On the Adequacy of Standardized Tests

Administered to Extreme Norm Groups

JULIAN C. STANLEY, JR.

George Peabody College for Teachers

The Problem

MOST STANDARDIZED TESTS are recommended by their publishers for use in more than one grade. Frequently, some convenient grouping corresponding to a prevalent type of school, such as the senior high, is suggested in the manual of directions. Quite a few tests are recommended for an even wider range, this being particularly true of intelligence scales. Thus presumably the Otis Quick-Scoring Mental Ability Test (9), Gamma Test, is equally useful anywhere from Grade 9 through Grade 16, while the California Test of Mental Maturity (2), Advanced Form, is designated for Grade 9-adult.

Thurstone found that "the factorial content of a test will change as it is given to populations that differ in age and schooling" (14, p. 43), and common sense long ago told us that IQ's based upon a children's test administered with a shortened time limit to adults probably do not have the same significance as they would for fifth graders. Perhaps among adults perceptual speed is the important variable, while for youngsters verbal ability may be more critical. Therefore, if the P and V factors are not very highly correlated, the person who at an early age earns a certain rating on a given test because at that level it demands chiefly verbal ability may score quite differently on the same scale years later even though his verbal "brightness" is unchanged.

Age scales, typified by Binet-type tests, are appropriately used in groups markedly heterogeneous with regard to age or grade. On the other hand, point scales (into which category nearly all group tests

fall) have no such a priori utility. Like the American Council on Education Psychological Examination for College Freshmen (1), they may be optimally suitable in only a single grade. Conversely, they may have been carefully standardized with respect to content and difficulty so as to be adequate for several grades; e.g., the Stanford Achievement Test (12), Intermediate Battery, was designed for Grades 4, 5, and 6. The present tendency is probably toward delimiting the field for use of a given test to a specified population, such as applicants for West Point, beginning graduate students, or entrants in the Pepsi-Cola scholarship contest.

When an author makes the content and difficulty level of his test fit a relatively homogeneous group and then recommends that the product be utilized outside this range, his customers may well be cautious. In addition to the possible change of factorial content, mentioned above, there is the even more obvious chance that the test will be too easy or too difficult for some of the "extrapolated" persons taking it, so that scores may pile up at one end of the scale and make differentiation among the persons tested less reliable than if an "adequate" instrument had been employed.[1]

Because test manuals usually contain norms for several grades and/or ages, very likely most users feel that the *construction* of the scale involved a similar variety of students. In many instances this assumption is not justified. Since norms are collected after the test is in its final form, they may be based upon scores earned by all sorts of students in different parts of the country. It is common practice for test authors to request that their customers send them data obtained in routine testing so that the norms can be expanded.

The Nelson-Denny Reading Test

As an illustration of the initial try-out of a test upon a relatively homogeneous sample with a subsequent recommendation that it be employed much more widely, the writer has chosen the Nelson-Denny Reading Test for Colleges and Senior High Schools (7), hereafter referred to as the NDRT. There are other equally appropriate examples. Professors Nelson and Denny have simply followed the prevailing

[1] See Goodenough's discussion (3) concerning the effects of insufficient range of difficulty upon test scores. She states: "There is good reason for believing that the form of distribution of most, if not all, mental abilities conforms at least roughly to that of the normal curve" (pp. 148-49). This is probably an oversimplification of the problem, however.

practice, possibly at the suggestion of their publisher. The NDRT was selected for scrutiny solely because suitable data concerning it was available.

The NDRT consists of two speed-plus-power subtests: 100 five-option multiple-choice vocabulary items, with a 10-minute time limit; and nine paragraphs, each accompanied by four five-option multiple-choice questions, and a 20-minute time limit. The items in each subtest are of graduated difficulty, beginning with easy questions. One point is credited for every correct answer to a vocabulary item and two for each correct paragraph response, so the maximum possible score is 100 + 2(36), or 172. Two forms of the test, Form A and Form B, are available. They were constructed in the same way and are said to be of equal difficulty.

In the preliminary tryouts by Nelson and Denny, 600 vocabulary items were "administered to approximately 390 students [who] represented all of the four college classes," presumably at the Iowa State Teachers College. "A number of high school seniors were also included" (8). From data obtained with these groups the authors constructed two forms of the vocabulary subtest, each containing 100 items. They used a similar procedure with the paragraph subtest, employing 450 students and 27 paragraphs, of which nine were later discarded.

