Schoolbook Simplification and Its Relation to the Decline in SAT-Verbal Scores

Donald P. Hayes, Loreen T. Wolfer, Michael F. Wolfe

Cornell University

Preprint. Appeared in *American Educational Research Journal*,
Summer 1996, Vol. 33, No. 2, pp. 489-508

The 50+ point decline in mean SAT-verbal scores between 1963 and 1979 is widely attributed to changes in the composition of the test-takers. Several inconsistencies in that explanation are identified. That explanation also ignores the pervasive decline In the difficulty of schoolbooks found by analyzing the texts of 800 elementary middle, and high school books published between 1919-1991. When this text simplification series is linked to the SAT verbal series, there is a general fit for the three major periods: before, during, and after the decline. Long-term exposure to simpler texts may induce a cumulating deficit in the breadth and depth of domain-specific knowledge, lowering reading comprehension and verbal achievement. The two time series are so sufficiently linked that the cumulating knowledge deficit hypothesis may be a major explanation for the changes in verbal achievement. Only an experiment can establish whether this is causal, so we describe a simple, low-cost experiment schools can use to test how schoolbook difficulty affects their students' verbal achievement levels.

Donald P. Hayes is a Professor of Sociology at Cornell University, Uris Hall, Room 346, Ithaca, NY 14853. His specializations are social interaction and interface with biology.

Loreen T. Wolfer is an Assistant Professor, 1501 Little Glouster Rd., Apt. C-09, Blackwood, NJ 08012. Her specializations are life courses and women and work.

Michael F. Wolfe is a Lieutenant in the U. S. Army. He will soon enter graduate school at Stanford.

The 1963—1979 decline in SAT-verbal achievement provoked many explanations; it was due to changes in family birth order, teacher abilities, school resources, curriculum, student composition, youth cultures, discipline problems, achievement motivation, nutrition, television, and lead poisoning. Some (e.g., the confluence theory, Zajonc & Bargh, 1980) had seemed plausible at one point, but they no longer fit the growing time series. The most widely accepted of the surviving explanations is that the decline was caused by changes in the composition of those taking the test. After 25 years of speculation and analysis, a scientifically adequate explanation for this decline continues to be elusive.

In 1963, an abrupt, unexpected, and unprecedented 16-year decline in U. S. verbal achievement began (Figure 1). Mean verbal scores fell from 478 to 424 in 1979. The early

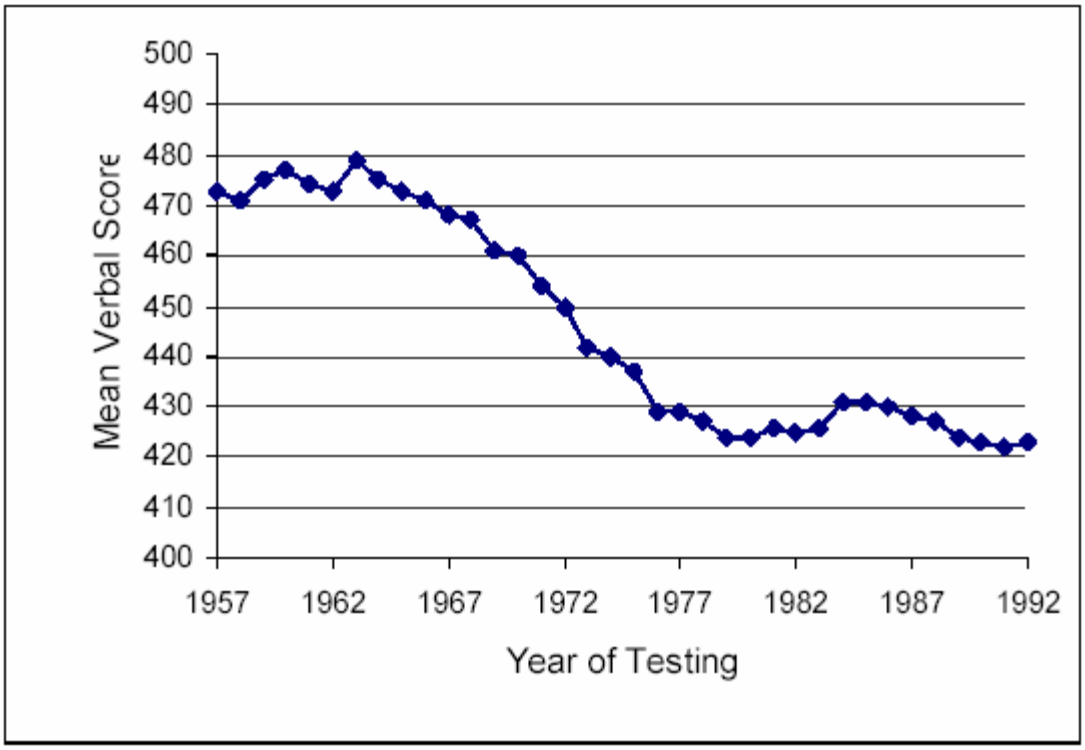

Figure 1: The SAT Verbal Time Series

decline in scores was treated as a regression effect—the 1963 mean had been an outlier, and consequently the next year's score was merely a return to normal. When scores continued to decline, the regression explanation was rejected and replaced by the assumption that the declining scores were a statistical artifact—the verbal (and math) subtests must have become harder. Modu and Stem (1977) tested this hypothesis and reported, to the contrary, that the verbal tests had actually become easier in the decade 1963-1973, by 8 to 13 points. When scale drift is combined with the official decline in scores, verbal achievement among American seniors nationwide fell *59 to 64 points in little more than one decade*. By early 1960 standards, Stedman (1994) estimates that students taking this test in the early 1990s were achieving at the 32nd percentile.

By this time, the explanation focusing on changes in the composition of the test-takers became more plausible. Between 1952 and 1963, the fraction of 17- and 18-years-olds remaining in school increased, as did the fraction of seniors taking the SAT. From being a test required for admission to select colleges only, the test was now serving a far wider range of colleges. That presumably diluted the talent pool, causing the mean score to fall.

