Journal of School Psychology
1981•Vol. 19, No. 4

0022-4405/81/1600-0350$00.95

THE FALLACY OF "TWO YEARS BELOW GRADE LEVEL FOR AGE" AS A DIAGNOSTIC CRITERION FOR READING DISORDERS

CECIL R. REYNOLDS

Texas A & M University

Summary: The use of grade equivalent scores contrasted against grade placement is widespread in the diagnosis of dyslexia and other reading disabilities. This method substantially overstates disabilities at upper grade levels while underestimating the severity of difficulties in the early grades. Other difficulties and distortions of this method are also pointed out. An accurate, alternative method for reliably determining aptitude/achievement discrepancies is presented and its use discussed.

Researchers and practitioners in the field of learning disabilities and its related specialties (e.g., neuropsychology, school psychology, special education, neurology, etc.) have long sought to quantify the severity of learning disorders with respect to the general intellectual level of the individual. That is, how much of a discrepancy must exist between an individual's intellectual level and his or her level of academic attainment in a specific subject matter area before the diagnosis of a learning disorder can appropriately be made? A variety of formulae and methods have been devised over the years, yet most have been found to be inappropriate, as was noted in the first set of federal rules on specific learning disabilities drawn in conjunction with the Education for All Handicapped Children Act of 1975 (P.L. 94-142).

One method that has gained rather widespread acceptance over the years is that of designating a child with normal intelligence who scores two years below grade level for age on a standardized achievement test as learning disabled. On the surface, this approach gives the appearance of being objective, stable, and easily quantifiable; it also seems to allow for replication of research by operationally defining "learning disabled" subjects in a clear, repeatable fashion. Since most all definitions of specific learning disabilities refer to what are basically normal levels of intelligence (e.g., Chalfant & Scheffelin, 1969; Gearheart, 1973; Johnson & Morasky, 1980; Mercer, 1979; Schroeder, Schroeder, & Davine, 1978; Wright, Schaefer, & Solomons, 1979), it seems to follow quite logically that a constant discrepancy between grade placement and grade equivalent scores would be a desirable feature in operationally defining dyslexic or reading disabled populations. Indeed, at the recent 1980 meeting of the International Neuropsychological Society, the operational definition of reading scores "two years below grade level for age" was a popular method of defining subjects for the study of dyslexia (e.g., Eskenazi & Diamond, Note 1). Reports of research in the professional literature on learning disabilities frequently refer to the identification of children as learning disabled[1] on the basis of

Requests for reprints should be addressed to Cecil R. Reynolds, Department of Educational Psychology, Texas A & M University, College Station, TX 77843.

[1]All too frequently the subject descriptions in studies of learning disabled children do not include the criteria used for diagnosis of a learning disability.

academic performance "below age-appropriate grade level" typically meaning 1 to 2½ years behind grade placement (Chapman & Boersma, 1979; Gottesman, 1979; Maloy & Sattler, 1979; Selz & Reitan, 1979; Smith & Rogers, 1978; Wallbrown, Vance, & Prichard, 1979; Leong, Note 2; Williams, Note 3). At the time this manuscript was prepared, the most recently published text in neuropsychology (Kolb & Whishaw, 1980), in defining learning disabilities, stated that "A commonly used criterion for establishing learning disability is that the child is at least *two years behind* [emphasis added] in a particular subject, such as reading or arithmetic, but not in others" (p. 456). An informal survey of school districts and clinicians conducted by the author revealed the two years below grade level for age criterion for a learning disabilities diagnosis to be quite prevalent. Among school districts, the "two years" definition was included either as the sole criterion or as one alternative criterion in a set of possible conditions that would indicate the presence of a specific learning disability.

While on the surface the two years criterion gives the appearance of constance and objectivity, this is not the case. The use of grade equivalent scores at a constant discrepancy level irrespective of actual grade placement produces considerable irregularity and distortion in the magnitude of aptitude/achievement discrepancies required for diagnosis of a learning disability across grade levels. The remainder of this paper will focus on an explanation of the difficulties involved in the use of grade equivalent scores in the evaluation of individual pupils, followed by the presentation of an alternative strategy designed to re-estabilsh the constancy and reliability of aptitude/achievement discrepancies across age.

