DOCUMENT RESUME

ED 141 411 TM 006 383

AUTHOR Lenke, Joanne M.; And Others
TITLE Differences Between Kuder-Richardson Formula 20 and Formula 21 Reliability Coefficients for Short Tests with Different Item Variabilities.
PUB DATE Apr 77
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Comparative Analysis; *Complexity Level; Diagnostic Tests; *Item Analysis; Mathematics; Secondary Education; Standardized Tests; *Statistical Analysis; *Test Items; *Test Reliability
IDENTIFIERS *Kuder Richardson Formula 20; *Kuder Richardson Formula 21; Stanford Diagnostic Mathematics Test

ABSTRACT

To investigate the effect of violating the assumption of equal item difficulty on Kuder-Richardson (KR) Formula 21 reliability coefficient, 670 eighth-and ninth- grade students were administered 26 short, homogeneous "tests" of mathematics concepts and skills. Both KR Formula 20 and KR Formula 21 were used to estimate reliability on each test. The 26 tests were sorted into a high item difficulty variability group and a low item difficulty variability group, and the magnitude of differences in KR20 and KR21 reliability coefficients were compared for the two groups. The difference in KR20 and KR21 reliability coefficients was significantly greater when the range of item difficulty values was .30 or more. Nevertheless, KR21 was a good estimate of KR20 when the range of item difficulty was relatively narrow. Implications for test selection are suggested. When KR21 has been used to estimate a test's reliability, the user should note that the test has a lower bound of internal consistency reliability, particularly when the item difficulty range is great. (Author/GDC)

Differences Between Kuder-Richardson Formula 20 and Formula 21 Reliability Coefficients for Short Tests with Different Item Variabilities*

Joanne M. Lenke, Barrie Wellens, and John H. Oswald
The Psychological Corporation

*Paper presented at the Annual Meeting of the American Educational Research Association, April 4-7, 1977, New York.

Differences Between Kuder-Richardson Formula 20 and Formula 21 Reliability Coefficients for Short Tests with Different Item Variabilities*

Joanne M. Lenke, Barrie Wellens, and John H. Oswald
The Psychological Corporation

Of the various statistical methods for estimating the internal consistency reliability of a test, the reliability estimates developed by Kuder and Richardson (1937) have been widely used by test makers. The use of the Kuder-Richardson reliability estimates requires only the administration of a single test and does away with any biases that might arise when a test is split any one of a number of ways, as in the split-half method. The two primary sources of error variance considered in the Kuder-Richardson method are content sampling and heterogeneity of the measured trait, and their assumptions call for test items of equal, or nearly equal, difficulty and intercorrelation.

The most accurate Kuder-Richardson formula, known as K-R 20, can be expressed as follows:

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - \Sigma pq}{\sigma_t^2}\right),$$

where n = the number of items in the test;

p = the proportion of correct responses to each item;

q = the proportion of incorrect responses to each item (1 - p); and

$\sigma^2 t$ = the variance of the distribution of the test scores.

An approximation to K-R 20, which assumes that all items in the test have approximately the same difficulty, calls for less information and is

*Paper presented at the Annual Meeting of the American Educational Research Association, April 4-7, 1977, New York.

therefore much easier to calculate by hand. The simpler formula, known as K-R 21, can be expressed as follows:

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2}\right),$$

where $\bar{p}$ = the average proportion of correct responses to each item; and

$\bar{q}$ = the average proportion of incorrect responses to each item $(1-\bar{p})$.

Although the reliability estimate obtained from K-R 21 is generally lower than that obtained from K-R 20 by $\frac{n^2}{n-1}\left(\frac{\sigma_p^2}{\sigma_t^2}\right)$ for a given set of items administered to a given group of examinees, test makers still report K-R 21 reliability coefficients since deviations from the assumption of equal item difficulty result in a reduction of the coefficient, and hence a lower bound.

