DOCUMENT RESUME

ED 109 246 TM 004 710

AUTHOR Hathaway, Walter E.
TITLE The Appropriate and Inappropriate Uses of Grade Level Equivalents in School Evaluation.
PUB DATE [Apr 75]
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975); Not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.76 PLUS POSTAGE. HC Not Available from EDRS.
DESCRIPTORS Academic Achievement; Achievement Tests; Comparative Analysis; Educational Assessment; Elementary Education; Evaluation Methods; *Grade Equivalent Scores; Measurement Techniques; *School Districts; Standardized Tests; Student Evaluation; Testing; *Testing Problems; *Testing Programs; *Test Interpretation; Test Results

ABSTRACT

The Portland Board of Education had requested that the Oregon Central Evaluation Department provide student achievement data so as to allow comparisons with other school districts by reporting national grade level equivalent (GLE) scores on standardized tests of reading and mathematics for grades 4 and 8. For years, the position of most research and evaluation personnel in Portland's district has been that national GLEs are an inadequate and misleading type of score for representing student achievement in the district. This position has been based on information about the discrepant meaning of GLEs from test to test and also upon certain technical characteristics of these scores that might make them unsuitable for research and evaluation purposes. This paper discusses the advantages, disadvantages, differences in variations, interpretations, interpolations and alternatives to reporting GLEs and other standardized scores. (Author/DEP)

The Appropriate and Inappropriate Uses Of Grade Level Equivalents In School Evaluation

By

Walter E. Hathaway
Portland Public Schools
631 N.E. Clackamas Street
Portland, Oregon
97208

Presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., April 1, 1975.

Like the discussion of the other topics in this Division H symposium the present review of the question of whether to report test data in terms of Grade Level Equivalents arose out of a situation in a school district which may find a parallel in the experience of some other members of AERA. It is hoped that this discussion will help toward the creation and sharing of workable solutions to common research and evaluation problems, including their real and important political and human dimensions.

The Problem

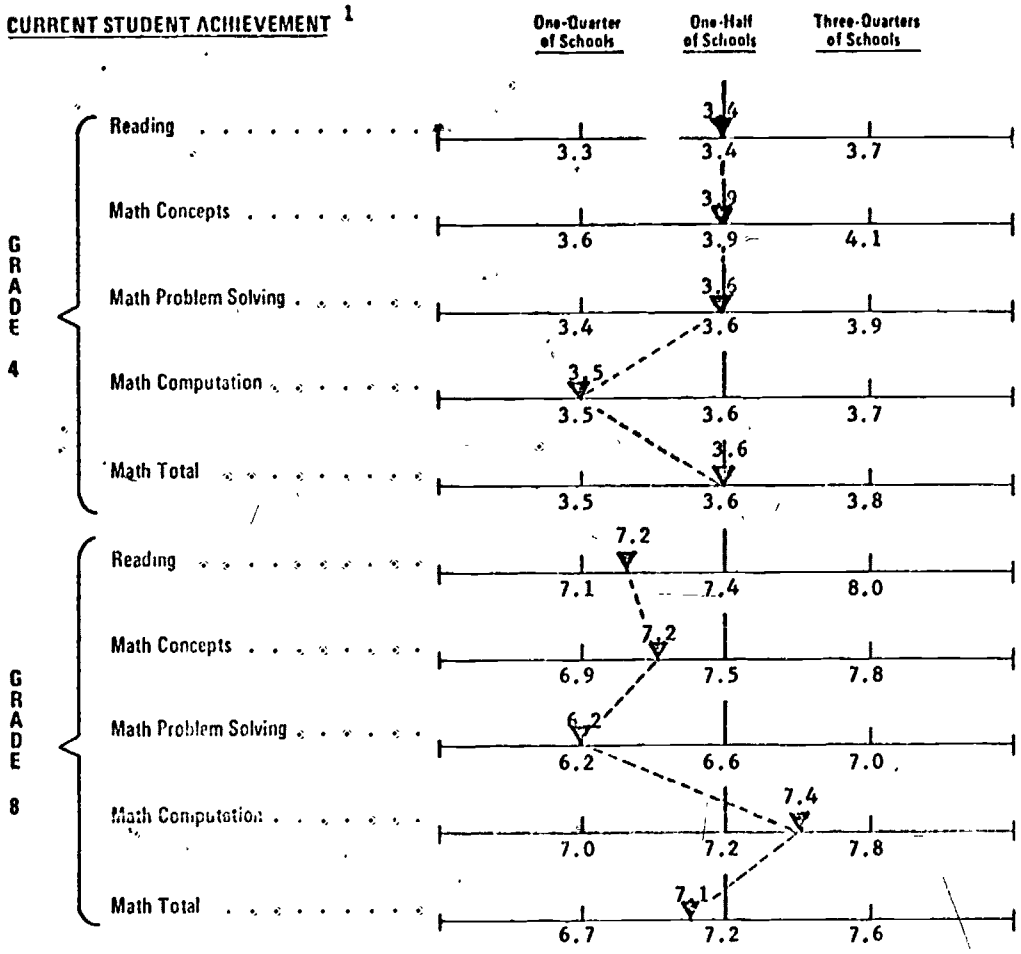
As late as 1973 the Portland, Oregon Central Evaluation Department found itself responding to the compellingly expressed need of its Board of Education for "data on student achievement allowing comparison with other school districts" by reporting national Grade Level Equivalent scores on standardized tests of Reading and Mathematics at grades 4 and 8 (see Figure 1). This occurred in spite of the fact that Portland had been one of the first cities in the country to move to locally developed and normed tests, having completed development of such a program well before 1970. It also transpired

