Introduction On Assessing Agreement With Continuous Measurement

Huiman X. Barnhart, Michael Haber, Lawrence I. Lin

1 Introduction

In social, behavioral, physical, biological and medical sciences, reliable and accurate measurements serve as basis for evaluation. As new concepts, theories and technologies continue to develop, new scales, methods, tests, and instruments for evaluation become available for measurement. As errors are inherent in every measurement procedure, one needs to ensure that the measurement is reliable and accurate before practical use. The issues related to "reliable and accurate measurement" have evolved over several decades dating back to Fisher (1925): from intraclass correlation coefficient (ICC) that measures reliability (Fisher, 1925; Bartko, 1966, Shrout and Fleiss, 1979; Vangeneugden et al., 2004), design of reliability studies (Donner, 1998; Dunn, 2002; Shoukri et al., 2004) to generalizability extending the concept of ICC (Cronback, 1951; Lord and Novick, 1968; Cronback et al., 1972; Brennan, 2001; Vangeneugden et al., 2005); from ISO's (International Organization for Standardization) (1994) guiding principle on accuracy of measurement (ISO 5725-1) to FDA's (Food and drug Administration) guidelines (1999) on bioanalytical method validation (1999), and to various indices to assess the closeness (agreement) of observations (Bland and Altman, 1986, 1995, 1999; Lin, 1989, 2000, 2003; Lin et al., 2002; Shrout, 1998; King and Chinchilli, 2001a; Dunn, 2004; Carrasco and Jover, 2003; Choudhary and Nagaraja, 2004; Barnhart et al., 2002, 2005; Haber and Barnhart, 2006). In the simplest intuitive terms, reliable and accurate measurement may simply mean that the new measurement is the same as the truth or agrees with the truth. Oftentimes, it is not practical to require the new measurement to be identical to the truth either because (1) we are willing to accept a measurement up to some tolerable (or acceptable) error or (2) the truth is simply not available to us because either it is not measurable or it is only possible to measure with some error. To deal with issues related to both (1) and (2), different concepts, methods, or theories have been developed in different disciplines. For continuous measurement, the related concepts are accuracy, precision, repeatability, reproducibility, validity, reliability, generalizability, agreement, etc. Some of these concepts, e.g., reliability, have been used across different disciplines. However, other concepts, e.g., generalizability and agreement, have been limited to a particular field and may have potential use in other disciplines.

In this introduction, we elucidate and contrast the fundamental concepts used in different disciplines and bring these concepts into one common theme: assessing closeness (agreement) of the observations. We focus on measurements that are continuous and summarize the methodological approaches on how these concepts and methods are expressed mathematically and for what data structures they are used for both cases with and without reference (or truth). Existing approaches for expressing agreement were divided in terms of following: (1) descriptive tools such as pairwise plots with a 45 degree line and Bland and Altman plots (Bland and Altman, 1986); (2) unscaled summary indices based on absolute differences of measurements, such as systematic bias, precision, limits of agreement (Bland and Altman, 1999), repeatability coefficient, mean squared deviation, coverage probability, and total deviation index (Lin et al, 2002); (3) scaled summary indices attaining values between -1 and 1, such as the intraclass correlation coefficient, concordance correlation coefficient, dependability coefficient, generalizability coefficient, and coefficients of individual agreement. These approaches were developed for one or more types of the following data structure: (1) two or more observers without replications; (2) two or more observers with replications; (3) one or more observer is treated as a random or fixed reference; (4) longitudinal data where observers take measurements over time; (5) covariates are available for assessing the impact of various factors on agreement measures. We discuss the interpretation of the magnitude of the agreement values on using the measurements in clinical practice and on study design of clinical trials. We identify gaps that require further research as well as future directions in assessing agreement. In section 2, we present definitions of different concepts used in the literatures and provide our critique. Detailed summary on statistical approaches can be found in our review paper under publication list.

