DOCUMENT RESUME

ED 295 145 CS 009 181

AUTHOR Brown, Sylvia J.
TITLE Assessment of First-Grade Reading Achievement: Spring 1986 Pilot Study. OEA R & D Report.
INSTITUTION New York City Board of Education, Brooklyn. Office of Educational Assessment.
PUB DATE [88]
NOTE 98p.
PUB TYPE Reports - Research/Technical (143) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Administrator Attitudes; *Beginning Reading; Grade 1; Listening Skills; Primary Education; *Reading Achievement; Reading Research; *Reading Skills; Speech Skills; Surveys; Teacher Attitudes; Writing Skills
IDENTIFIERS Metropolitan Achievement Tests; New York City Board of Education

ABSTRACT

To compare the progress of New York City children with those of other first graders across the country, as well as to establish baseline data for evaluations of early childhood programs and provide direction for second-grade reading instruction, a pilot study examined approximately 5,000 randomly selected first-grade students from monolingual general education classes in New York City public schools. In the spring of 1986, students were given the reading subtests of the Metropolitan Achievement Test (MAT) Primary 1 level. The study also included a teacher "checklist" evaluation of students' performance on 30 communication arts skills. Teachers and administrators were asked to complete a questionnaire on their opinions about the MAT, the "checklist" and their general views on assessing first graders. Results revealed wide variation in reading test scores among first graders, children who participated in early childhood education programs performed significantly better on the standardized tests than children with no school experiences before first grade. Nearly 37% of the sample was reading at or above grade level. Checklist results showed that over two-thirds of the children could perform basic reading skills, though fewer than half could routinely perform more complex reading skills. Survey results indicated that teachers and administrators felt that the checklist was not comprehensive enough to be the only measure of reading performance. (One figure and 20 tables of data are included, and appendixes consist of student surveys for monolingual and bilingual classes, teacher's guide to completing the student survey, administrator survey, and teacher survey.) (MM)

Assessment of First-Grade Reading Achievement:

Spring 1986 Pilot Study

Research Report
Office of Educational Assessment
Research and Development Section
Jane Canner, Administrator

Assessment of First-Grade Reading Achievement:

Spring 1986 Pilot Study

Report Prepared by

Sylvia J. Brown,
Research Specialist

New York City Board of Education
Office of Educational Assessment
Richard Guttenberg, Director

It is the policy of the Board of Education not to discriminate on the basis of race, creed, national origin, age, handicapping condition, sexual orientation, or sex, in its educational programs, activities, and employment policies, as required by law. Any person who believes he or she has been discriminated against should contact: Carole Guerra, Local Equal Opportunity Coordinator, Office of Educational Assessment, 110 Livingston Street, Room 743, Brooklyn, New York 11201. Inquiries regarding compliance with appropriate laws may also be directed to: Mercedes A. Nesfield, Director, Office of Equal Opportunity, 110 Livingston Street, Room 601, Brooklyn, New York; or the Director, Office of Civil Rights, U.S. Department of Education, 26 Federal Plaza, Room 33-130, New York, New York 10278.

SUMMARY OF THE SPRING 1986 PILOT STUDY
OF FIRST-GRADE READING ACHIEVEMENT

In 1986, the New York City Board of Education decided to expand its citywide testing program in reading to include first- grade students for the first time. The impetus for this decision was threefold: (1) to compare the progress of New York City children with those of other first graders across the country, (2) to establish baseline data for evaluations of early childhood programs, and (3) to provide direction for second-grade reading instruction. However, due to concerns about the appropriateness of standardized testing for first-grade children, the testing was limited to a randomly selected pilot sample of approximately 5,000 first graders.

In the Spring of 1986, children in the pilot sample were given the reading subtests of the Metropolitan Achievement Test (Primary 1 level), a standardized reading achievement test designed to test first graders. The pilot assessment also was expanded to include a teacher "checklist" evaluation of students' performance on 30 communication arts skills. Teachers and administrators were then asked to complete a questionnaire which elicited their opinions on the MAT and "checklist" as well as their more general views on assessing first graders. The results of this pilot study were used to assess the potential value of the MAT and the checklist to serve the three purposes described above and to obtain teacher and administrator recommendations regarding citywide assessment based on their experiences in this pilot study.

The pilot sample was chosen by randomly selecting 350 monolingual general education classes from classes throughout the city*); the final general education sample for whom both MAT and checklist data were obtained included about 4,200 children. Many of the classes selected for the pilot sample test first graders with the MAT reading subtests anyway, for their own district or evaluation purposes. In some cases, however, districts did test sample classes solely for pilot test purposes.

In other instances, districts declined to participate in the pilot sample and alternative classes were chosen from comparable schools in cooperating districts.

The sample was randomly chosen and thus it was hoped it would be representative of the reading achievement of first graders citywide. However, since a comparison of average second-grade reading scores in pilot schools and all city elementary schools showed that second graders in pilot schools had somewhat lower scores, the first grade results from this study probably underrepresent the true reading achievement of all first graders in New York City. Since 36.9 percent of the first-grade sample was reading at or above grade level, this suggests that first graders' reading achievement is slightly lower

*Smaller samples of bilingual and special education children were obtained and are reported on in the full report.

than but comparable to that of second graders in New York City, which was 42.3 percent reading at or above grade level in the spring of 1986.

The MAT pilot-study results also revealed wide variation in reading test scores among first graders, which is not surprising given the range of developmental progress in children this age. What was particularly interesting, however, was that not only was there a larger than predicted group who performed below the national average, there was also a larger than expected group of children whose reading achievement was much higher than the national average. In addition, children who participated in early childhood education programs performed significantly better on the standardized test than did children with no school experience before first grade, especially if those children had both pre-kindergarten and kindergarten experience. It is likely that factors related to pre-school experience, such as home environment, also contributed to this difference in achievement.

The potential value of the MAT in providing citywide information on first graders' reading performance was adversely affected by some strong teacher and administration concerns regarding the difficulty and length of the test and its inappropriateness for first graders. The value of the test as an objective measure of student reading achievement must thus be balanced by the perception that it was an inappropriate measure of student progress.

The checklist results revealed that over two-thirds of the children could perform "most of the time" basic reading skills, such as recognizing initial and final sounds and letters or associating letters and sounds. Fewer than half (40%) could routinely perform the more complex skills, such as using contextual clues when coming upon unknown words. Perhaps more valuable than these general checklist findings for all students together was the use of a checklist for each child. Teachers reported that the checklist helped them to focus on and assess the individual child's strengths, weaknesses and progress during the school year. The pilot study revealed a trade-off, however, between the information obtained by using checklists and the approximately three hours it took a teacher to complete checklists for an entire class.

Teachers and administrators would like to see a teacher checklist included as part of a citywide first-grade assessment process. Most felt, however, it was not comprehensive enough to be the only measure of reading performance. Although many would like to include a standardized test as part of the assessment program, there were serious reservations about using the MAT for that purpose.

The following recommendations are made as a result of the pilot study:

- A citywide assessment program for the first graders should reflect the need for diverse types of assessment instruments to suit diverse purposes, such as identifying students' strengths and weaknesses, comparing of New York City children with national norms, and evaluating early childhood programs.