We have identified three inconsistencies between this composition explanation and the evidence. The first is that SAT verbal scores should have declined during the 1952—1963 period because that was when the composition of the test-takers was changing most (from about 5 to over 50% of the senior cohort). To the contrary, ETS reported that scores fluctuated within a narrow range—between 472 and 478. While the average test-taker became less elite (academically) and more diverse in class, race and ethnic background, mean verbal scores did not fall.The second inconsistency concerns the prediction that when the changes in test-taker composition slowed, in the 1963-1979 period, mean verbal scores should have leveled off and remained relatively stable. Throughout this period, the fraction of the senior cohort taking this test was stable at just over 50%. it was in this period that verbal scores fell from a high of 478 to 424 (just 2 points above its lowest level ever, 422, recorded in 1991). That is, virtually the entire change in post-World War II American verbal achievement levels — from an initial high plateau of about 475 in the 1950-early 1960s to a much lower plateau of scores from 1979 through 1994 — *occurred within this single 16-year period*.

The third inconsistency between the evidence and the composition explanation stems from its assumption that as more lower scoring students took this test, the fraction scoring over 600 and 700 would necessarily decline (so long as top students continued to take the test). The absolute numbers of students scoring over 600 should have remained about the same.

The evidence is different: *The entire distribution of verbal scores from top to bottom, shifted to lower levels.* There was not only a proportional decline in top scorers but an absolute decline in the number scoring over 600. There are now 35% fewer scoring over 600. The number scoring over 700 fell from 17,500 in 1972 to just over 10,000 in 1993 — even as the number taking this test grew (Shea, 1993). Highly selective colleges and universities report mean verbal declines on the order of 40 points. The composition hypothesis does not predict this outcome. An acceptable explanation for the SAT-verbal decline must account for this huge decline in the performance of the academic elite.

Evidence for the composition hypothesis is based primarily on multivariate analyses of academic and demographic data which has become available only since 1976. Consequently, those analyses excluded (a) the first 13 years of the mean verbal decline (when scores fell nearly 50 points — 478 in 1962-1963 to 429) and (b) the scale drift of 8 to 13 points between 1963 and 1973 (Austin & Garber, 1982; Carson, Huelskamp, & Woodall, 1993; Murray & Herrnstein, 1992; Stedman, 1994). This hypothesis extrapolates from these analyses (which have less than 1/6th the range of variation) to an earlier period in which virtually all of the change from a high, relatively stable plateau to a low, relatively stable plateau of verbal achievement had already occurred.

In an effort to control important cognitive and demographic changes in test-taker composition, Bracey (1991) analyzed only White test-takers with one or more college-educated

parent. Those in the SAT norming population of 1941 had a mean verbal score of 500, while their modern counterparts scored 454, implying that this decline was real, not an artifact. His most important finding for our analysis, however, was the discovery that the two sets of White students (separated by 50 years) had comparable SAT-math scores, suggesting that *factors unique to verbal achievement* caused the verbal scores to fall. It is unlikely that compositional change alone will be able to explain the differentials in the math and verbal times series.

Changes in the composition of SAT test-takers remain a plausible contributing factor to the verbal decline, but, if they are not likely to be the principal explanation, what is? Jeanne Chall (1977) attributed the verbal decline to the widespread acceptance of an educational philosophy in vogue after World War II. She saw this development as encouraging publishers and educators to make schoolwork more accessible by reducing the pace of instruction, lowering student academic workloads, assigning fewer written papers, and using new and simplified schoolbooks. This post-1945 round of reduced academic demands on students 492 should, however, be viewed in the context of American educational history. As education has extended beyond the economic and religious elite, the trend has been to reduce academic demands. What had once been a brief, heavily sectarian and privately supported education became a more leisurely, less religious, universal and free education. Schoolbooks became less demanding after the Civil War and after both World Wars.

Unlike the composition explanation — which absolves publishers, administrators, and classroom teachers of responsibility for this nationwide verbal decline — in effect, Chall held them directly responsible for producing it.

The Mechanism

What causal mechanism lies behind this hypothesis that simplification of schoolbooks was responsible for the decline in verbal achievement? A major objective in the design of third through eighth grade readers is to promote the development of the child's knowledge based on those subjects deemed important by educators — for example, political and natural history, literature, and mathematics. For several years, the students' readers are their primary source for expanding the breadth and depth of this uncommon, academic knowledge. Unless students were compelled to read these texts, those topics might well be ignored. When simulators were developed to, perform functions as human experts perform them, it became evident that the simulator must have an extensive and appropriate knowledge base to perform well. Such a knowledge base is not built-in — it must be acquired.

According to this hypothesis, when publishers simplified their schoolbooks, they reduced the breadth and depth of the students' domain-specific knowledge. Domains are conceptual clusters (such as baseball, rock music, and genetics). Domains are differentiated from one another by their unique combination of concepts, objects, relationships, events, and domain-specific lexicons (Damashek, 1995). In baseball, when a catcher shouts, "Squeeze bunt!" each defensive player must have acquired an understanding of that expression's referents, which activate specific acts, contingent on how the play develops. Someone with *no* prior experience with baseball would have an empty baseball domain, with no knowledge of its concepts, relationships, events, objects, or its domain-specific lexicon (and would not have a clue as to what is about to happen).

Failure to recognize a text's words, especially its uncommon and rare words, is particularly damaging for understanding. Such words not only carry a disproportionate share of the passage's information load but also have fewer collateral meanings and are far harder to guess from context than the more common (and far more polysemous) words (Hayes, 1987). A well-designed vocabulary test estimates the breadth and depth of a child's knowledge base, estimates the child's understanding of the grammatical and substantive uses to which specific words may be put, and estimates how extensively concepts (and the words referring to them) are linked in the child's mind.

In simplifying schoolbooks, publishers after World War II deliberately reduced their use of the rarer, domain-specifying words. Ceteris paribus, daily use of a series of simplified textbooks across 11 years of elementary and secondary schooling should have produced a cumulating deficit in students' knowledge base and advanced verbal skill — as compared with those instructed with more challenging textbooks. This reduced-knowledge base directly affects the child's text comprehension.

The hypothesis of a cumulating knowledge deficit implies a time-lagged relationship between the SAT-verbal time series and the levels of school-work between the 1st and 12th grades. It predicts that, because the readers were simplified, verbal achievement levels would have declined. It predicts the approximate times for when the decline should have begun, how long it will last, when it should end, and its level into the next century.

The Dataset

We found our readers in the archives of 18 college and university libraries. In the East, the largest such archive is at Teachers College, Columbia University. The texts we sampled are readers explicitly designed to accomplish many educational objectives, simultaneously. Each publisher's series of readers begins with a primer (used for early reading instruction), one for each grade through Grade 8.