2 Concepts

2.1 Accuracy and Precision

In Merriam Webster's dictionary, *accuracy* and *precision* are synonyms.The meaning of *accuracy* is defined as "freedom from mistake or error" or "conformity to truth or to a standard" or "degree of conformity of a measure to a standard or a true value". The meaning of *precision* is defined as "the quality of being exactly or sharply defined" or "the degree of refinement with which a measurement stated". The "degree of conformity" and "degree of refinement" may mean the same thing. The subtle difference between these two terms may lie in whether a truth or a standard is required or not.

Accuracy

Historically, accuracy was used to measure systematic bias and precision was used to measure random error around the expected value. Confusions in using these two terms continue till today because different definitions exist and sometimes these two terms were used interchangeably. For example, FDA (Food and Drug Administration) guidelines on bioanalytical method validation (1999) defined *accuracy* as the closeness of mean test results obtained by the method to the true value (concentration) of the analyte. The deviation of the mean from the true value, i.e., systematic bias, serves as the measure of accuracy. However, ISO (the International Organization for Standardization) in 1994 used accuracy to measure both systematic bias and random error. In ISO 5725 (1994), the general term *accuracy* was used to refer to both trueness and precision where "trueness" refers to the closeness of agreement between the arithmetic mean of a large number of test results and the true or accepted reference value and "precision" refers to the closeness of agreement between test results. In other words, accuracy involves both systematic bias and random error because "trueness" measures the systematic bias. The ISO 5725 (1994) acknowledged that

"The term accuracy was at one time used to cover only the one component now named trueness, but it became clear that to many persons it should imply the total displacement of a result from a reference value, due to random as well as systematic effects. The term bias has been in use for statistical matters for a very long time, but because it caused certain philosoph-

ical objections among members of some professions (such as medical and legal practioners), the positive aspect has been emphasized by the invention of the term "trueness"".

Despite the ISO's effort to use one term, accuracy, to measure both systematic and random errors, it remains popular today so that the literature continues to use accuracy for measuring the systematic bias and precision for measuring the random error in medical and statistical research. For this reason, we will use *accuracy* to stand for systematic bias in this paper, where one has a "true sense of accuracy" (systematic shift from truth) if there is a reference and a "loose sense of accuracy" (systematic shift from each other) if there is no reference used for comparison. Thus, the "true sense of accuracy" used in this paper corresponds to FDA's accuracy definition and the ISO's trueness definition. Ideally and intuitively, the accepted reference value should be the true value because one can imagine that the true value always exist and true value should be used to judge whether there is an error. However, in social and behavior sciences, the true value may be an abstract concept, for example, characteristic or construct such as intelligence, that may only exist in theory, and one may not be able to measure it directly. In biomedical sciences, the true value may be measured with so called "gold standard" that may also contain small amount of systematic and/or random error. Therefore, it is very important to report the accepted reference, whether it is the truth or subject to error (including the degree of systematic and random error if known). In this paper, we only consider the case where the reference or gold standard is measured with error.

Precision

FDA (1999) defined *precision* as the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogeneous sample under the prescribed conditions. Precision is further subdivided into within-run, intra-batch precision or repeatability, which assesses precision during a single analytical run, and between-run, inter-batch precision or repeatability, which measures precision with time, and may involve different analysts, equipment, reagents, and laboratories.

ISO 5725 (1994) defined *precision* as the closeness of agreement between independent test results obtained under stipulated conditions. ISO defined repeatability and reproducibility as precision under the repeatability and reproducibility conditions (see section 2.2), respectively.

The key word here is "under the prescribed conditions" or "under stipulated conditions". Therefore, it is very important to emphasize the conditions used when reporting precision. Precisions are only comparable under the same conditions.

2.2 Repeatability and Reproducibility

Repeatability and reproducibility are two special kinds of precision under two extreme conditions. If precision is expressed by imprecision such as standard deviation, repeatability will be always smaller than or equal to reproducibility (see below for definition).

Repeatability

FDA (1999) used term *repeatability* for both intra-batch precision and inter-batch precision. ISO defined *repeatability* as the closeness of agreement between independent test results under repeatability conditions that are as constant as possible, where independent test results are obtained with the same methods on identical test items in the same laboratory by the same operator, using the same equipment within "short" intervals of time.

