# *19. RELIABILITY AND SEPARATION*

Validity and reliability have been key concepts in measurement for eighty years. These two topics command Chapters 1 and 2 of the Standards for Educational and Psychological Testing (1985). The Standards define reliability as "the degree to which test scores are free from errors," (1985, p.19). The "errors" referred to are measurement errors. The magnitude of these errors and the specification of their source are necessary in order to determine the efficacy of a measuring instrument. The reliability coefficient is the traditional statistic intended to quantify reliability. Coefficients are commonly reported for test-retest, multiple form and split-half replications. The purpose of this primer is to discuss how these topics are dealt with in Rasch measurement and how this improves and, hence, supersedes traditional methods.

## TRADITIONAL RELIABILITY

The KR20 for dichotomous responses (or its generalization, coefficient alpha) are estimates based upon a single administration of a test assumed to have homogeneous items. These coefficients are intended to be an estimate of the test's reliability with respect to a single attribute postulated to underlie all the test items. However, what any particular reliability actually refers to can only be whatever attribute the test items actually define. Sufficient time to answer the items is assumed (timed tests produce spuriously high coefficients). The KR20 and its variants (coefficient alpha and KR21) are calculated by comparing a numerator based on sampled item p-values with a denominator based on the sampled persons' raw scores, computed from the same response matrix of persons and items.

The statistics outlined in Figure 19.1 bring together the two contrasting elements which make up the KR20. One element summarizes the test items in terms of $pq$ in which $p$ comes from the sampled item p-values (where p = proportion correct) and $q = 1 - p$. Each item $pq$ is the variance of a response to that item for a "person" for whom that p-value is their probability of succeeding on that item. Since the $p$-value for an item is the sample mean of the dichotomous person responses to that item, this $p$-value is what we expect of an "average person" from that sample on that item.

The $p$-value for an item describes a "sample average person's" probability of success on that item and can be used to estimate an "average" sample response variance for that item. When these variances are summed over the items they yield a score variance for a "person" who has exactly those $p$-values. This "average" test score variance is the numerator in KR20.

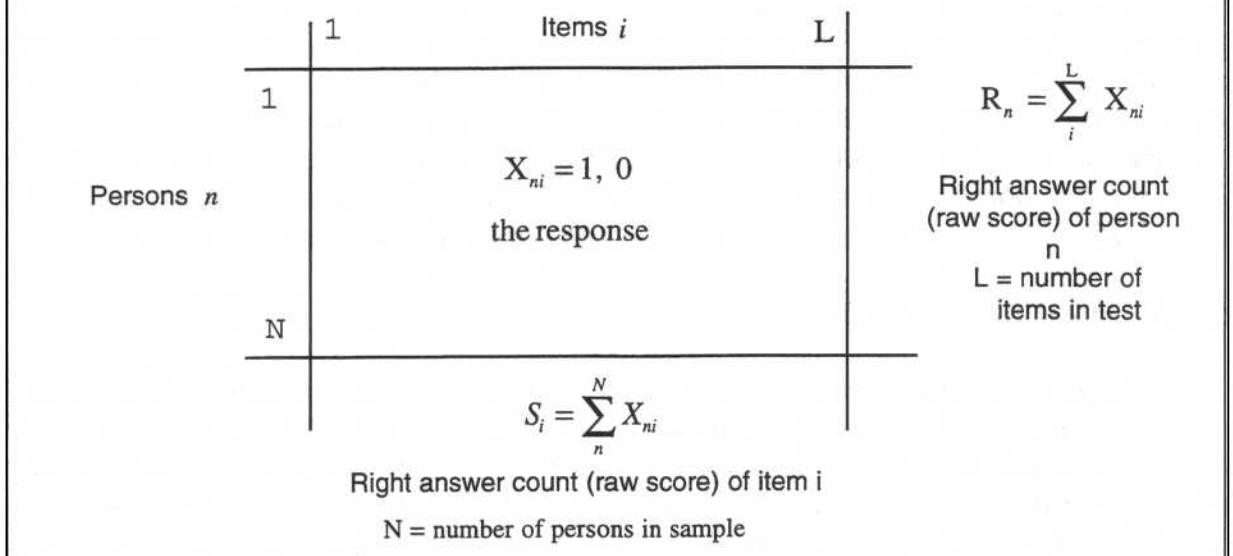
The KR20 denominator is the observed sample variance of person scores. Thus the KR20 combines a "test" characteristic for a "typical" person sampled, based on item p-values, with a "sample" characteristic from the observed sample variance of person raw scores.

