CSA TN: **205937**

Maxcost :**CSU**

10 August 2010

ODY /

IL:67940116

Patron: Meier, Nicholas

MB@ ILLiad 36185

Shipping Address:
CSU Monterey Bay
Library
100 Campus Center
SEASIDE CA 93955-8001

Borrower: **MB@**
Lending String: CPO,CSB,*CSA,CSH,CLA

Location: 3S
Call #: The National elementary principal.

Journal Title: The National elementary principal.

Vol.: 59 No: 2 Mon/Yr: January 1980
Pages: 77-80

Article Title: Cook, A. Mack, H & Meier, D; New York's Great Reading Score Scandal
Ariel: 198.189.249.141
Odyssey: 198.189.230.235

Interlibrary Services OFFICE USE ONLY

Search Info	Date:	Initials:	
Not on Shelf	☐	Not Yet Received	☐
Missing/Lost	☐	Lacking	☐
At Bindery	☐	Poor Condition	☐

Notes:

PROBLEM REPORT

If you have a problem with the delivery of the requested item, please contact us with the following information.
Please return via Ariel (130.86.12.241)
or fax (916) 278-7089.

☐ Pages missing or other pp._____ to _____

☐ Edges were cut off pp._____ to _____

☐ Illegible copy – please resend entire item

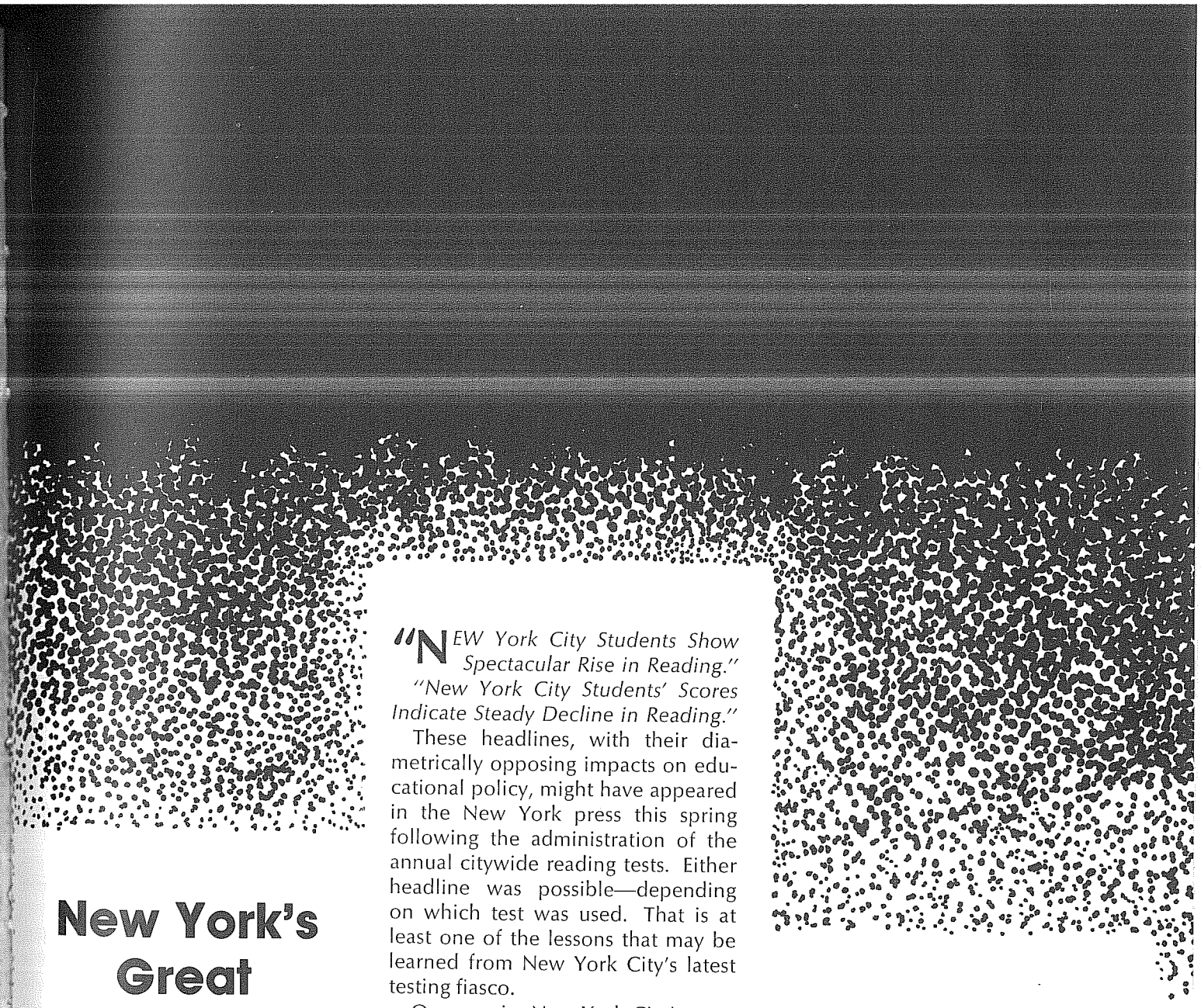
☐ Incorrect article sent

☐ Other (please explain):

