Quality Control in Testing

Mark H. Stone, Ph.D.

Adler School of Professional Psychology

Introduction

Quality assurance in testing has been approached in two very general ways. The first has been to assure that test materials are only available to those with the appropriate level of education and training. Most publishers of assessment tools and tests use criteria to assure that only appropriately trained persons have access to their instruments. The second has been to recommend test development procedures following the *Standards for Educational and Psychological Testing* (AERA, 1999). Both approaches are necessary, but not sufficient to assure that quality can be maintained in testing. Neither one provides assurance that test variables are adequately built and maintained. Neither method meets quality assurance standards. To do so requires specific attention to quality control (Stone, 2000). Interestingly, the index to the *Standards* does not include quality or quality control as headings, but statistical quality control has long been employed to assure the highest standards in manufacturing goods.

The earliest and most systematic exposition of quality control was given by Walter Shewhart of Bell Laboratories (1931, 1986). His efforts have been propagated through the lectures and writings of W. Edwards Deming also well-known for his work in quality control. The problem of quality control in testing has been frustrated by several fundamental conceptual issues. The first is addressed by Deming in his introduction to the reprint of Shewhart's *Statistical Method* (1986). Deming writes:

> There is no true value of anything. There is, instead, a figure that is produced by application of a master or ideal method of counting or measurement. This figure may be accepted as a standard until the method of measurement is supplanted by experts in the subject matter with some other method and some other figure. (p. ii)

Deming goes on to point out that all values and constants are in error because they are conditioned by the methods of their determination. "Every observation, numerical or otherwise, is subject to variation" (1986, p. ii). However, there is useful information in variation. The issue is not just error, but control over error. The second issue raised is the need for a method for establishing quality control. Error must be brought under control if the resulting values are to have any practical use. Shewhart's model for statistical control over error requires answers to these five questions:

1. How are the observations to be made?
2. How are the samples to be drawn?
3. What is the criterion for control?
4. What action will be taken as a consequence?
5. What quantity of data is required?

He arranged these questions into a dynamic model (see next page):

Figure 1. Shewhart's Dynamic Model

SPECIFICATION -----> PRODUCTION ------> INSPECTION

(Their arrangement is best seen as a circle — or even better, as a continuing spiral)

Accuracy and Precision

Quality control in testing requires addressing accuracy and precision. The concepts of accuracy and precision in testing and measurement are called validity and reliability. The first and most important matter is to determine what any concept means. Bridgman (1928) specified what has come to be known as an "operational definition." Such a definition serves as the mechanism for understanding what a concept means. He indicated that, "The concept is synonymous with the corresponding set of operations" (p. 5). Concepts are defined, explicitly or implicitly, by a methodology. A concept equals the method that describes it and vice versa. Concepts without methods are nonsense and the bantering about of concepts without considering methods is irresponsible and not scientific. Therefore, concepts such as accuracy and precision, validity and reliability cannot be separated from the methods of their determination. To be specific, we cannot speak of validity or reliability, but only of some method of determining validity or reliability specified about some occasion.

A specific example of the confusion over concepts can be observed when reading reports of a test's "validity." The determination of validity is situational and not extensive. The validity of the test is conditioned by a point in time and the setting in which it took place, i.e., the methodology and sample producing the value(s). It is an inferential leap to assume that what occurred in one circumstance has any application to another circumstance and it is even less likely to expect it apply to every other instance or to all other circumstances. Concepts such as validity and reliability require more careful inquiry. Specifically, a determination of validity or reliability needs to be operationally decomposed into two important aspects: the contribution from the items, and contribution from the persons. Typically, and all too often, studies of test validity and reliability fail to provide any coefficient resulting from use of the sample.

For determining reliability, the KR20 is often calculated for items, but almost never for persons. Hoyt (1941) recognized both approaches, saying that "extended examination of the 'among items' variance would make it possible to decide on the heterogeneity of the respective difficulties of the items while a more extended examination of the 'among students' variance would make it possible to answer certain pertinent questions regarding the individual differences among students" (p. 41). His good advice is almost never followed. Jackson (1939), Hoyt (1941), Alexander (1947), and Guilford (1954) have all proposed an analysis-of-variance approach to estimate reliability. The advantage of this strategy is that "test reliability" can be decomposed into the variance due to examinees, the variance due to items and the remainder or error variance. This more complete analysis is in keeping with a quality control process in testing.

