PhUSE 2011

SP06

Embedding equivalence t-test results in Bland Altman Plots visualising rater reliability

Jim Groeneveld, OCS Consulting, 's Hertogenbosch, Netherlands

SUMMARY / ABSTRACT

In determining **rater (instrument) reliability** many techniques are available. A main discrimination between methods is made by the level of measurement (binary, categorical, continuous). Most techniques, such as the Mean Absolute Difference (MAD) or the Intraclass Correlation Analysis (ICC), apply to continuous data. Two other analysis methods are **equivalence t-tests** (quantitatively determining significance of equivalence) and **Bland Altman Plots** (qualitatively judging the scatter in the plot). Here the **integration** of both methods is proposed and discussed. T-test results, especially the (in)significance of equivalence are visualised in the Bland Altman Plots and can be interpreted clearly, next to the qualitative judgement of the scatter in the Bland Altman Plot. This extends the value and usefulness of the Bland Altman Plotting technique.

INTRODUCTION

RATER RELIABILITY

Rater reliability determination is determining the **quality** of a **measuring instrument**, to determine to what extent it is precise[i], the consistency, reproducibility or repeatability, the degree to which repeated measurements on **same objects** under unchanged conditions show the same results. Measuring instruments can be equipment devices (e.g. a thermometer or a voltmeter), respondent's answers (e.g. on a questionnaire or psychological test), humans (e.g. who read equipment devices or who judge screen images), or a combination of those. In case of humans instruments are called **'raters'**. In the remainder of this document the term 'rater' will be mainly used to refer to all kinds of measuring instruments.

REPEATED MEASUREMENTS

Repeated measurements (judgements by raters) on **same objects** are needed to determine reliability. A temperature or voltage can be measured twice (or more often), a questionnaire or test can be conducted again after some time[ii], screen images can be judged multiple times. Situations where the repetitions are measured by the **same** or by **another**, but very similar instrument or rater can be discriminated. Another thermometer or voltmeter may be applied, another "parallel" questionnaire or psychological test (measuring the same quality or quantity) may be conducted, and another rater may judge images.

INTRA VS INTER

When same instruments or raters are used to repeat measurements the reliability that is determined is called **intra-rater** or **within-rater** reliability. When different, but comparable instruments or raters apply the reliability is called **inter-rater** or **between-rater** reliability. Intra-rater reliability represents the quality, the precision of individual raters, while inter-rater reliability applies to the quality of exchangeable raters. Reliability measurements and analysis may be designed to determine both types of reliability (multiple raters each rating multiple times).

APPLICATION AND VALUE

Determining reliability of instruments or raters may be important in at least the following two cases:

a. for **certification** purposes **before** starting a study to determine the usefulness of an instrument or rater in general. Generally representative objects (e.g. already available images) are used to be re-measured by raters;

[i] The complementary terms **accuracy** or **instrument validity**, representing the extent to which an instrument measures what it is intended to measure as specific as possible, are not further discussed here; neither are **sensitivity** and **specificity** of binary classifications.

[ii] Influencing, disturbing factors, like changed circumstances over time, people's memory of the past questionnaire or test (learning effects, for which parallel tests have been designed) are not discussed here.

b. for **QC** purposes **after** having conducted a study to determine the quality of the instruments or raters in a certain study and with that to be able to trust (or not) the quality of the measurements, the data and the analysed results of the study. This involves re-measuring (a sample of) certain study data.

RELIABILITY ANALYSIS

DATA TYPES AND ANALYSIS METHODS

Determining reliability of raters producing data can be done in a number of different mathematical ways. Based on the data level various groups of methods have to be discriminated:

1. **categorical** data can be binary (dichotomous), nominal or ordered (ordinal). Examples of statistical analysis methods are:
   a. Cohen's Kappa analysis for binary data (e.g. to determine consistency of missing vs non-missing data);
   b. Fleiss Kappa analysis for nominal data;
   c. McNemar's test for binary data;
   d. McNemar-Bowker's test for ordinal data.
2. **continuous** numeric data can be of interval or ratio level. Even discrete, ordinal data with many levels may be regarded and approached as continuous data. Examples of analysis techniques are:
   a. Mean Absolute Difference (MAD) of pairs of data points;
   b. Intraclass Correlation Coefficient (ICC) analyzing pairs or larger sets of matched data;
   c. Equivalence t-test, quantitatively judging data pairs as significantly (or not) equal;
   d. Bland Altman Plots to be interpreted qualitatively.

