 - \mu_2)^2 + (\sigma_1 + \sigma_2)^2 + 2(1 - p)\sigma_1\sigma_2
\]

where \( \rho \) is the Pearson Correlation Coefficient “also represents the expected squared perpendicular deviation from the 45 degree line, multiplied by 2” (Lin, 1989, p. 257). Perfect agreement would result in a value of zero, with the scale indexed between -1 and 1. The Concordance Correlation Coefficient determines to what degree a pair would fall on the 45 degree line using a bias correction factor (Lin, 1989).

Atkinson and Nevill (1997) argued that the Concordance Correlation Coefficient is “highly sensitive to sample heterogeneity…[which] is very common in agreement studies…[and] can lead to wrong conclusions…when compared to more appropriate limits agreement methods” (Atkinson & Nevill, 1997, p. 775). They claimed this method suffers from the same disadvantage of any type of correlation in that any amount of bias in the measurements is not being uncovered, despite Lin’s claims to the contrary. They concluded a high level of disagreement cannot be discerned to be the result of highly heterogenic samples, systemic bias or random error.
t-Test

The t-Test (also known as Student’s t Test) “must be counted among the best-known statistical procedures in current use” (S. S. Sawilowsky & Blair, 1992, p. 352). As with many parametric statistical tests, certain assumptions regarding the data must be met in order for the result to be accurate. It is necessary for the samples to have equal variances, and the distributions should resemble a normal distribution (S. S. Sawilowsky & Blair, 1992). The t-Test would be used to reject the null hypothesis that there is no difference between the two devices (unlike equivalency testing where the alternative hypothesis is that there is no difference between the two devices) (Blair & Cole, 2002). In the case of testing two measures that are likely to be correlated, the t-Test for paired sample is the most appropriate (Zaki et al., 2012). The formula for the paired sample t-Test is shown in Equation 4 (Ha & Ha, 2012, p. 149).

\[ t_{obtained} = \frac{\overline{D} \cdot \overline{D}}{S_D \cdot \sqrt{n(n-1)}} \]  

(4)

where \( \overline{D} \) is the mean difference score and \( S_D \) is the standard deviation of the differences.

In the case of comparing two measurement devices, the null hypothesis would be that there is no difference between the means of the devices’ measurements. The conclusion could be reasonably made that the devices measure in a similar enough manner that they could be used interchangeably. The alternate hypothesis would be that one device measures higher or lower than the other, leading to a two-tail test. The nominal alpha level is then typically set to 0.5. The test statistic is then calculated, and if the observed value of \( t \) is greater, then the null hypothesis is rejected in favor of the alternative, that a significant
difference between the devices exists (Hinkle, Wiersma, & Jurs, 2003) and that they may not be used interchangeably.

Bland and Altman (1986) stated when measuring agreement, a test of significance would not be relevant – that the test will demonstrate a relationship between the measurements, but “it would be amazing if two methods designed to measure the same quantity were not related” (Bland & Altman, 1986, p. 308). Zaki et al. (2012) stated that as the paired sample t-Test measures the difference between means, a non-significant result may actually be significant if an extreme value has affected one of the means: “It is possible that poor agreement can be hidden in the distribution of the differences, and thus the two methods can appear to agree” (Zaki et al., 2012, p. 5). Bland and Altman tackled this difficulty by plotting the measurements both by mean and difference to reduce error.

**Regression**

Ordinary least squares regression has also been frequently used in agreement studies (Zaki et al., 2012). Twomey and Kroll (2008) reported “[Linear regression] methods try to determine the best linear relationship between data points while correlation coefficients assess the association (as opposed to agreement) between the two methods” (2008, p. 529). If the relationship between two samples is linear, a line can be drawn in the middle between them (the regression line). This can be used to describe the amount of variance between the data sets that is explained by the regression equation ($r^2$). Error in the $r^2$ can result if the sample data are not evenly distributed (or restricted), if the sample is too small, if that samples are not independent, the presence of outliers, or if the relationship is non-linear (Twomey & Kroll, 2008). It is also necessary to have an adequately powered
procedure in order to detect any differences in the data sets. Total allowable error is, like the Bland-Altman plot, determined prior to the analysis. Typically, 95% of the measurements should be within 95% error limits (Twomey & Kroll, 2008).

Agreement between the two devices is determined by examining the slope and intercept of the regression equation to see if they fall within their respective 95% limits of agreement. The F test is used to test the null hypothesis (that there is no difference in how either device measures).

Linear regression however suffers from the same limitations as an extension of correlation in determination of agreement; among them that “slope and intercept estimates significantly different from 1 and 0, respectively, can result from sample size limitations, leading to misinterpretation” (Luiz & Szklo, 2005). Also, “the coefficient of determination ($r^2$) being related to the correlation coefficient relies on a similar concept, and is thus not suitable for measuring agreement” (Zaki et al., 2012).

**Bland-Altman Method**

Bland and Altman devised a graphical plot to demonstrate agreement between two methods, specifically using a line of agreement. If two methods were to measure in precisely the same way, all the measurement points would fall evenly on that line, also called the bias line. (1986). Both the difference and average between the two measures are calculated, and plotted using a scatterplot. A limit of agreement (both high and low) is then computed, typically using a 95 percent confidence interval. Should the majority of data points fall within the levels of agreement, the conclusion may be drawn that the devices are measuring interchangeably (Bland & Altman, 1986; Bunce, 2009). Bunce (2009) stated
there should be no observable relationship between discrepancies and the level of measurement.

Difference and mean are plotted (as opposed to strictly the difference) in order to observe true value and measurement error relationships (Bland & Altman, 1986) as the true value is unknown (Bland & Altman, 1986; Myles & Cui, 2007). This is often called the bias – if the two devices are measuring precisely the same, the line would be at zero (Bland & Altman, 1986; Myles & Cui, 2007). The levels of agreement are used to determine if the two measurement methods are reasonably close in agreement (Myles & Cui, 2007). “When the plotted differences represent the new method minus the established method, the bias quantifies how much higher (i.e., positive bias) or lower (i.e., negative bias) values are with the new method compared with the established one” (Hanneman, 2008, p. 226).

![Figure 1. Sample Bland Altman Plot.](https://www.unistat.com/guide/bland-altman-plot/)

In the sample plot (Figure 1), very few of the data points fall along the line of agreement. However, most of the measurements fall between the levels of agreement and there does not appear to be any systematic bias between the measures. It is likely the researcher/clinician upon examination of such a plot would determine there is no significant difference between the measurement methods.

