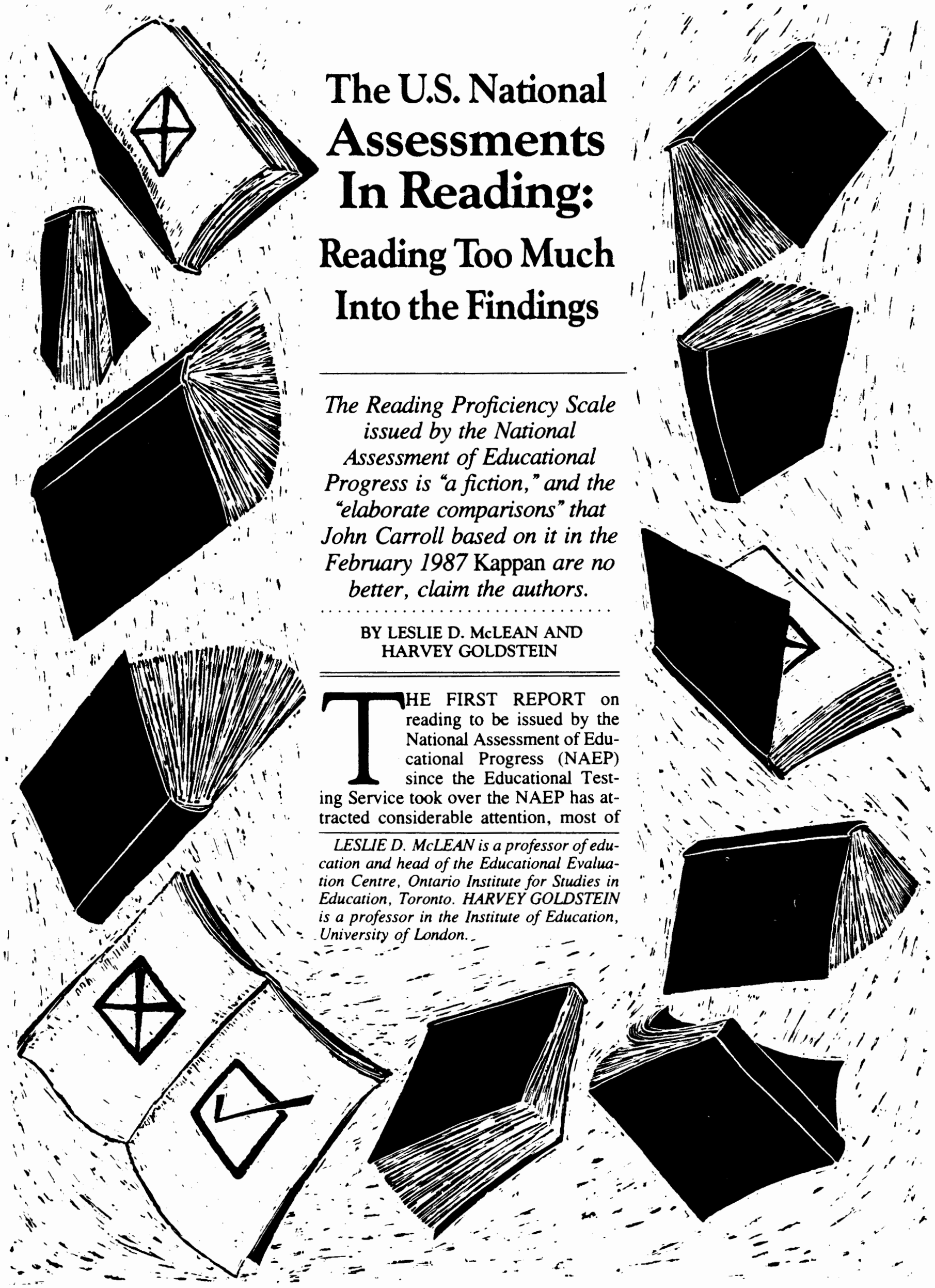

The U.S. National Assessments In Reading:

