

**Readability and Performance on the  
New York State Grade Four and Grade Eight  
English Language  
Arts Examinations, 1999-2000**

G. DeMauro,  
Office of State Assessment  
January, 2001

Overview

Before January 1999, New York State English Language Arts assessments for elementary and intermediary grades included use of the Degrees of Reading Power (c.f. Koslin, Zeno, & Koslin, 1987). The assessment employs cloze methodology, which embeds reading questions in the text. Difficulty of items, then, derives from readability. As a consequence, scale values depended heavily on estimates of passage readability.

In the development of the new generation of State tests, readability formulae are used to screen items and to select passages, but scale scores are determined first from extensive field testing and then from the operational test itself. The very structure of the tests reduces any role that readability can play in scaling because readability is now only a proxy measure of difficulty, which is directly measured through field testing and operational testing, question by question. The process of test development in the new generation of tests includes: use of focus groups and committees of New York State teachers and educators to identify the important characteristics of the new tests in terms of the New York State Learning Standards, and to develop appropriate curriculum guides; identification of reading passages by the contractor using readability as one variable for inclusion; review of reading passages by the State Education Department specialists; review of proposed reading passages by content and sensitivity committees of New York State educators; review of surviving passages with their associated items by New York State educators; extensive field testing of surviving passages and items; statistical analyses of field

test results for difficulty, discriminability, guessing, and bias; standard setting focused on the appropriate score values to discriminate proficiency groups; final selection of items and passages using both contractor and State Education Department experts.

After operational administration of the examinations, item analyses and expert reviews were applied to the operational results, as well, to assure that the items and passages were appropriate and the proficiency levels discriminated the four levels of proficiency. Simply put, even though readability is a tool for passage selection and item design, it matters little whether or not the readability formulae, alone, identify the reading passages as accessible to the children when the field and operational test administrations among thousands of children, provide the direct empirical answer to that question.

Nevertheless, readability continues to be a concern among constituencies that were familiar with the previous generation of tests. To address these concerns, it is important to demonstrate that the construct measured by the tests is unaffected by the variation in readability in those passages that survive all of the field testing. The most direct evaluation of this, then, is to analyze the relationship of the items associated with each passage to the test as a whole, a direct measure of construct fidelity, and to relate, in turn, that fidelity to readability estimates of those passages.

## Methods

Correlational analyses were made to estimate the relationship between readability and fidelity to the construct. Several measures for readability were employed. For both examinations, the Dale-Chall readability estimate was used as well as the midpoint of the range of readability estimates of each passage using the Coleman-Liau, Farr-Jenkins-Paterson, Flesch

Reading Ease, Flesch Kincaid, Fry Graph, Gunning Fog, and Smog indices. For grade four, the Spache estimate was also used. To provide a measure of fidelity to the construct of the examinations, performance on each of the 28 grade four multiple choice items and each of the 25 grade eight multiple choice items was correlated to the total raw score totals which included both multiple choice and open-ended questions. We hypothesize that, if readability of reading passages interferes with measure of the construct, then these point biserial correlation coefficients for the items related to those passages should be sensitive to that contamination. That is, a significant correlation between point biserial values and readability would lead us to reject the hypothesis that there is no relationship between these variables.

The readability midpoints and the associated multiple choice questions were as shown in Table 1.

**Table 1**

**Midpoint Passage Readability Estimates  
And Associated Multiple Choice Question Numbers,  
New York State Grades Four and Eight English Language Arts  
Examinations, 1999-2000**

		<u>Grade 4</u>	<u>Grade 8</u>
Passage 1	questions	1 – 7	1 – 6
	mid readability	4.33	7.15
	Dale-Chall	5.06	6.52
	Spache	2.92	-----
Passage 2	questions	8 – 12	7 – 11
	mid readability	6.56	12.28
	Dale-Chall	5.04	7.48
	Spache	3.07	-----
Passage 3	questions	13 – 18	15 – 16
	mid readability	7.96	too short for measure
	Dale-Chall	6.02	too short for measure
	Spache	3.74	-----
Passage 4	questions	19 – 23	17 – 20
	mid readability	7.40	9.22

	Dale-Chall	6.31	7.97
	Spache	4.00	-----
Passage 5	questions	24 – 28	21 – 25
	mid readability	5.45	9.75
	Dale-Chall	6.29	6.82
	Spache	3.41	-----

To perform the analyses, the point biserial correlation coefficients (item to raw total correlations) were transformed to z-scores, an equal interval scale. The average values were converted back to point biserial correlation coefficient values. The means point biserial values are given in Table 2.

**Table 2**

**Mean Point Biserial Values  
By Passage, for New York State  
Grades Four and Eight  
English Language Arts Examinations, 1999-2000**

<u>Grade</u>	<u>Passage</u>	<u>Mean Biserial</u>
Four	One	.432
	Two	.407
	Three	.332
	Four	.483
	Five	.423
Eight	One	.412
	Two	.401
	Three	-----
	Four	.394
	Five	.472

## Conclusion

The correlation between the point biserial values and the readability formulae values are given in Table 3 below.

**Table 3**

**Correlations between Readability Estimates  
And Item to Whole Point Biserial  
Discriminability Estimates**

<u>Grade</u>	<u>Midpoints</u>	<u>Dale-Chall</u>	<u>Spache</u>
4	-0.208	0.028	0.007
8	-0.025	-0.238	-----

None of these correlations reached significant levels. The analyses fail to discern any relationship between readability and fidelity to the construct being measured either for the Grade four or the Grade eight examinations in English Language Arts. Moreover, review of the values shows that there is no monotonic or progressive relationship between the point biserial values and the readability values. That is, higher point biserial correlations were not associated with higher or lower readability values on any of the scales. Clearly, in the range of readability values for the two examinations, there appears to be no contamination of the measure related to readability.

*Reference*

Koslin, B. L., Zeno, S., & Koslin, S. *The DRP: An Effectiveness Measure in Reading*. (New York, N.Y.: The College Entrance Examination Board, 1987).