DOCUMENT RESUME

ED 123 590 CS 002 676

AUTHOR Tucker, Elizabeth Sulzby
TITLE Grade Level Expectations and Grade Equivalent Scores in Reading Tests.
PUB DATE 75
NOTE 23p.; Paper presented at the Annual Meeting of the Virginia State Reading Association (Richmond, March 6-8, 1975)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Elementary Education; *Expectation; *Grade Equivalent Scales; *Grade Equivalent Scores; Reading Comprehension; *Reading Level; Reading Material Selection; *Reading Tests; Silent Reading; Standardized Tests; Test Validity

ABSTRACT

Statistical methods and test writing for reading comprehension have been based on the assumption that certain reading tasks or levels are appropriate at one age level that would be too difficult at another. A clear-cut determination of grade levels for reading materials has, however, not been defined. Grade and age equivalent scores on silent reading tests, readability scores attached to children's reading matter, and reading grade expectancy scores are investigated in light of their usefulness. The history of the grade equivalent scores used in standardized tests and in readability scores can be perceived as reflecting a circular, or skyhook relationship between these scores and curricular material. A testing procedure consisting of items beginning at a basal level, where a student was convinced of his or her mastery, and continuing until a ceiling of error is reached can eliminate the inattention factor and the guess factor. In addition, such a system can preserve the efficiency of the group format and make test results more interpretable to the specialist, as well as more understandable to the child. (KS)

GRADE LEVEL EXPECTATIONS AND GRADE EQUIVALENT SCORES

IN

READING TESTS

Elizabeth Sulzby Tucker
1975

Grade Level Expectations and Grade Equivalent Scores in Reading Tests

In 1924 E. A. Lincoln declared: "It is probable that we shall never know the real beginning of the use of standard tests" (p. 12). The movement began perhaps as early as 1575, according to Linden & Linden (1968, p. 2), and included, much later, Rev. Fisher's and Dr. Rice's desires for scaled measures of school children's progress year by year. Tests were needed in the early "scientific" days of the twentieth century, claimed Monroe (1917, p. 71), because of the inconsistency and inaccuracy of teacher marks or grades. Certainly this desire for more accurate and consistent quantification of children's performances has not lessened. The field of reading research has been filled with test constructors and test interpreters attempting to reach this accuracy and consistency to guide the improvement of children's reading ability.

Early test constructors, like Simon and Binet, wished to devise scales that increased gradually in difficulty so that children's performances could be ranged on a continuum. In mental measurement, Simon and Binet

thought that they could arrange tasks by levels appropriate to one age group and not to a younger age group. They used 67-75 percent accuracy as their criterion for such a stepwise age level discrimination (Linden & Linden, p. 17). Reading test experimenters were led in a similar fashion to devise reading scales, first oral, then silent. Thorndike and others invented reading scales using passages graduated in difficulty so that a reading age or reading grade might be established for a child. A colleague of Thorndike, W. A. McCall, was so enamoured with the idea of accurate and consistent measurement on a graded basis that he surveyed entire school populations and advised the promotion and demotion of children until their average grade scores (using conversion to his standard T score) for the entire academic curriculum (the testable part, that is) fit his range of expectations (McCall, 1927, p. 67ff.).

Statistical manipulation and test writing became more sophisticated, but still the idea clung in the minds of educators and test experts that certain reading tasks or levels are appropriate at one age level that will be too difficult at another. The problem has remained: a clearcut determination of grade levels for reading materials and reading tasks has not yet been defined.

This paper will investigate three uses for grade and age equivalent expectations and scores: 1) grade

and age equivalent scores on silent reading tests; 2) readability scores attached to children's reading matter including textbooks; and 3) reading grade expectancy scores. Second, the paper will discuss the history of the grade equivalent scores used in standardized tests and in readability scores and will describe the argument that a circular, or skyhook relationship exists between these two scores and curricular material. Finally, the paper will propose possible directions for test constructors in the future.