We use the ISO's definition on repeatability in this paper. To define the term more broadly, *repeatability* is the closeness of agreement between measures under the "same condition", where "same condition" means that nothing changed other than the times of the measurements. The measurements taken under the "same condition" can be viewed as true replicates.

Sometimes the subject does not change over time, such as x-ray slides or blood samples. However, in practice it may be difficult to maintain the "same condition" over time when the measurements are taken. This is especially true in social and behavior sciences where characteristic or construct change over time due to learning effect. It is important to make sure that observers (if they are human beings) are blinded to their earlier measurements of the same quantity. Oftentimes we rely on *believable assumptions* that the "same condition" is maintained over a short period of time when the measurements are taken. It is very important to state *what assumptions* are used when reporting repeatability. For example, when an observer uses an instrument to measure a subject's blood pressure, the "same condition"means the same observer using the same instrument to measure the same subject's blood pressure where the subject's blood pressure did not change when multiple measurements were taken. It is hard to believe that the subject's blood pressure remains constant over time. However, it is believable that the true blood pressure did not change over a short period time, e.g., a few seconds. Therefore, blood pressures taken in successive seconds by the same observers using the same instrument on the same subject may be considered as true replicates.

It is important to report repeatability when assessing measurement because it measures the purest random error that is not influenced by any other factors. If true replicates can not be obtained, then we have loose sense of repeatability that is based on assumptions.

Reproducibility

FDA in 1999 guideline defined *reproducibility* as the precision between two laboratories. It also represents precision of the method under the same operating conditions over a short period of time. ISO in 1994 defined *reproducibility* as the closeness of agreement between independent test results under reproducibility conditions under which results are obtained with the same method on "identical" test items, but in different laboratories with different operators and using different equipment.

We use the ISO's definition on reproducibility in this paper. To define the term more broadly, *reproducibility* is the closeness of agreement between measures under the "all possible conditions" on "identical" subjects on which the measurements are taken. "All possible conditions" are any conceivable situations under which a measurement will be taken in practice. The conceivable situations include, but not limited to, for example, different laboratories, different observers, etc. However, if multiple measurements on the same subject can not be taken at the same time, one needs to ensure that the subject under measurement, e.g, a subject's blood pressure, does not changed over time when the measurements are taken in order to assess reproducibility.

2.3 Validity and Reliability

The concepts of accuracy and precision were originated from physical science where measurement can be taken directly. Similar concepts, validity and reliability, existed in social science where a reference is required for validity and a reference is not necessarily required for reliability. As elaborated below, Validity is similar to "true sense of agreement" with both good "true sense of accuracy" and precision. The reliability is similar to "loose sense of agreement" with both good "loose sense of accuracy" and precision. Historically, the validity and reliability are assessed via scaled indices.

Validity

In social, educational and psychological testing, *validity* refers to the degree to which evidence and theory support the interpretation of measurement (AERA, 1999). Depending on the selection of the accepted reference (criterion or "gold standard"), there are several types of validity such as *content, construct, criterion* validity (Goodwin, 1997; AERA, 1999; Kraemer et al., 2002; Hand, 2004; Molenberghs, et al., 2007). The *content* validity is defined as the extent to which the measurement method assesses all the important content. The *face* validity is similar to the *content* validity that is defined as the extent to which the measurement method assesses the desired content at face. The face validity may be determined by judgment of experts in the field. The *construct* validity is used when we are trying to measure something that is a hypothetical construct that may not be readily observed such as anxiety. The *convergent* and *discriminant* validity may be used to assess construct validity by showing that the new measurement is correlated with other measurements of the same construct and that the proposed measurement is not correlated with the unrelated construct, respectively. The *criterion* validity is further divided into *concurrent* and *predictive* validity where the criterion validity deals with correlation of the new measurement with a criterion measurement (such as gold standard) and the predictive validity deals with the correlation of the new measurement with a future criterion, such as clinical endpoint.