Figure 1

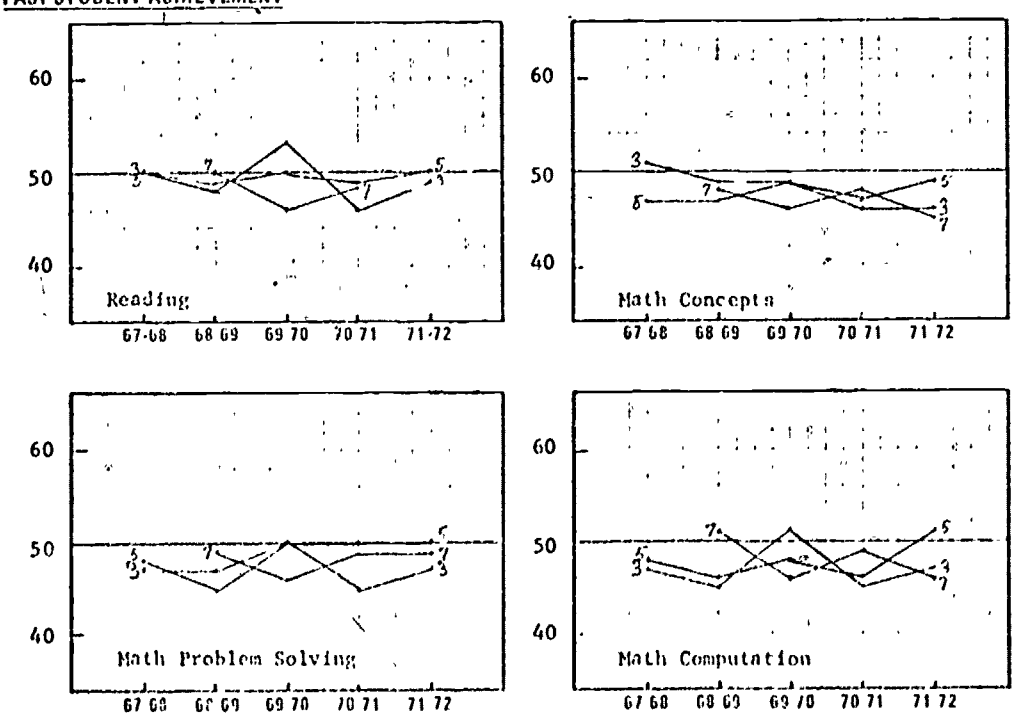
Area III — Elementary School Profile — 64.