THE TROUBLE WITH GRADE EQUIVALENTS

While grade equivalents never really qualify as a form of "standard scores," many seem to interpret them almost as though they have the psychometric qualities of standard scores. One should not make such interpretations, since grade equivalents have a number of major difficulties that distort their meaning. Nearly all of these difficulties can be traced to one of two factors: (a) grade equivalents ignore the dispersion of scores about the mean, and (b) the regression between grade and test score is not equivalent across grades or school subjects. The following discussion of the many interpretive problems these two difficulties and others present with respect to grade equivalents is taken substantially from Angoff (1971), Flanagan (1951), Gulliksen (1950), and Thorndike and Hagen (1977). The distortion in interpreting discrepancies between grade equivalent scores and grade placement are readily apparent in Table 1. Table 1 was developed from data available in the norms or technical manuals of the Wide Range Achievement Test (WRAT), Peabody Individual Achievement Test (PIAT), Woodcock Reading Mastery Test (WRMT), and the Stanford Diagnostic Reading Test (SDRT). As is typical of grade equivalent scores, some occasional interpolation was necessary to derive the exact values in Table 1. It is apparent from this table, however, that a third grader who reads two years below grade level for age has a much more severe problem than say a seventh or eighth grader reading two years below grade level. In fact, a twelfth grader with an IQ of 90 reading two years below grade level for age has no reading problem at all, but rather reads at a slightly higher than expected level. The rather dramatic changes

Table 1
Standard Scores and Percentile Ranks Corresponding to Performance Two Years Below Grade Level for Age on Four Major Reading Tests

Grade Placement	Two Years Below Placement	Wide Range Achievement Test		Peabody Individual Achievement Test[a]		Woodcock Reading Mastery Test[b]		Stanford Diagnostic Reading Test[b]	
		SS[c]	%R[d]	SS	%R	SS	%R	SS	%R
1.5	Pk.5	65	1	—		—		—	
2.5	K.5	72	3	—		—		—	
3.5	1.5	69	2	—		64	1	64	1
4.5	2.5	73	4	75	5	77	6	64	1
5.5	3.5	84	14	85	16	85	16	77	6
6.5	4.5	88	21	88	21	91	27	91	27
7.5	5.5	86	18	89	23	94	34	92	30
8.5	6.5	87	19	91	27	94	34	93	32
9.5	7.5	90	25	93	32	96	39	95	37
10.5	8.5	85	16	93	32	95	37	95	37
11.5	9.5	85	16	93	32	95	37	92	30
12.5	10.5	85	16	95	37	95	37	92	30

[a] Reading Comprehension subtest only.

[b] Total Test.

[c] All standard scores in this table have been converted for ease of comparison to a common scale having a mean of 100 and a standard deviation of 15.

[d] Percentile rank.

in percentile ranks (from 1 to 37) apparent in Table 1 underscore the magnitude of the changes in the distribution of grade equivalent scores as grade placement increases. These changes are not just inherent to limited or specific areas of reading skill but rather, as Table 1 shows, are true for word recognition, reading comprehension, and the more general reading process. Standard scores are by far more accurate representations of an individual's achievement level than grade equivalents, since they are based not only on the mean at a given age level but also on the distribution of scores about the mean. Thus, in the case of deviation standard scores, such as the Wechsler IQs, the relationship between standard scores is constant across age.