Grade equivalents for total scores of 1 (GE = 3.0) through 99 (GE = 16.6) are given in the manual of directions (8), the equivalents of scores below 20 (GE = 6.0) and above 96 (GE = 16.0) having been obtained by extrapolation. Subtest and total (per)centile norms based upon "senior high school students" and "college and university students" in unspecified institutions are compiled for Grades 9-16, with N's varying from a high of 5236 for college freshmen to a low of 407 for college seniors. Among ninth graders the total score corresponding to the 50th centile is 42, vocabulary being 18 and paragraph 24.

The Sample

By accident the writer discovered a "typical" beginning ninth grade in a New England public senior high school. On the California Short-Form Test of Mental Maturity, Advanced Form, these 161 boys and girls had median IQ's, both language and non-language, of approximately 101, with the total-IQ standard deviation just short of 15, which statistics agree well with norms for this grade furnished in the manual

of directions. The scores of these students on the NDRT should therefore be helpful in determining whether or not it is too difficult for average ninth graders.

Vocabulary Subtest Scores

Although the NDRT and California Test of Mental Maturity norms are based upon different pupils, still for this "typical" group the mean vocabulary subtest score on Form A, 18.9, corresponded to the 53rd centile, approximately the same as for the IQ average. The range was from 5 (fourth centile of ninth grade) to 52 (54th centile of college seniors), the standard deviation being 8.1. The standard error of measurement of the vocabulary subtest, computed by Rulon's method (10), is approximately three points.

As anticipated, scores piled up at the low end of the scale. A chi-square test of the goodness of fit of the obtained curve to a theoretical normal curve based upon the same data gives a $\underline{P}$ of .0008, confirming graphic evidence that the scores were not drawn from a normally distributed population.

Quite probably the vocabulary subtest, despite its split-halves reliability coefficient of .86,[2] is too difficult for the bottom two-thirds of this group. If the papers had been scored with a R-(W/4) formula, it is likely that a considerable number of scores would have approached zero or even been negative. By considering the lowest 27%, middle 46%, and highest 27% of the group, as recommended by Kelley (4) in another context, we obtain the figures shown in Table I[3]. There it is

TABLE I
STATISTICS CONCERNING THE NELSON-DENNY VOCABULARY SUBTEST AT THREE DIFFERENT LEVELS OF AN AVERAGE NINTH GRADE

LEVEL	σ	$\underline{N}$	Mean	σ_{meas}	Reliability Coefficient
Lowest 27%	1.9	43	10.5	2.4	-.66
Middle 46%	2.5	75	17.7	3.2	-.68
Highest 27%	7.2	43	29.3	3.2	.80
Entire Group	8.1	161	18.9	3.0	.86

[2] This reliability coefficient, obtained by a method essentially the same as correlating scores on the halves and "stepping up" that $\underline{r}$ with the Spearman-Brown prophecy formula, is open to the usual criticisms directed against reliability coefficients secured in this manner from speeded tests (13, p. 112), as is σ_{meas} also.

[3] Actually, because of the rounding-off process these levels are the lowest and highest 26.71% and the middle 46.58%.

obvious that the standard deviation increases drastically from only 1.9 for the low group to 7.2 for the top 27%, while the standard error of measurement shows no such marked or significant increase (11, pp. 250-251). Since we have artificially restricted the range of obtained scores by our truncating procedure, the standard error of measurement at two levels exceeds the standard deviation, and consequently the corresponding reliability coefficients are negative (-.66 and -.68). Since reliability coefficients cannot in theory be other than positive (5), this finding deserves attention in a later section of the article.

Paragraph Subtest Scores

Scores on the paragraph subtest ranged from 8 (1st centile of ninth grade) to 58 (83rd centile of college seniors), the mean being 28.1 (70th centile of ninth grade), significantly above the NDRT manual's 50th centile (24) beyond the 1% level of confidence. The standard deviation was 10.8, the standard error of measurement 4.4, and the reliability coefficient .83. A theoretical normal curve fits this data better than it did the vocabulary scores, but P is only .03, so we again reject the hypothesis of normality. Low scores still predominate, though the positive skewness is less than before.