However, as this conclusion remains subjective due to a lack of a test statistic, it could be incorrect (Bunce, 2009). Nonetheless as demonstrated in Zaki’s (2012) study, the Bland Altman plot is among the most popular techniques for assessing method agreement.

An example of the use of the Bland Altman method is a study that compared prostate volumes using the standard procedure (TRUS – a transrectal ultrasound) with a less invasive procedure (TAUS – transabdominal ultrasound) (Jandaghi et al., 2015). Both methods were performed on a sample of patients. The acceptable clinical difference was determined prior to the analysis. Despite high correlation, the Bland-Altman plot showed low agreement between the two methods, thus it was determined that the two ultrasound techniques were not interchangeable. Jandaghi et al. (2015) noted in either case the true values are unknown, therefore the “accuracy of the newer method could not be established” (p. 37). However, “this method is currently the most commonly used approach for measuring agreement” (p. 36). They rejected the t-Test in this type of study because it is “possible to observe almost the same value for the mean of the two methods [when there is] a large variation in the distribution of data” (p. 37). They rejected correlations because the association between the measurements would not reveal differences in agreement.
Lim and Kelly (2010) also employed the Bland-Altman method: here, as in the ultrasound study, the case was being made to use a less intrusive, potentially safer method of determining blood gas volumes in chronic obstructive pulmonary disease. A meta analysis was conducted and the data analyzed with a Bland Altman plot. The results indicated good agreement. Lim and Kelley (2010) recommended further study because the studies they found did not describe “how [this] agreement impacted on treatment decision making and clinical outcomes in COPD patients with exacerbations in the ED” (Lim & Kelly, 2010, p. 247).

Although the Bland-Altman plot is widely used for method comparison, it is not without its critics. Its use in device validation and calibration is erroneous according to Hopkins (2004). It is apt to show a positive bias for larger values of the difference between the calibrated instrument and the calibration standard. There would be difficulty in discerning false bias from real bias (Hopkins, 2004): “Evidently the BA plot fails to correctly diagnose the lack of bias even in the simple case of measures differing only in the extent of random error” (Hopkins, 2004, p. 46).

Similarly, Fernandez and Fernandez (2009) suggested the Bland-Altman plot is only effective when the measurements are largely homogenous with low positive error. Regression analysis would show that such bias doesn’t actually exist (Fernandez & Fernandez).

Zaki (2012) stated “Various statistical methods have been used to test for agreement…which method is best is still open to debate” (p. 1). A potential solution that
gives both a statistical test (to reduce subjectivity) and a graphical display is the Kolmogorov-Smirnov (K-S) test.

**Kolmogorov-Smirnov (KS test) (test of general differences)**

The KS test is a non-parametric procedure used to test two samples for differences between the population distributions (Neave & Worthington, 1988). The differences could be the location, spread, or shape of the distributions. Neave and Worthington (1988) indicated generally differences in averages are the primary interest (1988). The test statistic employed is the maximum vertical difference. The KS test is nonparametric so the distribution is not necessarily an issue, yet the distribution could be ascertained through the graphical aspect of the KS test (Kirkman, 1996). Fit indices such as the RMSEA, GFI, and CFI could also be used to test for model fit in order to determine device interchangeability.

**Construction of the KS test**

This hypothetical involves testing two devices that are each designed to test fasting blood sugar levels. The devices were both used on the same patient at the same time for each pair. The assumption is both devices are considered valid. The question is if both devices are similar enough to be used interchangeably. The data are shown in Table 1.
Table 1.

**Sample Blood Pressure Readings**

<table>
<thead>
<tr>
<th>Device A</th>
<th>Device B</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>82</td>
</tr>
<tr>
<td>81</td>
<td>85</td>
</tr>
<tr>
<td>84</td>
<td>95</td>
</tr>
<tr>
<td>91</td>
<td>97</td>
</tr>
<tr>
<td>94</td>
<td>99</td>
</tr>
<tr>
<td>98</td>
<td>101</td>
</tr>
<tr>
<td>100</td>
<td>116</td>
</tr>
</tbody>
</table>

Neave and Worthington (1988) noted the cumulative distribution function (CDF) is a “step function that rises by 1/n at each observation” (Neave & Worthington, 1988, p. 150). The CDFs for devices A and B are shown in Table 2 below.

**Table 2**

**Blood Sugar Reading CDFs**

<table>
<thead>
<tr>
<th>Device A</th>
<th>Device B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/7</td>
<td>1/7</td>
</tr>
<tr>
<td>2/7</td>
<td>2/7</td>
</tr>
<tr>
<td>3/7</td>
<td>3/7</td>
</tr>
<tr>
<td>4/7</td>
<td>4/7</td>
</tr>
<tr>
<td>5/7</td>
<td>5/7</td>
</tr>
<tr>
<td>6/7</td>
<td>6/7</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The data and the CDFs can then be arranged in letter sequence as shown in Table 3.
Table 3

*Letter Sequence and CDF*

<table>
<thead>
<tr>
<th>Device A CDF</th>
<th>1/7</th>
<th>2/7</th>
<th>2/7</th>
<th>3/7</th>
<th>3/7</th>
<th>4/7</th>
<th>5/7</th>
<th>5/7</th>
<th>5/7</th>
<th>6/7</th>
<th>6/7</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter Sequence</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

| Device B CDF | 0 | 0 | 1/7 | 1/7 | 2/7 | 2/7 | 2/7 | 3/7 | 4/7 | 4/7 | 5/7 | 5/7 | 6/7 | 1 |

The formula for calculating $D^*$ is shown in Equation 5

$$D^* = n_A n_B D$$  \hspace{1cm} (5)

where $D$ is the maximum difference between the data sets, $n_A$ is the first data set, and $n_B$ is the second data set.

Using this example, $D^* = 7 \times 7 \times \frac{3}{7} = 21$. In order to reject the null hypothesis that the two distributions are equal, $D^*$ must be greater than the critical value. Referring to the KS tables with an alpha of .05, the critical value is 42. With the test statistic falling below the critical value, the null hypothesis is retained (Neave & Worthington, 1988). The graph for the sample is shown in Figure 2. Although the graph is helpful in giving a visual determination of the similarity of the distributions, the test statistic gives the objective indication of similarity.
**Figure 2.** KS Results Plot

**Root Mean Square Error of Approximation (RMSEA)**

Approximate fit indices fall into categories of either goodness of fit where 1 indicates a perfect fit, or badness of fit where the best fit is 0 (Kline, 2011). RMSEA is a badness of fit index which “decreases as there are more degrees of freedom (greater parsimony)” (Kline, 2011, p. 205). A relatively simple model with few parameters would be considered parsimonious (Arbuckle, 2012).