We drew samples from all major and most minor publishers' series. The corpus is relatively complete for the period 1963 through 1991, but there are many omissions for the period 1919 through the early 1960s. Libraries, chronically short of space and staff, usually retain only the very old and the most recent series.

Once a reader was found, a stratified sample (the number depending on the size of the text) of 10-30 pages was taken. Every text sample in this corpus has at least 1,000 words. Each sample page was photocopied and laser-scanned. OCR software converted the bit-map back into text, and the text was then edited to correct OCR-induced and other errors. Every sample text now conforms to a common transcription standard. The sample corpus was taken from more than 800 readers and contains over 1.14 million words.

Because these measurements on a sample text's "lexical difficulty" will be unfamiliar, the LEX measure and the reasons for its choice need to be explained. Text difficulty is a complex, multidimensional concept, and much is not yet understood. It has at least three major components: syntactic, semantic, and lexical (Zakaluk & Samuels, 1988). The syntactic and semantic features, however, cannot yet be measured by software. Until they too can be

conceptualized and measured, LEX alone serves as a well-validated tool for estimating the lexical difficulty for these texts.

We do not use Readability Indices (such as Flesch/Kincaid or Gunning's FOG). While widely used to estimate text "difficulty," psycholinguists (e.g., Holland, 1981) have many strong reservations about these measures. In calculating a readability score, texts are not first edited to a common standard. The scores must be renormed periodically (making comparisons across time difficult), and the two constituents (fraction of long words and sentence length) of the indices are treated as positively correlated, whereas the association varies in natural texts from +.7 to -.7. The strongest objection to these indices, however, is that they are atheoretical.

We use LEX as our measure of a text's difficulty. LEX is based on a theoretical statistical distribution — the log-normal (Ahren & Hayes, 1990; Dewey, 1923; Gordon, 1985; Hall, 1989; Hayes, 1988; Hayes & Ahren, 1988; Just & Carpenter, 1987). Like its close relation, the normal distribution, it is found throughout the sciences. Mathematicians and statisticians first discovered that word choice in large natural texts closely fits the log-normal (Carroll, 1971; Dewey, 1923; Herdan, 1956, 1966; Yule, 1944). Our discovery was that word choice in every natural text is one or another *variant* of this statistical distribution — that is, word choice in all texts is skewed to one or another degree, toward or away from use of the most common words (Hayes, 1992). Figure 2 shows three variants on the log-normal distribution: an undemanding set of texts

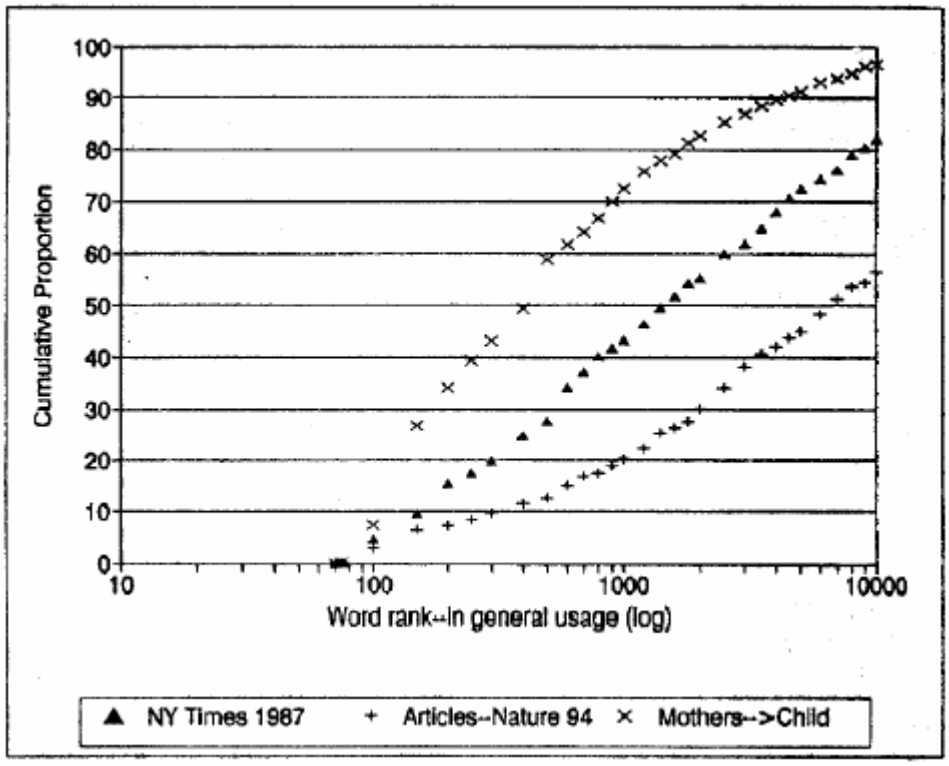

Figure 2. Patterns of open class (content) word use in three texts

(mothers talking with their 3-year-old children), a text which closely fits the theoretical model (a sample from the *New York Times*), and the pattern of word choice in articles taken from the journal *Nature*. LEX describes the pattern of word choice from the over 600,000 *open class* (content) word types available in English. A separate statistic (not reported here) shows the pattern of word choice from the 100 *closed class* (grammatical) words of English.

LEX has the following properties: The measure satisfies the requirements of a ratio scale; its theoretical limits are -123 to +85; the, log-normal distribution is equal to 0.0 LEX, which is the value newspaper samples closely approximate; natural texts range from -80.5 (a pre-primer by Houghton Mifflin) to the most difficult text yet found, an article in *Nature* (at LEX = +58). 0.0 is the LEX value of a text if it fits the lognormal perfectly. Texts with (+) LEX scores are more difficult than newspapers, texts with (-) LEX scores are less difficult than newspapers. The larger the numeric score, the harder (or easier) the text. The standard error for the LEX statistic is 0.5 LEX in a theoretical range of 208 units. This estimate is based on a sampling distribution of 22 independent samples (each consisting of 24 subsamples of 100 words each, from the same large text). If still greater precision is required, the number of subsamples and their sample size must be increased. The most interesting property of LEX is its stability across the centuries: Samples from English and American newspapers have grown more difficult at the rate of about 1 LEX/century since 1665. LEX values for several classes of text, and several analyses in this article, are reported in Table 1.