## CHARACTERISTICS OF THE KR20 STATISTIC

1. The item response variance used is that of an "average" person sampled. This is not the same as an average of the persons' test score error variances. If the sample score distribution is not symmetric, then the error variance of an "average" person must be different from the average of individual persons' error variances.

*Figure 19.1*

*Traditional analysis of a response matrix.*




**Items:** P-value of item *i:*

$$p_i = S_i / N \qquad q_i = (1 - p_i)$$

variance for item *i*: $p_i q_i = [S_i / N][(N - S_i) / N]$

**Person raw score variance for a "person" for whom the sample item p-values are their probabilities of success:**

$$VR = \sum_i^L p_i q_i$$

Persons: Sample raw score mean: $R = \sum_n^N R_n / N$

Sample raw score variance: $VS = (\sum_n^N R_n^2 / N) - R^2$

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

2. While $pq$ provides a test score error variance for an "average" person, we know that the sampled people vary, i.e., the variance of their raw scores is greater than zero. Persons with high or low scores have less score error variance than those with scores near fifty percent correct where the score error variance is maximum. Since the "average" person variance used in the KR20 formula is always larger than the lower score error variance of persons with extreme scores, it must always overestimate their score error variances.

3. If we want to anticipate reliability for a proposed application, a previously reported KR20 cannot be used as is, unless we know that the proposed sample will have the same score distribution as the sample used for the reported KR20. This is quite unlikely.

4. The use of raw scores as the data for calculating the sample variance is misleading to the extent that raw scores are not linear representations of the variable they are intended to indicate. Proof that raw scores cannot be linear representations can be seen by plotting the raw scores from a hard test against the raw scores from an easy test measuring the same attribute.

Figure 19.2 shows that the relationship between this pair of raw scores must be curvilinear. As a result, neither set of raw scores can be linear indicators of what they purport to represent. But the calculation of means and variances necessary to estimate reliabilities assumes linearity in the numbers used. Therefore, the calculation of these statistics from raw scores is always incorrect to some unknown degree.

If we expressed person measures in a linear, rather than curvilinear, form, then the sample variance estimates would be improved.

If person error variances were averaged instead of using the error variance of an "average" person, the information about sample test error conveyed by the reliability coefficient would also be improved.

## RASCH RELIABILITY

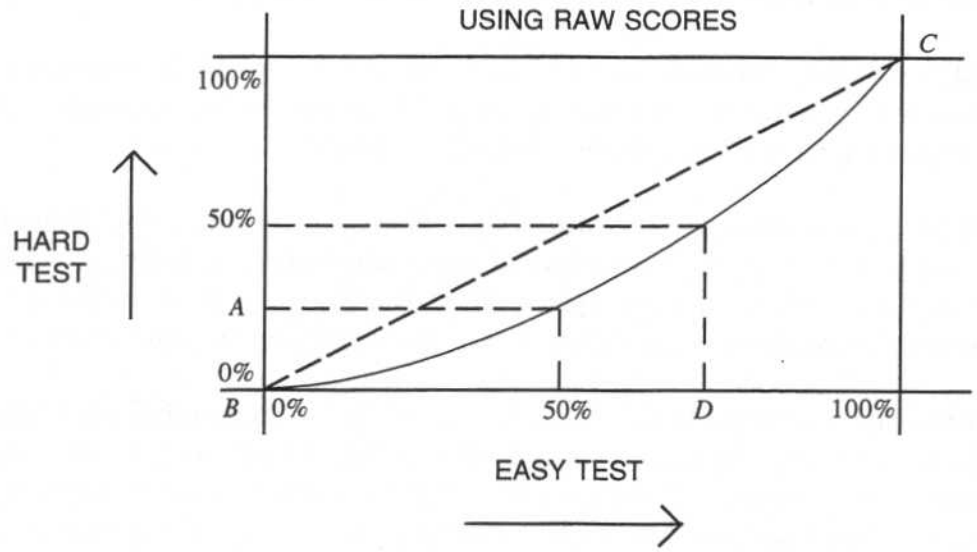
These shortcomings in KR20, or any other reliability coefficients based on raw scores, are remedied when a Rasch measurement analysis is made of the same data and reliability calculated from Rasch results. Rasch measurement produces a measure of each person's ability on a linear scale calculated from a logistic transformation of their raw score. The result is a linear comparison of the Hard and Easy tests as shown in Figure 19.3. These linear ability measures are numerically suitable for calculating sample variances.