The last two analysis methods, the equivalence t-test and the Bland Altman Plots, are discussed and integrated in this paper.

EQUIVALENCE T-TEST

The equivalence t-test that applies in reliability analysis is the **paired equivalence t-test**, testing paired differences to be equal or around zero. **Limits** must be specified within which a (small) difference is considered to be virtually zero. The test actually consists of two one-sided non-inferiority t-tests **(TOST)**, each with the same confidence level (e.g. 95%) as a single test, and each testing a sample mean for being significantly larger than the lower bound and at the same time smaller than the upper bound. Once **both** partial tests show significance the overall TOST is regarded to show significant equivalence.

Usually the lower and upper limits are symmetrical, i.e. they have the same absolute value; the lower bound is negative and the upper one is positive. The TOST in such a case tests equivalence to 0. However, it is possible to test equivalence to any other (positive or negative) value, unequal to 0. In that case the lower and upper limits for the TOST are not symmetrical, but the principle of the TOST does not change at all. The partial one-sided non-inferiority t-tests are conducted independent from each other, each with their own test value, which is one of the value limits.

In any case the result of a TOST is **significance or no significance**, i.e. a conclusion on (in)significant equivalence of two paired sets of data, where their difference is regarded 0 (or on an (in)significant difference of any size unequal to 0 if so desired). The (in)significance of course depends on the data, but it also depends in the confidence level (e.g. 95%) and the chosen lower and upper limits. Actually the TOST, like any other statistical significance test, also yields a range in which a mean sample difference may lie in order to yet be regarded significantly equal to 0; this range is the **Confidence Interval of the Mean** (CI) and will be discussed later in this paper where the integration with the Bland Altman Plots is introduced.

BLAND ALTMAN PLOTS

In order to visualize possible equivalence between two paired sets of data (from repeated measurements) the data might be plotted against each other in a usual **scattergram of Y against X**. If the data pairs indeed represent about the same values (or their difference is around 0) the scattergram should show all points lying in the neighbourhood of the line Y=X. However, it is rather difficult to visually judge the scattered points for equivalence and reliability.

To overcome this disadvantage of a usual scattergram Bland and Altman [1, 2] developed a way to qualitatively judge about equality and reliability using a graph with scatter of points other than the basic paired data points. In such a graph they plot the **difference of the data pairs** [$Y=v_1-v_2$] against the **mean of the data pairs** [$X=(v_1+v_2)/2$] for each (non-missing) data pair and in addition they draw the horizontal lines representing the **Confidence Interval** limits of the (calculated) **points** (as they occur) as well as the horizontal line representing the **mean difference** (that ideally should be close to 0). An example of a usual Bland Altman Plot is given in figure 1, where the green line is the mean difference and the blue lines both define the CI of the points.

In any case the Bland Altman Plots are to be interpreted and judged **qualitatively** rather than quantitatively by a statistician or researcher. This is an important limitation and drawback. In order to add some firm visual, but quantitative information from the equivalence t-tests to the plots the integration of both methods is introduced in this paper.

INTEGRATION

The integration of the equivalence t-test (TOST) with the Bland Altman Plot consists mainly of the conventional Bland Altman Plot, but without the horizontal lines representing the CI of the (calculated) points. Instead the horizontal lines representing the **CI of the mean difference from the TOST** are added as well as the horizontal lines representing the **specified range limits** for the TOST. (Note that the CI of the mean sample difference is much smaller than the CI of the points from the Bland Altman Plot.) In this way the TOST result is visualized inside the (somewhat stripped) Bland Altman plot.