Table 1. The Spectrum of Natural Texts and Their LEX Levels

Text Source	**Date/N**	**LEX**
Nature — an article on transhydrogenese	1960	58.6
New England Journal of Medicine — articles	1990	33.3
Scientific American — articles	1991	14.3
Popular Science — articles	1994	4.6
Time — articles	1994	1.6
Newspapers: English, N = 61 International	1665-1994	0.0
National Geographic — articles	1984	-0.6
Sports Illustrated — articles	1994	-10.3
Adult books — fiction, USA	N = 34	-15.8
The funnies — in newspapers	1982	-21.6
Nancy Drew mystery series	N = 69	-23.4
Comic books — GB & USA	N = 37	-23.7
Children's books age 10 — 14, GB	N = 261	-24.3
TV — cartoon shows	N = 26	-28.6
Children's books age 9 —12, USA	N = 94	-29.0
TV — reruns — popular with children	N = 33	-35.3
TV — primetime shows	N = 44	-36.4
Preschool books read to children	N = 31	-37.0
Mother's talk to children, age 5	N = 32	-45.8
Dairy farmer talking to his cows	1988	-56.0
Pre-primer — Scott, Foresman	1956	-80.5

Results

Elementary and Middle-School Readers

The principal findings regarding the simplification hypothesis are summarized in Table 2 and Figures 3 and 4.

Table 2. The LEX Levels of American and British School Readers

ERA	STATISTICS	PRIMER	1ST	2ND	3RD	4th	5TH	6TH	7TH	8TH
1946 to 1962	LEX mean	-51.8	-47.3	-42.4	-40.1	-29.6	-21.8	-18.4	-13.1	-13.9
	SD	8.6	8.2	6.3	5.0	4.0	6.3	6.8	7.2	4.9
	No. texts	11	26	14	21	11	10	11	10	11
	No. words	7,280	20,155	14,365	36,228	16,176	17,370	19,921	23,440	26,631
	MLU	6.1	7.7	10.0	12.2	15.0	19.0	18.9	20.7	20.4
1946 to 1962	LEX mean	-67.9	-59.7	-53.1	-40.1	-31.7	-28.3	-25.0	-21.5	-19.7
	SD	7.5	6.7	6.4	7.3	10.7	4.4	5.2	4.0	4.3
	No. texts	10	14	10	24	12	11	10	10	8
	No. words	7,136	13,104	16,020	37,282	19,492	21,736	17,951	19,700	18,128
	MLU	5.1	6.1	8.5	11.5	11.5	12.1	11.9	14.1	13.2
(BRITISH) 1946 to 1962										
	LEX mean		-44.40							
	SD		6.50							
	No. texts		14							
	No. words		11,654							
1963 to 1991										
	LEX mean	-53.0	-50.3	-43.2	-37.2	-33.2	-28.2	-25.0	-23.0	-22.2
	SD	12.2	10.8	8.3	6.8	5.2	5.4	5.4	534	6.4
	No. texts	32	75	58	58	37	40	34	19	20
	no. words	17,886	85,050	82,208	80,340	62,841	72,827	70,618	46,044	47,720
	MLU	6.6	7.4	8.9	10.3	10.5	11.6	12.7	12.9	14.6