Interlibrary Services **41600101**
Sacramento State University Library
2000 State Univ. Dr., East
SACRAMENTO CA **95819-6039**

LIBRARY MAIL
RETURN POSTAGE GUARANTEED

CSU Monterey Bay
Library
100 Campus Center
SEASIDE CA 93955-8001



New York's Great Reading Score Scandal

ANN COOK
HERB MACK
DEBORAH W. MEIER

"NEW York City Students Show Spectacular Rise in Reading."
"New York City Students' Scores Indicate Steady Decline in Reading."

These headlines, with their diametrically opposing impacts on educational policy, might have appeared in the New York press this spring following the administration of the annual citywide reading tests. Either headline was possible—depending on which test was used. That is at least one of the lessons that may be learned from New York City's latest testing fiasco.

Once again, New York City's mass testing program has been the subject of major scandal. In 1977, the scandal was sufficient to delay the reporting of scores for nearly half a year while a court battle waged over the reliability of the results. Last year, an urgent memo from the chancellor's office ordered principals to secure the results under lock and key and to recall any test data released. A new test was subsequently administered to the more than half a million students.

Ann Cook is an educational consultant and children's book author.
Herb Mack is director of a curriculum development and teacher training program.
Deborah W. Meier is an elementary school principal in New York City.

Even before the chancellor's edict, however, some principals and teachers had already begun to question the validity of the 1979 testing program. They were baffled by wide discrepancies between the format, content, and scoring system used in the 1979 test and those of previous tests—including other tests by the same publishers. Specifically:

☐ The 1979 tests were harder to look at—the print was smaller and the lines packed more closely together.

☐ The amount of time young children were required to spend reading silently, without assistance or interruption, was almost twice as long as it had been (forty minutes,

compared to twenty-three in previous tests).

☐ The children had to read eleven stories rather than six, with a total of 1,100 words rather than 300.

☐ The stories themselves were staggeringly harder—they involved bigger and longer words, and longer and more complex sentences and ideas. (See Sample Questions.)

☐ Children had to choose among four rather than just three multiple choices.

☐ The multiple choices themselves were considerably more complex and longer. For example, compare this item from the old MAT: *The fireman who showed the class around might be called ( ) Shorty ( ) Slim ( ) Red;* with this one from the new test: *The merchant loaded his donkey with sponges because he ( ) had to deliver the sponges to a buyer ( ) had run out of salt ( ) wanted to break the donkey of a bad habit ( ) thought sponges would make the donkey more comfortable.* And yet the children were expected to get more, not fewer, answers right!

"In fact, New York educators were asking, why give standardized norm-referenced tests at all if they don't provide comparative information accurately?"

The test for third and fourth graders was similarly disturbing. There were more items, the stories were more complicated, the questions were more wordy, and the type face was smaller. And here, too, the number of right answers needed to score "on grade level" was at least as high as in previous tests and, in the case of the fourth grade, significantly higher. A fourth-grade child now needed fifty-two out of sixty right answers, versus only thirty out of forty-five on the old test.

One school administrator, disturbed by these facts, called fellow teachers and principals around the city to check their perceptions. She found that some had been aware of how hard the 1979 test was but hadn't actually checked over the scoring system. Others had angrily resigned themselves to seeing a drop in their students' scores.*

Further investigation turned up the following: this test was the new, updated, revised, "hot-off-the-press" 1978 Metropolitan Achievement Test (MAT) in reading. Five or six cities had used the test in the fall of 1978. It was, like the old MAT Achievement Test, published by Harcourt Brace Jovanovich under the imprint of its wholly owned subsidiary, Psychological Corporation.

But why should the revised Metropolitan, published under an alias, be significantly harder than the 1970 version?