- Because of the generally negative reaction to using standardized citywide tests at this grade level, a sampling approach to citywide testing should be considered. A carefully chosen sample would give the data needed for citywide program evaluation without requiring that every student be tested, or that results be reported for every school and district.

- Since teachers and administrators reacted more negatively to the MAT than to the concept of standardized testing of first graders in general, an alternative might be to seek a more acceptable standardized test to administer to the selected sample. The benefits of any alternative test would have to be weighed against the benefits of a uniform citywide testing program from grade to grade, since the MAT is used citywide at grade 2.

- Those districts that continue to use the MAT for their own evaluation purposes should offer appropriate staff development in the use and interpretation of the test results, particularly in light of the pilot respondents' strong concerns about the test at this grade level.

- If the checklist continues to be used either as an option or as part of a citywide assessment program, aspects of the checklist such as which skills it measures and what king of rating scale it should have need to be re-examined.

- The MAT test results should be shared with the Division of Curriculum and Instruction so they are aware of the large number of first graders performing below the national average on the skills assessed by this test.

In sum, these results have implications for both test developers and public education decision-makers, particularly regarding the attitudes of teachers toward standardized testing of young children and appropriate ways to gather standardized reading achievement data for the first graders.

CONTENTS

CONTENTS (Continued)

CONTENTS (Continued)

TABLES

TABLES (Continued)

FIGURE

I. INTRODUCTION

The New York City Board of Education planned to expand its citywide reading testing program to include first-grade students for the first time in the spr[illegible], 1986. This was done in order to compare the progress of New York City children with those of other first graders across the country, to establish baseline data for evaluations of early childhood programs and to provide direction for second-grade reading instruction. However, various groups of parents, teachers and early childhood personnel strongly opposed testing first graders on a citywide basis; this opposition was based on the belief that first-grade children are too young to be tested reliably and that the testing experience is traumatic, yielding results which are not a true reflection of achievement.* In response to these objections it was decided to limit the administration of the first-grade reading tests to a representative research sample of approximately 5,000 students, rather than to test all first graders citywide. In addition, a "checklist" evaluation of students' communication arts achievement was also completed by teachers for students in this sample. Teachers and administrators who participated in the pilot program were also given questionnaires which asked for their opinions about the standardized test and checklist as well as their overall suggestions for selecting appropriate ways to assess first-grade students' achievement in reading.

The testing, which took place in the spring of 1986, had the following purposes:

---

*It is interesting to note, however, that at the time this opposition was voiced, 28 of the 32 local community school districts already tested some or all of their first graders in reading for their own district purposes.

- To describe first-grade reading achievement levels on both the standardized test and the teacher checklist.
- To examine the relationship between scores on the standardized test and those on the teacher checklist.
- To analyze the relationship between first-grade reading achievement and previous educational experience, i.e., all-day kindergarten and pre-kindergarten experience.
- To understand administrators' and teachers' opinions of the standardized test and the checklist and to obtain their suggestions for assessing first-grade students' reading achievement in general.

The ultimate goal of the study was to provide information to policy-makers regarding the most appropriate means of assessing first-grade students' reading achievement. In addition, the results were to be used to judge the usefulness and limitations of test scores for individual children in early grades, to provide information on the effectiveness of early childhood programs and to provide baseline data on first-grade reading achievement.

II. METHODOLOGY

MEASURES OF READING ACHIEVEMENT

Metropolitan Achievement Test (MAT)

The test selected as the citywide reading test for grade two (the Degrees of Reading Power test was selected for the grades where it was available, i.e., grades 3-12), after extensive review by technical and curriculum experts at the Board of Education, was the 1986 edition of the Metropolitan Achievement Test (MAT). It was therefore decided to administer an appropriate (Primary I) level of this test to first graders for this study. This level of the test included three subtests: Vocabulary, Word Recognition and Reading Comprehension.

Communication Arts Checklist

The checklist used was adapted from a form developed by the Bank Street College of Education and revised by the Early Childhood Unit of the New York City Board of Education's Division of Curriculum and Instruction (C. and I.). The checklist required teachers to evaluate students' communication arts skills, in the areas of listening, speaking, reading, and writing, on a scale of "1" (not yet) to "3" (most of the time). Children who were bilingual were judged on their communication arts skills in their native language. (A copy of the monolingual and bilingual checklist and the directions sent to teachers appears in Appendix A.)

MEASURES OF ADMINISTRATOR AND TEACHER OPINION

Two questionnaires were developed by the Office of Educational Assessment (O.E.A.) to gather information on the opinions of administrators and

teachers regarding the two measures of reading achievement as well as towards a citywide first-grade assessment program in general. Each questionnaire included closed-ended and open-response items. A copy of each appears in Appendix B.

THE TEST SAMPLE

Since the decision was made not to test first graders on a citywide basis, a random sample of 5,000 first graders was to be selected for testing instead. However, it is interesting to note that many more than 5,000 first graders were actually tested with the MAT in the spring, 1986. First, 17 of the 32 community school districts opted to administer the MAT to all of their first graders as part of their district-wide testing programs. In addition, 119 schools from 25 districts throughout the city which were part of the "Reduced First-Grade Class Size" program evaluation also tested all their first graders. Thus, when the pilot sample of 5,000 students was selected, many of the classes selected were already planning to test their first graders, for one of the two reasons described above. Some were not already being tested; these classes were tested solely for the purpose of the pilot program.

A total of 42,771 first-grade students attended classes that were part of the MAT testing program, either because of their district-wide testing programs, the "Reduced Class Size" evaluation, or the pilot test of first-grade reading achievement. Although approximately 8,000 of these 42,771 did not take the test because they were absent the day of testing or were exempt from testing, the almost 35,000 students who were tested comprised close to half of the approximately 73,000 first graders in the New York

City school system. It is important to reiterate, however, that the 5,000 students in the sample were chosen to be representative of all New York City first graders whereas the larger group was not. Thus, this report presents results only from the sample.

The sample was chosen by randomly selecting 350 monolingual general education classes (which included about 9,000 students) from all such classes throughout the city. Whole classes were sampled both for practical purposes and because student performance was to be analyzed on a group rather than an individual basis. In addition to assessing monolingual general education children, the pilot study also sought to examine the reading achievement of bilingual and special education students. A random sample of 30 bilingual classes from 292 first-grade bilingual classes in the city was selected for participation. These children were not tested with the standardized instrument because they would ordinarily have been exempt from citywide testing since English was not spoken at home and they had spent less than two years in an English-language school system. They were instead assessed only with the Communication Arts Checklist. In order to ensure data on special education children, ten classes were randomly chosen from MIS IV classes* throughout the city. Children diagnosed as having emotional problems, learning disabilities, or speech and language problems were not selected for testing at this time. Students in MIS IV classes were given the MAT with any modifications that were indicated on their Individual Education Plans and were also assessed by the checklist.

---

* First graders placed in MIS IV classes are readiness-delayed learners.