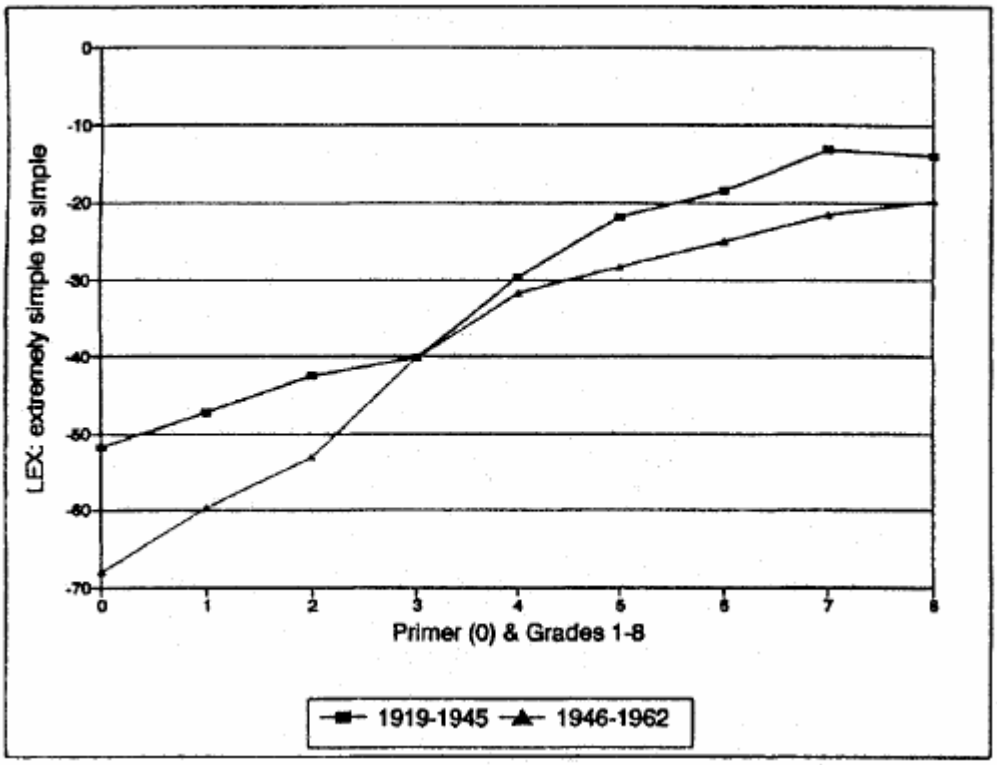

Figure 3. Mean LEX levels for school readers: 1919-1945, 1946-1962

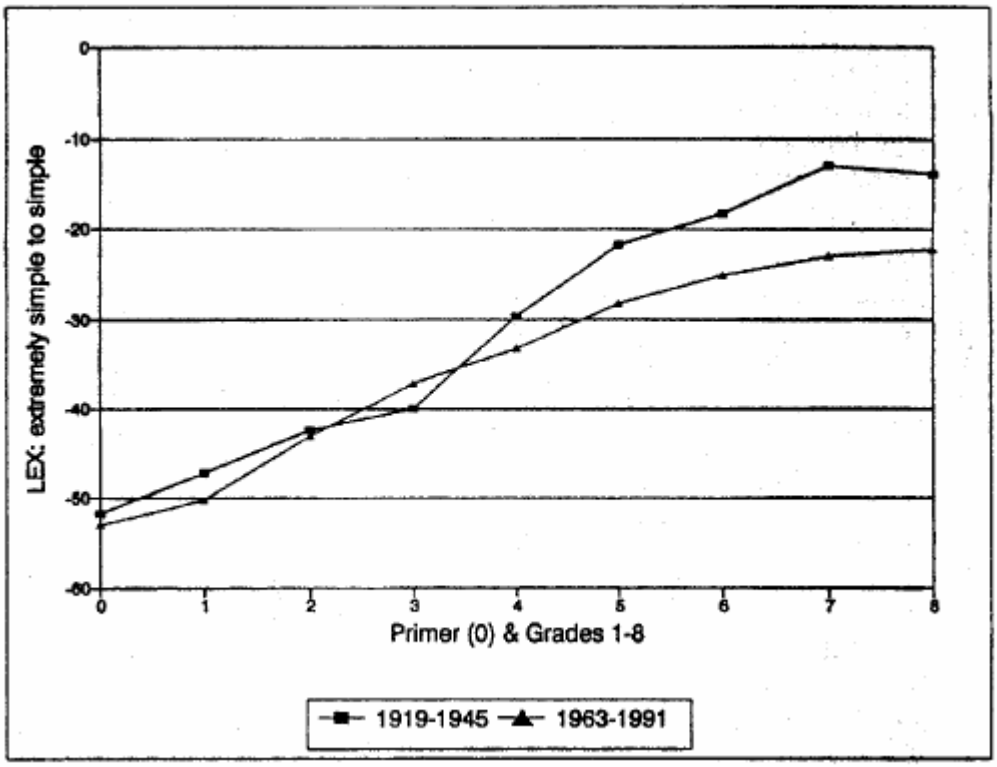

Figure 4. Mean LEX levels for school readers: 1919-1945, 1963-1991

a. Between 1860 and 1991, American publishers produced readers for the same grade at widely divergent LEX levels (e.g., in 1968, first grade readers were available between -68 and -31).

b. The most difficult readers were generally published before 1918. By modern standards, Professor McGuffy's pre- and post- Civil War readers were very difficult.

c. After World War I, mean reader LEX levels for all grades were generally simplified.

d. After World War II, mean levels of readers for all grades but third became even simpler. These were the books used by the Baby Boomers and successive cohorts.

e. School publishers in Great Britain did not simplify heir first grade readers after World War II, implying that there was no compelling educational reason for American publishers to further simplify their readers.

f. Today's mean sixth, seventh, and eighth grade readers are simpler than fifth grade readers were before World War II.

g. Sentences were also shortened, from about 20 words before World War II to about 14 words now in Grades 4—8.

h. Texts for grade x + 1 are typically more difficult (lexically) than those for grade x. When the early readers in a series are simplified by a publisher, so too (generally) are the readers for later grades.

American readers were rewritten after World War II for many reasons: to modernize their content and graphics; to incorporate principles from research in education, psychology and linguistics; to emphasize new kinds of social relations; and to respond to television, which was fast becoming an unprecedented rival for students' attention. Those considerations required that readers be changed but not that they be simplified. Chall (1967) described the justifications given at the time for simplifying schoolbooks; they were to increase their accessibility for students and raise the level of reading success.

The extent of this simplification is notable. Apart from one extraordinarily simple text (LEX = -75), most first graders in the Baby Boom era used readers whose LEX levels were between -53 and -65; that is, on average, first grade readers were 12 LEX units less difficult than the readers used by their parents and grandparents between World Wars I and II and 15 LEX less difficult than the texts used by students in the same grade in Britian — where the students were a year younger in age. Since the empirical range for LEX is about 140 units, a 12 LEX simplification represents 9% of the empirical range. Workbooks are also designed by the same publisher as the readers and are used in conjunction with them. Given the workbooks' format, we cannot measure their LEX levels. It is unlikely the publishers would make the workbooks harder than their own reader for the same grade level.

To insure that our eastern reader samples are not a biased sample, we went to the nation's most comprehensive archive of schoolbooks at the Center for Research Libraries in Chicago and were able to find 7 more third grade readers for the period 1919-1945 and 12 more for 1946-

1962. While the combined numbers are small, they represent every major, and most of the minor, third grade series published over that period. Adding these new samples changed the mean third grade levels from -39.0 to -40.1 for the era 1919-1945 and from -40.3 to -40.1 for the 1946-1962 era; that is, the additions reduced the mean era difference from 1.1 to 0.0 LEX. There is no apparent bias in our eastern sample.

In simplifying the readers' patterns of word choice, the publishers also shortened sentence length (Figure 5). From the fourth grade on, the average sentence length was contracted by about