The value of RMSEA is the point estimate of $\varepsilon$ and the lower bound of a 90 percent confidence interval should equal 0 (Kline, 2011). In the close fit hypothesis (where the null
hypothesis is $\epsilon_0 \leq .05$) the value of .05 may be considered a good fit (Kline, 2011). RMSEA is also considered sensitive to model size as smaller models may be penalized for complexity (Kline, 2011).

**Goodness of Fit Index (GFI)**

As a goodness of fit index, the indication of perfect fit is 1 and estimates the amount of variance that the model explains (Kline, 2011). Kline (2011) points out that GFI is less sensitive to sample size than RMSEA, however “the mean values of the GFI tend to increase along with the number of cases” (Kline, 2011, p. 207).

**Comparative Fit Index (CFI)**

CFI is an incremental fit index that is compared to a baseline model where value of 1 would indicate a good fit (Kline, 2011). It is important to be assured of a plausible baseline model as a better fit over a baseline model may not impart much meaning (Kline, 2011).
CHAPTER 3 METHODOLOGY

The Bland Altman method is one of the most widely used in determining the interchangeability of two measurement devices (Zaki et al., 2012). The researcher/clinician makes a determination prior to the study how much difference between the devices is tolerable before the resulting measurements are considered too different to safely interchange (Bland & Altman, 1986). The Bland Altman method however has been shown to display bias where no bias actually exists (Hopkins, 2004) and the potential resultant subjectivity due to the lack of a test statistic could lead to an error in the determination of the interchangeability of the devices.

The use of the Bland Altman method with the more objective Kolmogorov-Smirnov (KS test) and such fit indices as the RMSEA (the root mean square error, a parsimony-corrected fit index), GFI (Goodness of Fit), and CFI (an incremental fit index) (Kline, 2011) will be examined using Monte Carlo simulation. Monte Carlo simulation is repeated sampling with replacement that uses many iterations to obtain a long range average (Sawilowsky & Fahoome, 2003a) or in this case, the fit of the data over many iterations. The procedure that demonstrates the most precision as well as objectivity will be considered the superior method for measurement device comparison over the Bland Altman method alone.

Samples

It will be necessary for the samples to be correlated as that is a general understanding with sample data from two similar devices (Bland & Altman, 1986). The following example demonstrates the creation of correlated data (although note this method
does not maintain the distribution properties, and, it only works for normally generated pseudo-random numbers). Suppose the data for A are given, and you wish to create B, which is correlated with A where \( r_{AB} = 0.8 \).

Draw a normally distributed random number \( C_1 \), and using \( A_1 \), put it in the following equation:

\[
B = rA + \sqrt{1 - r^2} C, \quad \text{where } r \text{ is the desired correlation of } A \text{ with } B.
\]  

(6)

This produces \( B_1 \), the first score to be paired with the first \( A_1 \) value. Repeat this process for the entire sample size of A. Therefore, to correlate A with B to, for example, \( r = 0.8 \):

\[
B = rA + \sqrt{1 - r^2} C
\]

\[
= 0.8A + \sqrt{1 - 0.8^2} C
\]

\[
= 0.8A + \sqrt{1 - 0.64} C
\]

\[
= 0.8A + 0.36C
\]

\[
= 0.8A + 0.6C
\]

The sample size would need to be sufficient to determine any significant differences between the two devices: “The investigator would be derelict in concluding that the test methods are interchangeable when the difference between methods would be significant with a larger sample size” (Hanneman, 2008, p. 225). Bunce (2009) recommended samples between 100 and 200 in order to discern differences: “Without large numbers, there is a very real potential for incorrectly finding a new method acceptable and for such methods to be recommended for widespread use without justification” (Bunce, 2009, p. 5). Moses et al (2009) used 80 paired values from 62 patients (Moses et al., 2009). Velez-Montoya et al. (2010) used 70 paired samples when comparing ultrasound devices (Velez-Montoya et
al., 2010). Jandaghi et al. (2015) used samples from 40 patients (Jandaghi et al., 2015). For the purposes of this study, samples of various sizes will be run.

**Fit Index Formulas**

**RMSEA**

RMSEA is a parsimonious fit index – preferring the model with fewer parameters (Arbuckle, 2012). The basis is the population discrepancy function $F_0$, the “value of the discrepancy function obtained by fitting a model to the population moments rather than to sample moments” (Arbuckle, 2012, p. 602). It is the degree of discrepancy between an ideal model and a researcher’s mathematical model (which is always an approximation) (Coffman, 2008). However, as $F_0$ will not necessarily distinguish between complex and simple models, the square root of $F_0$ divided by the degrees of freedom is used (Arbuckle, 2012). The formula RMSEA is shown in below (Arbuckle, 2012):

$$
\text{Population RMSEA} = \sqrt{\frac{F_0}{d}}
$$

(7)

where $F_0$ is the population discrepancy function and $d$ is the degrees of freedom used.

**GFI**

The GFI is a goodness of fit index, looking at the division of $\hat{F}$ (discrepancy function minimum value) by $\hat{F}_b$, which is where $F$ is the fitting function that is evaluated with $\sum^g = 0, g = 1,2, \ldots, G$. (Arbuckle, 2012).

The formula is given below (Arbuckle, 2012):
GFI = 1 - \frac{\hat{F}}{\hat{F}_b} \quad (8)

CFI

CFI is a comparative fit index, compared against a baseline model. The formula is given in Equation 9 (Arbuckle, 2012):

\[ CFI = 1 - \frac{\max(\hat{C} - d, 0)}{\max(\hat{C}_b - d_b, 0)} = 1 - \frac{NCP}{NCP_b} \quad (9) \]

where \( \hat{C}, d, \) and \( NCP \) are the discrepancy, degrees of freedom, and noncentrality parameter estimate (for model being evaluated), and \( \hat{C}_b, d_b, \) and \( NCP_b \) are discrepancy, degrees of freedom, and noncentrality parameter for the baseline model (Arbuckle, 2012). The noncentrality parameter estimate how far the test statistic’s sampling distribution mean is from the mean when the null is true (Minitab 17 Support Manual, 2015).