Grade equivalents are also inappropriate for use in any sort of discrepancy analysis of an individual's test performance for the following reasons:

1. The growth curve between age and achievement in basic academic subjects flattens out at upper grade levels. This can also be observed in Table 1 where it is seen that there is very little change in standard score values corresponding to two years below grade level for age after about grades 7 or 8. In fact, grade equivalents have almost no meaning at this level since reading instruction typically stops by high school and grade equivalents are really only representing extrapolations[2] from earlier grades. An excellent example of the difficulty in interpreting grade equivalents beyond about grade 10 and 11 has been provided by Thorndike and Hagen (1977) using an analogy with age equivalents. Height can be expressed in age equivalents just as reading can be expressed as grade equivalents. But, while it might be helpful to describe a tall first grader as having the height of an 8½ year old, what happens to

[2]Extrapolation is a sophisticated term for "fancy guesswork."

the 5'10" 14-year-old female since at no age does the mean height of females equal 5'10". Since the average reading level in the population changes very little after junior high school, grade equivalents at these ages become virtually nonsensical, with large fluctuations resulting from a raw score difference of 2 or 3 points on a 100-item test.

2. Grade equivalents assume that the rate of learning is constant throughout the school year and that there is no gain or loss during summer vacation.

3. As partially noted above, grade equivalents involve an excess of extrapolation at the upper and lower ends of the scale. However, since tests are not administered during every month of the school year, scores between the testing intervals (often a full year) must be interpolated on the assumption of constant growth rates. Interpolation between sometimes extrapolated values on an assumption of constant growth rates is at best a highly perilous activity.

4. Different academic subjects are acquired at different rates and the variation in performance varies across content areas so that two years below grade level for age may be a much more serious deficiency in math, for example, than in reading comprehension.

5. Grade equivalents exaggerate small differences in performance between individuals and for a single individual across tests. Some test authors (e.g., Jastak & Jastak, 1978) even provide a caution on test record forms that standard scores only, and not grade equivalents, should be used for comparison purposes.

THE ADVANTAGES OF STANDARD SCORES

The primary advantage of standardized or scaled scores lies in the comparability of score interpretation across age. By standard scores, of course, we refer to scores of the Wechsler Deviation IQ genre and not to ratio IQ types of scales employed by the early Binet and current Slosson Intelligence Test. Ratio IQs or other types of quotients have many of the same problems as grade equivalents and should be avoided for many of these same reasons. Standard scores of the deviation IQ type have the same percentile rank across age since they are based not only on the mean but the variability in scores about the mean at each age level. For example, a score that falls ⅔ of a standard deviation below the mean (90 on a Wechsler scale) has a percentile rank of 25 at every age. A score falling ⅔ of a grade level below the average grade level has a different percentile rank at every age.

Standard scores are more accurate and precise. When constructing tables for the conversion of raw scores into standard scores, interpolation of scores to arrive at an exact score point is typically not necessary, whereas the opposite is true of grade equivalents. Extrapolation is also typically not necessary for scores within three standard deviations of the mean, which accounts for more than 99% of all scores encountered.

AN ALTERNATIVE METHOD FOR ACCURATELY DETERMINING APITITUDE/ACHIEVEMENT DISCREPANCIES

It may not be possible to arrive at a constant score differential that represents a real or reliable aptitude/achievement discrepancy in any area of academic achievement. This is due to the fact that the size of the score discrepancy required to indicate a real difference in performance is a function of the standard deviations of

the tests involved and, more importantly, the reliability of the two tests involved. It is possible, however, to answer the question "Does this individual's performance on a test of reading skill (or other subject matter area) significantly differ from his/her performance on a test of aptitude or intelligence?" provided that an hypothesis testing approach is adopted. The hypothesis testing approach requires some mathematical work, but is worthwhile since it reveals whether or not the difference in scores between an aptitude and an achievement measure is real or due to chance and errors of measurement. Fortunately, the statistical formulations necessary for determining the reliability of the discrepancy between two scores for a single individual have been worked out previously by Payne and Jones (1957).