Grade or age equivalent norms were one way of departing from the reporting of raw score data and attempting to cast an interpretation on test scores. These norms were used to describe the average score obtained by children taking a test (Anastasi, 1968, p. 16; McLaughlin, 1960, p. 6). While the grade or age equivalent scores for children above and below the mean were usually determined by extrapolation and interpolation, some newer tests have been normed on children one grade above and one grade below the intended level (Robeck & Wilson, 1974, p. 370).

In development of the scores, the age score was first developed as a parallel to the mental age derived from mental tests. W. A. McCall referred to Thorndike's

wish to continue his open-ended scale:

> But measurement continued to be a matter for experts because scale scores were difficult to compute and were generally incomprehensible.
>
> To overcome this difficulty the writer developed and popularized a plan for having all tests yield comparable and easily understood age scores such as reading age, arithmetic age, educational age, mental age, promotion age, and the quotients . . . .
>
> Later the writer invented the grade scale yielding G scores. These proved to be so popular that they came into almost immediate use on most tests from New York to Nanking.

McCall went on to boast that his "objectively-scorable tests yielding age scores or G scores gave measurement to the millions and provided the large profits" [emphasis mine] (1939, p. 25).

Grade and age scores lost their separate identity, however, so that in later years reading experts have proposed formulas ignoring any difference by adding or subtracting 5.0 from whichever score was not in comparable form (Della-Piano, 1968, p. 41). The grade equivalent grew to be more preferred and was used in formulas for calculating gains in reading ability (Bleismer, 1970), for predicting and describing ranges of achievement (MacGinitie, 1973), and in selecting candidates for remedial reading classes (Harris, 1971).

The grade equivalent seemed easy to interpret, even though test experts found areas for criticism of this

uneven interval, non-standard score measurement. Test experts criticized the use of the grade equivalent score not only because its range does not include actual scores of real children (Spache, 1963, p. 360), but also because its range does not describe how the content which is tested is actually taught in schools. Anastasi pointed out that all subjects are not given equal emphasis from grade to grade and that an estimated "grade equivalent" may mask the emphasis or neglect of that area in certain grade levels (p. 61). Thorndike and Hagen suggested that a comparison with a child's own age group may be more relevant, especially if multiple sets of norming populations are used and described (1960, p. 220).

The criticisms above have vascillated between a desire for norming populations above and below the grade level for which the test is intended and a desire for a test which will measure only the curriculum taught at that age level "ordinarily," whatever ordinarily may mean, given the range of abilities in any classroom.

Thorndike, and later writers, felt that gradedness attached to the passages used for reading, not just to the child's obtained score. Greene & Jorgensen (1929, pp. 14-15) and Monroe (1917, pp. 71-72) referred to the extensive (for that time) experimental re-ordering of the passages in response to children's performances with the tasks.

Validity studies will be dealt with in the second major section of this paper; however, one criterion test should be mentioned in the first section. Part of the oral history of reading instruction has been that a set of reading materials, the McCall-Crabbs Standard Test Lessons in Reading, were so good that tests were standardized using the lessons as criterion. Joyce Kamons, Test Editor for Teachers College Press (Columbia), confirmed just the opposite: that the Test Lessons used the Thorndike-McCall Reading Scale with its grade scores as criterion and that grade scores were plotted and assigned to the Lessons from the Scales. Ms. Kamons stated that more complete information was not available (1974, letter).

These Standard Test Lessons in Reading, with their grade scores derived from comparison with the Thorndike test, were used as a basis for the assignment of the first readability scores and many others:

> The readability formulas developed by Lorge, Flesch, and Dale and Chall all used the McCall-Crabbs test lessons as a criterion. (Harris, 1974, p. 2)

Readability scores, in turn, were used to estimate the difficulty of reading textbooks, reading tests, library books, et cetera. Harris & Jacobson used passages from many basal readers as criteria for their basic revised scale, but also used correlation with the

McCall-Crabbs for additional validity (Revised . . . Formulas, p. 5). These writers indicated that the formulas must be interpreted with a correction factor for realistic use by children in reading.