"Scores" on norm-referenced standardized tests do not represent the "percent of right answers" or "number of right answers" as they typically do on spelling tests or other teacher-made tests. Instead, standardized test scores are simply a way of presenting comparative information, or rank order data. They compare children both with others around the nation and with themselves over time. Test makers have always assured us that it shouldn't be easier for a child to get a different "score" if he or she switched to another reputable test, and further that changes over time on any such test are assumed to be of significance. In fact, this is one way of checking on the technical validity of this kind of test. Tests like these *should* produce about the same scores. If they don't, then both tests can't be right, unless they purport to measure quite different skills. But the story comprehension tests included in the old MAT and the new MAT were surely intended to measure the same skill, so how could they produce different results? It made no sense.

In fact, New York educators were asking, why give standardized norm-referenced tests at all if they don't provide comparative information accurately? That is, after all, the advantage they can and do claim over ordinary, inexpensive teacher-made or school-made tests and other informal inventories.

Uneasiness about this test caused one educator to consult with several top-notch testing experts at a well-known independent testing corporation. Around a table strewn with as many old and new Metropolitan test manuals as could be obtained, the test experts pored over the material. They began with a strong belief that the tests were valid. Well-established test makers like Harcourt Brace Jovanovich might make bad and even biased tests, but surely the two tests—the old and the new MAT—couldn't produce substantially different results.

"With a simple stroke of a pen, the test publishers had increased the number of severely retarded readers."

After concentrated study, however, it was evident that on the face of it, the new test seemed clearly much more difficult in every respect—amount of reading, type of reading, and test format. Moreover, the scoring system was not devised to take this increased difficulty into account. As many right answers (and in some cases considerably more) were required to get the same grade level score as could be obtained on the much simpler earlier test. It therefore followed that children *should*

*Since scores were not reported, it is difficult to give an overall picture of the scoring pattern in New York City, but it appears that scores went down in lower-income areas, and up in upper-middle-income areas. There were some outstanding exceptions to this rule, perhaps explained by a principal's keen desire for high scores. An interesting pattern seems to emerge when schools change principals. Scores drop the first year after a new principal is appointed, and rise (sometimes drastically) the second year of a principal's tenure.

score considerably worse this year.

The manner in which grade-level equivalents—long a disgrace in the testing field—were manipulated in this test was particularly disturbing. For no sound reason, the levels had been set to result in more children scoring both two years below and two years above grade level than was the usual practice. This was an apparently arbitrary artifact of the way the new test was scored. With a simple stroke of a pen, the test publishers had increased the number of severely retarded readers.

In fact, one official responsible for the development and distribution of this Harcourt Brace test had acknowledged that the company does publish different tests that are known to produce different results. This official pointed out that the company provides that information to potential customers in order to help them select a test.

The widely disparate methods of reporting scores and arriving at so-called grade equivalents are also tailored to meet the latest needs of the customers. Thus the latest Harcourt Brace test has the widest range of scores (that is, many more very high and very low ones) of any test on the market. Harcourt Brace defends itself—at least privately—by contending that it has long urged the abandonment of grade equivalents since they have no scientific or standardized meaning, but uses them only because school systems demand them.

Given all these differences between the old and the new MAT, the experts considered the available hypotheses to explain them.

It was suggested that perhaps Harcourt Brace had found that children throughout the nation were all reading a lot better in 1979 than they had in 1970, when the last test was produced. Not only would that justify making the revised test a much harder one, but it would be major news indeed.

The experts also hypothesized that Harcourt Brace might have used a different sample population (what's technically called the "norming" or "standardization" group) to try out the two tests. If so, which sample population was more accurate—the old one or the new one? A third grader, for example, could get a "post high school" score on the new test, while in the past, the score was a grade equivalency.

Still another possibility raised was that it didn't matter a bit how hard the actual reading passages are. Perhaps the children would actually do the same on both these tests—the old and the new—despite their widely disparate levels of difficulty. If that were the case, what could the school possibly be measuring under the guise of testing reading skills? There were three possibilities which were intriguing. Perhaps all three were partially true.

First, maybe, after all, despite the national hue and cry over declining standards, children actually were reading better, so that a more difficult test was needed to produce the same distribution of comparative scores.

"Did the . . . administration want the system managed by their predecessors to look as bad as possible so that future test scores would demonstrate that reading had improved under their direction?"

The second possibility, that the sample population had been changed, also bore further investigation. A quick look at the information available in the new manual suggested that the 1978 Metropolitan test was pretested on a population group in which metropolitan areas were significantly underrepresented. It wasn't even clear if the sample included any large urban centers. While this seemed odd for a test that was being used in New York City, was it also true of the old MAT?

As for hypothesis three, perhaps the tests didn't really test reading. Critics have long argued that the actual skill of reading has never been the main determinant of success or failure on tests like these. Maybe New York had simply demonstrated the truth of this disturbing idea.