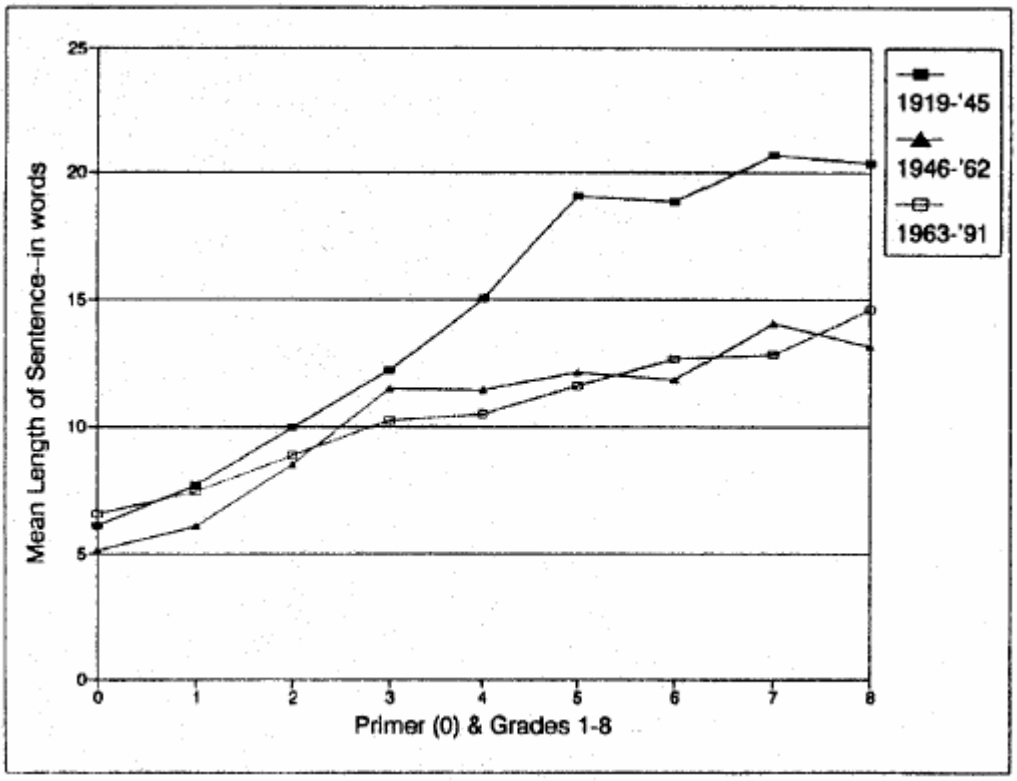

Figure 5. Mean length of sentence in school readers: 1919-1991

6 words — this is the equivalent of dropping one to two clauses from every sentence. This reduced the students' experience in working out the meanings of more complex sentences.

Widespread objections from teachers and parents, and media publicity over the simplified, often dull, Dick, Jane, and Spot-like readers, caused most publishers to acknowledge that they had oversimplified their first and second grade readers. In the 1963 —1991 era, the mean LEX levels for those grades was restored to pre-Worid War II levels — and without ill effects. While they made those early texts harder, the publishers made the fourth, fifth, sixth, seventh, and eighth grade readers even easier — to the point where these are now at their lowest level in American history.

Analyses of High School Textbooks

Since the average primer and reader for Grades 1— 8 had been simplified, we would have expected those same publishers to have simplified texts for Grades 9 —12, as well. This analysis

is not comparable in scale to those made on the readers for elementary and middle schools, but, as a test case, we examined all the required texts used in English and science classes in the local (Ithaca, NY) high school.

We found that LEX levels of required English texts used in 9th through 12th grades differ little and that the texts used in the five tracks (the highest being AP) were not systematically more difficult. Whatever the grounds for a text's choice in the curriculum, it is not the texts' lexical difficulty. The average literature text required in 12th grade English classes is nearly 10 LEX simpler than the average 7th or 8th grade reader published before World War II. There were, however, large differences in LEX between English and science texts. The ninth grade earth science text (LEX = -16 was the simplest, but it was still harder than the average book required in honors senior English. Texts for biology, chemistry, and physics were 20 or more LEX higher than English hooks for the same grade. Science texts used for AP chemistry, biology, and physics were the only high school texts with positive LEX scores — as high as +9.9.

Analyses of Materials Read Outside of Class

Readers are not the only academic source from which children build their knowledge base and verbal skills. Readers are normally supplemented by workbooks, readings from literature, and out-of-class, leisure- time reading. We have tested the assumption that these supplementary books are more challenging than modern readers.

The data for these analyses include 516 nonacademic books and periodicals popular with children between the ages of 9 and 14. Since girls at every age read more books than boys, a special analysis is included on a series popular with girls: the first 50 from the Nancy Drew series and 36 books from Enid Blyton's comparable British series. To estimate the level at which periodicals read by children are pitched, our analysis includes 38 periodicals popular with school-age children, and 37 British and American comic books and the newspaper funnies. The results were reported in Table 1.

The average supplementary book chosen and read outside of school by American 9- to 12-year-olds has a LEX of -29. This is comparable to a pre-World War II fourth grade reader but nearer to today's fifth grade readers. The 246 most popular books read by a nationwide probability sample of 7,839 British students in 1971 (Whitehead, Capey, Maddren, & Wellings, 1977) were pitched to a higher level (-24), but those children were both older (10—14) and further along in school than their American counterparts. These analyses are highly aggregated and overrepresent book reading by two categories of students: the younger (who read more books than older students) and those performing at higher levels in school (who read more books, at slightly more difficult levels, than other students). Book reading declines rapidly between 9 and 14 years (cf. Hayes, Ahrens, Roschke, Tsay, & Wolfer, 1994, for a disaggregated analysis of book reading). As expected, self-chosen supplementary books are pitched at levels close to a child's schoolbooks.

The Nancy Drew series was for most of this period the most popular supplementary reading with American girls. We tried to determine whether a publisher producing books designed for school-age children (but not their schoolbooks) would also have simplified these books after

World War II. There are two ways to test this. Nineteen Nancy Drew books were written between 1930—1945 (mean LEX: -21.3). The other 31, published between 1946 — 1979, were written at -24.4, establishing that a modest simplification occurred after World War II and that those books are slightly harder than the average child's leisure-time book. Being more difficult did not prevent them from being popular. A second test capitalizes on a rare opportunity: After World War II, the publisher of the Drew series re-issued 12 books written between 1930-1945. If schoolbooks were being simplified in that period, were these re-issued versions also simplified? Book title, plot, and characters remained the same — but the text's cover was changed, and the text was shortened somewhat. The re-issued versions were on average 1.4 LEX simpler than the originals, and sentence length was shortened (by an average of 1.3 words). Text simplification after World War II was evidently not confined to schoolbook publishers. In Britain, Enid Blyton's books were extraordinarily popular with girls but strongly disliked by their teachers (in part, because they are written at such low levels: -36, comparable to a typical American third-to-fourth grade reader, today). Analysis of Periodical Reading

Whitehead et al. (1977) report that, while book reading declines rapidly toward the end of elementary school, most children report an increase in periodical reading as they grow older. We find that newspapers, magazines, comics, and funnies aimed at young adults and adolescents are pitched to a higher level than American schoolbooks — most are comparable in level to today's average seventh or eighth grade reader (Table 1). The drift away from book to periodical reading may be due, in part, to the latter's greater relevance to children's lives (e.g., in their coverage of sports and music personalities, personal care, romance, youth cultures, and life styles). While periodicals are pitched to a higher level than the students' schoolbooks, their difficulty evidently does not affect their popularity.

In short, estimates of children's self-chosen, out-of-school reading suggest that they are at least comparable to, or in many cases pitched to, a substantially higher level than their schoolbooks. Given a choice, students select more demanding texts than they are assigned at school.