Readability formulas were designed to help children select material on a gradually increasing difficulty level as their reading ability increases. Thorndike was described as ordering his passages by difficulty from empirical evidence, though modern writers have criticized the extreme difficulty of his passages (Tuinman, 1971, p. 197). Thorndike's test was used as criterion for the McCall-Crabbs passages which were used as criterion for readability scales. Now Harris and Jacobson (p. 6) wish to re-standardize the McCall-Crabbs passages, hoping that

> . . . the average comprehension scores of children on them will provide another and perhaps better criterion for validating or improving our readability formulas.

Reading grade expectancy scores involved the use of grade or age scores from one or more areas other than reading. Harris (1971, pp. 113-120) described a number of such formulas. These formulas use calculations involving such measures as chronological age, mental age, reading age, arithmetic age, and years in school. Grade scores were used rather than standard scores. O'Connor (1972, pp. 78-79) cautioned against lumping such scores and measures and ignoring the standard error

of measurement between the scores. It is the opinion of this writer that such formulas may occasionally be justified if their only use is to admit children to a remedial reading class for instruction; if, however, such a formula is used to measure attainment of a program goal, such as that of the 1973 Virginia Standards of Quality, the formula appears to be invalid. (Program goal, Standards of Quality:

> The average achievement level of the student population in reading and mathematics as measured by standardized tests will equal or exceed the average ability level of the student population as measured by scholastic aptitude tests.)

Conversion of the scores to standard scores would lend more validity to the formulas, as would considering the standard errors of measurement. (The quality of the test for the purpose intended is deliberately ignored at this point in the argument.)

The basis for the reading grade expectancy score, however, was that a child could be expected to read at a level commensurate with his ability, usually interpreted as mental ability. Some writers, including Spache & Spache (1969, pp. 64-69) have questioned such a one-to-one relationship between mental ability and potential for reading ability. O'Connor (1972, pp. 78-79) criticized the idea that mental age, in particular, could be used as a measure toward which a child could aim in reading.

Since the early twentieth century writers have commented on the correlation between reading tests and intelligence tests (Davis, 1972, p. 635) and Thorndike boldly entitled his 1917 article: "Reading as Reasoning." Intelligence tests have been used to predict potential for reading ability for children measured by reading tests which correlate highly with scores on intelligence tests. Certainly, the relationship between reading and intelligence has not been ascertained. The possible relationships have implications for the total question of validity, both for reading tests and for the readability scores derived from them.

The APA Standards for educational and psychological tests and manuals (1966) redefined the kinds of validity from predictive, concurrent, construct, and content validity to content validity, criterion-related validity, and construct validity. A review of tests in reading has been published at intervals since 1938. Inferences about the validity considerations of early reading tests have been drawn from O. K. Buros' Mental Measurements Yearbooks (re-published in one volume with dates preserved, Reading Tests and Reviews).

In the first Yearbook, 1938, statistical validation, such as scores fitting the normal curve, was discussed. Predictive, or criterion-related validity, was considered

when some tests were compared to the later college point averages of students. The other type of criterion-related validity was mentioned when tests were compared with other tests. It was not until the second Yearbook in 1940 that reading tests were criticized for not describing how items were arrived at (pp. 1571-1576, 1556) and for not describing how the test in question distinguished between good and poor readers (p. 1578). Considerations of construct and content validity were second considerations, it would appear.

The major criterion mentioned in all the yearbooks was another test (or tests). One of the earliest tests used as a criterion was the Thorndike-McCall Reading Scale. Thorndike's original test had open-ended questions, rather than the multiple-choice format that was chosen by his colleague McCall for the Test Lessons that were used as criterion for the readability scales (Thorndike, p. 425). Once Thorndike's type of questioning (called reasoning by himself and others) was converted into multiple-choice format, his test and its descendents lent themselves to being a ready source for what Ronald P. Carver (1973, p. 52) described as "face validity, discriminate reliability among individuals at any given level, and level-to-level group increments." Carver's overall criticism was that Thorndike's test began a circular relationship between

reading tests and tests of reasoning or intelligence. The criticism of Davis (1972, p. 635) was repeated.