CRESTON

CURRENT STUDENT ACHIEVEMENT [1]



PAST STUDENT ACHIEVEMENT [2]



NOTES ON ACHIEVEMENT DATA:

1. Fall '72, Ntl. G.E. Ave. (CTBS, Norm) 4.0 & 8.1 for grades 4 & 8 respectively.
2. 5-Yr. period '67-68 thru '71-72, Local (Norm-base Yr. '69-70). P-score Ave. 50.

in the face of continuing efforts to inform board members and other district leaders of the limitations of national Standardized Tests in general and Grade Level Equivalents as a means of reporting their results in particular.

Drs. Mazer and Hansen have already reviewed some of the reasons urged against national Standardized Tests which led the district to return in 1974 to reporting standard scores on locally developed and normed tests for our district wide testing program (see Figure 2). And you are all familiar with the limitations and merits of Grade Level Equivalents since they have been well and frequently documented (Flanagan, 1951; Coleman, 1970; Thorndike, 1971; Davis, 1974). Nevertheless, having this information recounted again in terms which helped one district toward a better testing system may help others in similar situations. And a report of some efforts to discover and develop even more responsive measuring and reporting systems than those currently available may be of even greater interest.

Method of Derivation of the Grade Equivalent Scale

The process of deriving a Grade Equivalent scale is commonly begun with a test, usually an achievement test, being given to large and hopefully representative groups of students in the consecutive grades for which it is desired to report the Grade Equivalents. The test is administered at the same time of year for all pupils, usually at the end of the year. The average raw score of each grade level is then found and plotted against grade level. Next, a curve is fitted and smoothed to connect the points thus plotted. Often the curve is extrapolated to cover upper and lower grades. Finally, tables of the raw scores paired with each tenth of a Grade are prepared.