Validity is historically assessed by the correlation coefficient between the new measure and the "reference" (or construct). If there is no systematic shift of the new measure from the reference or construct, this correlation may be expressed as the proportion of the observed variance that reflects variance in the construct the instrument/measurement method was intended to measure (Kraemer et al, 2002). For validation of bioanalytical method, FDA (2001) provided guideline on full validation that involves parameters such as (1) accuracy, (2) precision, (3) selectivity, (4) sensitivity, (5) reproducibility, and (6) stability when a reference is available. When the type of validity is concerned with closeness (agreement) of the new measure and the reference, we believe that an agreement index is better suited than the correlation coefficient in assessing validity.

Reliability

The reliability concept has been evolved over several decades that was initially developed in social, behavior, educational and psychological disciplines and was later widely used in other disciplines such as physical, biological and medical sciences (Fisher, 1925; Bartko, 1966, Lord and Novick, 1968; Shrout and Fleiss, 1979; Müller and Büttner, 1994; McGraw and Wong, 1996; Shrout, 1998; Donner, 1998; Shoukri et al., 2004; Vangeneugden et al., 2004). Rather than reviewing everything in the literature, we provide our point of view in its development. *Reliability* is originally defined as the ratio of true score variance to the observed total score variance in classical test theory (Lord and Novick, 1968; Cronbach et al, 1972). It is interpreted as the percent of observed variance explained by the true score variance. It was initially intended to assess the measurement error if observer takes the measurement repeatedly on the same subject under identical conditions or to measure the consistency of two readings obtained by two different instruments on the same subject under identical conditions. If the true score is the construct, then reliability is similar to the criterion validity. In practice, the true score is usually not available and in this case, reliability represents the scaled precision. Oftentimes, reliability is defined with additional assumptions. The following three assumptions are inherently used and usually are not stated when reporting reliability.

(a) True score exists but not directly measurable

(b) The measurement is the sum of the true score plus a random error where random errors have mean zero, are uncorrelated with each other and with the true score (both within and across subjects)

(c) Any two measurements for the same subject are *parallel* measurements.

The *parallel* measurements here mean that any two measurements for the same subject have the same means and same variances. With assumptions (a) and (b), reliability, defined above as ratio of variances, is equivalent to the *square of the correlation coefficient* between the observed reading and the true score. With assumptions (a) through (c), reliability defined above is equivalent to the correlation of any two measurements on the same subject. This correlation is called *intraclass correlation* (ICC) that dates back to Fisher (1925) from the study of fraternal resemblance in genetics. Parallel readings are considered to come from the same class and they can be represented by a one way ANOVA model (see Section 3). The reliability expressed in terms of ICC is the most common parameter used across different disciplines. Different versions of ICC for assessing reliability existed and were advocated (Bartko, 1966; Shrout and Fleiss, 1979; Müller and Büttner, 1994; McGraw and Wong, 1996) when different ANOVA models were used in place of assumptions (b) and (c).

2.4 Dependability and Generalizability

Recognizing that the assumptions (b) and (c) in the classical test theory are too simplistic, generalizability theory (GT) (Cronbach et al., 1972; Shavelson et al, 1989; Shavelson and Webb, 1981, 1991, 1992; Brennan, 1992, 2000, 2001) has emerged. GT is widely known and used in educational and psychological testing literature. However, it is barely used in the medical research even though there were many efforts attempted for its broader use since the introduction by Cronbach et al. (1972), This may be due to the overwhelming statistical concepts involved in the theory and limited number of statisticians who have worked in this area. Recently, Vangeneugden et al. (2005) and Molenberghs et al. (2007) have presented linear mixed model approaches to estimate reliability and generalizability in clinical trial setting.

GT extends the classical test theory by decomposing the error term into multiple sources of measurement errors and thus relax the assumption of parallel readings. The concept of reliability is then extended to the general concept of generalizability or dependability within the context of GT. In general, two studies (G-study and D-study) are involved where the G-study is aimed at estimating the magnitudes of variances due to multiple sources of variability through an ANOVA model and the D-study uses some or all sources of variability from the G-study to define specific coefficients which generalizes reliability coefficient depending on the intended decisions. In order to specify a G study, the researcher needs to define the universe of generalizability *a priori*. The universe of generalizability contains the factors with several levels/conditions (finite or infinite) so that the researchers can establish interchangeability of these levels. For example, suppose there are J observers and the researcher wants to know whether the J observers are interchangeable in terms of using a measurement scale on a subject. The universe of generalizability would include observer as a factor with J levels. This example corresponds to the single facet design. The question of reliability among the J observers thus become the question of generalizability or dependability of the J observers. To define the generalizability coefficient or dependability coefficient, one needs to specify a D study and the type of decision. The *generalizability coefficient* involves the decision based on relative error, i.e., how subjects are ranked according to J observers regardless of the observed score. The *dependability coefficient* involves the decision based on absolute error, i.e. how the observed measurement differs from the true score of the subject. To specify a D study, the researcher would also need to decide how he or she is planning to use the measurement scale, e.g., does he/she want to use the single measurement taken by one of J observers, or the average measurement taken by all J observers. Different decisions will result in different coefficients of generalizability and dependability.