Each district superintendent was sent a letter which explained the purpose of the testing program and which identified classes in that district which had been randomly selected for inclusion in the sample. In response to this request for participation, 17 districts who were already testing all their first-grade classes for their own purposes and two other districts who were not testing any of their first graders agreed to allow the selected classes to be a part of the sample. Ten districts who were part of the Reduced First-Grade Class Size Program would permit participation only for those classes that were to be tested anyway as part of the evaluation of the Reduced Class Size Program. Two districts which were not testing first graders for either their own district purposes or for the Reduced Class Size Program refused to allow testing in any of the classes selected for the sample. Because 12 districts permitted either limited or no testing in the selected pilot sample classes, a total of 111 classes from 89 schools originally selected as part of the sample had to be replaced.

In most cases, substitute schools were selected from the same boroughs as the original school in order to maintain the geographical balance of the sample. However, the major criterion for choosing a substitute school was that the median grade equivalent for second graders tested with the California Achievement Test in the spring of 1985 was similar to that of the originally selected school. In all but a few cases, where a difference between median scores of a month occurred, perfect matches on this criterion were made. If more than one possible substitute school had the same median grade equivalent, then the school with both the closest overall school percentage of students performing at or above grade level and New York City-wide rank was selected.

PROCEDURES

The MAT was administered to monolingual first-grade students on the same day as the citywide reading tests for grades 2, 3, 4 and 7, i.e., April 21 and 22, 1986. Those districts not testing district-wide received test materials for selected classes in the sample schools or for all classes in schools that were in the Reduced First-Grade Class Size Program evaluation. In addition to the student information that was usually collected, teachers of first-grade students in the pilot sample were asked to provide on the answer document information regarding students' previous school experience and language spoken at home.

All testing materials were provided by the Office of Educational Assessment (O.E.A.) to the schools. The same procedures were followed as for any citywide testing program (e.g., in such areas as production, packaging, and delivery of test materials; retrieval of answer documents for scoring; retrieval of all test materials after the administration; and transmittal of answer documents to O.E.A.'s Testing Section).

Teachers in both monolingual and bilingual pilot classes were also asked to voluntarily complete a Communication Arts First-Grade Checklist for some or all of their students. In bilingual classes, teachers were asked to rate students' communication arts skills in their native language. Checklists were to be completed for at least every third student (from an alphabetical class list) in each class in the sample; those teachers who wished to complete a checklist for all the students in their class were encouraged to do so. As an incentive, districts were reimbursed based on the number of checklists completed and returned to O.E.A. Since the

checklist was developed by the Division of Curriculum and Instruction's Early Childhood Unit, the administration of the checklist was the joint responsibility of the district test liaison and the early childhood liaison.

Finally, teachers of participating classes were also asked to complete a questionnaire which asked for their reactions to the test, the checklist and first-grade assessment in general. Likewise, administrators of schools with participating classes were asked to complete a survey for their opinions on the test, the checklist and assessment of first graders in general.

III. RESULTS

DESCRIPTION OF SAMPLE

As indicated earlier, almost 43,000 first graders were in classes that were part of the spring, 1986 testing program. Many more students were tested than were in the pilot sample because they were being tested for other purposes; only data from students in the pilot sample are reported here since that sample was chosen to be representative of the city. Of all the students tested, O.E.A. received and analyzed 6,936 MAT answer documents for the purposes of this study.

Of the 350 monolingual classes asked to participate in the pilot study, 311 teachers returned a total of 5,544 completed checklists. (This includes nine special education classes out of the ten who were asked to participate.) In addition, 27 bilingual class teachers sent in a total of 375 completed checklists.

Although there were 6,936 pilot study students who took the MAT, only 5,544 students had checklists completed. There were fewer checklists than test answer documents because some teachers filled out checklists for every third student in the class. The actual number of students for whom there was both a checklist and a MAT score was even lower, i.e., 4,243, because some of the students for whom there was a checklist were absent the day of the MAT or were classified as "Limited-English proficient" and thus were not tested.

Results presented in this report are based on the sample of students who had both MAT and checklist data unless otherwise noted. This makes it possible to examine relationships between reading achievement as measured

by a standardized reading achievement test and by a teacher-completed observational checklist. Statistical analyses support the decision to use this sample: the MAT test results for the "matched" sample did not differ significantly from those of the rest of the pilot study sample. (See Table 1).

Table 1

Comparison of MAT Scores for "Matched" Sample and Total Sample Excluding "Matched" Sample

	"Matched" Sample	Total Sample Excluding "Matched" Sample
Mean Scaled Score	507.0 (n=4198)	506.8 (n=2669)

$t = .19$, $p < .05$

Within the matched sample of 4,243 students, 4,198 were general education students and 45 were special education students. The discussion that follows is based on results for the group of 4,198 general education students who have both MAT and checklist data. Findings for the 45 special education students and for the 375 bilingual students are presented later in the report.

The pilot sample was carefully chosen to provide a random sample of [illegible] first graders in the city. The fact that the vast majority of classes who agreed to participate did so suggests that the results from this pilot sample could be generalized to all first graders. However, random selection does not ensure representativeness; and some replacements had to be made. Therefore, additional analyses were done to judge how well the pilot sample results reported here reflect the reading level of all first graders in the city.

One way to address this question was to compare the pilot sample's performance on the MAT with that of the close to 32,000 first graders citywide who took the MAT. However, that larger group was biased in two ways. It included a higher proportion (one-third) of children from Reduced Class-Size schools than would be found if all first-grade children were included. In addition, two districts which typically out-perform most other districts did not test their children with the MAT. Thus, the MAT results for the larger city sample of 32,000 would be expected to be lower than if all first graders had been tested. The pilot sample, therefore, should have scored higher on the average than the citywide sample of first graders tested, since the pilot sample was designed to be representative of all first-grade children.

A comparison of the pilot sample and the citywide sample confirms that the pilot sample had higher average MAT scores than the city group and thus a greater proportion of children reading at or above grade level. These results suggest the pilot sample was a more representative sample than the citywide sample of first graders who took the MAT. However, further analyses were conducted to try to judge how well the pilot sample represented all first graders, including the close to 40,000 children not tested.

Table 2

Comparison of MAT Results:
Citywide Sample vs. Pilot Sample

	n of Students	Mean MAT Score	Median MAT Score	Percent At or Above Grade Level
Citywide Sample	31,839	50.3	44	32.1
Pilot Sample	4,198	52.8	48	36.9

Since first-grade test scores for all first graders were not available for a complete comparison of sample and population results, it was decided to instead compare second-grade MAT test results for the pilot sample schools versus all second-grade scores. The assumption was made that first-grade and second-grade children within a school would have similar levels of reading achievement. Thus, if the pilot sample school's average second-grade MAT scores were found to be similar to the average MAT score for all second graders in the city, one could infer that the first-grade pilot students had MAT scores similar to all first graders in the city and were us a representative sample.

A comparison of average second-grade reading scores in pilot schools and all city elementary schools (see Table 3) showed that second graders in pilot schools had somewhat lower scores. These data suggest that, in spite of efforts to choose a representative sample, the pilot sample results may reflect a lower level of reading achievement than if all first graders were tested. This conclusion is supported by the additional finding that there was a higher proportion of reduced-class-size children in the pilot sample (20 percent) than in the population of all first graders (15 percent).