Did the New York City School Board and Chancellor Machiarolla know any of these facts when they selected the new test? Were they conceivably hoping for worse scores? Did the new, incoming Livingston Street administration want the system managed by their predecessors to look as bad as possible so that future test scores would demonstrate that reading had improved under their direction? Did they actually want that *New York Times* headline spotlighting still another failure for New York City, its schools, and its children? Or was it simply a case of ignorance and bureaucratic blundering?

These speculations suggest another: is it possible to shop around for a test that will produce whatever result a school system wants to demonstrate?

The testing industry—which is also the textbook publishing and, thus, the curriculum development industry—is a major force in shaping our schools and our children. As is now evident from the lobbying efforts against New York State's "truth in testing" legislation, the testing industry probably represents the most cohesive and well-financed lobby in the field of education. These tests themselves are part and parcel of that industry.

Tests have now been introduced as a new "subject matter," added to the usual classroom fare and gradually replacing such "frills" as music, art, history, and science. There are new textbooks and course outlines aimed at teaching children (even quite young ones) how to succeed on tests. A staggering amount of public funds are spent on testing, especially if we calculate the time spent in preparation, administration, and distribution of tests and testing information. And finally, policy decisions about curriculum and education

and the public's perception of schools both rest to a considerable extent upon the published results of such tests.

We have placed our faith in these "objective" tests because we did not trust the "subjective" parents and teachers who know our children best. We forgot that the nameless "theys" out there can also be subjective, biased, or just plain stupid. The difference is that their subjectivity masquerades as "science" and is even harder to combat.

If Harcourt Brace's test—the outcome of ten years of preparation by one of the most prestigious test publishers in the country—has serious flaws, how do we account for the fact that five major school systems used the same test without reporting any problems?

And what leads us to believe that the alternate test used in New York City in June (McGraw-Hill's California Achievement Test) does not have similar or other serious flaws?

As a school administrator, how would you find answers to questions like these regarding your own school's testing program? You're in much the same dilemma as the concerned citizens who worry about their nearby nuclear reactor. Who do *they* go to to get the "truth"—the Nuclear Regulatory Commission?

Perhaps administrators are cynical and acknowledge that they select those tests that will produce the results they find most useful—high scores to qualify for federal monies earmarked for the gifted; low scores for learning disability funds; or, as in the case of Harcourt Brace's recent wonder test, scores that manage to do both at once.

Questions about New York's reading score scandal remain unanswered. The city selected a particular test for reasons we don't know. They recalled the test scores for reasons we don't know. They selected another test for reasons we don't know.

Moreover, despite a subsequent investigation into the results of the test, the public remains uninformed. Despite statements that the board would not pay the almost half million dollars owed to Harcourt Brace Jovanovich, no further information has been released about whether this sum has been paid or not.

SAMPLE TEST QUESTIONS

Are these stories comparable? Try them out on your seven-year-olds.

A. David's grandpa is a forest ranger. His home is on a mountain far from the city. Trees are all around Grandpa's home. Many animals live in the forest. David calls the animals his friends.

B. Miss Matthews' class visited the fire station in their city last week. When they arrived at the station, a tall, thin fireman welcomed them and showed them around. The boys and girls were especially interested in the bright red hook-and-ladder truck, and some of them asked how long its ladder was. "More than 100 feet," the fireman answered.

C. Carol used to ride to and from school with Rita and John every day. In the morning, Rita's father would take the three friends to school in his big old truck. They would all crowd into the front. Then, at three o'clock John's mother would take the children home in her red station wagon. John's baby brother would come along too.

Two weeks ago, Carol moved away. Now getting to school is hard work, and Carol is not happy because she has to walk there by herself. But, in time, she will meet new friends to go to school with again.

D. Headed for the local market, a prosperous merchant loaded his donkey with some bags of salt. On the way, they had to cross a broad stream on a board that spanned the water. in the process, the heavily burdened donkey accidentally slipped and toppled into the stream. It was soaked, but its load was obviously lighter, for the water had dissolved the salt. The next day, laden with salt again, the clever donkey purposely fell into the water to make its load lighter. Disgusted with the donkey, the merchant was determined to prevent it from repeating this behavior. The following day, he loaded the beast with sponges. The enterprising donkey again threw itself into the water, but now it really had to exert itself to get out since the sponges had absorbed the water. For the remainder of the trip, the exhausted donkey staggered under its extremely heavy burden. It had apparently learned its lesson, for it made no additional attempts to outsmart its master.

Passages A and B are the second and last stories that children had to deal with in the old MAT test. They reflect the range of difficulty that seven-year-olds were then expected to handle.

Passages C and D are the second and last stories that children had to deal with in the new MAT test. They reflect the range of difficulty that seven-year-olds were expected to handle on this new test.