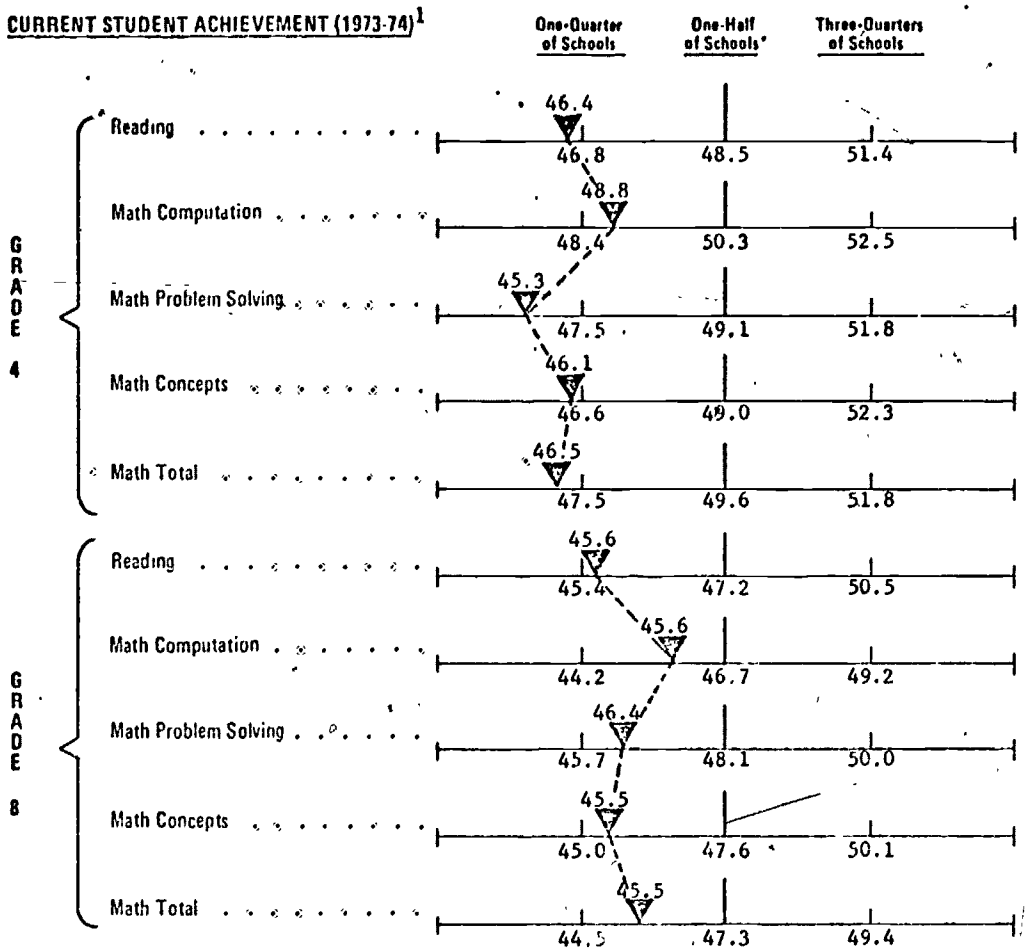
Disadvantages

Many of the possible and actual limitations of Grade Level Equivalents

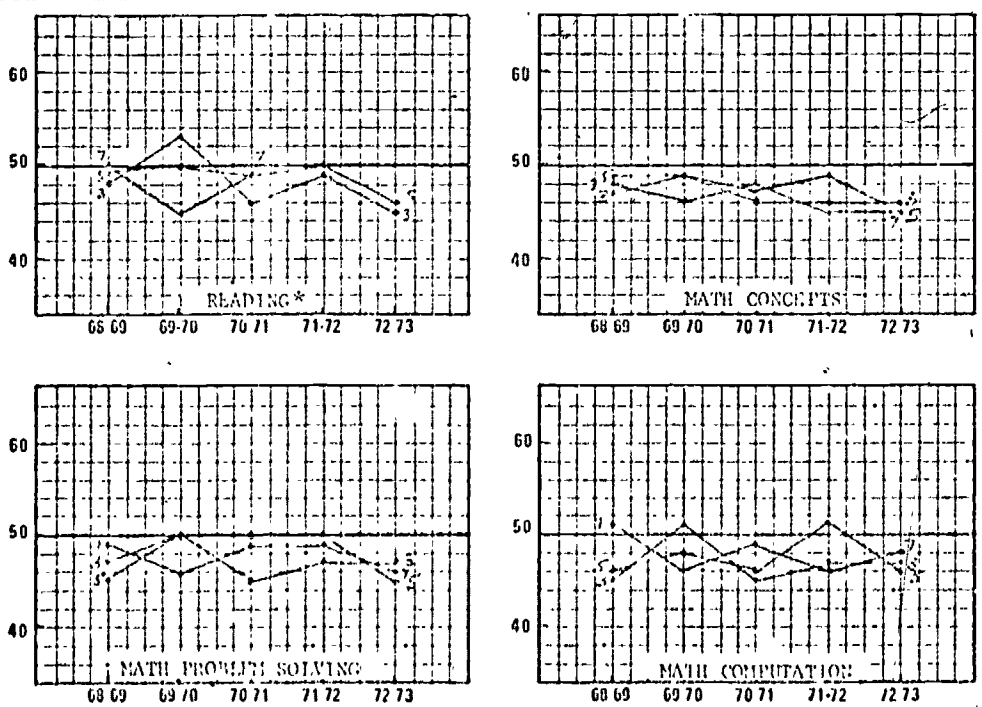
Figure 2

Area III — CRESTON — Elementary School Profile 77.

CURRENT STUDENT ACHIEVEMENT (1973-74)[1]



PAST STUDENT ACHIEVEMENT (1968-69 -- 1972-73)[2]



NOTES ON ACHIEVEMENT DATA:

1. Fall '73, Metropolitan Area Ave. M-score 50.
2. 5-yr. period '68-69 thru '72-73, Local (Norm-base yr. '69-70) P-score Ave. 50.

* Gr. 7 Metropolitan Advanced Reading Test, Form Bm, not administered after '70-71.

arise from the way in which these scales are commonly established, and still other limitations arise from the nature of the scale itself. The limitations result in such disadvantages as the five listed below.

1. Interpretation - The naive interpretation is often wrong, e.g. a sixth grader scoring at the eighth grade level is probably not "performing" at the level of an average eighth grader in the sense that he or she knows about the same things about as well as the more advanced student. He or she is, however, probably performing exceptionally well on the items dealing with sixth grade matter.