R. T. Lennon (1970, p. 123) took a generous view toward the validity question for reading tests; he noted that, when test makers cannot know or determine the universe they are attempting to measure, they look for internal consistency among other tests. Robeck & Wilson pointed to improvements in recent standardized reading tests (1974, pp. 367-371), but concluded that:

> The problem of validity plagues all of the test producers since there is no external criterion against which a reading test can be validated (p. 376).

Test constructor (or re-constructionist)

J. R. Bormuth (1970, p. 9) criticized test authors for ignoring an external criterion. He pointed to the circularity in validation, one reading test's being correlated against another against yet another (as described in the circle with the Thorndike test, other reading tests, and even intelligence tests). Bormuth reiterated Carver's accusation against test writers, that face validity and obvious discrimination in the items was more important to them than was the question of whether the test does, in fact, measure what it was intended to measure. Bormuth's solution was for test writers to examine actual instruction in the classrooms, then to apply certain transformational-generative grammar rules to develop test items; however,

his basic criticism--that test authors have ignored the search for an external criterion--seems to be directly related to this paper. His accusation of circularity in validation has been echoed by many other writers.

Davis (1972) conducted factor analyses for over 30 years, attempting to isolate skills and sequences of skills in reading comprehension. He criticized test authors for being economy-conscious in item selection (cf. McCall, 1939, p. 9). Pyrczak (1972, p. 64) noted that many students were able, far beyond chance levels, to mark test items on reading tests without reading the accompanying passages. Many items could be marked from general knowledge and others could be marked because items were interrelated (i.e., reading one question was a hint to the answer of a second question). In "The Assessment of Change," Davis suggested choosing the best among the existing tests, using cautions from a knowledge of possible error of measurement, and being willing to depend upon tentative data (1970, pp. 326-339).

Carver (p. 53) described the ultimate in circularity of validation: ETS's National Anchor Test Equating Study in Reading, in which seven existing norm-referenced tests in reading were to be made statistically more equal. Combine this equation with the relationship between reading

tests and readability scales--and the child who has not been able to achieve in current curricular material will be promised little. The readability scale used to devise and measure his textbook is tied to his reading test which is tied to other reading tests; from his tests and the readability scales come other textbooks, and sometimes even his library books. (And his potential for reading is estimated from intelligence tests to which, it might be supposed, the Anchor Test Equating method might be tied.)

Critics like Davis and Carver have alerted educators to the problem of test circularity. Robeck & Wilson, along with Buros, have described ways of devising better reading tests which attempt to measure tasks which require reading in addition to reasoning or intelligence tasks which can be done without reading the passages on the tests.

Criterion-referenced test writers like Wayne Otto (1974) have attempted to examine the curriculum, to obtain teacher and developmental psychologist opinions, and to validate skill sequences so that tests will be reading tasks discretely. Suggestions that "reading" can be divided into bits and pieces have been objected to by other writers, to the extent that Kenneth Goodman (1973) suggested a reversion to the "inefficiency" of older tests. Goodman implied that, if validity is desired,

reading may have to be evaluated in the natural setting, perhaps even in a one-to-one situation like that used in the informal reading inventory (pp. 30-33).

Kenneth Goodman's suggestion seems more persuasive to this writer because criterion-referenced test writers have surveyed the instruction in reading comprehension and have not yet been able to agree upon what is being taught nor about what is appropriate (learnable) for each age/grade/ability level, except perhaps at the lower levels of word attack skills. (Even the so-called word attack skills have been tested in widely divergent sequences and formats.) The research of Davis and his cohorts has not yielded fruitful, isolatable skills and skill sequences. Thorndike, in 1917, did realize that he had to use reading passages, rather than isolated items, that were progressively more difficult. This strategy, of progressively more difficult passages, is incorporated into the informal reading inventory (Betts, 1946) that Kenneth Goodman mentioned. It was mentioned earlier that Thorndike's original test used open-ended questions; this question format presented problems for group testing, quite obviously.

Whatever test is used, however carefully the passages are arranged in order of difficulty, and however closely the passages duplicate real-life reading situations,

test writers will wish to have a more realistic appraisal of the appropriate reading tasks for each age/grade level, able to be conducted as a group test and able to meet the requirements for validity, reliability, and understandable norming data.