Table 3

Comparison of MAT Spring 1986 Scores for Second Graders:
Pilot Sample Schools vs. All Schools*

	n of Schools	Raw Score Mean	Raw Score Median
All schools	610	67.6	67.3
Pilot sample schools	181	65.2	63.9

* Based on aggregate school scores.

In sum, this additional analysis of second-grade MAT scores, though not conclusive, provided the best possible evidence on representativeness. They suggest the following first-grade results must be viewed with caution: it is likely the sample under-represents the reading achievement of all first graders. If all first-grade children were tested, the scores would probably be higher than those reported here.

Table 4

Ethnic Representativeness

	Pilot Sample Schools	Citywide
	%	%
Hispanic	40	38
Black	43	37
White	12	20
Asian	5	6
American Indian	0	0

Other information was collected to try to better understand the characteristics of the pilot test sample. The pilot sample represented ethnic groups in roughly the same proportion as they exist in first-grade classes throughout the city. The only difference was that the pilot sample included a slightly higher proportion of black students and relatively fewer white students. Thus, the random sampling resulted in a pilot group that fairly represented the ethnic diversity in the city.

For approximately half of the sample (n=2,228), information was also available on students' home language. Close to three-fourths of this group had English as a home language, but a large minority (about 20 percent) were from a Spanish-language background. In decreasing order, students in the sample were also from Chinese (2 percent), Haitian-Creole (1 percent), Greek (.5 percent), or "other" language environments (4 percent).

Prior educational experience also varied among students in the pilot sample. For the 2,153 children for whom information was provided, the majority (n=1,472) had only kindergarten experience, 442 had kindergarten and pre-kindergarten experience and 239 had neither. Comparisons of the two major language groups, English and Spanish, reveals a significant association between home language and pre-school experience (x^2=42.32, df=2, $p < .001$). More children with English as a home language had pre-school experience than those where Spanish was the primary language spoken at home.

RELIABILITY AND VALIDITY OF THE MAT

The MAT is a nationally standardized test. The internal consistency reliability (Kuder-Richardson 20) of the MAT was high for the national standardization sample as well as for the New York City sample (see Table 5 below).

Table 5

MAT6 Kuder-Richardson Reliability: National and N.Y.C. Data

	Vocab.	Word. Rec.	Reading	Total
National	.92	.88	.91	.96
N.Y.C. Pilot	.90	.88	.89	.96

According to the MAT Preliminary Technical Manual, "The most critical aspect of validity in relation to an achievement test is content validity -- the extent to which test content constitutes a representative sample of the skills, knowledge, and understanding that are the goals of instruction." The designers of the MAT sought to develop a test that best represented curriculum across the country. In choosing a test series to measure achievement of New York City students across all grades, various review committees unanimously chose the MAT as the series that provided the closest match to the New York City curriculum across all grade levels for both mathematics and reading. Although some problems with the MAT were noted, particularly at the lowest grade levels, it was thought overall to represent the best match of the tests offered. The content validity of the MAT specific to the N.Y.C. first-grade curriculum can best be determined by a careful comparison of the test content with the curriculum.

Evidence of criterion-related validity was also gathered by the test developers. They found a high correlation between scores on the MAT and the Otis-Lennon School Ability Test. The technical manual reports that earlier editions of the MAT yielded correlations with other achievement tests regularly in the .60-.85 range, i.e., scores on tests measuring similar content are strongly related to MAT scores.

RELIABILITY AND VALIDITY OF THE CHECKLIST

Unlike the MAT, the checklist was devised for the purpose of this study and thus there were no published data on its reliability and validity. In order to judge these properties of the checklist, a number of analyses to provide data on how reliable and valid the checklist was for the pilot sample were conducted.

As mentioned earlier, the checklist divided communication arts and skills into four categories: listening, speaking, writing, and reading. One way to examine first graders' communication arts skills would be to separately analyze performance in each of these four categories. In order to judge whether it was appropriate to create separate checklist scores for each of the four skills subsections, factor analysis was applied to examine the pattern of relationships amongst items. The factor analysis did not support the multidimensionality of the checklist. Only one factor, which included items from each of the four subsections, appeared to be operative. Thus, the checklist was viewed as assessing one generalized concept, communication arts skills, which included listening, speaking and writing in addition to reading.

Once it was decided that to derive only a single checklist score using all of the checklist items, a total communication arts score was created for each student by adding the rating of "1" (not at all), "2" (sometimes) or "3" (most of the time) for each of the 30 checklist items. The minimum score was 30 (all "1"s) and the maximum was 90 (all "3"s). The reliability of this total checklist score was examined by calculating a measure of internal consistency, the coeffiecient alpha. The internal consistency reliabiity estimate for the checklist scores in the pilot sample was very high, .98.

Since the checklist score is based on teacher rating of students' performance rather than student performance itself, it is possible that ratings may be affected by factors other than student achievement, such as different teachers' standards or varying interpretation of the checklist

items or rating scale. Thus, it was necessary to judge how consistent the checklist ratings were from one teacher to another, i.e., inter-rater reliability. Fortunately, there are some classes in the city system which are team taught, i.e., two teachers work with the same class, so that teachers could separately rate the same group of children. In all, 66 students from six team-taught classes were evaluated on the checklist. Each child was separately rated by the two teacher who worked as a team.

Inter-rater reliability was examined in two ways. First, for each item on the checklist, teachers' ratings were compared to determine how often two teachers agreed and gave a child the same rating ("1" - "not yet", "2" - "sometimes", or "3" - "most of the time"). The higher the "percent of agreement," the more reliable the measure. This analysis revealed on average rate of agreement for all items of 74 percent. While this result shows considerable consistency in ratings, it also reveals there is some variation in teachers' judgments about children. The two items with the lowest percent of agreement were item 2, "retells a simple stoy in sequence", (53.8 percent agreement) and item 6, "looks at pictures and demonstrates understanding of content", (56.9 percent agreement). Particular caution should be used in examining data based on these items since teachers are less consistent in judging these skills.

A second way to look at inter-rater reliability is to consider the child's total checklist score (the sum of ratings on all items). In all but one of the six team-taught classes, there was a near perfect correlation (Spearman rank correlation $\geq$ .94) between how children were rank ordered using one team teacher's total checklist rating and the other team member's

total rating of the same child. Thus, the overall checklist assessment of communication arts is very reliable. While overall ranking of students is approximately the same for team teachers, various combinations of items can result in the same ranking position. Given the average agreement on ratings of individual items of only 75 percent, caution should be used in interpreting students' ratings on individual items.

To encourage uniform teacher judgement (high inter-rater reliability), each teacher was given a Teacher's Guide to the Checklist. The Guide included a description and illustrations for each of the 30 items. (A copy of the Guide is in Appendix A). The district test liaison and the district early childhood liaison were also strongly encouraged by C. and I. and O.E.A. to plan an orientation session for teachers on how to complete the checklist.