TABLE II
STATISTICS CONCERNING THE NELSON-DENNY PARAGRAPH SUBTEST AT THREE DIFFERENT LEVELS OF AN AVERAGE NINTH GRADE

LEVEL	σ	N	Mean	σ_{meas}	Reliability Coefficient
Lowest 27%	3.3	43	16.0	4.1	-.58
Middle 46%	3.9	75	26.7	4.9	-.61
Highest 27%	6.2	43	42.6	3.5	.68
Entire Group	10.8	161	28.1	4.4	.83

Table II reveals a pattern already noted in Table I, though now it is less accentuated; standard deviations again have the same upward trend. The standard errors of measurement are not markedly unalike in magnitude, differences among them barely missing being significant at the 5% level. Scores of the 161 students on the vocabulary and paragraph subtests correlate fairly well, with $r = .66$.

Total Scores

For the two subtests combined the range of scores obtained was

from 19 (GE = 5.9; 7th centile of ninth grade) to 109 (70th centile of college seniors), with a mean of 46.9 (GE = 9.7), a standard deviation of 17.2, a standard error of measurement of 5.3, and an reliability coefficient of .90. The obtained curve is skewed to the right, with a large number of low scores and more very high ones than would ordinarily occur if the distribution were normal. In testing the fit of a theoretical normal curve to this data we obtain a chi square of 70.24 with 6 d.f., corresponding to a P of .0000, so the hypothesis of normality in the parent population from which these scores were drawn randomly can be rejected with considerable assurance.

TABLE III
STATISTICS CONCERNING THE NELSON-DENNY READING TEST
TOTAL SCORE AT THREE DIFFERENT LEVELS OF AN AVERAGE NINTH GRADE

LEVEL	σ	N	Mean	σ_{meas}	Reliability Coefficient
Lowest 27%	4.4	43	29.3	5.3	-.47
Middle 46%	6.3	75	43.9	5.4	.26
Highest 27%	12.9	43	69.8	4.7	.87
Entire Group	17.2	161	46.9	5.3	.90

For the whole test the standard deviations at the several levels shown in Table III vary systematically. Discrepancies among the standard errors of measurement are slight and unreliable. The SE_{meas} of the lowest 27% still exceeds the SD, causing the reliability coefficient to be negative (-.47).

Discussion

We have seen that *in this sample* the Nelson-Denny Reading Test for Colleges and Senior High Schools fails to discriminate well among the less able testees. With a below-average ninth grade it would probably have been even less discriminating, while for superior pupils the test should be more effective. This is not a *general* criticism of the NDRT, which many persons consider quite satisfactory. Rather, it is simply an indication that for average or slow ninth graders—and probably for some tenth and eleventh graders, too—an easier scale, one designed specifically for the level at which it is to be used, is preferable.

Reliability

The overall reliability of the NDRT in this ninth-grade group appears

to be quite high, reliability coefficient for the total score being .90 and PE_{meas} 3.6. In their manual of directions the authors cite a *comparable-forms* reliability coefficient of .914, with N = 171 college freshmen, SD about 22, and PE_{meas} approximately 4 points (more precisely, 4.3). Since the ratio of the two variance errors of measurement (F) is 1.5, with d.f. = 170 and 160, there might be ample reason for doubting that the test is equally reliable in the two ranges if it were not for the fact that the reliability coefficients resulted from different procedures: split halves for the ninth grade, comparable forms for the college freshmen. The split-halves technique tends to give a higher estimate of the reliability coefficient than results from the correlation of scores on two comparable forms administered with separate time limits (13), particularly when a speeded test is involved. If the standard deviation of the ninth graders had been 22 instead of 17.2, theoretically the reliability coefficient would have equaled .94.