Of course, both of the Kuder-Richardson formulas presented here are based on the assumptions that a uni-factor trait is being measured and that the test is made up of parallel items. Studies carried out to determine the robustness of K-R 20 under violation of the uni-factor assumption indicated that K-R 20 showed little bias when the tests were relatively long (more than 18 items) and when the item intercorrelations were low (less than .6) (Brogden, 1946). Although Kuder and Richardson (1937) assumed in the derivation of their formulas that inter-item correlations were equal, Jackson and Ferguson (1941) claimed that the only necessary assumption was that the average covariance between parallel items be equal to the average covariance between nonparallel items. For the most part, K-R 20 is considered to be a satisfactory estimate of internal consistency reliability, if not a lower bound.

Given the indications that K-R 20 is quite robust under violations of its assumptions, the purpose of the present study was to investigate the robustness of K-R 21 as an estimate of K-R 20 under violations of the additional assumption of equal item difficulty. The results of the study may also help to shed some light on test selection procedures when such procedures include the examination of reliability coefficients.

Method and Data Source

The data for the present study were obtained as part of the Equating of Forms Phase of the National Standardization Research Program for the 1976 edition of Stanford Diagnostic Mathematics Test (SDMT) conducted in the fall of 1975. Five school systems participated in this research phase and administered both of two parallel forms, A and B, of SDMT to the same students within a three-week period. Since the order of administration of the two forms was counterbalanced by classroom to obviate practice effect, the administration of the two parallel forms to the same students can be thought of as one long test. The data presented in this paper is limited to the approximately 625 eighth-grade and high school students completing the Blue Level of this "long" test.

Stanford Diagnostic Mathematics Test is designed to measure competence in the basic skills and concepts that are important in daily affairs and prerequisite to the continued study of mathematics and, as such, can be used as an effective instructional tool. Therefore, in addition to the reporting of scores on each of its three subtests, Number System and Numeration, Computation, and Applications, scores are also reported on mutually exclusive groups of items within each subtest. These groups of items are referred to as Concept/Skill Domains. There are 13 such domains on each form of the test, or 26 in all. Since scores are routinely reported for each of these domains and educational decisions made on the basis of these scores, it was imperative that reliability estimates be obtained for these "short" tests. Alternate-forms reliability for the domains has been discussed elsewhere (Oswald, Wellens, Lenke, 1977). Internal consistency reliability coefficients estimated by means of Kuder-Richardson Formulas 20 and 21 are reported in Table 1, along with the range of item difficulties, for each domain. It is interesting to note in this table that short tests with as few as six items can have reliabilities in the 70's.

In order to investigate the extent to which the reliability coefficient determined by means of K-R 21 is a good estimate of that obtained using K-R 20 under violations of the assumption of equal item difficulties, each Concept/Skill Domain was identified as having "equal" item difficulties or "unequal" item difficulties. It was decided that domains having a range of item difficulties of .30 or more be designated as having "unequal" item difficulties; those having a range of less than .30, as having "equal" item difficulties. Therefore, on both Forms A and B, domains 1.1, 1.2, 3.1, 3.2, and 3.3 were designated "unequal," as were domains 2.8 on Form A and 2.5 on Form B; the remaining domains were designated "equal."

All K-R 20 and 21 reliability coefficients were transformed to Fisher's z coefficients and differences between these z coefficients determined for the "equal" and "unequal" sets of domains. A t-test of the difference between the means of these K-R 20 - K-R 21 differences for the two groups of domains was carried out, resulting in a t of 3.63 with 24 degrees of freedom. This t value is significant at the .01 level.