Procedure

The Bland-Altman Limits of Agreement (LoA) will be constructed at a 95 percent confidence interval (which is typical of this procedure) (Bland & Altman, 1986). The alpha for the KS test will be studied at .05 and .01. The Limits of Agreement for the Bland Altman method will be compared to the maximum difference of the KS test, and the KS test statistic will be computed. The fit indices will be studied at various levels of correlation and alphas (depending on the type of fit index).

Samples will be obtained using ISML – International Mathematical and Statistical Libraries (1982) including subroutine rnund. The Monte Carlo simulation will be conducted using Compaq Fortan 6.6c program. The samples are based on a normal
probability distribution to fit the requirements of the statistical procedures, however it is important to get a wide range of values in order to test better fit with a more heterogeneous sample. 10,000 iterations will be conducted.

The null hypothesis for this research question is that the results will indicate no greater level of precision or objectivity between any of the procedures or indices. The alternative hypothesis is that either the KS test or the indices will show greater precision and objectivity in conjunction with the Bland Altman method.

The tests will be conducted at varying levels of correlation and alpha levels (.05 and .01) and varying sample sizes in order to determine rejection levels at the given correlations. As reliability goes up, the rejection rate of the preferred statistic would go down.

In this case the preferred statistic will be the one that rejects most frequently at any given appropriate correlation. Some fit indices will reject with a p-value less than .05 (p <.05). Some will reject if the index is too high (p>.95). The rejection would indicate significance in the difference between the data sets and make interchangeability questionable, even if the Bland Altman method shows consistent levels of agreement. The superior statistic will not reject when rejection isn’t appropriate, and reject when it is.
CHAPTER 4 RESULTS.

Unintended Findings

Programming Platform

The initial programming was conducted in Fortran as indicated in Chapter 3. Coding the correlation algorithm and the procedure to run the Kolmogorov-Smirnov test (KS test) was completed. The program results were checked for accuracy against a worked correlation algorithm result, and KS test results in SPSS. It was then necessary to code the fit indices. Many attempts were made to locate clear, worked examples of the fit index formulas with no success. At that point the decision was made to switch from Fortran to R where the fit indices were available (in the R package Lavaan).

The Correlation Algorithm

To determine the accuracy of the algorithm used to create correlated data, two datasets were created with correlations of .05, .10, .20, .30, .70, .80, .90, and .99. Sample sizes were set at $n_1 = n_2 = 20, 30, 40, 50, 100, \text{ and } 200$. The number of repetitions was initially set at 1,000,000 to generate a precise value. Then, the number of repetitions was reduced to 10,000. The obtained correlation was compared with the value obtained with the higher number of repetitions to ensure the smaller number of repetitions was sufficiently precise.

Contained in Appendix A are the correlations for the various sample sizes at 10,000 and 1,000,000 repetitions. The algorithm to produce them was depicted in Equation 6. Note the correlations were unstable at low magnitudes and sample sizes. This remained true
regardless if the number of repetitions was 10,000 or 1,000,000. The correlations did not begin to stabilize until the magnitude was at least .70 and the sample size at least 50. Therefore, correlations of at least .70 were used for the balance of the study. The smaller sample sizes, however, were retained to determine how the statistical procedures performed at lower sample sizes.

**Change of Fit Index**

With the lower magnitudes of correlations eliminated, it necessitated removing the removal of the RMSEA, because it is a badness of fit index. Provided that correlation would not be problematic in and of itself, smaller degrees of freedom (DF) (and the model for this study had low DF) have resulted in the RMSEA values indicating poor fit regardless of how parsimonious the model may be (Kenny, Kaniskan, & McCoach, 2015). It was replaced with the Normed Fit Index (NFI), a relative fit index which compares empirical data with a null (baseline) model. The formula for the NFI is given in Equation 10.

\[
\text{NFI} = 1 - \frac{\hat{C}}{\hat{C}_b},
\]  

(10)

where \(\hat{C}\) is the discrepancy of the model being evaluated, and \(\hat{C}_b\) is the discrepancy of the baseline model.

**Results for the Kolmogorov-Smirnov (KS) Test**

The KS Test was conducted at various levels of correlation, and led to a rejection of the null hypothesis at each level (i.e., the distributions were statistically different). To confirm that the KS Test was not appropriate for correlated data in the absence of a shift in
location in the context of this study, two data sets were constructed, one with zero correlation (Figure 3) and the other with .80 correlation (Figure 4). Regardless of the two levels of correlation, D (maximum distance) was always equal to 1, and the p-value was less than 0.05.

**Figure 3.** Uncorrelated Data for KS Test

**Figure 4.** Correlated data for KS Test
The KS test was then employed for various levels of shift in location. Values of x were set equal to y, and the KS test gave an accurate result of the maximum distance (i.e., \( D = 0 \)) with a non-statistically significant p-value. Increasing shift in produced distances and corresponding p-values, eventually indicating a statistically significant p-value.

Two valid machines measuring the same thing will always be related (Bland & Altman, 1986). Hence, the determination was made that the KS test would not be a suitable approach to testing the Bland & Altman hypothesis, and was not given further consideration in this study.

**Monte Carlo Simulations of Fit Indices**

The literature on SEM modeling was reviewed in order to ascertain the most accurate and appropriate structural equation model, particularly as programed in R. There was a paucity of description regarding data extraction. Relatively nondescript examples included “data were simulated using a combination of SAS macro, SAS BASE, SAS PROC IML. Multivariate normal data were simulated by using on the matrix decomposition procedures” (Fan & Sivo, 2007); and “Proc CALIS (under its standard defaults) was used to generate the various statistics of interest under the confirmatory factor model” (Garrido & Abad, 2016, p. 736). This dearth of documentation obviated replicability, which was surprising considering the caliber of the peer-reviewed journals in which those studies were published.

Themessl-Huber (2014) published well documented R code for how the models were fitted, and the data simulation. Thus, that R code informed the programming for this
study, specifically in the regression of the variables related to the factors and the setting of the factors to zero.

Using only one data pair at one level of correlation in a structural equation model resulted in an overly identified -model with no degrees of freedom. An over identified model has “as excess of identifying information” (i.e., no parameters to solve) (Bollen, 1989, p. 89) and thus no degrees of freedom. Degrees of freedom are necessary to compute the chi-square statistic that the fit indices (NFI, CFI, and GFI) are based on (Bollen, 1989). Restrictions in a model are important in order to not over identify it; this can be done by adding variables to the model. It can also by setting elements of the model to zero (Bollen, 1989).