We also have, for each person measured, an accompanying standard error of measurement. These individual errors can be squared and summed to produce a correct average error variance for the sample. When these results are substituted for those in the traditional KR20 formula, the result is a new formula which, while equivalent in interpretation, gives a better estimate of reliability than KR20, coefficient alpha, or any other reliability based on nonlinear raw scores.

When terms are replaced in this way, a better reliability coefficient results because (1) the numerical arguments are now linear rather than curvilinear, and (2) the actual average error variance of the sample is used instead of the error variance of an "average" person (see Figure 19.5).

*Figure 19.2*

*Comparing scores from easy and hard tests.*




Legend:

Pt. A = less than 50% on Hard Test.

Line BC = presumed linear relation between Easy and Hard test scores.

Arc BC = actual non-linear relation between Easy and Hard test scores.

Pt. D = more than 50% on Easy Test.

Explanation:

0% on the Easy Test implies 0% on the Hard Test (Pt. B).

100% on the Hard Test implies 100% on the Easy Test (Pt. C).

But 50% on the Easy Test implies *less* than 50% on the Hard Test (Pt. A)

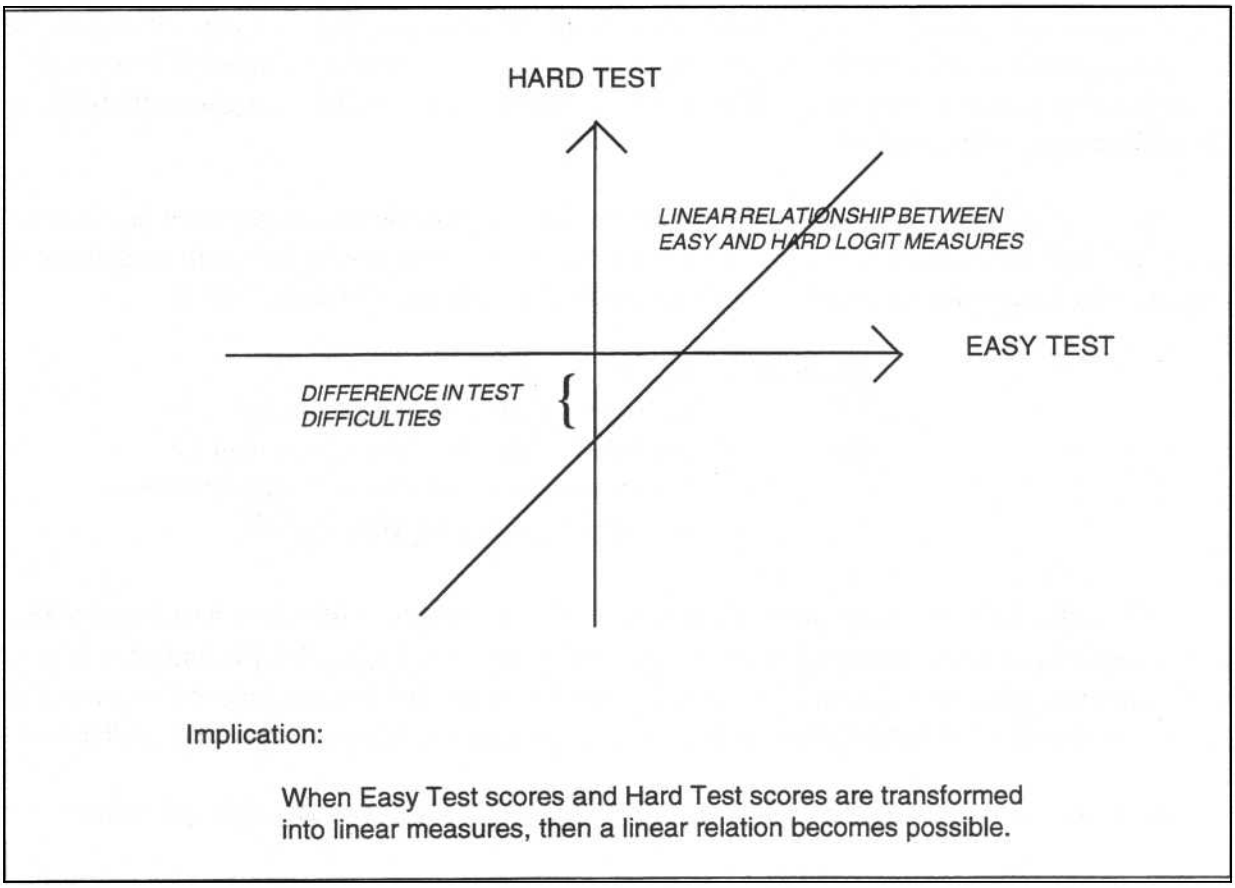
and 50% on the Hard Test implies *more* than 50% on the Easy Test (Pt. D).