Wright and Stone (1999) have demonstrated that these matters can be even better accomplished using Rasch measurement techniques which are explained in *Best Test Design* (Wright &

Stone, 1979, pp. 151-166) and all of the analyses discussed below can be produced using WINSTEPS (Linacre, 2000). The shortcomings of using raw scores are remedied when a Rasch measurement analysis is made of the same data and reliability is calculated from Rasch values. In addition, Rasch measurement provides the standard errors for every person and item. These individual errors can be squared and summed to produce a correct average error variance for the sample or any subset of persons and for the items or any subset of items. When these results are substituted for those in the traditional KR20 formula, the result is a new formula, equivalent in interpretation, but giving a better estimate of reliability than any other value produced by using raw scores. Deming's adage of progressive improvement by better methods in quality control is clearly demonstrated through applying these methods. Shewhart (1986) also spoke of predication as an important aspect of quality control. "Every meaningful interpretation involves a prediction" (p. 92) and "Knowledge in this sense is a process or a method of predicting an ideal" (p. 104). The element of prediction makes scientific results useful. In the application of a test, it is the characteristics of the new sample to which we intend to apply the test, rather than simply the description of a previous sample, that is our focus. We want to know how the test will work with the new samples who are about to take it, not old history. We want a relevant reliability coefficient which applies to the people we intend to test, not one that only describes the people who were previously tested. But we can actually predict the reliability for a new sample if we postulate the mean and variance for that sample. One can use these statistics and the Rasch targeting formula to calculate the reliability of the test in its new application. (See Wright & Stone 1979, 129-140.)

Deming, as quoted above, indicated that new methods can supplant old ones when they provide better methods and values. The Rasch separation index is such a method for producing a more useful value. Correlation-based reliability coefficients are nonlinear. The increase in reliability from .5 to .6 is not twice the improvement in reliability from .9 to .95. In fact, the increase from .9 to .95 is actually about twice the improvement in precision of the other. The Rasch Separation Index (G) is the ratio of the unbiased estimate of the sample standard deviation to the root mean square measurement error of the sample. It is in a ratio scale in the metric of the root mean square measurement error of the test for the sample postulated. The Separation Index quantifies "reliability" in a more direct way with a clear interpretation.

Separation G = SDT/SET

SDT = The expected SD of the target sample

SET = The test standard error of measurement for such a sample, which is almost always well approximated by SET = 2.5 / □ L

SET can also be estimated as SET = □(C/L) where L is the number of items in the test and C is a targeting coefficient (see Wright & Stone, 1979, pp. 135-136). A figure given below expedites applying this procedure (see pp. 22-23 for remaining figures).

The *Standards* (AERA, 1999) in Section 13.14 recommend that "score reports should be accompanied by a clear statement of the degree of measurement error associated with each score" (p. 149). Rasch measurement analysis routinely provides standard errors for every possible test measure along the variable that fully meets this recommendation. If reliability, as defined by the *Standards*, is the degree to which test scores are free from errors of measurement, then it follows that every ability measure should be accompanied by a standard error as an index of the degree to which this criterion is met for that measure. Not to do so is to ignore the *Standards*.

The Rasch measurement standard errors satisfy this recommendation by providing individual errors of measurement for every observable measure. Where a collective index of reliability is

desired, the Rasch Separation Index is even more useful than the traditional indices of reliability. Figure 2 describes the Rasch analysis of a response matrix and Figure 3 describes the computation of the Rasch person separation index. The targeting coefficient C varies between 4 and 9 depending on the range of items difficulties in the intended test and the target sample's expected average percent correct on that test. Figure 4 gives some values of C for typical item difficulty ranges and typical target sample mean percents correct. However, it is not the algebraic and statistical similarity of the KR20 and the Separation Index C that is of major importance. Instead it is the decomposition of these single indices into their constituent parts leading to a more detailed and more useful management of information. Quality control is now operating.