The results of this study indicate, therefore, that the difference between K-R 20 and K-R 21 is significantly greater when the range of item difficulty values is .30 or more than when the range is less than .30. Nevertheless, it does appear that, even for short tests, K-R 21 is a good estimate of K-R 20 if the range of item difficulties is relatively narrow. (A quick glance at Table 1 may not bring this out quite so clearly since some of the K-R 20 and K-R 21 coefficients "look" sufficiently close, despite quite a broad range of item difficulty values. Since correlation coefficients do not represent an interval scale, absolute differences between them are not comparable throughout the -1 to +1 range. The closer the correlations are to -1 or +1, the greater the difference between them really is. It is for this reason that differences between correlations must be examined using Fisher's r to $\underline{z}$ transformation.)

As for the practical implications of this study, test users should take note of the method used to estimate a test's reliability, particularly if "high" test reliability is a major criterion for test selection. If K-R 21 has been used to estimate a test's reliability, the fact that it is a lower bound of internal consistency reliability must be considered, particularly when the range of item difficulty is great. Table 2 demonstrates that different test selection decisions can be made purely on the basis of the particular formula used to estimate test reliability.

References

Beatty, L.S., Madden, R., Gardner, E.F. and Karlsen, B. Stanford Diagnostic Mathematics Test (New York: Harcourt Brace Jovanovich, 1976).

References (Cont'd)

Brogden, H.E. The effect of bias due to difficulty factors in product-moment item intercorrelations on the accuracy of estimation of reliability. Educational and Psychological Measurement, 1946, 6, 517-520.

Kuder, G.F. and Richardson M.W. The theory of the estimation of test reliability, Psychometrika, 1937, 2, 151-160.

Oswald, J.H.; Wellens, B., Lenke, J.M. A comparison of the tetrachoric correlation coefficient and the Pearson product-moment correlation coefficient on the same variables. A paper presented at the Annual Meeting of the American Educational Research Association, New York, 1977.

Table 1. Kuder-Richardson Formula 20 and 21 Reliability Coefficients for Forms A and B of *Stanford Diagnostic Mathematics Test* Concept/Skill Domains for the Blue Level Equating of Forms Sample (N = 626).

Concept/Skill Domain	Number of Items	K-R 20 Reliability Coefficient	K-R 21 Reliability Coefficient	Range of Item Difficulties
FORM A				
Number System and Numeration				
1.1	18	.75	.71	.34-.92
1.2	12	.78	.76	.33-.68
1.3	6	.69	.66	.39-.65
Computation				
2.2	6	.65	.64	.76-.88
2.3	9	.71	.70	.72-.88
2.4	9	.80	.80	.65-.83
2.5	9	.78	.77	.54-.80
2.6	6	.80	.79	.54-.78
2.7	3	.58	.54	.47-.66
2.8	6	.73	.68	.46-.76
Applications				
3.1	12	.72	.70	.43-.79
3.2	9	.70	.64	.38-.89
3.3	12	.61	.57	.39-.70
FORM B				
Number System and Numeration				
1.1	18	.76	.72	.27-.92
1.2	12	.77	.74	.31-.71
1.3	6	.64	.62	.36-.59
Computation				
2.2	6	.72	.72	.77-.87
2.3	9	.71	.70	.73-.89
2.4	9	.80	.79	.60-.78
2.5	9	.81	.80	.48-.81
2.6	6	.79	.78	.52-.77
2.7	3	.56	.52	.42-.63
2.8	6	.74	.68	.42-.70
Applications				
3.1	12	.72	.70	.38-.80
3.2	9	.74	.70	.39-.80
3.3	12	.62	.52	.23-.81

Table 2. Contingency Tables Showing the Number of <u>Stanford Diagnostic Mathematics Test</u> Blue Level Concept/Skill Domains Accepted and Rejected on the Basis of Various Reliability Selection Criteria.

Criterion: $r_{tt} \geq .80$

	K-R20	K-R21
Accept	4	2
Reject	22	24

Criterion: $r_{tt} \geq .75$

	K-R20	K-R21
Accept	10	7
Reject	16	19

Criterion: $r_{tt} \geq .70$

	K-R20	K-R21
Accept	19	16
Reject	7	10