Thus a scheme is needed for testing comprehension in a group setting. Such a test would need to avoid the guess factor; it would also need to avoid the inattention factor that may come from a test's being too long. Such a test might contain passages arranged, like Thorndike's and his successors', in order of increasing difficulty. The passages should be written reflecting the best research on interest and degree of complexity leading to comprehension. The passages could be divided into 15-20 minute "sittings," packaged as separate booklets, scored and normed separately.

A student could begin the test at the level he was convinced would be very easy for him. The test score could be calculated beginning with the student's highest basal level (this might be defined as the level at which he answers 7 of 8 consecutive questions correct, for example). The test sittings could continue until the student met a ceiling, based upon research of the best possible level of error consistent with preventing the generalization of frustration. (This level might be 5 errors in 8 answers, more or less, but research might support allowing certain

students to continue above the ceiling, for further diagnostic purposes.)

The use of the ceiling and basal scheme would, hopefully, eliminate the inattention factor at the lower level (by removing the child from the task between sittings as well as asking the child to start at a level estimated to be appropriate to his independent reading level). Often a good reader is penalized at the lower end of the scale by being required to answer questions which are far too easy and to which he does not adequately attend. The basal scheme might also help eliminate the guess factor at the higher level for the good reader who may be fatigued by the time he gets to the end of the test battery. It should also eliminate the guess factor for the child who is a poorer reader who nevertheless is encouraged to try to answer as many questions "as you can" in the time limit of traditional tests. The ceiling and basal scheme (well-established in intelligence testing history, as well as in the history of the informal reading inventory) should isolate a body of reading material with which the child is comfortable; this scheme should make the "grade level" designation more appropriate to everyday reading tasks of the child. It would preserve the efficiency (and Mr. McCall's profits) of the group test format. To be less trivial, it might make group test results more interpretable for the classroom teacher and

for the specialist wishing to use the results in a more complete case study. The scheme should also be more understandable to the most important person involved in the testing, the child.

BIBLIOGRAPHY

Ahmann, J. S. & Glock, M. D. Evaluating pupil growth: Principles of tests and measurements. 4th ed. Allyn & Bacon, 1971.

Anastasi, A. Psychological testing. 3rd ed. New York: The Macmillan Company, 1968.

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. 2nd. ed. Washington: American Council on Education, 1971.

Applebee, A. N. Silent reading tests: What do they measure? School Review, 1971, 80, 86-93.

Berg, P. C. Evaluating reading abilities. In W. H. MacGinitie (Ed.), Assessment problems in reading. Newark, Del.: International Reading Association, 1973.

Betts, E. A. Foundations of reading instruction. New York: American Book Co., 1946.

Bliesmer, E. P. Evaluating progress in remedial reading programs. In R. Farr (Ed.), Measurement and evaluation of reading. New York: Harcourt, Brace, and World, 1970.

Bloom, B. S. Testing cognitive ability and achievement. In N. L. Gage (Ed.), Handbook of Research on teaching. Chicago: Rand McNally & Co., 1963.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Bormuth, J. R. Reading literacy: Its definition and assessment. Reading Research Quarterly, 1973-1974, 9, 7-66.

Buros, O. K. Reading tests and reviews. Highland Park, N. J.: Gryphon Press, 1968.

Carver, R. P. Reading as reasoning: Implications for measurement. In W. H. MacGinitie (Ed.), Assessment problems in reading. Newark, Del.: International Reading Association, 1973.

Cronbach, J. Essentials of psychological testing. 3rd. ed. New York: Harper & Row, 1970.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Assessment problems in reading. Newark, Del.: International Reading Association, 1973.

Davis, F. B. The assessment of change. In R. Farr, (Ed.), Measurement and evaluation of reading. New York: Harcourt, Brace, and World, 1970.

Davis, F. B. Psychometric research on comprehension in reading. Reading Research Quarterly, 1972, 8. 628-678.

Della-Piano, G. M. Reading diagnosis and prescription: An introduction. New York: Holt, Rinehart, & Winston, 1968.

Goodman, K. S. Testing in reading: A general critique. In R. B. Ruddell (Ed.), Accountability and reading instruction. Urbana, Illinois: National Council of Teachers of English, 1973.