2. Uniform Growth and Emphasis - Within a subject the units of measurement do not represent reasonably equivalent amounts of subject matter being measured, e.g."a gain from a grade-equivalent score of 6.9 to 7.9 (on a test of Arithmetic Computation) indicates that a student has improved about thirteen times as much as a grade-equivalent score of 1.9 to 2.9"[1] Moreover, to the extent that the assumption that the same curriculum and consistent emphasis is shared within a subject by the norm and test groups is violated, any comparisons between these two and among test groups using Grade Equivalents is invalidated.

3. Differences in Variation - From subject to subject the same Grade Level Equivalents mean different things, e.g., a fifth grade student receiving a Grade Equivalent score of 7.0 on a test of arithmetic may stand at the ninety-fifth percentile relative to his grade group. Whereas, the same result on another test, say of reading, where correlation between grade and test score is lower, may indicate only a standing at the sixtieth percentile.

---

1. Davis, Frederick B. Educational Measurements and Their Interpretation. Belmont, California: Wadsworth Publishing Company, 1964, p. 40.

4. Interpolation - It is common to interpolate between testing groups by fitting and smoothing a curve between the plotted points. This process involves the application of questionable assumptions about the nature and course of learning. Grade norms are most appropriate only for elementary school subjects which are studied continuously at fairly commonly increasing levels of difficulty over the grades. Grade Equivalents should never extend beyond the ninth grade since there is little continuous and systematic instruction beyond that grade for the subjects taught in elementary school.

5. Extrapolation - It is also common to extrapolate from the curve to low and high grades. To the extent that this is the case, reported scores are almost worthless due to unreliability and invalidated judgment.

Such disadvantages as those listed below have led the authors and editors of Standards for Educational and Psychological Tests to make some strong warnings about the use of Grade Equivalents. These include:

D5.23 "Interpretive scores which lend themselves to gross misinterpretations such as mental age or grade equivalent scores, should be abandoned or their use discouraged." (italics added)

AND J5.2 "Test users should avoid the use of terms such as I.Q., I.Q. equivalent, or grade equivalent where other terms provide more meaningful interpretations of a score." (italics added)

An analysis by Dr. George Ingebo (Evaluation Specialist in Portland's Area III) of the recent report of the ANCHOR Test Study provides a means of verification of the impact of the technical and practical limitations of Grade Level Equivalent scores. Figure 3 is a table showing the discrepancies between the Grade Level Equivalents reported among four well known and

Figure 3

Grade 5

Vocabulary				Comprehension			
CTBS	ITBS	MAT	SAT	CTBS	ITBS	MAT	SAT
3.5	- .2	0	+ .1	3.5	+ .4	+ .3	+ .3
4.0	- .2	- .1	- .1	4.0	+ .3	+ .4	+ .2
4.5	- .1	- .1	- .4	4.5	+ .2	+ .3	- .1
5.0	- .1	0	- .4	5.0	+ .1	+ .1	- .3
5.5	- .1	- .1	- .6	5.5	- .1	+ .2	- .5
6.0	- .3	0	- .8	6.0	- .3	0	- .6
7.0	- .4	+ .3	- .8	7.0	- .8	- .5	-1.0
8.0	- .8	+ .1	-1.1	8.0	-1.1	- .7	-1.4
9.0	-1.1	+ .8	-1.2	9.0	-1.8	-1.2	-2.0

widely used standardized tests. In the table is reported the discrepancy between the Grade Equivalents for the Fifth Grade California Test of Basic Skills (CTBS) and the Iowa Test of Basic Skills (ITBS), The Metropolitan Achievement Test (MAT) and the Standford Achievement Test (SAT). It seems apparent from this data that foreknowledge of even very roughly where a majority of students might score (low, medium, high) would allow an unscrupulous test director to improve his or her district's apparent performance by as much as two Grade Equivalents.

Advantages

One positive thing is occasionally said about Grade Level Equivalents. Even though test users are constantly misinterpreting Grade Equivalents in the ways we have been describing, nevertheless they like these scores because of their apparent familiarity, simplicity and directness of meaning. Grade Level Equivalents seem, in short, more easily understood.