We have already noted that negative reliability coefficients may result when the range of obtained scores is restricted, since the standard error of measurement is fairly constant throughout the levels. In its simpler form the reliability coefficient is the complement of a variance ratio:

$$\text{reliability coefficient} = 1 - \frac{(SE_{meas})^2}{(SD)^2}$$

This formula rests upon several assumptions, such as that errors of measurement are uncorrelated with the individual's hypothetical "true" score on the test. When the standard deviation of obtained scores is artificially reduced, for example, to zero, so that only those individuals with *identical* obtained scores on a given test are considered, the split-half procedure for computing reliability coefficients is obviously invalid. If we secure two scores for each such testee by any split whatsoever and correlate these half-scores, no result other than -1.00 can possibly occur if there is any variability in either of the distributions. Furthermore, in this group errors will be perfectly correlated negatively with true scores—the smaller the signed error the larger the true score, and vice versa.

The variance of the obtained scores is usually considered to equal the variance of the true scores plus the variance of the chance errors, so since variances are squared quantities and therefore cannot be nega-

tive, in ordinary unrestricted testing situations the error variance is never greater than the obtained variance except by "chance."

Errors of Measurement

None of the differences among standard errors of measurement were significant at the 5% level, though for the paragraph subtest they approached that point very closely ($P<.06$). Vocabulary subtest SE_{meas}'s showed a similar trend ($P = .09$), in accord with Mollenkopf's finding (6) that (in an essentially non-chance situation) the standard error of measurement is constant throughout the entire range of scores only if the distribution is mesokurtic and has negligible skewness. All three of the curves described above were positively skewed.

On the NDRT a person who knows nothing whatsoever and merely marks items randomly without even reading them will have a *true* score equal to the number of items he attempts (n) divided by 5. Obtained scores will then fluctuate according to the rules of chance, with an expected standard deviation of $\sqrt{n(.2)(.8)}$, or $.4\sqrt{n}$. Thus, if a group of such testees attempt 49 items, they will presumably average 9.8 right, with a SD of 2.8. In order to secure a SD of 1.86, that earned by the lowest 27% on the vocabulary subtest, uninformed persons need to guess at just 22 items. But then their mean would be only 4.4, well below the 10.5 actually earned by these students. So we may surmise that at least some of these 27% knew the answers to a few questions or could figure them out.

Summary

In order to make their tests more salable, a considerable number of authors have recommended them for use in grades below or above those for which the tests were initially designed. Thus questions concerning changing factorial content and difficulty level arise. As an illustration of a test too hard for the lowest grade suggested by its constructers, the writer arbitrarily selected the Nelson-Denny Reading Test for Colleges and Senior High Schools, on which there was data available. This instrument was found to be of unsuitable difficulty for approximately the lower half of a typical ninth grade (161 pupils) in a New England public coeducational senior high school. During the analysis several negative reliability coefficients were secured. This statistical anomaly and theoretical issues related to it are discussed briefly.

Bibliography

1. *American Council on Education psychological examination for college freshmen.* Educational Testing Service, 20 Nassau St., Princeton, N.J.
2. *California test of mental maturity.* California Test Bureau, 5916 Hollywood Blvd., Los Angeles 28, Calif.
3. Goodenough, F. L. *Mental testing.* New York: Rinehart, 1949.
4. Kelley, T. L. The selection of upper and lower groups for the validation of test items. *J. educ. Psychol.*, 1939, **30**, 17-24.
5. Kelley, T. L. The reliability coefficient. *Psychometrika*, 1942, **7**, 75-83.
6. Mollenkopf, W. G. Variation of the standard error of measurement. *Psychometrika*, 1949, **14**, 189-229
7. *Nelson-Denny reading test for colleges and senior high schools.* Houghton Mifflin Co., 432 Fourth Ave., New York 16, N.Y.
8. *Nelson-Denny reading test, manual of directions.* New York: Houghton Mifflin, (no date).
9. *Otis quick-scoring mental ability tests.* World Book Co., 313 Park Hill Ave., Yonkers 5, N.Y.
10. Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. *Harv. educ. Rev.*, 1939, **9**, 99-103.
11. Snedecor, G. W. *Statistical methods.* Ames, Iowa: Iowa State Coll. Press, 1946 (4th ed.).
12. *Stanford achievement test.* World Book Co., 313 Park Hill Ave., Yonkers 5, N.Y.
13. Thorndike, R. L. *Personnel selection: test and measurement techniques.* New York: Wiley, 1949.
14. Thurstone, L. L., & Thurstone, T. G. Factorial studies of intelligence. *Psychometr. Monogr.*, No. 2. Chicago: Univ. of Chicago Press, 1941.