To determine the significance of the difference between two scores for an individual on two tests, the following equations must be solved consecutively:

$$Z_x = \frac{(X_i - \bar{X})}{\sigma_x} \quad (1)$$

$$Z_Y = \frac{(Y_i - \bar{Y})}{\sigma_y} \quad (2)$$

$$Z_D = \frac{|D_z|}{\sqrt{(1 - r_{XX}) + (1 - r_{yy})}} \quad (3)$$

Where:

X_i = individual's score on test X
Y_i = individual's score on test Y
$\bar{X}$ = mean standard score on test X
$\bar{Y}$ = mean standard score on test Y
σ_x = standard deviation of scores on test X
σ_y = standard deviation of scores on test Y
$|D_Z|$ = the absolute value of $(Z_x - Z_Y)$
r_{xx} = internal consistency reliability coefficient for test X
r_{yy} = internal consistency reliability coefficient for test Y

Once these equations have been solved, the test statistic from equation (3), Z_D is referred to a table of the normal curve. Since, typically, referred individuals are believed to have specific academic deficits, a directional test of the hypothesis may be appropriate. The values of Z_D required at three commonly employed levels of significance for a one-tailed test are: 1.65, $p = .05$; 2.33, $p = .01$; 3.08, $p = .001$. For a two-tailed test, the values are: 1.96, $p = .05$; 2.58, $p = .01$; 3.28, $p = .001$.

AN ILLUSTRATION

A nine-year-old boy was referred for evaluation. The teacher suspected that the child was learning disabled, since he seemed to be far behind other children in the acquisition of most basic reading skills. As one part of the evaluation, a standardized aptitude test was administered (having a mean of 100 and *SD* of 15) along with a test

of reading comprehension (having a mean of 100 and *SD* of 20) and a variety of other achievement and adaptive behavior scales. The child earned an IQ of 90 on the aptitude scale and a 75 on the reading comprehension test. Mental retardation was subsequently ruled out. To determine whether the aptitude/achievement discrepancy with IQ and reading was real or due to errors of measurement, the above formulae were applied. The IQ test (test X) had a reliability coefficient of .95 at age 9, while the reading test had a reliability coefficient of .84 at this age. Substituting in the equations we have:

$$Z_X = \frac{(90 - 100)}{15} = -.667 \qquad (1)$$

$$Z_Y = \frac{(75 - 100)}{20} = -1.25 \qquad (2)$$

$$Z_D = \frac{|(-1.25) - (-.667)|}{\sqrt{(1 - .95) + (1 - .84)}} = \frac{.583}{.458} \qquad (3)$$

$$Z_D = 1.27$$

Since the value of Z_D does not reach the minimal value of 1.65 required for statistical significance at the .05 level of probability, one can say with confidence that this child's reading skill does not deviate significantly from his overall level of intellectual functioning, and that, unless other problems were located, the diagnosis of a learning disability and subsequent special placement or implementation of other remedial strategies for a reading disorder are not indicated.

While these computations are not complex, they can become time-consuming for the practitioner. Yet, if one frequently employs the same tests, a table of score discrepancies and their corresponding probabilities can easily be developed from the above formulae and a table of the normal curve (available in most statistics books). However, some rules of thumb that will be applicable to most situations can serve a useful purpose *if applied cautiously*. Rather than work through the equations each time a new test is given, one can be relatively certain that a significant difference exists between scores when:

1. the reliability of each test falls between .85 and .90 and a difference of at least one standard deviation exists between the two test scores;
2. the reliability of one test is between .85 and .90 and the other is between .90 and .95 and a difference of at least .75 standard deviations exists between the two test scores;
3. the reliability of each test falls between .90 and .95 and a difference of at least .66 standard deviations exists between the two test scores.

In order to apply these rules of thumb, the test scores must be expressed on a common scale. That is, the standard deviations must be equated. For this purpose, Table 2 was developed. This table can quickly equate standard scores for the vast majority of tests in use today. To equate scores and ascertain the number of standard deviations that the scores are apart, the table should be entered for each test on the appropriate column (corresponding to the mean and SD of the scale) and then read

Table 2
Conversion of Standard Scores Based on Several Scales
to a Commonly Expressed Metric