Reading Too Much Into the Findings

The Reading Proficiency Scale issued by the National Assessment of Educational Progress is "a fiction," and the "elaborate comparisons" that John Carroll based on it in the February 1987 Kappan *are no better, claim the authors.*

BY LESLIE D. McLEAN AND HARVEY GOLDSTEIN

THE FIRST REPORT on reading to be issued by the National Assessment of Educational Progress (NAEP) since the Educational Testing Service took over the NAEP has attracted considerable attention, most of

LESLIE D. McLEAN is a professor of education and head of the Educational Evaluation Centre, Ontario Institute for Studies in Education, Toronto. HARVEY GOLDSTEIN is a professor in the Institute of Education, University of London.

Illustration by Claudia Tantillo

it favorable. The report brings together data from in-school assessments of 9-, 13-, and 17-year-olds, conducted between 1971 and 1984.[1] Statisticians on the staff of the Educational Testing Service (ETS) created a Reading Proficiency Scale to describe the findings, on the assumption that a single human characteristic or trait – reading proficiency – suffices to account for responses to all reading tasks. They tested that assumption within the framework of item response theory[2] and claimed to have found a satisfactory fit. The age range covered by the scale was extended through the addition of a household-based sample of 21- to 25-year-olds.[3]

The scale was then interpreted by "reading experts." At five points along the scale, they examined the test items answered correctly by 80% of the examinees whose scores fell at those points.* The point on the scale achieved by only the top 2% or 3% of examinees was labeled "Advanced." Those items that were answered correctly by 80% or more of these high scorers were found to require such skills as extension and restructuring of ideas presented in specialized and complex texts. The items answered correctly by 80% of the *lowest* 2% or 3% of examinees were labeled "Rudimentary" and described as follows:

> Readers who have acquired rudimentary reading skills and strategies can follow brief written directions. They can also select words, phrases, or sentences to describe a simple picture and can interpret simple written clues to identify a common object. *Performance at this level suggests the ability to carry out simple, discrete reading tasks.*[4]

These functional and notably context-free descriptions account in large part for the popularity of the recent NAEP report. However, they failed to satisfy the distinguished psychologist, John Carroll. He ventured the opinion in the February 1987 *Kappan* that "it is now possible for the first time to gain a nearly complete view of the state of literacy in the U.S., at least up to age 25."[5] But Carroll warned that "there is a tendency to misread the findings – or not to examine them carefully enough."[6]

Making some further assumptions, Carroll created graphs showing the range of scores for each age group for each of the four assessments. In the process, he revealed that the modest average gains over time reported by ETS masked uneven gains – perhaps even losses – in different subpopulations and at different ability levels.[7] If the second highest level on the NAEP's Reading Proficiency Scale ("Adept") is considered roughly equivalent to a 12th-grade level of literacy, the NAEP data appear to show that only 39.2% of 17-year-olds had reached this level. Carroll drew from his findings a host of provocative inferences, including that "for the most part, we must confront the possibility that somehow the nation will have to accommodate itself pretty much to the levels of reading skill now attained by various segments of the population."[8]

Our purpose here is to argue that such inferences give too much meaning to scores on the NAEP Reading Proficiency Scale. There is more danger that the NAEP findings will be *overread* than that they will be *misread*, because scores on the NAEP scale are not connected to processes of teaching and learning. That gap was vividly illustrated in three *Kappan* articles that accompanied Carroll's.[9] Indeed, Kenneth Cadenhead explicitly argued that the whole concept of reading level stands in the way of effective instruction.

Moreover, the procedure by which the NAEP reading scores were derived ignores most modern thinking about the nature of language and language learning – especially the factor of human intentionality, which becomes all-important in adult learners. Carroll's own words suggest that he noticed this fact: "It is tempting, though probably unpopular, to say that to a considerable extent the reading assessments are assessments of national verbal intelligence."[10] But he did not pursue this insight. The sophisticated mathematics used to derive the NAEP scale tends to distract us from the naive ideas about human literacy on which the scale is based.