The content validity of the checklist is supported by the process by which the checklist was developed. The checklist was developed by staff members of the Division of C. and I., who are also responsible for development of the city curriculum in early childhood education. Since their intent was to use the checklist results to assess the results of their curriculum, it is reasonable to assume that they closely matched the checklist content with the curriculum.

CHECKLIST FINDINGS

Teachers used the checklist to judge the degree to which a child had developed each of thirty communication arts skills. Teachers' observations and ratings of the child were based on each child's classroom performance over a period of time. No separate "testing" situation was created asking

children to perform each of the thirty skills. Instead, teachers' knowledge of student performance day-to-day in their classroom setting formed the basis for the assessment.

Teachers rated whether children accomplished each of thirty communication arts skills "most of the time", "sometimes" or "not yet". It is important to remember that this rating system does not provide information on the quality of the child's performance. Also, since the pilot sample probably under-represents the level of performance of all first graders, the following checklist results are a very conservative estimate of all first graders' communication arts achievement. Close to 85 percent of both monolingual and bilingual general education first-graders in the matched pilot sample could usually demonstrate such basic skills as writing upper and lower-case letters or establishing left to right and top to bottom directionality on a printed page, according to their teachers (See Table 6). Reading skills which at least half the children could perform "most of the time" include: recognition of initial and final sounds and letters; identification of sight words; associating letters with their sounds; reading experience charts; and reading and understanding a variety of mathematical symbols. The communication skills least likely to be mastered were the more complex ones, such as writing simple stories with minimal assistance from adults, using texts to find answers to questions posed by adults, following written directions, and using contextual clues when coming upon unknown words.

Close to 15 percent of the students in the matched pilot sample could not perform at all one or more of the reading skills on the checklist. Many students could only "sometimes" demonstrate some skills beyond the

Table 6

Teacher Checklist Ratings for General Education Students in "Matched " Sample (N=4198)

Item	Not Yet %	Some-Times %	Most of the Time %
A) LISTENING SKILLS			
1. Listens to others reading aloud with interest and pleasure.	3 (6)*	26 (27)*	72 (67)*
2. Retells a simple story in sequence	7 (12)	32 (35)	61 (53)
3. Perceives the main idea of a story	8 (14)	33 (35)	58 (51)
4. Follows oral directions	5 (5)	31 (29)	64 (67)
5. Recognizes rhyming words aurally	7 (12)	29 (37)	64 (51)
B) SPEAKING SKILLS			
6. Looks at picture and demonstrates understanding of content	2 (5)	24 (27)	74 (67)
7. Relates own experiences, ideas, and feelings.	6 (10)	29 (34)	65 (56)
8. Ask questions.	10 (18)	37 (40	53 (42)
9. Reveals understanding through replies and reactions to questions.	6 (9)	31 (39)	63 (53)
10. Expresses thoughts clearly enough to be understood.	5 (8)	24 (33)	71 (60)
11. Predicts next probable event in sequence.	8 (16)	34 (37)	58 (47)
C) WRITING SKILLS			
12. Writes upper and lower-case letters.	2 (2)	12 (8)	87 (90)
13. Uses invented spelling.	16 (24)	37 (45)	48 (31)
14. Writes simple stories with minimal assistance from adults.	24 (35)	35 (43)	41 (22)

* Ratings on bilingual students in parentheses.

Table 6 (continued)

Item	Not Yet %	Some-Times %	Most of the Time %
D) READING SKILLS			
15. Distinguishes between realism and fantasy	2 (5)*	16 (21)*	82 (74)*
16. Establishes left to right and top to bottom directionality on a printed page.	2 (4)	13 (17)	85 (79)
17. Identifies sight words, print in the environment, and signs and labels.	5 (8)	25 (37)	69 (55)
18. Reads experience charts.	9 (13)	30 (37)	61 (50)
19. Reads and understands a variety of mathematical symbols, e.g., numerals, clocks, calendars.	4 (6)	27 (25)	69 (69)
20. Follows written directions.	16 (23)	37 (44)	47 (33)
21. Recognizes initial sounds and letters.	3 (6)	18 (25)	79 (69)
22. Recognizes final sounds and letters.	5 (9)	21 (29)	74 (62)
23. Associates letters of the alphabet with their sounds.	3 (6)	18 (29)	78 (64)
24. Reads aloud to and with others from books and own stories.	14 (18)	31 (38)	55 (44)
25. Sounds out words.	11 (20)	35 (35)	54 (45)
26. Uses contextual clues when coming upon unknown words.	19 (28)	41 (49)	39 (22)
27. Reads high-frequency words easily in any format or context.	14 (26)	33 (38)	55 (36)
28. Uses texts to find answers to questions posed by adults.	16 (39)	42 (44)	41 (17)
29. Makes inferences from materials read.	16 (28)	42 (48)	41 (24)
30. Recognizes the sound of different consonant clusters (e.g., bl. tr).	15 (25)	31 (35)	54 (40)

* Ratings on bilingual students in parentheses.

very basic ones, such as recognizing the sound of different consonant clusters or sounding out words. It is interesting to note that the bilingual children, who were rated on communication arts performance in their home language, generally performed similarly to monolingual children but at a somewhat lower level. Average ratings on each of the 30 communication arts skills for both monolingual and bilingual children in the pilot samples also appear in Table 6.

A total checklist score was created by summing up the ratings for each checklist item to arrive at a total score which could range from 30 (not yet able to demonstrate any skills) to 90 (able to demonstrate all skills most of the time). The results (see Table 7) show that the average checklist score for the monolingual general education children is quite high, i.e., 76 out of a maximum score of 90. Bilingual children had a slightly lower average rating of 71. While this suggests a high level of accomplishment, the large standard deviation also suggests that not all children

Table 7

Average Total Checklist Scores of
Monolingual (Regular Education and Resource Room)
and Bilingual Students

	N of Cases	Mean Score	Standard Deviation
Monolingual	4115	75.9	14.04
Regular education	4074	76.1*	13.9
Resource room	41	60.7*	15.2
Bilingual**	375	71.0	15.4

* $t = 7.01$, $p < .01$
** Rated on performance in native language.

were doing well. In fact, about 15 percent of the monolingual children had a total checklist score of 60 or less. Many of these lower scores were from children in resource rooms, who scored significantly lower than children in regular classrooms.

Early Childhood Experience

Children who have participated in early childhood education programs perform better on the checklist than those without such experience. As the data in Table 8 show, the average total checklist score is higher for children with both pre-kindergarten and kindergarten than for children with just kindergarten. However, both groups perform better than first graders who have not previously attended school.

Table 8

Checklist Scores for Children With
Different Amounts of Early Childhood Education

	N	Mean
Pre-Kindergarten and Kindergarten	439	78.9
Kindergarten Only	1,440	75.8
No Pre-School	235	71.6

$F = 22.9$, $p < .001$

Although differences between each of these three groups was statistically significant, a more dramatic difference (close to half a standard deviation) was found between the children with two years experience prior to first grade and the children with no formal educational experience before first grade.

When evaluating these findings, it is important to consider how much of the difference in scores is due to early childhood education versus other related factors, such as home environment. In this pilot test, we do not have data to answer this question. Other studies, however, emphasize the critical role of home environment.