Relating the SAT and Reader Simplification Times Series

The cumulative knowledge deficit hypothesis predicts: (a) Cohorts whose education was with pre-World- War-II-level texts would have scored at the high plateau levels of the SAT-verbal throughout the middle 1950s/early 1960s. (b) The first cohorts to have lower SAT scores should have been those who used the harder schoolbooks throughout most of their education but encountered these less challenging schoolbooks toward the end of their secondary education. (c) The more years a cohort was exposed to the simplified levels of schoolbooks, the lower their verbal scores should be. (d) The 16-year decline in verbal scores should have ended and stabilized once cohorts used simplified schoolbooks throughout their entire primary and secondary education.

The absence of proprietary sales data on each publisher, by year and region, makes it impossible to perform cross-lagged correlations on these series. The predicted effects depend on the relative sales of the high, medium, and low LEX level series. To perform such crossed-lagged correlations would also require precise data on the length of two lags. The first was caused by the delay in the introduction of new reader series after World War II. The first two

new series to appear after the war (by Scott, Foresman, and Silver Burdett) were introduced in 1947. The majority of the new series did not appear until after 1950. Many students continued to work with reprints of the harder pre-war and war-time readers into the mid-1950s, but, to set the appropriate length of this lag we need detailed sales data, by publisher, which is not available. This lag is important for the cumulating knowledge deficit hypothesis because it indicates that the transition from the high verbal plateau (about 475) to the new and low plateau (about 425) should have been continuous, not abrupt. The second lag is caused by a school district's cycle of book purchasing — as lost, worn out, and damaged books are replaced by new and updated series. The length of that cycle varies greatly between school districts. A publisher may have produced a series in year x, but it might not have been adopted until x + 4, since 4 to 7 years is the normal cycle. That adds imprecision to the lag of the SATverbal series on the reader simplification series.

In the absence of vital data, the predictions from the hypothesis can only fit loosely with the SAT-verbal time series. As predicted from the changes in reader difficulty, during the entire 40-year period (1955— 501 1994), there should have been, and there was, only one major discontinuity in the verbal series. Its beginning, its length, its ending, and levels since 1979 are also consistent with the assumption that 11+ years of a cohort's exposure to the older and to the newer simplified readers and supplementary in- and out ofclass readings affected the children's breadth and depth of domain specific knowledge.

Specifically, all cohorts whose entire education was with pre-World-War II-level readers achieved mean SAT-verbal scores associated with the high plateau. The first of the 16 consecutive cohorts to have a falling score were children born in 1946 who entered first grade in 1952. What texts would those children have used? Most would still have been using the older/harder readers, but some were now using the new, simplified readers. Eleven years later, about half of that cohort took the SAT in 1963. That was the first cohort to show the decline in verbal scores. The last cohort to have a declining score was the one which still had some experience with those pre-World War-II-level texts. They entered first grade in 1967. By that time, many of their first and second grade readers would have had been restored to pre-1946 levels. But, from the fourth grade on. their texts were the simplest in American history. That cohort took the SAT in 1978 and was the first to reach the new low plateau of SAT-verbal achievement. Every cohort since then has been exposed to simplified schoolbooks from fourth grade on, and their mean verbal scores have consistently remained at the low plateau level. Unless there is a significant increase in the level of schoolbooks, this hypothesis predicts that achievement will continue at this low level through the rest of this century and into the first decade of the next, because it will take time for the publishers to produce texts at the new higher level and for the cycle of school district purchasing to have acquired those texts for every grade.

The fit between these two time series is not tight (it could not be without precise sales data on each reader series and the two lags), but the overall pattern — the initial high plateau; a single, very gradual decline; and the new low plateau of achievement — is generally compatible with the measured changes in LEX levels of the students' readers and their other books and periodicals, in and out of class. While these data suggest a causal relationship, we cannot prove that simplification of textbooks caused a cumulative deficit in the breadth and depth of student knowledge which resulted in a decline of verbal achievement. Next, we show how this hypothesis can be turned into a causal relationship.

An Experiment on the Cumulative Knowledge Deficit Hypothesis

Our analysis of 800 readers provides a plausible case that simplifying readers had an unexpected and undesirable side-effect — it may have reduced the breadth and depth of children's knowledge base and affected the children's verbal achievement levels — detected by the SAT and ACT. There is more at stake here, then, than a test of an academic hypothesis — at stake is the conception of how children learn.

Before publishers began to simplify readers after World War II, they and most educators undoubtedly considered (and rejected) Chall's and our hypothesis that simplification might have negative side-effects and that an entire generation of American students would be adversely affected by it. They found the cumulative knowledge deficit hypothesis implausible, implying that they held a different model for how children learn — specifically, a different model for how a knowledge base is constructed and how domain-specific knowledge relates to text comprehension and to verbal achievement. Chall challenged their model, but most professional educators rejected her challenge.

The experiment we propose is a simple and inexpensive way to test both the professional consensus, which asserts that text simplification has no important effects on student verbal achievement, and our cumulative knowledge deficit hypothesis, which asserts that it has. The design, which any large school district can adopt at negligible cost and effort, is described In Table 3. The experiment begins a year before readers are replaced, and only half the students get new readers in the first—second year.

Discussion and Implications

The policy of making schoolbooks more accessible by reducing their relative use of the uncommon and rare English words and by shortening sentences may not have been cost-free after all. Our evidence establishes beyond reasonable doubt that a major simplification of schoolbooks occurred after World War II and that, beyond third grade, current levels have never been lower in American history. Any explanation for the decline In SAT-verbal scores will now have to introduce statistical controls for this time series of reader simplification.

We have reached a point in the analysis of the SAT-verbal decline where epidemiology was 30 years ago, when heavy cigarette smoking was found to be correlated with higher rates of lung cancer and a diet rich in certain types of cholesterol was found to be related to higher rates of heart disease. Given those correlations, did it follow that reducing cigarette smoking (or high cholesterol) would reduce rates of lung cancer (and heart disease)? Does the apparent association between reader simplification levels and SAT verbal score also imply a causal relationship? Transforming correlations into causal relations requires experimentation. After 25 years of continuous speculation and analysis, it is about time.

Table 3. A Design for Testing the Cumulative Knowledge Deficit Hypothesis

1. Many pairs of same-grade classes should be formed, each from a different school, by matching (e.g., on mean reading levels—above or below grade level). The school, district must be large because single-year effects should be small, requiring many pairs to detect the effects. A cumulating small effect, over 11 years, may become large.