Greene, H. A., & Jorgensen, A. N. The use and interpretation of educational tests. New York: Longmans, Green, & Com., 1929.

Hambleton, R. K., & Novick, R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, A. J. A comparison formulas for measuring degree of reading disability. In R. E. Leibert (Ed.), Diagnostic viewpoints in reading. Newark, Del.: International Reading Association, 1971.

Harris, A. J. Some new developments in readability. Paper presented at the Fifth IRA World Congress on Reading, Vienna, Austria, August 1974.

Harris, A. J., & Jacobson, M. D. Revised Harris-Jacobson readability formulas. Paper presented at the meeting of the College Reading Association, Bethesda, Md., October 1974.

Kamons, J. Letter to writer, December 2, 197 .

Lennon, R. T. Assumptions underlying the use of content validity. In L. R. Aiken, Jr. (Ed.), Readings in psychological and educational testing. Boston: Allyn & Bacon, 1973.

Lennon, R. T. What can be measured? In R. Farr (Ed.), Measurement and evaluation of reading. New York: Harcourt, Brace, and World, 1970.

Lincoln, E. A. Beginnings in educational measurement. Philadelphia: J. B. Lippincott Co., 1924.

Linden, K. W., & Linden, J. D. Modern mental measurement: A historical perspective. Boston: Houghton-Mifflin Co., 1968.

McCall, W. A. How to measure in education. New York: Macmillan, 1927.

McCall, W. A. Measurement. Revision of How to measure in education. New York: Macmillan, 1939.

MacGinitie, W. H. (Ed.), What are we testing? In his Assessment problems in reading. Newark, Del.: International Reading Association, 1973.

McLaughlin, K. F. Interpretation of test results. DE-25038. Washington: U. S. Government Printing Office, 1964.

McLaughlin, K. F. Understanding testing: Purposes and interpretation for pupil development. U. S. Dept. of Health, Education, and Welfare, Office of Ed. O-E-25003. Washington: Government Printing Office, 1960.

National assessment of educational progress: Reading objectives. 72-111034. Ann Arbor, Michigan: National Assessment of Educational Progress, 1970.

O'Connor, E. F., Jr. Extending classical test theory to the measurement of change. Review of educational research, 1972, 42, 73-97.

Otto, W., Chester, R., McNeil, J., & Myers, S. Focused reading instruction. Reading, Mass.: Addison-Wesley, 1974.

Otto, W. Thorndike's "Reading as reasoning": Influence and impact. Reading Research Quarterly, 1971, 6, 435-442.

Pyrczak, F., Jr. Objective evaluation of the quality of multiple-choice test items designed to measure comprehension of reading passages. Reading Research Quarterly, 1972, 8, 62-71.

Robeck, M. C., & Wilson, J. A. R. Psychology of reading: Foundations of instruction. New York: John Wiley & Sons, Inc., 1974.

Smith, H. L, & Wright, W. W. Tests and measurements. New York: Silver, Burdett & Co., 1928.

Spache, G., & Spache, E. B. Reading in the elementary school. Boston: Allyn & Bacon, Inc., 1969.

Spache, G. Toward better reading. Champaign, Illinois: Garrard Publishing Co., 1963.

Standards for educational and psychological tests and manuals. Washington, D. C.: American Psychological Assn., Inc., 1966.

Stauffer, R. G. Thorndike's "Reading as reasoning": A perspective. Reading Research Quarterly, 1971, 6, 443-448.

Thorndike, E. L. Reading as reasoning: A study of mistakes in paragraph reading. Journal of Educational Psychology, 1917, 8, 323-332. Reprinted in Reading Research Quarterly, 1971, 6, 425-434.

Thorndike, R. L., & Hagen, E. Measurement and evaluation in psychology and education. 3rd. ed. New York: John Wiley & Sons, 1969.

Tuinman, J. J. Thorndike revisited--some facts. Reading Research Quarterly, 1971, 6, 195-202.

Wood, D. A. Test construction: Development and interpretation of achievement tests. Columbus, Ohio: Charles E. Merrill, 1960.