Score Points	SCALES									
	$\bar{X}=0$ $SD=1$	$\bar{X}=10$ $SD=3$	$\bar{X}=36$ $SD=6$	$\bar{X}=50$ $SD=10$	$\bar{X}=50$ $SD=15$	$\bar{X}=100$ $SD=15$	$\bar{X}=100$ $SD=16$	$\bar{X}=100$ $SD=20$	$\bar{X}=500$ $SD=100$	Percentile Rank
	2.6	18	52	76	89	139	142	152	760	>99
	2.4	17	51	74	86	136	138	148	740	99
	2.2	17	49	72	83	133	135	144	720	99
	2.0	16	48	70	80	130	132	140	700	98
	1.8	15	47	68	77	127	129	136	680	96
	1.6	15	46	66	74	124	126	132	660	95
	1.4	14	44	64	71	121	122	128	640	92
	1.2	14	43	62	68	118	119	124	620	88
	1.0	13	42	60	65	115	116	120	600	84
	.8	12	41	58	62	112	113	116	580	79
	.6	12	40	56	59	109	110	112	560	73
	.4	11	38	54	56	106	106	108	540	66
	.2	11	37	53	53	103	103	104	520	58
	0.0	10	36	50	50	100	100	100	500	50
	- .2	9	35	48	47	97	97	96	480	42
	- .4	9	34	46	44	94	94	92	460	34
	- .6	8	33	44	41	91	90	88	440	27
	- .8	8	31	42	38	88	87	84	420	21
	-1.0	7	30	40	35	85	84	80	400	16
	-1.2	6	29	38	32	82	81	76	380	12
	-1.4	6	28	36	29	79	78	72	360	8
	-1.6	5	26	34	26	76	74	68	340	5
	-1.8	5	25	32	23	73	71	64	320	4
	-2.0	4	24	30	20	70	68	60	300	2
	-2.2	3	23	28	17	67	65	56	280	1
	-2.4	3	21	26	14	64	62	52	260	1
	-2.6	2	20	24	11	61	58	48	240	<1

to the extreme left column. Once the two scores have been located in the extreme left column, one need only find the difference between the scores now located in the left column. This difference is expressed in standard deviations. For example: Test X has a mean of 100 and *SD* of 15. Test Y has a mean of 36 and *SD* of 6. Suppose an individual earns a score of 88 on test X and 28 on test Y. How far apart are these two scores? Reading to the extreme left column, a score of 88 on test X yields a score of -0.8 and a score of 28 on Test Y yields a score of -1.4. The difference value is 0.6. This means that these two scores are 0.6 standard deviations apart. Applying the rules above, it is highly unlikely that these scores are significantly different unless the reliability coefficients of the two tests are at about .95 or higher. If one of the reliabilities falls above .95, then one should probably solve equations 1, 2, and 3 to determine the exact probability of these scores representing real differences in performance levels.

A CAUTION

The method proffered above is insufficient in and of itself for the complete and accurate diagnosis of a reading disorder or other learning disability. It only answers

the question of whether there is a real difference in performance on two tests. Another question of interest should be the frequency of occurrence of a given discrepancy. Knowing that a discrepancy score represents a real difference in abilities does not indicate how often this occurs. The frequency of ability differences is somewhat surprising to most. On the WISC-R, for example, a Verbal-Performance IQ discrepancy of 12 points represents a statistically significant difference ($p < .05$) on these two scales, yet approximately one of three normal children show such a discrepancy (Kaufman, 1979). If the correlation between two measures is known, the distribution of difference scores can be estimated (Payne & Jones, 1957). The method of estimation breaks down at the extreme of the distribution however (Reynolds, 1979), just the area where exactness is of greater necessity in this case.

The method described above is quite superior to the two years below grade level for age and similar criteria however. Using the two years method with a random sample of 9th, 10th, 11th, or 12th grade pupils on three of the four tests listed in Table 1 results in more than 30% of these children being considered learning disabled in reading. In other subject matter areas, the percentages may be even greater. The frequency of the difference will also need to be considered in conjunction with the reliability of the difference. This relationship is not well-researched, though some very limited regression data are available on highly selective samples. Large-scale studies that would enable the development of frequency distributions of score discrepancies for individuals on major individually administered tests have not been conducted. The hypothesis testing approach described herein does have the advantage of providing a constant level of confidence in the appraisal of aptitude/achievement discrepancies as reflecting real differences in ability as opposed to errors of measurement, chance, or inappropriate methods of scaling (e.g., grade or age equivalents).