With Rasch measurement analysis, we are able to obtain the standard error of calibration for each individual item as well as the standard error of measurement for each person ability. With traditional methods, a single standard error of measurement is provided and only for measures at the group mean of person ability. The standard error specific to each item or person statistic is far more useful than any single sample or test average.

The location of each item and person on a line representing the variable together with their standard errors provides definition and utility to the test variable. The definition of the variable is specified by the location of the items. The utility of a test variable for measuring persons is quantified by the standard error that accompanies each person measure.

A variable can be thought of as a straight line. To measure successfully we must be able to locate both items and persons along this line. A simple example is given in Figure 5. Items are located by the number of persons getting a specific items correct. Persons are located by how many items they were able to answer correctly. Items to the left side of the line are easier than those to the right while persons to the left have less ability than others to the right.

It is necessary to locate persons and items along the line of the test variable with sufficient precision to "see" between them. Items and persons must be separated along this line for useful measurement to be possible. Separation that is too wide usually signifies gaps among item difficulties and person abilities. Separation that is too narrow, however, signifies redundancy for test items and not enough differentiation among person abilities to distinguish between them. Items must be sufficiently well separated in difficulty to identify the direction and meaning of the test variable. To be useful, a selection of items, i.e., a test, must separate relevant persons by their performance. The item locations are the operational definition of the variable of interest while the person locations are the application of the variable to measurement. Such an approach meets Bridgman's requirements for an operational definition.

Conclusion

Item and person separation statistics in Rasch measurement provide analytic and quality control tools by which to evaluate the successful development of a variable and by which to monitor its continuing utility. Successful item calibration and person measurement produces a map of the test variable (Stone, Wright, & Stenner, 1999). The resulting map is no less a ruler than the ones constructed to measure length. The map indicates the extent of content, criterion, and construct validity for the test variable. Empirical calibration of items and measures of persons should correspond to the original intent of item and person placement. Changes must be made when correspondence is not achieved. Rasch measurement provides the quality control necessary in testing.

There should be continuous dialogue between the plan for the test, the items calibrations, and person measures. Test variables are never created once and for all. Continuous quality control is required in order to keep the map coherent and up-to-date. Support for reliability and validity does not rest in coefficients, but in substantiating demonstration of relevance and stable indices for items and measures. Such procedures assure quality control in maintaining the test variable and assuring its relevance.

References

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA. Author.

Alexander, H. (1947). The estimation of reliability when several traits are available. *Psychometrika*, *12,* 79-99.

Bridgman, P. (1928). *The logic of modern physics*. New York: Macmillan.

Guilford, J. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6,* 153-160.

Jackson, R. (1939). Reliability of mental tests. *British Journal of Psychology*, *29,* 269-287.

Linacre M. (2000). *WINSTEPS*. Chicago: MESA.

Shewhart, W. (1931). *Economic control of quality of manufactured products*. New York: Van Nostrand.

Shewhart, W. (1986). *Statistical method from the viewpoint of quality control*. New York: Dover.

Stone, M. (2000). *Establishing quality control in testing*. Second International Congress on Licensure, Certification and Credentialing of Psychologists, Oslo, Norway, July 20, 2000.