In order to impose such a restriction, following Themessl-Huber’s (2014) code as an example, each level of correlation was used as its own specific factor with each data set serving as the variables. Each factor regressed the y data on the x data for each correlation. This method created degrees of freedom and thereby the construction of a model that allowed for the computation of the fit indices to test the fit of the data sets.

The estimation method used was Maximum Likelihood (ML) and Monte Carlo simulations were run. The correlation magnitudes studied are compiled in Table 4:
### Table 4

*Order of Experiments*

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Correlation Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.5 and .7</td>
</tr>
<tr>
<td>2</td>
<td>.55 and .7</td>
</tr>
<tr>
<td>3</td>
<td>.60 and .7</td>
</tr>
<tr>
<td>4</td>
<td>.65 and .7</td>
</tr>
<tr>
<td>5</td>
<td>.7 and .7</td>
</tr>
</tbody>
</table>

### Fit Index Results

Listed in Table 5 are the fit index mean values at each level of correlation.

### Table 5

*Fit Index Results*

<table>
<thead>
<tr>
<th>Sample Sizes</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NFI</td>
</tr>
<tr>
<td></td>
<td>.5 and .7</td>
</tr>
<tr>
<td>n1 n2</td>
<td>.5 and .7</td>
</tr>
<tr>
<td>20 20</td>
<td>NFI</td>
</tr>
<tr>
<td></td>
<td>GFI</td>
</tr>
<tr>
<td></td>
<td>CFI</td>
</tr>
<tr>
<td>30 30</td>
<td>NFI</td>
</tr>
<tr>
<td></td>
<td>GFI</td>
</tr>
<tr>
<td></td>
<td>CFI</td>
</tr>
<tr>
<td>40 40</td>
<td>NFI</td>
</tr>
<tr>
<td>Sample Sizes</td>
<td>Correlations</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>$n_1$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>GFI</td>
<td>.906</td>
</tr>
<tr>
<td>CFI</td>
<td>.885</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>GFI</td>
<td>.966</td>
</tr>
<tr>
<td>CFI</td>
<td>1</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>GFI</td>
<td>.973</td>
</tr>
<tr>
<td>CFI</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>GFI</td>
<td>.987</td>
</tr>
<tr>
<td>CFI</td>
<td>1</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>GFI</td>
<td>.999</td>
</tr>
<tr>
<td>CFI</td>
<td>1</td>
</tr>
</tbody>
</table>

Results at a higher correlation, which produced similar patterns, are displayed in Table 6.
Table 6

*Correlations Starting at .9*

<table>
<thead>
<tr>
<th>Sample Sizes</th>
<th>Correlations</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>n1</td>
<td>n2</td>
<td>Fit Index</td>
<td>.5 and .9</td>
<td>.55 and .9</td>
<td>.60 and .9</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>NFI</td>
<td>.905</td>
<td>.908</td>
<td>.911</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFI</td>
<td>.965</td>
<td>.965</td>
<td>.965</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CFI</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>NFI</td>
<td>.959</td>
<td>.960</td>
<td>.962</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFI</td>
<td>.966</td>
<td>.966</td>
<td>.966</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CFI</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>NFI</td>
<td>.998</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GFI</td>
<td>.999</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CFI</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

None of the fit indices employed in the structural equation modeling could be labeled superior to the others. (Gerbing & Anderson, 1992). An ideal fit index would have properties such as a disassociation with sample size, fall between 0 (little to no fit) and 1 (perfect fit), and have distributional characteristics such that confidence intervals can be calculated (Gerbing & Anderson, 1992).

The Normed Fit Index, an incremental fit index, is used to compare the model to what is known as a null (or baseline) model that represents the worst possible fit (all the measured variables are uncorrelated) (Hooper, Coughlan, & Mullen, 2008). The NFI was
the most sensitive of the three indexes to the conditions presented, because it responded
(gave a differing average value) to both sample size and levels of correlation (Table 5). The NFI is typically regarded with some caution in structural equation modeling due to its sensitivity to small sample sizes (Hooper et al., 2008). That sensitivity was evidenced here. It started at .694 for a sample size of 20 at a correlation pairing of .5 and .7, and .997 for the same correlation pairing when the sample size was 200. The NFI rose with each incremental increase in data correlation (Figure 5):

![Increase in NFI Mean Value by Correlation and Sample Size](image)

**Figure 5.** Increase in NFI based on Sample Size and Correlations

The NFI increased at a greater percentage when starting with both a lower sample size and correlation. With a sample size of 20, the percentage increase between the correlation grouping of .5 and .7 was 16.6. With a sample size of 100, that percentage increase was 0.8, and at 200, 0.1.
Starting at higher correlation values of .5 and .9 (Table 6), the NFI started at a high value even with a low initial correlation pairing and a small sample size. As with the correlation pairings that start at .7, the NFI had the most increases with smaller sample sizes and greater correlation pairings (Figure 6):

![Increase in NFI Mean Value by Correlation and Sample Size](image)

**Figure 6.** NFI Increases with Higher Correlations

With a sufficiently high correlation (i.e., at least one correlation in the pair at .9), sample size appeared to have little effect on the NFI result. This sensitivity to changes in correlation could not be observed in the resulting Bland-Altman Plots. Contained in Figure 17 is a Bland-Altman plot with a sample size of 50 and a correlation pairing of .5 and .7:
Figure 7. BA Plot, Sample 50, Correlations .5 and .7

For this sample size and level of correlation, the NFI mean value was .918. Contained in Figure 8 is a Bland-Altman plot with the same sample size but a correlation pairing of .7 and .7:

Figure 8. BA Plot, Sample 50, Correlations .7 and .7
There were no observable differences between the results in the Bland-Altman plots when the correlations were increased. However, the mean value of the NFI increased from .918 to .941 indicating an improvement in fit that would not be detected by visual inspection of the Bland-Altman plot.

The Goodness of Fit index is an absolute fit index and measures the discrepancy between the sample and predicted covariances (Gerbing & Anderson, 1992; Hooper et al., 2008). The GFI also varies by sample size and is therefore dependent on it (Tanguma, 2001). Where the NFI would increase with each increase in correlation pairings, the GFI proved impervious to changes in correlation, remaining at the starting value for each sample size (Figure 9):

![Figure 9. GFI Mean Values by Sample Size and Correlation](image)

Using the same example of the sample size 50 presented above with the Bland-Altman plots, the NFI showed improvement in fit with each rise in correlation where the
GFI stayed flat (Figure 10). (Note the axis in Figure 10 was narrowed to more clearly demonstrate the change in the NFI.)