Since the NAEP scaling procedures – and hence the validity of the NAEP analyses – are crucially dependent on the assumption that reading is a one-dimensional trait or factor, it is also important to understand the meaning of the statement that a unidimensional model provides a "satisfactory fit" with the data. That statement means that ETS did not find a multidimensional model to explain the pattern into which examinees' reading scores fell. There are at least two possible explanations for this fact: 1) perhaps a unidimensional model *can* satisfactorily describe the population from which the sample data were selected, or 2) perhaps ETS did not try very hard to find – or lacked the resources to search for – a multidimensional model.

Space does not allow us to present a

*The five points chosen were the mean (average) plus one and two standard deviation units above and below the mean in the combined distribution of 9-, 13-, and 17-year-olds in 1984.

"What a sex ed class we had today! Jennifer Weedsport went into labor!"

detailed critique of the ETS analyses or to explore in depth the problems associated with the use in educational assessment of simplistic models that spring from item response theory. However, an examination of the ETS technical report on the NAEP analyses[11] allows us to focus on a few salient points.

First, both the school-based sample and the household-based sample were heterogeneous, a situation that would tend to enhance the impression of reading as a single dimension or factor. Moreover, in its analysis of dimensionality, the NAEP made no attempt to use any kind of statistical adjustment to eliminate the effects of such variables as race or parents' educational level – variables that might cause a single common factor to appear more influential than it actually is. The most appropriate technique for studying dimensionality (and the one used by ETS) is the so-called *full information factor analysis*. Unfortunately, because of cost, that technique was applied to only 42 of the 100 items intended for grade 8. Across those 42 items, a one-factor solution accounted for just 39% of the variance, with different groups of items having different-sized loadings on this factor. This is hardly strong support for unidimensionality.

For this reason and for more fundamental reasons having to do with the changing relevance of any given item over time,[12] the attempt by the NAEP to provide an "absolute" measurement scale – in order to measure overall trends in performance over time – remains suspect. A reasonable verdict on the NAEP analyses would have to be, "Must try harder."

LIMITED USEFULNESS OF THE SCALE

The ETS Reading Proficiency Scale serves as a single measure with which to compare groups that may differ widely in age and in socioeconomic background. As we noted earlier, one experienced observer suspects that the Reading Proficiency Scale actually measures verbal intelligence. Even if we were willing to accept verbal intelligence as a surrogate for *literacy*, what can we do with the information that the scale provides? That question is important because the NAEP was initiated not only to monitor progress, but also to contribute to improvement.

> The purpose of national assessments of reading is to give the education community and the general public a total view of the state of reading literacy in the nation, *with enough detail to enable both groups to draw inferences about what steps might be taken to improve that state*. (Emphasis added)
>
> * * * * *
>
> As it stands, the NAEP scale is of little help in determining what this level [functional literacy] might be, or ought to be.
>
> * * * * *
>
> The authors of the NAEP reports suggest that reading skills in the upper ranges of the distribution (i.e., beyond the level of functional literacy) could be improved by training in "higher-order reading skills." What such skills are, or what that training would consist of, no one has yet spelled out.
>
> * * * * *
>
> Moreover, it is also possible that upper-level reading skills can be improved to some significant degree if students and adults will simply devote more time and effort to reading and if better instruction is made available to them.[13]

To produce general statements from such a limited exhibition of competencies requires a breathtaking set of assumptions.

What a revelation. Devote more time and effort to reading, and improve your skills! But why separate "more time and effort" from "better instruction"? Better instruction is largely a matter of motivating students to read more and to devote more effort to that activity. Frank Smith summarized that very argument in his new book,[14] and more and more people are coming to see that measures of individual traits will never yield insights on how to improve the level of literacy. According to cognitive scientist David Olson, whose specialty is literacy, there used to be a trait – call it verbal ability – that explained things. "It's gone," Olson says. "At least, it's gone for me."[15] It is "gone" because a single trait has no power to explain how people do such things as read, understand spoken language, or write comprehensible prose.