When we consider that this apparent understandability is in fact largely merely apparent and that the choice of the Grade Equivalent scale is often a choice to "misunderstand in comfort" rather than to make the additional effort necessary to understand correctly then even this sole positive thing to be said about Grade Level Equivalents doesn't seem very compellingly in favor of their use.

Alternatives

Traditional alternatives to Grade Level Equivalents have included percentile rank within grade score, Z-scores, K-scores, stanines, etc. All of these scores with the aid of good reporting techniques are capable of being rendered as apparently understandable as the Grade Level Equivalent without the dangers of misinterpretation inherent in that form of conversion

(refer again to Figures 1 and 2 for a comparison of the understandability of Grade Level Equivalents and standard scores when embedded in a well designed graphic reporting format).

In Portland exploration of another alternative is underway, an approach to testing based upon the Rasch model. That model may provide for interval scaling of both test scores and individual test items on the underlying trait being measured. Work is currently in progress toward the building up of a pool of items calibrated by the model through the cooperation of a number of districts in the Northwest Evaluation Association and toward a simultaneous verification of the validity of the model. The existence of such a pool of calibrated items related to the comprehensive set of learning outcomes developed by the Tri-county Goal Development Project would allow accurate reporting of student progress toward goals set at the classroom and individual student level, thus meeting the instructional purposes of measurement.[2] It would simultaneously permit comparable reports of aggregate student performance at the building, area and district levels, thus satisfying the administrative and management uses of testing. Moreover, although the Rasch approach does not provide norms itself, the capability to equate test results through this technique makes it possible to take advantage of available norming information when and if such information should also be required for further administrative and management purposes.

---

2. Doherty, Victor W. and Walter E. Hathaway, Designing behavioral goals, K-12. Oregon Association for Supervision and Curriculum Development Curriculum Bulletin Volume 27, No. 320, December, 1973.

Conclusion

There are very few cases where the numerous assumptions which must be met in order for Grade Level Equivalents to be free of serious distortion are in fact satisfied. In view of this it seems best to avoid the use of these conversions entirely. With a little care existing derived scales which are relatively free from at least some of the dangers inherent in Grade Level Equivalents can be rendered similarly "understandable" to users. Current explorations of such promising approaches as the Rasch model may lay the groundwork for valid comparisons among locally autonomous programs while at the same time providing needed information on the progress of individuals and groups of students toward attaining the specific learning outcomes sought within those programs.

References

Ahman, J.S. & Glock, M.D. Evaluating Pupil Growth. Boston: Allyn & Bacon, Inc., 1967.

Coleman, James S. Measures of School Performance. Rand Corp., 1970

Cronbach, Lee J. Essentials of Psychological Testing, third edition. New York: Harper and Row, 1970.

Davis, Frederick B. Educational Measurements and Their Interpretations. Belmont, California: Wadsworth Publishing Co., 1964.

Davis, Frederick B. (Chair) Standards for Educational and Psychological Tests. Washington, D.C.: American Psychological Association, Inc., 1974.

Doherty, V.W. & Hathaway, W.D. Designing behavioral goals, K-12. Oregon Association for Superivision and Curriculum Development Curriculum Bulletin, December, 1973, 27(320).

Ebel, Robert L. Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentiss Hall, Inc., 1972.

Flanagan, J.C. Units, scales, and norms. In E.F. Lindquist (Ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1951.

Hansen, J.B. Content referenced vs. norm refereneed testing. Paper presented at the annual meeting of The American Educational Research Association. Washington, D.C.: 1975.

Rasch, G. An Individualistic Approach to Item Analysis. Lazarsfeld and H.W. Henry (Eds.) Readings in Mathematicsl Social Science. Chicago: Science Research Associates, 1966, pp. 89-108.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.

Thorndike, R.G. Educational Measurement, second edition. Washington, D.C.: American Council on Education, 1971.

Whitely, S.E. & Davis R.V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, Fall, 1974, 11(2)

Wright, B. & Panchapakesan, N. A procedure of sample-free item analysis. Educational and Psychological Measur[illegible] 1969, 29, pp. 27-40.

Wright, B. Sample-free test calibration a[illegible]asch free measurement proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1967, pp. 85-101.

WH:ep
3/20/75