Figure 10. Comparison of GFI and NFI Mean Values

The GFI also started at a higher mean value than the NFI even at the smallest sample of 20; the GFI had a mean value of .913 (exceeding the acceptable fit range) over the NFI’s .694. This indicated that the GFI would be helpful in determining fit when the NFI is suppressed by a low sample size.

The Comparative Fit Index was designed to be less sensitive to sample size (Hooper et al., 2008) and Tanguma’s (2001) study showed that to be true. The mean values of the CFI produced in that study varied only slightly, however it was concluded it still had sample size sensitivity (Tanguma, 2001). Hooper (2008) described the CFI as a less sample size sensitive revision of the NFI and as such compares model data to a null (baseline, worst possible fit) model (Hooper et al., 2008).
With the initial run with a sample size of 20 and a correlation pairing of .5 and .7, the GFI produced a perfect fit of 1, which was stable through all correlation pairing increases. With the sample size of 30, the GFI decreased to .928 and increased incrementally with each correlation pairing increase. The mean value of the GFI remained at 1 with sample sizes of 50 and above regardless of correlation pairings. The increases of the CFI at sample sizes 30 and 40 are depicted in Figure 11:

![Increase in CFI Mean Value by Correlation and Sample Size](image)

**Figure 11.** CFI Mean Values by Correlation and Sample Size

Unlike the NFI and GFI which yielded higher starting mean values in accordance with higher sample sizes, the starting mean value for the CFI was lower for the sample of 40 (.885) than the sample of 30 (.928). The p-value associated with the obtained Chi-squared result for the sample of 30 was not statistically significant (p = .176). With a sample of 40 it was not quite statistically significant (0.06).
With the sample size set at 50, the CFI indicated a perfect fit (1) from the first correlation pairing (Figure 12).

![GFI, CFI, NFI Mean Values at Sample Size 50](image)

*Figure 12. GFI, NFI, CFI Comparisons*

The NFI results were the most sensitive to not only sample size but to correlation pairings, which would make it the most helpful statistic in for determining the interchangeability of biomedical devices, because small changes in the data lead to statistical significance.

**Bland-Altman Plot Results**

The Bland-Altman plots at the same sample sizes and correlations constructed with 95% levels of agreement (LoA) are located in Appendix B. The number of standard deviations used to calculate the LoA are meant to be determined in advance of the study. A typical choice is two standard deviations. The case made by Bland and Altman (1986) that correlation does not necessarily indicate agreement was demonstrated here. The data
did not vary from low to high correlation pairings. Note that the LoA is narrow when agreement is good (Myles & Cui, 2007).

Depicted in Table 7 are the mean, standard deviation, and the LoA at each sample size for correlations of .7 and .7.

Table 7

*Bland Altman Computations*

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Lower Limit</th>
<th>Mean of Differences</th>
<th>Standard Deviation of Differences</th>
<th>Upper Limit</th>
<th>Interval Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>-4143.30</td>
<td>-4141.21</td>
<td>1.07</td>
<td>-4139.12</td>
<td>4.18</td>
</tr>
<tr>
<td>30</td>
<td>-4144.60</td>
<td>-4141.5</td>
<td>1.58</td>
<td>-4138.40</td>
<td>6.62</td>
</tr>
<tr>
<td>40</td>
<td>-4144.37</td>
<td>-4141.29</td>
<td>1.57</td>
<td>-4138.22</td>
<td>6.15</td>
</tr>
<tr>
<td>50</td>
<td>-4143.73</td>
<td>-4141.34</td>
<td>1.22</td>
<td>-4138.96</td>
<td>4.77</td>
</tr>
<tr>
<td>100</td>
<td>-4144.17</td>
<td>-4141.50</td>
<td>1.36</td>
<td>-4138.83</td>
<td>5.34</td>
</tr>
<tr>
<td>200</td>
<td>-4144.26</td>
<td>-4141.45</td>
<td>1.44</td>
<td>-4138.63</td>
<td>5.63</td>
</tr>
</tbody>
</table>

As sample sizes increased from 20 to 200, the lower limit decreased by .96. The change from 20 to 200 resulted in a .49 increase for the upper limit, leading to a .47 decrease -indicating only slightly better agreement. The best agreement occurred with the sample of 30, with the lowest standard deviation, and an interval width of 4.18. The least agreement occurred with the sample size of 40 with the highest standard deviation and an interval width of 6.62. There was a moderate negative correlation (−.57) between the mean
of the differences and the interval width (as the mean increased the interval width tended to decrease).

Increases in the sample sizes also led to more data points falling outside the levels of agreement. Depicted in Figure 13 are the plot with a sample size of 20. The mean was included for comparison purposes.

![Figure 13. Bland-Altman Plot, Sample Size 20](image)

All but two of the data points fell within the LoA, and one of those outliers was only just outside the lower limit (Figure 13).
Figure 14. Bland-Altman Plot, Sample Size 200

There was a greater spread in the data points (Figure 14) with the sample size of 200 and a less precise fit than with the smaller sample. However there was a reasonably even spread of data points (above and below the bias/mean line) which indicated that there was no consistent bias between the two measurements (Hanneman, 2008). Taken in conjunction with the fit index results, a good fit was indicated. However, improvements in fit with stronger positive correlations could not be observed with the Bland-Altman plots, necessitating the use of the fit indices in general and the NFI in particular (with its greater sensitivity to changes in correlation).
CHAPTER 5 DISCUSSION AND CONCLUSION

Discussion

The underlying problem with the Bland-Altman method is the subjectivity of the method itself. There is no test statistic to rule out serious differences in the distributions of the devices. Hence, the purpose of this study was to test various other methods that could potentially provide an objective method of determining interchangeability between two devices.

Guidelines for proper use of the Bland-Altman plot were developed over time in an endeavor to make it at least more in line with good scientific practice, such as determining Levels of Agreement LoA in advance per defined clinical criteria, a scatterplot of the mean and the differences, and a pattern of relationship between the mean and differences (Chhapola, Kanwal, & Brar, 2015). However, studies showed many of the general guidelines are not followed when using Bland-Altman plots (Dewitte, Fierens, Stockl, & Thienpont, 2002). Chhapola et al. (2015) found an evaluation of the pattern of relationships was only discussed in 28% and the LoA predetermined in 74% of studies reviewed.

