

Take measurement, reliability and validity seriously

Steve Kan, STSM, Technical Manager, iSeries Software Quality

April 2005

There are a set of desirable attributes that distinguish good metrics from bad ones. **Reliability** and **validity** are the two most important issues of measurement quality. These two issues should be well thought through before a metric is proposed, used, data being gathered and results analyzed.

- **Reliability** refers to the consistency of a number of measurements taken using the same measurement method on the same subject. If repeated measurements are highly consistent (or even identical), then there is a high degree of reliability with the measurement method or the operational definition. If the variations among repeated measurements are large, then reliability is low.

For example, if an **operational definition** of a body height measurement of children includes the following and actual measurement taking is done right, it is likely that the measurement data this obtained is **reliable**:

- specifications of the time of the day to take measurements
- the specific scale to use
- who takes the measurements
- checking the scale before taking measurement
- whether the measurements should be taken barefooted

On the other hand, if the operational definition is very vague in terms of these considerations, or the actual implementation (measurement taking) is poorly done, the data **reliability** may be low. Measurements taken in the early morning may be greater than those taken in the late afternoon as children's bodies tend to be more stretched after a good night's sleep and become somewhat compacted after a tiring day. Indeed, recent studies by the National Institute of Health (NIH) indicated that for males and females under the age of 45, one's body height could vary by as much as half of an inch within the day, depending on the time of the day. By the same token, different scales used, trained versus untrained personnel, with or without shoes on, etc., are factors that can contribute to the **variations** of the measurement data.

- **Validity** refers to whether the measurement or metric really measures what we intend it to measure. In other words, it refers to the extent to which an empirical measure adequately reflects the real meaning of the concept under consideration. In cases where the measurement involves no higher level of abstraction, for example, the measurements of body height and weight, validity is simply accuracy. However, validity is different from reliability. Measurements that are reliable may not necessarily be valid, and vice versa. For example, a new bathroom scale for body weight may give identical results upon many consecutive measurements and therefore it is reliable. However, the measurements may not be valid to reflect the person's body weight if the offset of the scale was at a number other than zero. For abstract concepts, validity can be a very difficult issue. For instance, the use of church attendance for measuring religiousness in a community may have low validity because religious persons may or may not always go to church. On the other hand,

the measurement of church attendance may be quite reliable because it is specific and observable, and people tend to give honest answers.

Researchers tend to classify validity into several types.

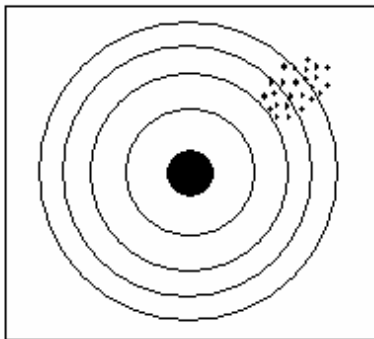
- **Construct validity:** The type of validity we have discussed so far is called construct validity, which refers to the validity of the operational measurement or metric representing the theoretical construct.
- **Predictive validity:** One other type of validity is predictive validity. For example, the validity of a written driver's test is determined by the relationship between the scores people get on the test and how well they drive. Predictive validity is also applicable to modeling, which we will discuss later on with regard to software reliability models or defect projection.

Given a theoretical construct, the **purpose of measurement is to measure the construct validly and reliably**. The figure below graphically portrays the difference between validity and reliability.

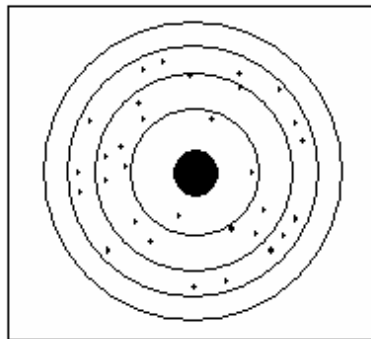
If the purpose of the measurement is to hit the center of the target, we see that reliability looks like a tight pattern regardless of where it hits, because **reliability is a function of consistency**.

Validity, on the other hand, is a **function of shots being arranged around the bull's eye**. In statistical terms, if the expected value is the bull's eye, then it is valid; if the variations are small relative to the entire target, then it is reliable.

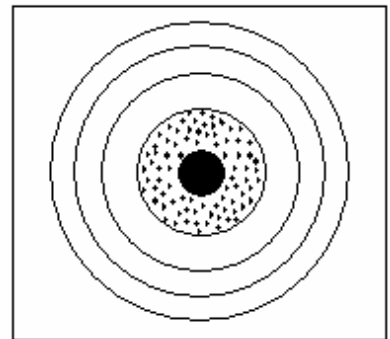
An Analogy to Validity and Reliability



Reliable but not valid



Valid but not reliable



Valid and reliable

Source: Kan, S. H. Metrics and Models in Software Quality Engineering (Addison-Wesley, 2002)

There are many problems in software metrics and measurements. Many of these problems are related to reliability and validity issues. To improve, we must ask ourselves:

What does this metric measure – what are we trying to accomplish?
Is the measurement data reasonably reliable for decision making?

Food for thought:

1. Discuss the reliability and validity issues associated with the LOC measurements. How do we improve LOC measurement reliability? What exactly do we want LOC to measure?
2. Pick one metric that is being used in your organization, evaluate the metric from the perspectives of reliability and validity and come up with a set of improvement recommendations.