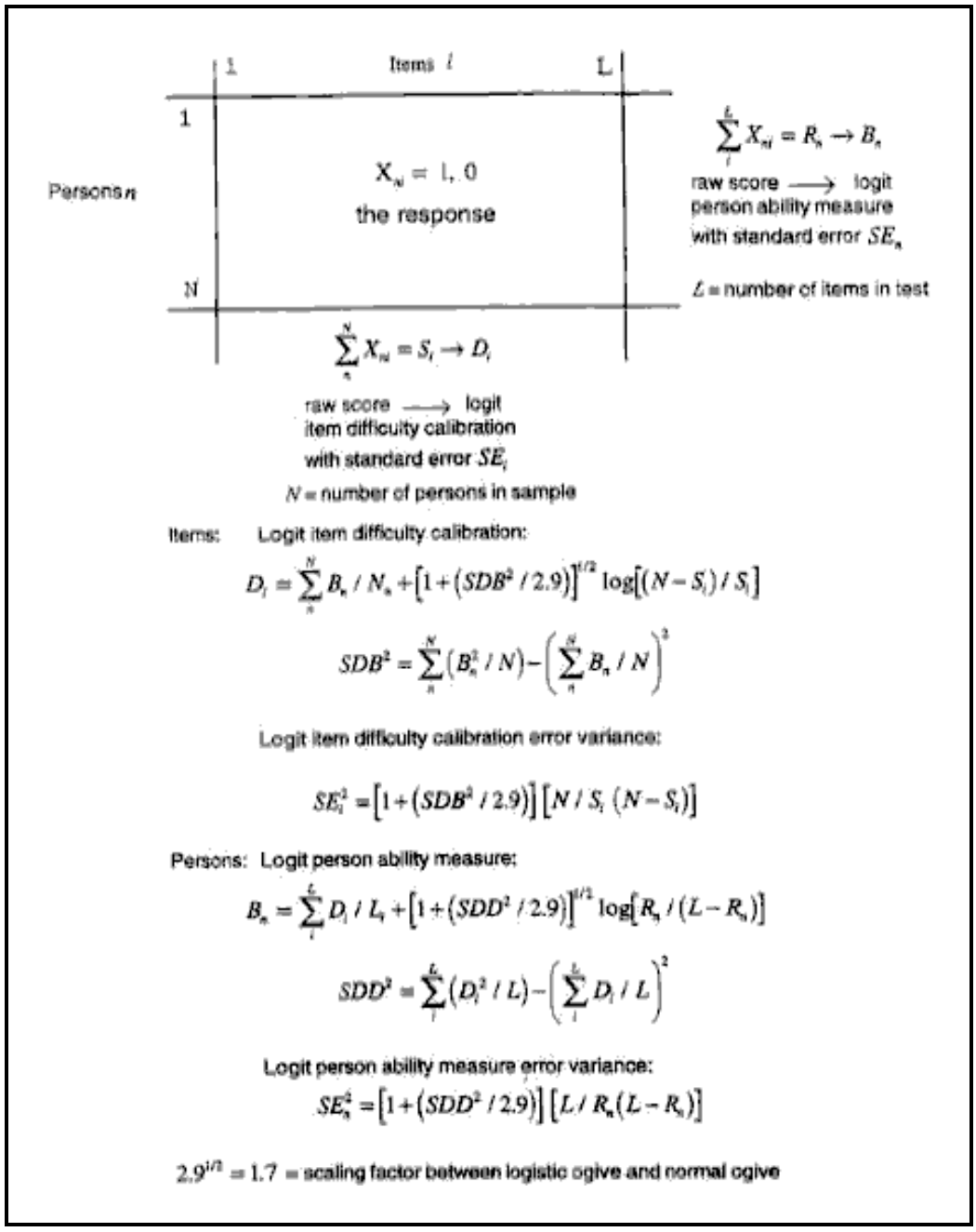
Stone, M., Wright, B., & Stenner, J. (1999). Mapping variables. *Journal of Outcome Measurement*, *3, (4),* 308-322.

Wright, B., & Stone, M. (1979). *Best test design*. Chicago: MESA.

Wright, B., & Stone, M. (1999). *Measurement essentials*. Wilmington, DE: Wide Range, Inc.

Figure 2

Rasch Analysis of Response Data



(See PROX estimation formulas, Wright and Stone, 1979, pp. 21-22)

Figure 3

Rasch Person Separation Index

$G = STB / RMSEB$ where

$STB^2 = SDB^2 - MSEB$

$$SDB^2 = \sum_{n}^{N} B_n^2 / N - \left(\sum_{n}^{N} B_n / N \right)^2$$

$$RMSEB^2 = MSEB = \sum_{n}^{N} SEB_n^2 / N$$

B_n = logit measure of person n

SEB_n = standard error of B_n

so $G^2 = R / (1 - R)$ and $R = G^2 / (1 + G^2)$

and $R = 1 - (MSEB / SDB^2)$ is

$= 1 - (VR / VS) = [(L - 1) / L] KR20$

with VR and VS as defined in Figure 19.1

note: $MSEB = C / L$ in which $4 < C < 9$
and $C = 5$ or 6 is typical.