INTERPRETATION

Next to the display of the Bland Altman points in the plots the **specified range limits** and **the CI of the mean difference** from the TOST are visualized as horizontal lines. It can clearly be seen how these lines lie in relation to each other. The interpretation of significant equivalence or not can be deduced as follows:

a. If the CI of the mean difference is lying **entirely within** the specified range limits the data represent **significant equivalence** and thus have acceptable reliability. An example of such a result is given in figure 2, in which the green line is the mean difference, the red lines represent the specified range limits and the blue lines define the CI of the mean sample difference from the TOST;
b. If the CI of the mean difference **overlaps** the interval given by the specified range limits or is lying **entirely outside** that interval then there is **no significant equivalence**. An example of such a result is given in figure 3, in which the green line is the mean difference, the red lines represent the specified range limits and the blue lines define the CI of the mean sample difference from the TOST.

ADVANTAGES OF THE INTEGRATION

The integration of both methods in a graph visualizing results can be regarded as an extension of (the qualitative value of) the Bland Altman Plots with a quantitative interpretation on equivalence (in)significance and, hence, on reliability. Equivalence (in)significance is clearly visualized among the scatter of the Bland Altman points and also clearly depends on the range limits. This **integration** of two different analysis techniques in one graphical result combines a **qualitative view and impression** with **firm, quantitative information on significance**.

EXTENSIONS

NUMBER OF REPEATED MEASUREMENTS

Both the paired t-test (in general) and the Bland Altman Plot technique work on data pairs, 2 measurements of one object. It, however, may occur that there are more than 2 measurements per object (either by the same or by different raters). In that case **all combinations** of data pairs have to be made and analysed. E.g. with 3 measurements there are 3 combinations of pairs, 3 times as much as with 2 measurements. With 4 or even more measurements there are many more pairs to generate and analyse separately. It is obvious that an **overall conclusion** from many separate analyses is **quite difficult** to make.

The number of generated pairs with 4 or more measurements may be somewhat reduced by data reduction. Instead of comparing each individual set of measurements with each other one, each individual set of measurements may be compared to the **matched means** (from the same objects) of **all other sets** of measurements. With 4 measurements per object this results in 4 **"pairs"** (consisting of one measurement set and the mean of all other sets) of data, each of the 4 against the rest and so on. This reduces the amount of analyses and facilitates an overall interpretation of the results.

ADDITIONAL FEATURES

The currently discussed methods assume pairwise differences that are **independent** of the individual values within the pairs. It may be desired to interpret differences **relative** to the absolute values of the measurements. This can be realized by preceding transformations of the standard [$Y=v_1-v_2$] values in the plots. Three relative transformations are proposed:

1. proportional difference with mean of both: $Y = (v_1 - v_2) / mean[v_1,v_2] = 2 * (v_1 - v_2) / (v_1 + v_2);$
2. (relative) proportion of both values, minus 1: $Y = (v_1 / v_2) - 1 = (v_1 - v_2) / v_2;$
3. proportion of one value of mean of both, minus 1: $Y = (v_1 / mean[v_1,v_2]) -1 = (v_1-v_2) / (v_1+v_2),$ half of the first one.

Of course limits have to be specified in terms of such transformed, proportional values and the analysis is performed on these values as well. The Bland Altman Plots show these transformed values rather than the untransformed ones.

REALISATION

A SAS macro (Concord) is currently under development in which these techniques already are supported and applied.

REFERENCES

1. Altman DG, Bland JM (1983). "Measurement in medicine: the analyis of method comparison studies". Statistician **32**: 307–317
2. Bland JM, Altman DG (1986). "Statistical methods for assessing agreement between two methods of clinical measurement". Lancet **1** (8476): 307–10

ANNEX

The tickmarks in the graphs below have not been automatically generated by SAS (based on arbitrary minimum and maximum values with many digits), but by the SAS macro Tickmark, also by the author. SAS just generates its plot range from the points, while, particularly for the Y-axis, the initial value range should also include the (calculated) CI of points and means and the user specified equivalence limits. This might make the initial range larger than just based on points.