Special Education Students

Teachers of special education classes for readiness-delayed learners completed checklists for 45 children. Since there were so few of these children in the pilot sample, any interpretation of their performance must be made with caution. Three-fourths of these children were able to write upper and lower-case letters. (See Table 9). The only other skills mastered by at least half of this group were: establishing directionality on a printed page; recognizing initial sounds; and associating letters of the alphabet with their sounds. Children were able to perform most of the skills "sometimes", a finding which is consistent with the classification of this group of children as readiness-delayed. The skills least likely to be achieved on any level were: writes simple stories with minimal assistance from adults; uses contextual clues when coming upon unknown words; uses texts to find answers to questions posed by adults; makes inferences from materials read; and recognizes the sound of different consonant clusters.

The mean total checklist score for readiness-delayed children in the sample was 62.5 (standard deviation = 16.0). Although, on the average, these children scored below the general education children, the mean score of 62.5 is comparable to that of resource room children. There was also

Table 9

Teacher Checklist Ratings for Special Education Students in "Matched" Sample (N=45)

Item	Not Yet %	Some-Times %	Most of the Time %
A) LISTENING SKILLS			
1. Listens to others reading aloud with interest and pleasure.	2	56	42
2. Retells a simple story in sequence	22	47	31
3. Perceives the main idea of a story	27	51	22
4. Follows oral directions	2	64	33
5. Recognizes rhyming words aurally	18	44	36
B) SPEAKING SKILLS			
6. Looks at picture and demonstrates understanding of content	4	58	38
7. Relates own experiences, ideas, and feelings.	11	47	42
8. Ask questions.	13	47	40
9. Reveals understanding through replies and reactions to questions.	9	60	31
10. Expresses thoughts clearly enough to be understood.	4	47	49
11. Predicts next probable event in sequence.	27	56	18
C) WRITING SKILLS			
12. Writes upper and lower-case letters.	0	[illegible]	73
13. Uses invented spelling.	49	18	33
14. Writes simple stories with minimal assistance from adults.	62	16	22

Table 9 (continued)

Item	Not Yet %	Some-Times %	Most of the Time %
D) READING SKILLS			
15. Distinguishes between realism and fantasy.	9	56	36
16. Establishes left to right and top to bottom directionality on a printed page.	7	36	58
17. Identifies sight words, print in the environment, and signs and labels.	27	31	42
18. Reads experience charts.	44	31	24
19. Reads and understands a variety of mathematical symbols, e.g., numerals, clocks, calendars.	18	47	36
20. Follows written directions.	42	36	22
21. Recognizes initial sounds and letters.	13	29	58
22. Recognizes final sounds and letters.	29	27	44
23. Associates letters of the alphabet with their sounds.	16	29	56
24. Reads aloud to and with others from books and own stories.	44	27	29
25. Sounds out words.	40	31	29
26. Uses contextual clues when coming upon unknown words.	56	27	18
27. Reads high-frequency words easily in any format or context.	40	44	16
28. Uses texts to find answers to questions posed by adults.	62	29	9
29. Makes inferences from materials read.	51	40	9
30. Recognizes the sound of different consonant clusters (e.g., bl. tr).	56	20	24

considerable variation in this group, i.e., some children have not yet mastered many skills and some have the skill level of children in regular classrooms. It is not surprising that some of these children would have performed as well as first graders in regular classes on this checklist assessment. Since this was a group of readiness-delayed learners, by the spring -- when this assessment took place -- a combination of instruction and maturational development could have led to grade-level performance.

Opinions on Checklist

A high proportion of teachers and administrators (see Table 10) returned questionnaires in which they expressed opinions on the checklist.

Table 10

Questionnaire Response Rate

	No. Sent Out	No. Returned	% Returned
Teachers	394	323	82
Administrators	216	154	71

For some questions, they were asked to choose among responses, e.g., the checklist was either "very useful", "moderately useful", "minimally useful" or "not at all useful". Other questions were open-ended, e.g.: "What do you see as the strengths of a checklist such as this?" Many teachers and administrators responded with detailed comments on the strengths and weaknesses of the checklist. Each comment was systematically categorized using a content analysis approach which classified statements with similar meaning into one category. Categories were developed and comments classified by two independent researchers to try to ensure a

reliable analysis. The number of statements in each category was added up to understand the degree of consensus on each of the the strengths and weaknesses of the checklist identified by respondents.

The survey responses provided important information, particularly from the teachers who were using the checklist for the first time. Quotes from teachers and administrators are included below to more clearly present the reactions of participants in the pilot sample.

Appropriateness for New York City curriculum and students.

Both administrators and teachers overwhelmingly responded "yes", the checklist "adequately covers the skills in the New York City first-grade communication arts curriculum." Indeed, one of the major strengths of the checklist cited by administrators (n = 53) and teachers (n = 62) is that it provides a comprehensive listing of communication arts skills to be taught in the first grade. This listing "helps to reinforce teacher objectives at the beginning of the year" and provides a clear guide as to the skills first graders should master. As one teacher states, "It crystalizes for the teacher those skills which are minimally essential for success by a first grader." A few administrators (n = 8) suggested that using the checklist as a curriculum guide was particularly valuable for new teachers.

On the other hand, a number of administrators (n = 28) and teachers (n = 59) in their open-ended comments expressed concern that the checklist is too general or does not include enough of the skills that should be mastered in first grade, such as word families, vowel diagraphs, blending

skills, sentence structure, and comprehension, etc. The sense of their comments was that the checklist may be used as part of a larger assessment process which takes into account the wider range of communication arts skills taught in first grade as well as individual student characteristics that affect reading performance. In other words, "the checklist as it is, is not comprehensive enough to be the only assessment" of the reading skills of first-grade students.

The vast majority of both administrators and teachers (over 80 percent) believe that the difficulty levels of the skills on the checklist adequately reflect the difficulty level of the first-grade curriculum. A small number caution that the checklist items may be too difficult and thus not reflect growth in children without kindergarten experience or those who are developmentally below level.

Format issues.

Virtually all administrators and teachers agreed with the statements that items are clearly defined on the Teacher's Guide and that directions for completing the checklist are understandable. Interestingly, comparatively few adminstrators (n = 2) and teachers (n = 11) commented in the open-ended section that "ease of use" was a strength of the checklist.

A number of concerns about the format were raised in comments about perceptions of weaknesses of the checklist. One concern expressed by 18 administrators and 51 teachers was that the response options of "not yet", "sometimes", and "most of the time" serve to "limit the person completing it in the range and quality of their response". Thus "a very wide range of children could rate all 3's on the checklist", "'Sometimes' can mean

once or twice or 85 percent of the time", and also, "in some cases, a child's ability can rest between categories." Furthermore, "the terms refer to how often a child performs a skill and not how well or poorly." Clearly, the rating scale requires serious review prior to any further use of the checklist.

A second issue raised about the checklist approach is the perception of inherent subjectivity of teacher ratings. This problem was identified by both administrators (n = 17) and teachers (n = 18) since "each teacher had different standards", "there is a tendency to rate children in relation to others in the class rather than to a universal standard" and some teachers may be biased by student personality factors. In sum, the checklist results are "only as accurate as the person who is doing it."