The NAEP survey of reading proficiency is attractive because of its simplicity. Yet thoughtful observers are forced to ask whether the findings are useful enough to justify their cost. Make no mistake about it; such surveys are costly – and the complex statistical analysis of the data by ETS added significantly to the price.[16]

How much harm will be done if policy makers accept Carroll's conclusion from the NAEP data: that "the nation will have to accommodate itself pretty much to the levels of reading skill now attained by various segments of the population"? By way of contrast, one has only to look at the Japanese schools, which produce middle to high literacy levels in virtually every pupil by age 11 – and that in the face of an initial task (learning the ideographs) that is much more difficult than learning to recognize English words.

In their attempts to make the ETS Reading Proficiency Scale meaningful, the designers have resorted to descriptions that make it trivial. Of the Rudimentary level, for example, they say, "Performance at this level suggests the ability to carry out simple, discrete reading tasks." The scale scores are derived from examinees' performances on a small number of tasks during an in-school assessment.

To produce general statements (such as the one I just quoted) from such a limited exhibition of competencies requires a breathtaking series of assumptions. We now have enough evidence regarding the importance of context in language assessment to make such assumptions untenable and the descriptions quite misleading, insofar as they can accurately tell us what the students can do.[17] To predict with any accuracy which reading materials an individual will be able to comprehend, we must know that person's prior knowledge and cultural experiences.[18] Moreover, there is a danger that the descriptive *labels* will become attached to individuals as simple summary descriptions of their reading ability. In reality, reading achievement is not unidimensional. People tend to exhibit different performances in different contexts, since interest, motivation, intention, and the like all play

a role. To place all examinees on a single aggregated scale is not a worthwhile end in itself, and the distortions that such a scale inevitably creates are a long-term disservice to us all.

THE POTENTIAL USEFULNESS OF NATIONAL ASSESSMENTS

In spite of the criticisms that we have aired here, we believe that national assessments can contribute to decisions on education policy. To have relevance for policy, however, such assessments must use measures that are connected to teaching and learning. In part, we have argued here that simple statistical models – however dazzling the mathematics they employ – do more harm than good when they separate the assessment of reading from classroom processes. The conclusion that Carroll drew from his interpretation of the ETS Reading Proficiency Scale – that the nation will somehow "have to accommodate itself pretty much to the levels of reading skill now attained by various segments of the population" – illustrates the danger.

Efficient and policy-relevant assessment of language achievement (including reading) is feasible, according to an independent appraisal of the language surveys carried out since 1979 in England, Wales, and Northern Ireland by the Assessment of Performance Unit (APU). Geoffrey Thornton, who conducted the appraisal, used his findings to suggest policy implications for such groups as decision makers, developers of public examinations, parents, government officials, employers, postsecondary educators, commentators, and textbook publishers.[19]

"Brother! Did you ever have one of those days? By the way, our principal said to tell you the school system's attorney will be contacting you."

The APU surveys were based on functional (communicative) language theories, which specifically reject the view of reading as a one-dimensional, task-independent, context-free trait. Variations in achievement due to variations in reading tasks, the amount of time allotted to them, and the ages of the readers were both expected and sought. Speaking and listening were also assessed.[20]

The NAEP design proceeded from two assumptions: 1) that a one-dimensional construct, *reading proficiency*, exists and 2) that a valid measurement scale can be derived by applying an item response model to right/wrong questions. Some evidence suggests that the decision to apply an item response model was dominant. For example, Rebecca Zwick noted that "some items were excluded from the reading and writing proficiency scales because of practical considerations or, in the case of reading items, because they were expected to produce violations of unidimensionality assumptions."[21]

The issue is far too complex to discuss in full here, but the discussion above should have made clear the fact that the NAEP and the APU are based on totally different concepts of reading. We strongly believe that a functional view (such as is taken by the APU and by a majority of language theorists today) yields measures connected to the processes of teaching and learning and hence can yield policy recommendations about those processes. The view of reading as a single trait has no link to classroom processes and hence cannot yield recommendations, whether related to policy or otherwise.

The NAEP Reading Proficiency Scale is a fiction, attractive for its simplicity and elaborate computer programs but lacking in the most essential commodity: validity. The elaborate comparisons that Carroll derived from it are no better, and it would be unwise to give them much weight.