The macro Tickmark automatically generates neat tickmarks for graphs based on a given minimum and maximum of an existing value range. The generated tickmarks will have 1 to 2 significant digits and equal intervals. Optionally the desired minimum and maximum number of tick marks and minimum percentage of coverage of the existing data range by the generated value range may be specified; their default values are: minimum=7, maximum=12, percentage coverage=80. From-, To- and By-values are returned via specified, existing macro variables or as a single return value (From-value TO To-value BY By-value).

The macro Tickmark is called from macro Concord with explicit, deviant values for the minimum number of tickmarks (7 for the X-axis, 6 for the Y-axis), for the maximum number of tickmarks (12 for the X-axis, 10 for the Y-axis) and for the percentage of coverage (initial value range divided by the generated value range as a percentage) (75 for both axes).

The (auxiliary) macro Tickmark can be obtained from: http://jim.groeneveld.eu.tf/software/SASmacro/TickMark.zip

CONTACT INFORMATION

Author Name :	Y. (Jim) Groeneveld	Work Phone :	+31 (0)73 523 6000
Company :	OCS Consulting	Fax :	+31 (0)73 523 6600
Address :	P.O. Box 3434	Email :	SASquestions@ocs-consulting.com
	5203 DK 's-Hertogenbosch	Web :	http://www.ocs-consulting.com/nl
	The Netherlands	Author's website :	http://jim.groeneveld.eu.tf

GRAPHS

Figure 1. Conventional Bland Altman Plot example (same data as in figure 2).