2.5 Agreement

Agreement measures the "closeness" between readings. Therefore, agreement is a **broader** term that contains both accuracy and precision. If one of the readings is treated as the accepted reference, the agreement is concerning validity. If all of the readings can be assumed to come from the same underlying distribution, then agreement is assessing precision around the mean of the readings. When there is a disagreement, one needs to know if the disagreement sources were systematic shift (bias) or random error. This is important because a systematic shift (inaccuracy) usually can be easily fixed through calibration, while a random error (imprecision) usually is a more cumbersome exercise of variation reduction.

In an absolute term, readings agree only if they are identical and they disagree if they are not identical. However, readings obtained on the same subject or materials under "same condition" or different conditions are not, in general, identical due to unavoidable errors in every measurement procedure. Therefore, there is a need to quantify the agreement or the "closeness" between readings. This is best based on the distance between the readings. Therefore, measures of agreement are often defined as functions of the absolute differences between readings. This type of agreement is called *absolute agreement.* The absolute agreement is a special case of the concept of *relational agreement*, introduced by Stine (1989). In order to define a coefficient of relational agreement, one first needs to define a class of transformations that is allowed for agreement. For example, one can decide that observers are in agreement if the scores of two observers differ by a constant and then the class of transformation consists of all the functions that add the same constant to each measurement. This corresponds to additive agreement. Similarly, in the case that the interest is in linear agreement, observers are said to be in agreement if the scores of one observer are a fixed linear function of those of the other. In most cases, however, one would not tolerate any systematic differences between observers. Hence, the most common type of agreement is absolute agreement. In this paper, we only discuss indices based on absolute agreement and direct the readers to the literature (Zeger, 1986; Stine, 1989; Fagot, 1993; Haber and Barnhart, 2006) for relational agreement.

We note that the concepts of agreement and reliability are different. As pointed out by Vangeneugden et al. (2005) and Molenberghs et al. (2007), the agreement assesses the degree of closeness between readings within a subject while reliability assesses the degree to differentiate subjects from a population. It is possible that in homogeneous populations, the agreement is high but the reliability is low, or in heterogeneous populations, the agreement is low but the reliability is high. However, when a scaled index is used to assess agreement, values based on the agreement and reliability indices can be very similar.

Assessing agreement is often used in medical research for method comparisons (Bland and Altman, 1986, 1995, 1999; Laurent, 1998, Dunn, 2004), assay validation and individual bioequivalance (Lin, 1989, 1992, 2000, 2003; Lin et al., 2002). We note that while the concepts of agreement and reliability are different, they are closely related. As indicated in our review paper, concordance correlation coefficient (CCC) is a popular index for assessing agreement that is scaled index of assessing difference between observations. In comparison of the CCC and the ICC, the CCC reduces to ICC under the ANOVA models used to define the ICC. Therefore, reliability assessed by ICC is a scaled agreement agreement index under ANOVA assumptions. However, if the agreement is assessed by unscaled indices, it is possible that in homogeneous populations, the agreement is high but the reliability is low, or in heterogeneous populations, the agreement is low but the reliability is high. This is because the scaled agreement indices often depend on between-subject variability and as a result they may appear to assess the degree to differentiate subjects from a population (Vangeneugden et al., 2005; Molenberghs et al., 2007).

In summary, the concepts used to assess reliable and accurate measurements have a common theme: assessing the closeness (agreement) between observations.