2. Within each class-pairing, assignment is at random to a treatment:

   (a) Old—remain with the current reader one more year, for classes assigned to the control condition
   (b) New—given the newly purchased readers, for classes assigned to the experimental condition

3. The 4-year-long design as follows:

			Year			
Grade	Treatment		1	2	3	4
2	Paired	Controls	old	old	new	new
2	Classes	Experimentals	old	new	new	new
3	Paired	Controls	old	old	new	new
3	Classes	Experimentals	old	new	new	new

4. Dependent variable: any reputable verbal achievement test

5. The predicted primary effects for the LEX levels of readers are conditional on the sign and magnitude of the differences between the old and new readers, in each grade—e.g., if the new reader is 6 LEX harder than the old, then mean verbal achievement should rise more than if the difference were 2 LEX.

6. The primary question is: Are these small effects lasting and cumulative? Years 3 and 4 provide that evidence.

7. Is such an experiment ethical? Professional consensus for nearly 50 years has been that simplification was harmless. If so, the children are not being put at risk. If there are such effects, then schools would be in a position to alter this unintended side effect of their reader choices.

Changes in the composition of the students is still a plausible explanation for part of the decline in SAT verbal levels—but few, if any, policy implications flow from that explanation. By contrast, the cumulative knowledge deficit hypothesis has clear, direct policy implications. In their next cycle of schoolbook purchases, schools have both a justification and the means for strengthening their students' verbal achievement levels—they can select more challenging readers from the many existing commercial alternatives.

In the long run, publishers may raise the levels of their series (as they did once before when it became evident they had overshot the mark in the 1946—1962 era). Before acknowledging they have overshot for Grades 4 through 12, the experimental evidence must be compelling. Publishers control these levels and monitor them with the Readability Indices (with which LEX is correlated in the low .70s—fractionally higher with the Flesch index, lower with the others). There are major risks for a publisher in raising LEX levels. It could reduce their market share if schools do not buy the new, harder series, so educators too must be convinced of this hypothesis if they are to buy such readers.

Finally, we find it anomalous that no drugs can be sold in the United States without first demonstrating, by experimental tests and clinical trials, their efficacy and safety, while publishers and schools can freely impose simplified readers and related schoolwork on children without having to produce experimental evidence on the efficacy or safety of their schoolbooks (e.g., evidence on the effects those texts have on students' breadth and depth of knowledge and on a major cognitive dimension: verbal achievement). A modest way to encourage the adoption of a more demanding curriculum would be to publish our measurements on modern series — at least schools would have that information. The better, more compelling way would be to publish results from large-scale, well-executed experiments showing there is a causal link between reader difficulty and verbal achievement.

Note

The authors wish to thank Lawrence C. Stedman for his close reading and valuable suggestions on an early draft.

References

Ahrens, M. G., & Hayes, D. P. (1990). Accommodations in word choice for audience and topic: experimental evidence (Sociology Technical Report Series No. 90-13). Ithaca, NY: Cornell University.

Austin, G. R., & Garber, H. (1982). *The rise and fall of national test scores.* New York: Academic.

Breland, H. M. (1977). The SAT score decline: a summary of related research (Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline: Appendixes to On Further Examination). New York: College Board.

Carroll, J. B. (1971). Statistical analysis of the corpus. In J. B. Carroll, P. Davies, & B. Richman (Eds.), *Word frequency book*. (pp. xxi—xl). Boston: Houghton Mifflin.

Carson, C. C., Huelskamp, R. M., & Woodall, T. D. (1993). Perspectives on education in America, standardized tests, *Journal of Educational Research*, 86, 267-272.

Chall, J. S. (1967). *Learning to read: the great debate*. New York: McGraw-Hill.

Chall, J. 5. (1977). An analysis of textbooks in relation to declining SAT scores (Report of the Advisory Panel on the Scholastic Aptitude Test score decline: Appendixes to On Further Examination). New York: College Board.

Damashek, M. (1995). Gauging similarity with n-grams: language independent categorization of text. *Science*, 267, 843-848.

Dewey, G. (1923). *Relative frequencies of English speech sounds*. Cambridge, MA: Harvard University Press.

Gordon, B. (1985). Subjective frequency and the lexical decision latency function: Implications for mechanisms of lexical access. J*ournal of Memory and Language*, 24, 631-645.

Hall, W. S. (1989). Reading comprehension. *American Psychologist*, 44,157-161.

Hayes, D. P. (1987). The effects of word polysemy and topic range on lexical acquisition. Unpublished manuscript, Cornell University, Ithaca, NY.

Hayes, D. P. (1988). Speaking and writing: distinct patterns of word choice. *Journal of Memory and Language*, 27, 572-585.

Hayes, D. P. (1992). The growing inaccessibility of science. *Nature*, 356, 739-740.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: a special case of "motherese." *Journal of Child Language*, 15, 395-410.

Hayes, D. P., & Ahrens, M. G., Roschke, S. H., Tsay, R., & Wolfer, L. T. (1994, August). On making your own environments: the effects of ability. Paper presented at the national meeting of the American Sociological Association, Los Angeles.

Herdan, G. (1956). *Language as choice and chance*. Groningen, Holland: Noordhoff.

Herdan, G. (1967). *The advanced theory of language as choice and chance*. New York: Springer-Verlag.

Holland, V. M. (1981). *Psycholinguistic alternatives to readability formulas* (Tech. Rep. No.12). Washington DC: American Institutes for Research.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and comprehension*. Boston: Allyn & Bacon.

Modu, C., & Stern, J. (1977). The stability of the SAT-verbal score scale (Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline: Appendixes to On Further Examination). New York: College Board.

Murray, C., & Herrnstein, R. J. (1992). What's really behind the SAT-score decline? *Public Interest*, 106, 32-56.

Shea, C. (1993). Fewer test takers get top scores on the verbal SAT. *Chronicle of Higher Education*, 39 (19), 29-33.

Stedman, L. C. (1994). The Sandia Report and U. S. achievement: An assessment. *Journal of Educational Research*, 87, 133-147.

Whitehead, F., Capey, A. C., Maddren, W., & Wellings, A. (1977). *Children and their books.* London: Macmillan.

Yule, G. U. (1944). *The statistical study of vocabulary.* Cambridge, England: Cambridge University Press.

Zajonc, R. B., & Bargh, J. (1980). Birth order, family size, and decline of SAT scores. *American Psychologist*, 35, 662-668.

Zakaluk, B., & Samuels, S. J. (1988). *Readability: Its past, present and future*. Newark, DE: International Reading Association.