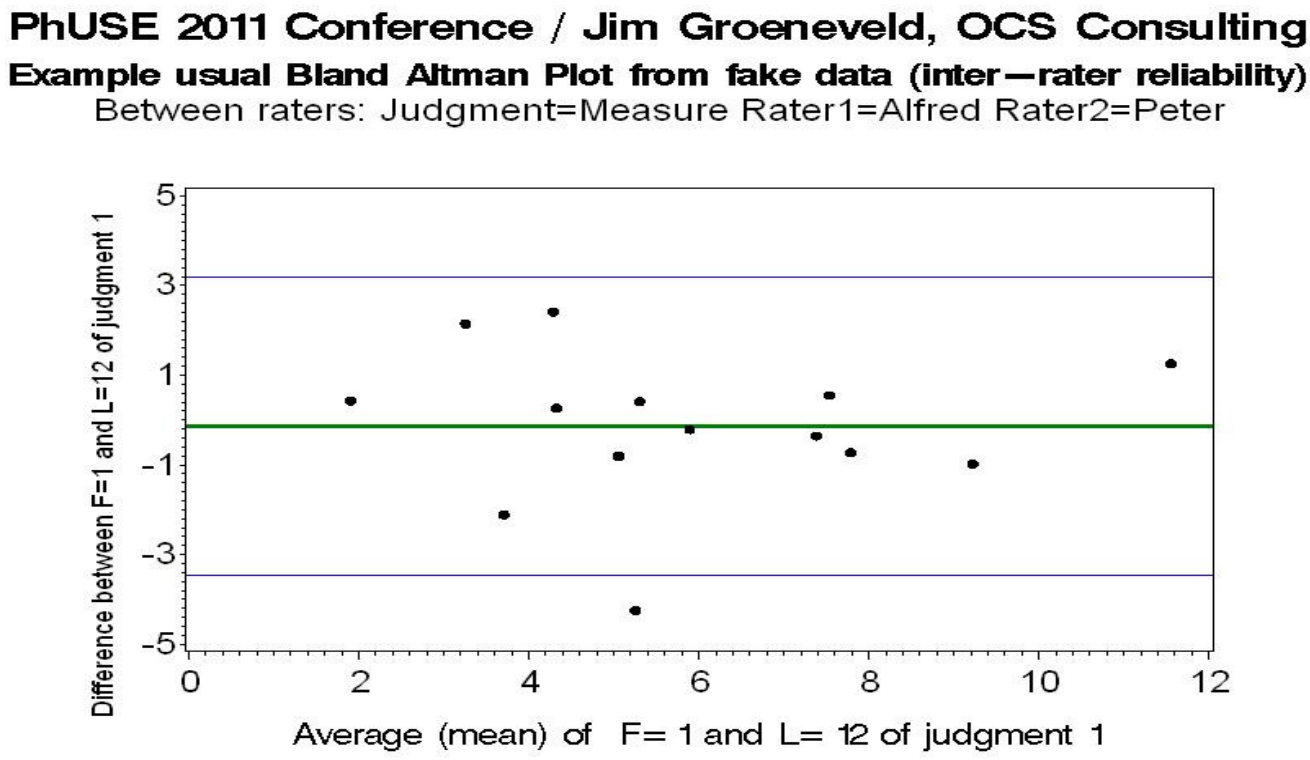

Figure 2. Integrated equivalence t-test result with Bland Altman Plot, example showing significance.

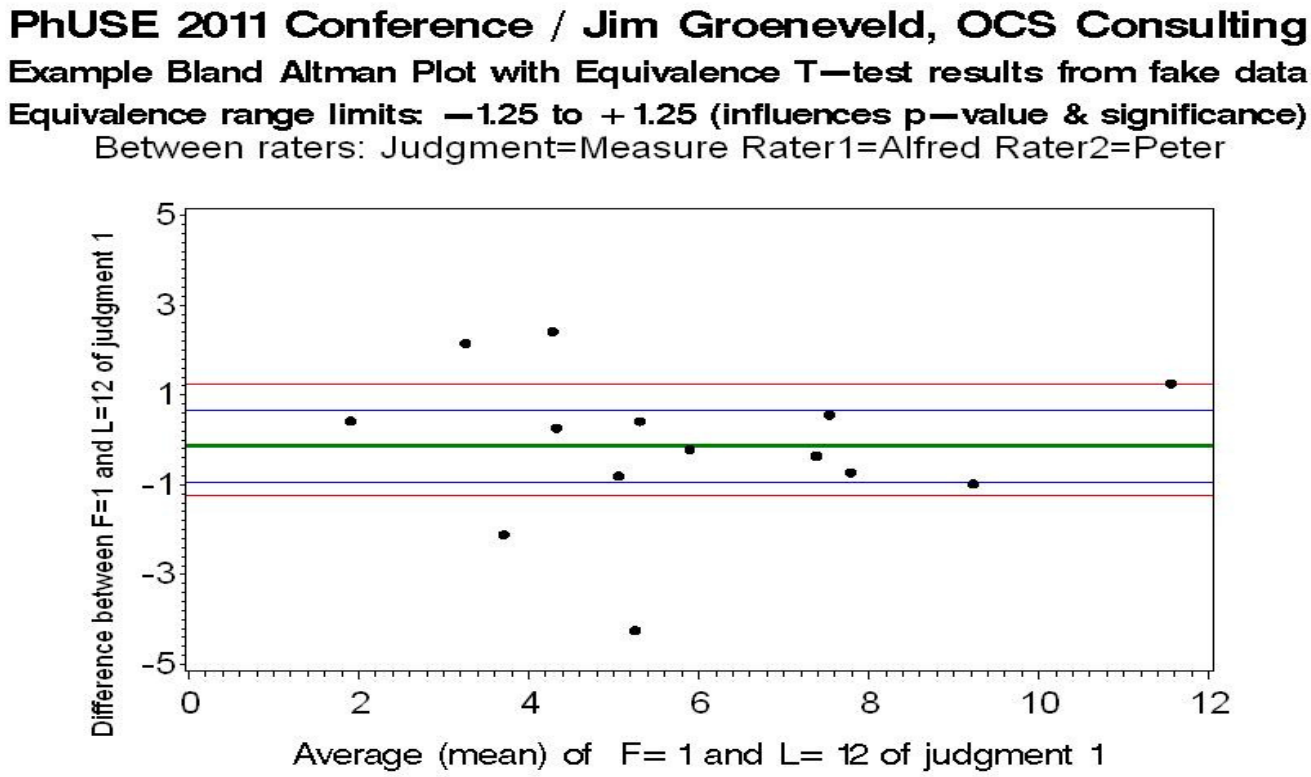

Figure 3. Integrated equivalence t-test result with Bland Altman Plot, example showing insignificance.