---

1. National Assessment of Educational Progress, *The Reading Report Card: Progress Toward Excellence in Our Schools; Trends in Reading over Four National Assessments, 1971-1984* (Princeton, N.J.: Educational Testing Service, 1985).
2. Frederic M. Lord, *Applications of Item Response Theory to Practical Testing Problems* (Hillsdale, N.J.: Erlbaum, 1980); and Albert E. Beaton, "The Design and Analysis of NAEP Data," paper presented at an invitational seminar sponsored by the National Foundation for Educational Research and the London University Institute of Education, London, 25-27 June 1987.
3. Irwin S. Kirsch and Ann Jungeblut, *Literacy: Profiles of America's Young Adults* (Princeton, N.J.: National Assessment of Educational Progress/Educational Testing Service, 1986).
4. National Assessment of Educational Progress, *The Reading Report Card. . . .*
5. John B. Carroll, "The National Assessments in Reading: Are We Misreading the Findings?," *Phi Delta Kappan*, February 1987, p. 424.
6. Ibid., p. 426.
7. Ibid., pp. 428-29.
8. Ibid., p. 430.
9. Marie Carbo, "Reading Styles Research: 'What Works' Isn't Always Phonics," *Phi Delta Kappan*, February 1987, pp. 431-35; Kenneth Cadenhead, "Reading Level: A Metaphor That Shapes Practice," *Phi Delta Kappan*, February 1987, pp. 436-41; and Denise Nessel, "Reading Comprehension: Asking the Right Questions," *Phi Delta Kappan*, February 1987, pp. 442-45.
10. Carroll, p. 429.
11. Albert E. Beaton, *Implementing the New Design: The NAEP 1983-84 Technical Report* (Princeton, N.J.: National Assessment of Educational Progress/Educational Testing Service, Report No. 15-TR-20, 1987). The profession is indebted to the NAEP, to ETS, and to the ETS statisticians working under Beaton for this thorough report detailing the extremely complex design of the NAEP. We are critical of the assumptions and the models they employed, but we admire their high level of professionalism in carrying out the design.
12. Harvey Goldstein, "Measuring Changes in Educational Attainment over Time: Problems and Possibilities," *Journal of Educational Measurement*, vol. 20, 1983, pp. 369-77.
13. Carroll, pp. 426-30.
14. Frank Smith, *Insult to Intelligence: The Bureaucratic Invasion of Our Classrooms* (New York: Arbor House, 1987).
15. David R. Olson, "Mining the Human Sciences," *Interchange*, vol. 17, 1986, p. 177.
16. Beaton, "The Design and Analysis of NAEP Data."
17. See, for example, Beth Davey and Carol Lasasso, "The Interaction of Reader and Task Factors in the Assessment of Reading Comprehension," *Journal of Experimental Education*, vol. 20, 1984, pp. 199-206; Keith E. Stanovich, "Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy," *Reading Research Quarterly*, vol. 21, 1986, pp. 360-407; Joanne M. Golden and John T. Guthrie, "Convergence and Divergence in Reader Response to Literature," *Reading Research Quarterly*, vol. 21, 1986, pp. 408-21; and Roger Fowler, *Linguistic Criticism* (Oxford: Oxford University Press, 1986), Ch. 7.
18. Peter Johnson, "Prior Knowledge and Reading Comprehension Test Bias," *Reading Research Quarterly*, vol. 19, 1984, pp. 219-39; and Margaret S. Steffensen et al., "A Cross-Cultural Perspective on Reading Comprehension," *Reading Research Quarterly*, vol. 15, 1979, pp. 10-29.
19. Geoffrey Thornton, *APU Language Testing 1979-1983: An Independent Appraisal of the Findings* (London: Department of Education and Science, 1986).
20. Greg Brooks, *Speaking and Listening: Assessment at Age 15* (Windsor, England: NFER-Nelson, 1987).
21. Rebecca Zwick, "Validity Issues in NAEP: Year 15 Reading and Writing Assessments," in Beaton, *Implementing the New Design. . .* , pp. 525-44.