Implication:

The score relationship between any pair of tests which differ in difficulty *cannot be linear.*

*Figure 19.3*

*Comparing measures from easy and hard tests using logit measures.*




## PREDICTING RELIABILITY

In the application of a test, it is the characteristics of the new sample to which we intend to apply the test, rather than a description of some previous sample, that is our real concern. We want to know how the test will work with the new people who are about to take it. We want a relevant reliability coefficient which applies to the people we intend to test, rather than an obsolete one describing people who were previously tested. But few practitioners know how to use an old KR20 to estimate a new KR20 for a new sample.

In fact, it is easy to predict the reliability for a forthcoming sample, if we are willing to postulate an expected mean and variance for this sample. From these statistics and the Rasch targeting formula we can calculate the reliability of the test for the new application without reference to any previous sample (Wright and Stone, 1979, 129-140).

## ADVANTAGES OF A SEPARATION INDEX

Correlation-based reliability coefficients, however, are also nonlinear in implication. For example, improvement of KR20 from .6 to .7 is not twice the improvement from .9 to .95. Although the difference in amount of reliability between .9 and .95 is half as much as the difference between .6 and .7. This half-as-much signifies twice the improvement in measurement precision. We can escape this

shortcoming of KR20 by replacing the traditional reliability coefficient with a Separation Index (G) (See Figure 19.5).

The 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 on a ratio scale in the metric of the root mean square measurement error of the test for the sample postulated. It quantifies "reliability" in a simple and direct way and has a clear interpretation. This expedites comprehension of what changes in reliability mean in terms of measurement precision.

The estimation of separations for new samples is easy. No reference to any previous samples is required. We need only estimate the expected standard deviation of our new target sample and then divide this estimate by the average standard error of the intended test for such a sample. As in:

| | |
|---|---|
| Separation: | $G = SDT/SET$ |
| *SDT:* | the expected *SD* of the target sample |
| *SET:* | the test standard error of measurement for such a sample, a value which is almost always well approximated by $SET = 2.5 / \sqrt{L}$ |

*SET* can be estimated more precisely as $\text{SET} = \sqrt{C / L}$ where $L$ is the number of items in the test and $C$ is a targeting coefficient (explained in Wright and Stone, 1979, pages 135-136 and tabled for most test and target relationships on pages 214-215). $C$ varies between 4 and 9 depending on the range of item difficulties in the intended test and the target sample's expected average percent correct on that test.

Here are some values of $C$ for typical item difficulty ranges and typical target sample mean percents correct:

*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 |

$\text{SET} = \sqrt{C / L}$

$L$ = Number of Items in Test

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

Thus, SET is easy to approximate well enough for the calculation of an expected target sample. Separation *G*: $G = SDT/SET$.

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

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

We use the Rasch Model in our example. But this Separation Index is applicable to any latent trait model. With it, one can predict the reliability of a test with any sample to be used in a study, if one can specify an expected sample mean and variance. No information about any previous samples is necessary.

The Standards (1985, page 22) recommend that, "Standard errors of measurement be reported at critical score levels." Rasch measurement analysis routinely provides standard errors for every possible test measure along the variable as shown in Figure 19.4. Thus, the Rasch approach meets this recommendation completely. 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.

The Rasch measurement errors satisfy this goal by providing individual errors of measurement for every observable measure. If a collective index of reliability is desired, the Rasch Separation Index is more useful in basis and numerical form than the traditional indices of reliability.

Figure 19.5 summarizes the calculation of the Separation Index.

*Figure 19.5*

*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

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

with $VR$ and $VS$ as defined in Figure 19.1

(see Wright and Masters, 1982, Figure 1, pp. 105-106)

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

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

# MEASUREMENT ESSENTIALS

## 2nd Edition

**BENJAMIN WRIGHT**

**MARK STONE**



WIDE RANGE, INC.
Wilmington, Delaware