Use of results.

When asked to rate how useful the checklist results would be for instructional planning, close to half of all administrators and teachers responded "moderately useful". A somewhat higher proportion of administrators (42 percent) than teachers (28 percent) rated the results as "very useful" or, conversely, more teachers thought the results would be minimally useful. Thus, the overall reaction to the checklist results was favorable though somewhat more so from the perspective of administrators.

Teachers were asked to provide more specific information about how they might use the checklist results. In response to the options provided, the following proportion of teachers said they would use the results for: assessing children's progress (75 percent); planning individualized activities (69 percent); grouping (67 percent); instructional purposes (60 percent); and curriculum planning (49 percent).

Responses to the open-ended question about the checklist's strengths are consistent with the ratings above. The major strength of the checklist from the point of view of both administrators (n = 69) and teachers (n = 168) is that the results are useful for evaluating and assessing students' strengths and weaknesses and, hence, needs and progress during the school year. The following comments are typical: "The pupil's abilities and difficulties come into sharper focus as an individual", "It would help to identify the strengths and weaknesses of the child.", and "It can be used to measure the children's progress."

In addition, a number of people used words like "it forces the teacher to take the time to think about an individual child's abilities in each skill area" or it helps the teacher to "zero in", "pinpoint" or "focus" on specific student strengths and weaknesses. In other words, use of the checklist supplements the teachers' ongoing student assessments and encourages an individualized and defined evaluation process.

The second most frequently mentioned strength of the checklist is that it is useful as a guide for classroom or individualized instruction. Although obviously closely related to the comments above on assessment, some people clearly emphasized the use of checklist results for instructional guidance and planning. For example, "it allows one to see the holes in one's instructional program", "As a teacher of 33 students, using a checklist of this kind in the fall would enable me to plan for grouping, individualized instruction, and curriculum planning", and "Helps me to better organize my instructional program". Checklist results are also useful for grouping students. Teachers (n = 34) are "able to categorize the children with certain weaknesses and work with them in groups."

In all, half the teachers and administrators felt the checklist results would be "moderately useful" for overall assessment of first graders' communication arts skills. Other administrators (42 percent) and teachers (32 percent) felt the results would be "very useful" and relatively few administrators (8 percent) though somewhat more teachers (17 percent) thought the results would be "minimally" or "not at all" useful. Fifteen teachers' comments suggested that the checklist is "unnecessary", since "every teacher already knows the children's strengths and weaknesses."

An issue that was raised largely by administrators (n = 18) was the possibility that using the checklist might have negative effects on teaching. One concern was that use of the checklist might "stifle" or "limit" teachers' creativity and "restrict" them to "teaching only those items on the checklist". They feared the checklist "may become the only sanctioned criteria, thereby locking staff into a particular mold." Only three teachers expressed similar concerns.

Issues in administration.

Teachers and administrators were asked whether fall, midyear, or spring would be the best time of year to administer the checklist. Almost half of the respondents indicated midyear and comments suggested this was when the checklist results could best serve as an assessment of progress and provide a guideline for remediation. About 20 percent of both administrators and teachers felt the checklist should be administered in the fall. Comments in the open-ended section revealed that a fall administration is viewed as best for early diagnosis and appropriate grouping. Another approximately 17 percent selected a spring administration and

comments indicated this was because the checklist would be useful in evaluating end-of-year progress as well as in placement for the following year. Over ten percent checked off more than one response, indicating assessments should take place more than once a year. This would enable student progress to be judged and appropriate instructional activities planned.

Another important administration issue is the time it takes teachers to complete the checklist and the reaction to adding this task to teachers' responsibilities. The number of checklists completed by the teachers in this pilot study varied considerably from a low of "1 to 5" completed by five percent of teachers to a high of "more than 25", completed by 21 percent of the teachers. The time reported to complete the checklists varied concomitantly from less than an hour reported by 16 percent of the teachers to at least three hours for 18 percent of the teachers. A positive relationship (Spearman correlation = .45, $p < .01$) was found between the number of checklists completed and the time spent in completing checklists. It is interesting to note that although about half the sample of teachers (56 percent) filled out at least 16 checklists, comparatively few (18 percent) spent more than three hours working on them. An inference could be made that it would take most teachers about 10 minutes to complete one checklist. For a class with 20 children, this could mean the teacher spending over three hours to complete checklists. The vast majority of teachers (86 percent) agreed with the statement that they would indeed need additional time to complete checklists for every child in their class.

In describing weaknesses of the checklist, 21 administrators and 42 teachers stated that completing the checklist is time-consuming. A couple of administrators added the proviso "but its value far outweighs its weaknesses". However, the teachers expressed greater concern about the additional burden completing checklists placed upon them. They felt that "for completion of a checklist such as this, ample time must be given to the teacher in order to make a fair and objective assessment for each child." Other concerns expressed were that checklist completion is "just more paperwork", "I do not think that evaluations must always be written.", "Taking time out to assess means taking time away from other meaningful activities." and, "The teacher knows all this already."

MAT FINDINGS

Description of the MAT

The reading achievement test administered, the Primary I level of the MAT6 (Form L) is made up of three subtests with a total of 103 items: Vocabulary (22 items), Word Recognition (28 items), and Reading Comprehension (53 items). Most test items are at the primer and first-grade level of difficulty although within the reading comprehension section, items increase in difficulty to grade-three level.

Test Administration Procedures

Practice tests were made available to schools prior to testing to help children become more familiar with the test format and the types of questions, and to give them practice in marking their answers in the test booklets. The MAT was administered to first graders during the same two-day period as the citywide reading tests for other grades, i.e., April 21 and 22, 1986. The Vocabulary and Word Recognition sections of the test were administered on April 21 and took a total of 35 minutes, excluding time for test distribution, collection, preparation of the answer document, and sample questions. The Reading Comprehension subtest was given the next day and the working time for that test was 43 minutes. The Directions for Administering the MAT recommend that no more than one subtest be administered in a single sitting and that no more than two sittings be given during any half-day.

Total Test Results

The mean raw score (or number of items correct) on the total test (103 items) for the New York City sample was 52.7, 2.7 points lower than that for the national norm group, as Table 11 shows.

Table 11

Total Reading Test Results:
New York City* and National Samples

	Mean Raw Score	Standard Deviation	Median Raw Score
New York City Sample	52.7	20.8	48
National Sample	55.4	21.4	56

* This sample includes only general education children.