(See Wright and Stone, 1979, pp. 134-136)

Figure 4

Values of the Targeting Coefficient C

Test Item Difficulty Range in Logits

Expected Percent Correct of Target Sample	1	2	3	4	5	6
50	4.0	4.4	4.8	5.3	5.8	6.8
60	4.4	4.4	4.8	5.3	6.2	6.8
70	4.8	5.3	5.3	5.8	6.8	7.3
80	6.2	6.8	6.8	7.3	7.8	8.4

$SET = \sqrt{C / L}$

L = Number of Items in Test

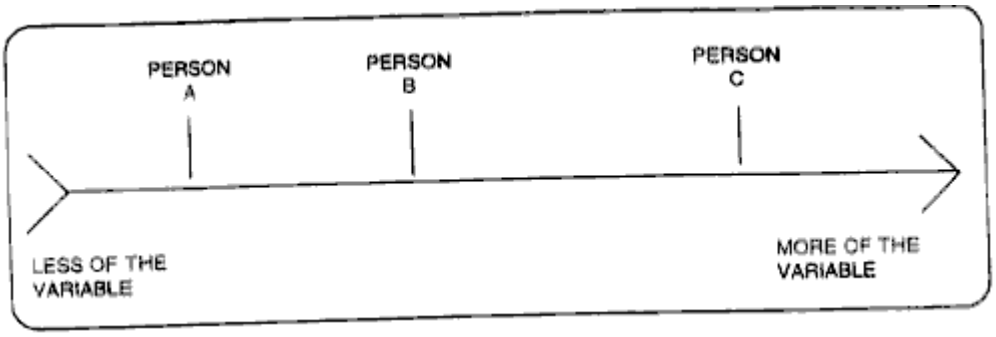
(See Wright and Stone, 1979, p. 214)

If an expected reliability is also desired, it can be obtained from: $R = G^2 / (1 + G^2)$.

Rasch Separation Indexes $G = \sqrt{[R/(1-R)]}$	*Corresponding Reliability Coefficients* $R = G^2/(1+G^2)$
1	0.50
2	0.80
3	0.90
4	0.94
5	0.96

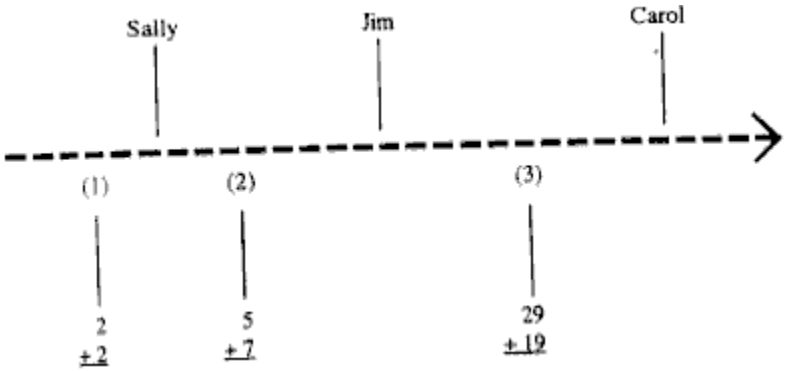
Figure 5

Positions of persons A, B, C on the line of the variable



Once the variable is constructed by the line of items, we can proceed to position students on this same line. Their probable positions can be specified initially by our best guess as to their ability to correctly answer the items which define the variable. The line of our variable shows both the positions of items and the positions of students. Eventually the positions of students will become more explicit and more empirical as we observe what items they correctly answer.

Consider this picture:



Sally's position on the variable is indicated by an expected correct response to Item 1 but expected incorrect responses to Items 2 and 3. Her differing responses to Items 1 and 2 locate her on the variable between two items that describe her ability in arithmetic computations. She can add 2 and 2 but not 5 and 7.

Jim's position is between Items 2 and 3 because we expect him to answer Items 1 and 2 correctly but not item 3. In Jim's case we have somewhat less precision in determining his arithmetic ability because f the lack of items between Items 2 and 3. If we had additional items in this region, we could obtain a more accurate indication of Jim's position on the variable as defined by his responses to these additional items.