This is problematic when the procedure is subjective, attempts to make it more rigorous are not followed, and problematic even when employed. Myles & Cui (2007) posed the question, “would a difference between measurement methods as extreme as that described by the 95% limits of agreement meaningfully affect the interpretation of the results?” (p. 309). Thus, it was proposed the Bland Altman plot should only be used in conjunction with a statistic that can more objectively determine differences in distributions that may not necessarily either appear in a scatterplot or be visually recognized as such.
The procedures tested in this study, with the exception of the Kolmogorov-Smirnov test, demonstrated potential usefulness when paired with the BA plot.

The KS test was eliminated from further consideration even though it is a fit statistic sensitive to a shift in location between two distributions, because it is not sensitive to fit between theoretical and empirical curves that differ only in terms of how they are correlated with each other.

Consider, for example, sample size of 30, where the Bland-Altman plot did not change under any of the correlation pairings as shown in both panels of Figure 15.

![Pairing at .5 and .7](image1.png) ![Pairing at .7 and .7](image2.png)

**Figure 15.** Comparison of BA Plots at Sample Size 30

However, these changes were discerned by the other fit indices. The NFI started at .834 for the correlation pairing of .5 and .7, and increased to .875 with the .7 and .7 pairing. The CFI started at .928 and increased to .948. Note an observation of the improved fit could not be made by the visual inspection of the BA plot.
Cut-off scores to assess good model fit have been developed (Table 8). However, considering generally accepted ranges of for the fit indices to indicate good fit will not be helpful, because the sensitivity of the index to changes in the data will prove useful in objective determinations of device interchangeability. It was noted in the literature that making recommendations regarding optimal fit index cut scores was made difficult by sensitivity to sample size, and “the inherent inability of a structural model to exactly account for the phenomena it seeks to describe” (Sivo, Fan, Witta, & Willse, 2006, p. 268). Sivo et al. (2006) found “optimal cut-off values may vary considerably depending on the sample size, with smaller sample size resulting in lower optimal cut-off values” (p. 276).

Table 8

Fit Index Cut Scores

<table>
<thead>
<tr>
<th>Fit Index</th>
<th>Value for Acceptable Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFI</td>
<td>.90 or Greater</td>
</tr>
<tr>
<td>NFI</td>
<td>.95 or Greater</td>
</tr>
<tr>
<td>CFI</td>
<td>.95 or Greater</td>
</tr>
</tbody>
</table>


Hence, the fit index that demonstrated the most sensitivity to correlation change (the NFI) was also the most sensitive to sample size and was more sensitive with small samples. Using the GFI and CFI for assessing the starting values would be helpful, because
they are less sensitive to smaller sample sizes. In the case of the sample size of 30, the GFI showed a starting value of .913, which exceeded the recommended cut score for fit as displayed in Table 8.

The recommendation to use all three indices is not an uncommon in SEM studies, because no single index reflects all aspects of model fit (Hooper et al., 2008). Given the sensitivity to changes in correlation that occurs with the NFI over the other two indices would make the NFI, in this case, the superior statistic.

**Limitations of the Study/Recommendations for Further Research**

An issue inherent to the production of the Bland-Altman scatterplot was the aspect ratio of the graphical display, which is “the ratio between its horizontal and its vertical extent” (Fink, Haunert, Spoerhase, & Wolff, 2013, p. 2326). Few (2013) stated “the aspect ratio of a graph’s data region greatly influences perception of the data” (Few, 2004, p. 249). Identical data may show potentially different results depending on the aspect ratio used. There may not necessarily be a preferred, ideal, or even standard aspect ratio for a Bland-Altman scatterplot, however it is important to understand that the plot could be visually manipulated and therefore compromise the interpretation of the results.

The correlation algorithm used in this study to create the comparison dataset was unstable at both lower levels of correlation and sample sizes. Hence, the study parameters were restricted to higher levels of correlation. The smallest sample size (10) was excluded in order to achieve accurate and reliable. Further studies in this area might include the use of more complex techniques to create correlated data sets. Headrick and Sawilowsky (1999), for example, provided procedures to create more precise correlated data; are useful
for modeling nonnormal data; and preserve the mean, standard deviation, skewness, kurtosis, and higher moments of the referent distribution into the target distribution.

Pairing correlation decisions were made with the algorithm consistency restrictions in mind, as well as limits where the indices would cease to give more detailed information. Also, as stated by Gerbing (1992), “the chief limitation of Monte Carlo research is that there are too many possibilities to effectively consider in any one study, so the results necessarily represent a compromise” (p. 138). With a more complex algorithm technique (as suggested above) it would be of interest to test the behavior of the indices at smaller sizes of correlation and varying sample sizes. The data in this study were simulated, which is appropriate for the purpose of this method comparison. Further research using actual data sets for such devices as used in bio-medicine or other Bland-Altman studies could prove beneficial.

The R code for the simulations was based on Themessl-Huber’s (2014) published code. Another publication by Finch and French (2015) also gives code in R for setting up structural equations. This became available after the running of the simulations for this study, but could prove useful for further research in this area. If a new model is created with more variables it may be worthwhile to test the RMSEA fit index to assess whether it would be useful in determining fit.

Conclusion

The Bland-Altman plots on their own are useful for determining bias around the mean (either above or below the mean; systematic or otherwise) and fit within the pre-determined LoA. However as they can only be inspected visually, changes within the
correlation values of a particular data set could not be detected and statistically significant changes would not be observed.

Although the Kolmogorov-Smirnov test results indicated it was not appropriate for this application, three fit indices typically used in structural equation modeling (NFI, CFI, & GFI) produced results that were sensitive to sample and correlation changes. The NFI was sensitive at almost all level of correlation within a given sample size. It is recommended, therefore, that the NFI combined with the CFI and GFI (for a more helpful look at starting values, particularly at lower sample sizes) be used in conjunction with the Bland-Altman plots in order to more accurately and objectively assess the results of device comparison.