REFERENCE NOTES

1. Eskenazi, B., & Diamond, S *An analysis of visual scanning strategies in dyslexic children.* Paper presented to the annual meeting of the International Neuropsychological Society, San Francisco, February 1980.
2. Leong, C.K. *An investigation of spatial-temporal information processing in children with specific reading disability.* Unpublished doctoral dissertation, University of Alberta, Edmonton, Canada, 1974.
3. Williams, N.H. *Arousal and information processing in learning disabled children.* Unpublished doctoral dissertation, University of Alberta, Edmonton, Canada, 1976.

REFERENCES

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

Chalfant, J., & Scheffelin, M. *Central processing dysfunctions in children: A review of research.* National Institute of Neurological Disease and Stroke Monographs, 1969, No. 9.

Chapman, J.S., & Boersma, F.J. Learning disabilities, locus of control, and mother attitudes. *Journal of Educational Psychology,* 1979, *71,* 250-258.

Flanagan, J.C. Units, scores, and norms. In E.F. Lindquist (Ed.), *Educational Measurement* (1st ed.) Washington, D.C.: American Council on Education, 1951.

Gearheart, B. *Learning disabilities: Educational strategies.* St. Louis: C.V. Mosby, 1973.

Gottesman, R.C. Follow-up of learning disabled children. *Learning Disability Quarterly,* 1979, *2,* 60-69.

Gulliksen, H. *Theory of mental tests.* New York: John Wiley, 1950.
Jastak, J.F., & Jastak, S. *Wide range achievement test.* Wilmington, Del.: Jastak Associates, 1978.
Johnson, S.W., & Morasky, R.L. *Learning disabilities* (2nd ed.). Boston: Allyn & Bacon, Inc., 1980.
Kaufman, A.S. *Intelligent testing with the WISC-R.* New York: Wiley-Interscience, 1979.
Kolb, B., & Whishaw, I.Q. *Fundamentals of human neuropsychology.* San Francisco: W.H. Freeman, 1980.
Maloy, C.F., & Sattler, J.M. Motor and cognitive proficiency of learning disabled and normal children. *Journal of School Psychology,* 1979, *17,* 213-218.
Mercer, C.D. *Children and adolescents with learning disabilities.* Columbus, OH.: Charles Merrill, 1979.
Payne, R.W., & Jones, H.G. Statistics for the investigation of individual cases. *Journal of Clinical Psychology,* 1957, *13,* 115-121.
Reynolds, C.R. Interpreting the index of abnormality when the distribution of score differences is known: Comment on Piotrowski. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 401-402.
Schroeder, C.S., Schroeder, S.R., & Davine, M.A. Learning disabilities: Assessment and management of reading problems. In B.B. Wolman, J. Egan, & A.O. Ross (Eds.), *Handbook of Treatment of Mental Disorders in Childhood and Adolescence.* Englewood Cliffs, NJ: Prentice-Hall, 1978.
Selz, M., & Reitan, R.M. Rules for neuropsychological diagnosis: Classification of brain function in older children. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 258-264.
Smith, M.D., & Rogers, C.M. Reliability of standardized assessment instruments when used with learning disabled children. *Learning Disability Quarterly,* 1978, *1,* 23-31.
Thorndike, R.L., & Hagen, E.P. *Measurement and evaluation in psychology and education.* (4th ed.). New York: John Wiley, 1977.
Wallbrown, F.H., Vance, H.B., & Prichard, K.K. Discriminating between attitudes expressed by normal and disabled readers. *Psychology in the Schools,* 1979, *16,* 472-477.
Wright, L., Schaefer, A.B., & Solomons, G. *Encyclopedia of pediatric psychology.* Baltimore: University Park Press, 1979.

Cecil R. Reynolds
Associate Professor
Department of Educational Psychology
Texas A&M University
College Station, TX 77843

Manuscript received: March 17, 1980
Revision received: October 16, 1980