Use of the fit indices in addition to the Bland-Altman plots further the existing science by providing a numeric reference as to whether or not enhancements, calibration, or other such changes applied to either of the devices increases or decreases the agreement of the measurements. This could potentially save the researcher/clinician resources in terms of sample sizes (more information available on potentially smaller samples), or the number of experiments necessary to establish agreement with a level of objectivity that has been heretofore unavailable.
### APPENDIX A: CORRELATIONS AND SAMPLE SIZES FOR ALGORITHM

Table 9

*Correlation Results and Sample Sizes*

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Sample Size</th>
<th>10,000 Repetitions</th>
<th>1 Million Repetitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>20</td>
<td>-0.354005</td>
<td>-0.352924</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>-0.417433</td>
<td>-0.414150</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>-0.188592</td>
<td>-0.187755</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.104921</td>
<td>0.103879</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-0.128662</td>
<td>-0.126978</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>7.880547E-03</td>
<td>7.836309E-03</td>
</tr>
<tr>
<td>.10</td>
<td>20</td>
<td>-0.111981</td>
<td>-0.111192</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.232912</td>
<td>0.234152</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.126230</td>
<td>0.125403</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.157217</td>
<td>0.156352</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.144044</td>
<td>0.142369</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>9.075584E-02</td>
<td>9.108227E-02</td>
</tr>
<tr>
<td>.20</td>
<td>20</td>
<td>0.270583</td>
<td>0.269004</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.377507</td>
<td>0.375540</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.309510</td>
<td>0.311722</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.412597</td>
<td>0.408580</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>9.934952E-02</td>
<td>9.957828E-02</td>
</tr>
<tr>
<td>Correlation</td>
<td>Sample Size</td>
<td>10,000 Repetitions</td>
<td>1 Million Repetitions</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.187769</td>
<td>0.187516</td>
</tr>
<tr>
<td>0.30</td>
<td>20</td>
<td>0.586466</td>
<td>0.588458</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.588209</td>
<td>0.588756</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.261060</td>
<td>0.262584</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.476732</td>
<td>0.475129</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.331054</td>
<td>0.331924</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.362360</td>
<td>0.364293</td>
</tr>
<tr>
<td>0.70</td>
<td>20</td>
<td>0.807734</td>
<td>0.811528</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.646361</td>
<td>0.645129</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.704732</td>
<td>0.704808</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.698596</td>
<td>0.692229</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.634279</td>
<td>0.629908</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.722305</td>
<td>0.727570</td>
</tr>
<tr>
<td>0.80</td>
<td>20</td>
<td>0.674316</td>
<td>0.681713</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.587241</td>
<td>0.588522</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.740689</td>
<td>0.745749</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.774291</td>
<td>0.769719</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.828101</td>
<td>0.817446</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.792843</td>
<td>0.795882</td>
</tr>
<tr>
<td>0.90</td>
<td>20</td>
<td>0.795308</td>
<td>0.796796</td>
</tr>
<tr>
<td>Correlation</td>
<td>Sample Size</td>
<td>10,000 Repetitions</td>
<td>1 Million Repetitions</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.95</td>
<td>20</td>
<td>0.893976</td>
<td>0.888906</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.99</td>
<td>20</td>
<td>0.994740</td>
<td>0.999175</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX B: BLAND-ALTMAN PLOTS

The horizontal lines in each plot represent the upper and lower Levels of Confidence.

Sample size 20

Figure 16. Sample size 20, Correlations .5 and .7
Figure 17. Sample size 20, Correlations .55 and .7

Figure 18. Sample Size 20, Correlations .6 and .7
Figure 19. Sample Size 20, Correlations .65 and .7

Figure 20. Sample Size 20, Correlations .7 and .7
Sample Size 30

Figure 21. Sample Size 30, Correlations .5 and .7

Figure 22. Sample Size 30, Correlations .55 and .7
Figure 23. Sample Size 30, Correlations .6 and .7

Figure 24. Sample Size 30, Correlations .65 and .7
Figure 25. Sample Size 30, Correlations .7 and .7

Sample Size 40

Figure 26. Sample Size 40, Correlations .5 and .7
Figure 27. Sample Size 40, Correlations .55 and .7

Figure 28. Sample Size 40, Correlations .60 and .7
Figure 29. Sample Size 40, Correlations .65 and .7

Figure 30. Sample Size 40, Correlations .7 and .7
Sample Size 50

Figure 31. Sample Size 50, Correlations .5 and .7

Figure 32. Sample Size 50, Correlations .55 and .7
Figure 33. Sample Size 50, Correlations .60 and .7

Figure 34. Sample Size 50, Correlations .65 and .7
Figure 35. Sample Size 50, Correlations .7 and .7

Sample Size 100

Figure 36. Sample Size 100, Correlations .5 and .7
Figure 37. Sample Size 100, Correlations .55 and .7

Figure 38. Sample Size 100, Correlations .6 and .7
Figure 39. Sample Size 100, Correlations .65 and .7

Figure 40. Sample Size 100, Correlations .7 and .7
Sample Size 200

Figure 41. Sample Size 200, Correlations .5 and .7

Figure 42. Sample Size 200, Correlations .55 and .7
Figure 43. Sample Size 200, Correlations .6 and .7

Figure 44. Sample Size 200, Correlations .65 and .7
Figure 45. Sample Size 200, Correlations .7 and .7
REFERENCES


doi:10.1097/01.AACN.0000318125.41512.a3


*Psychological Test and Assessment Modeling, 56*(3).


DOI 10.1371/journal.pone.0037908
ABSTRACT

OBJECTIVE STATISTICS FOR THE MEASUREMENT OF AGREEMENT

by

ELIZABETH PERKIN MOEN

December 2016

Advisor: Dr. Shlomo S. Sawilowsky

Major: Education Evaluation and Research

Degree: Doctor of Philosophy

The prevailing method for biomedical device interchangeability is a scatterplot of means and differences bounded by levels of agreement (Bland-Altman Plot) which results in subjective judgements made in the discernment of differences in data distribution and fit. The purpose of this study was to test four statistics - the Kolmogorov-Smirnov test, and three fit indices (NFI, GFI, & CFI) in order to identify a more objective statistic for device interchangeability. The Kolmogorov-Smirnov test proved incompatible with the data structure. The Normed Fit Index (NFI) proved most sensitive to correlation shifts where the Goodness of Fit Index (GFI) and Comparative Fit Index (CFI) would be useful for starting values. The result is the recommendation to report all three indexes in conjunction with the construction of a Bland-Altman plot.
AUTOBIOGRAPHICAL STATEMENT

Elizabeth Perkin Moen was born on October 15, 1964 in Detroit, Michigan. She received a Bachelor of Arts in Communication from Wayne State University in 1987, and a Masters of Education in Educational Research and Evaluation in 2008. She is currently employed as a data analyst in the College of Nursing at Wayne State University, and as an adjunct instructor in the College of Education.