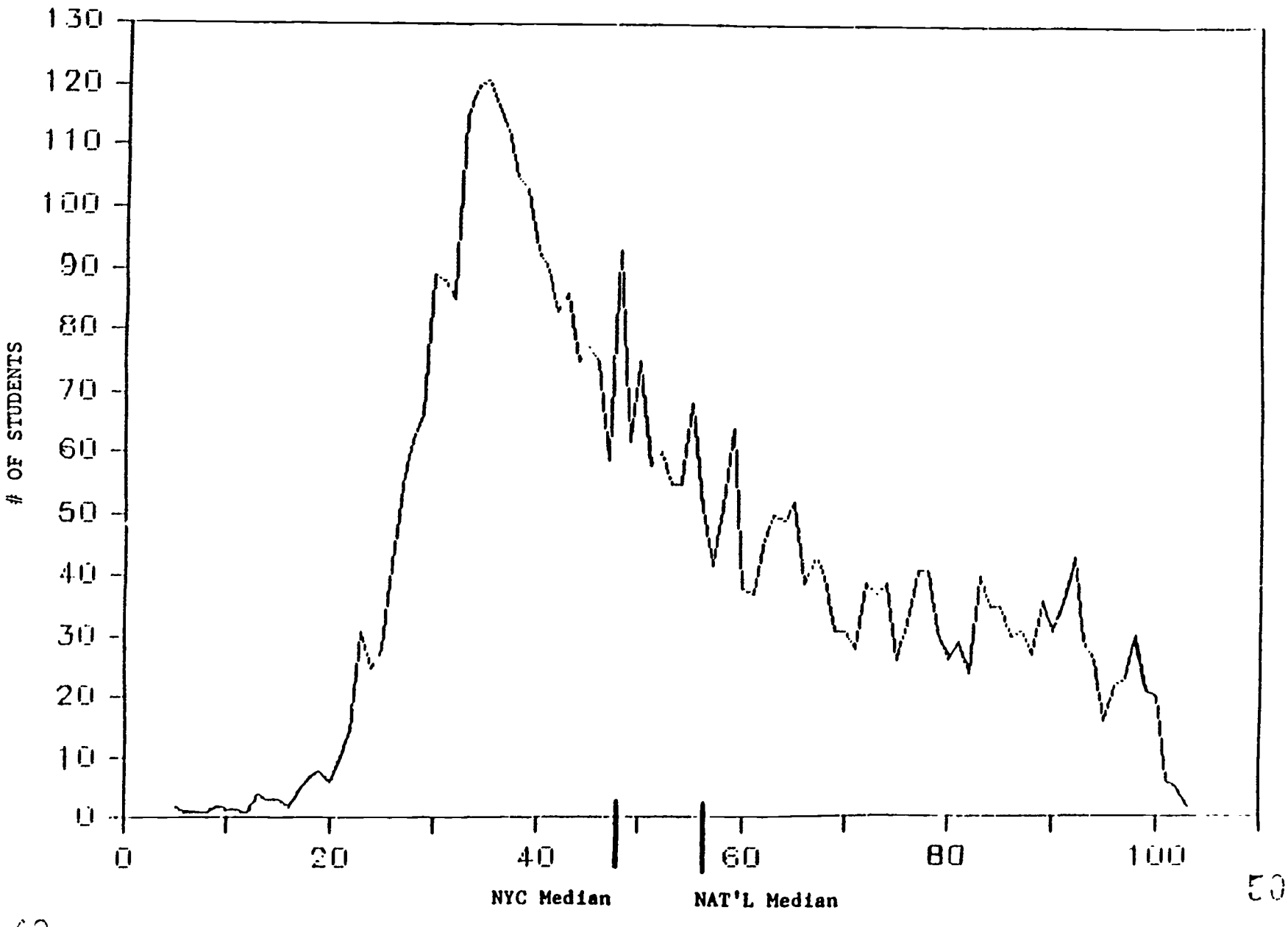
The median raw score however, that is, the middle point in the distribution of all scores, was eight points lower for the New York City sample than for the national sample. This means that the distribution of scores for the New York City sample differed from that of the national sample. The implications of this difference become clearer when a graph of the New York City distribution is analyzed.

Figure 1 graphically shows the frequency distribution of raw scores for the New York City sample. It shows few children with raw scores below 20, a large cluster of students with raw scores in the 30's and 40's and a slowly decreasing number of students obtaining raw scores of 50 and above. The mean score is higher than the median because it is influenced by the unexpectedly high number of first graders who did very well on the test.

The frequency distribution for the national group is not available. However, a comparison of the New York City and national samples' median scores on the distribution in Figure 1 illustrates that a greater proportion

FIGURE I

TOTAL RAW SCORE ON MAT6 TOTAL READING TEST



of the national norms sample had higher raw scores than the New York City sample. Note, however, that the New York City and national norms samples have mean scores that are almost identical. This implies that there was also a greater proportion of students in New York at the higher levels of reading achievement than in the national norms group: given the group of lower test scores in New York City a greater number of high test scores than in the national group would be necessary to raise the New York City test score mean to the level of the national mean. In sum, the graph suggests: (1) a large group within the New York City sample of first graders read somewhat below the national average; (2) a larger than expected group of New York City students in the sample read at the higher levels. These and all other test results must be tempered by the probability that the pilot sample performance is less than that of all first graders in New York City.

The graph also depicts the wide variation in scores for the New York City sample: there were sizable numbers of students getting each of the raw scores from 20 to 99. This considerable variation in scores among students manifested itself in the high standard deviation (20.8). That standard deviation implies that close to two-thirds of New York City students had scores between 32 and 73. This wide range in scores is not surprising given the large developmental differences in young children as well as the strong influence of varying home environments at this age. A very similar level of variability was found in the national norm group (standard deviation = 21.4).

Another way to compare the New York City sample with the norms sample shows that the mean performance of the New York City sample was better than 45 percent of individuals in the norm sample. The median performance of New York City children was better than 35 percent of the children in the national sample.

A little more than one-third of New York City general education first graders in our sample (36.9 percent) were reading at or above grade level, i.e., performing at or above the 50th percentile (see Table 12).

Table 12

Quartile Distributions:
New York City and National Norms Samples

Quartile	First (1-24)	Second (25-49)	Third (50-74)	Fourth (75-99)
National Norm Sample	25%	25%	25%	25%
New York City Sample	38.5%	24.6%	15.2%	21.7%

A disproportionate number of children (38.5 percent) were reading in the bottom quartile, i.e., the level at which the lowest 25 percent of the national sample are reading, and fewer New York City students than students in the national sample performed in the top two quartiles. However, when the top quartile was analyzed more closely, it became clear that there was also a larger than expected group of children with very high reading scores. As Table 13 shows, more New York City students scored in the top decile (90-99), in the top five percentiles (95-99), and in the top percentile (99) than students in the national sample.

Table 13

Performance in Top Decile:
New York City and National Norm Sample

	90-99	95-99	99
National Norm Sample	10%	5%	1%
New York City Sample	13.0%	8.9%	4.8%

When one compares these results to those of second graders, the proportion of students reading at and above grade level are about the same. The proportion of second graders reading at or above grade level as of spring, 1986 was 42.3 percent. This figure includes 6.9 percent of second graders who were limited-English-proficient (LEP) and, hence, assumed to score below grade level. The proportion of first graders reading at or above grade level in the pilot sample was 36.9. However, if all first graders were tested, the true percent reading at or above grade level would be higher, i.e., more like the scores of second graders.

Influence of early childhood education on reading achievement. One of the reasons for testing first graders was to provide information on the reading achievement of children with different amounts of early childhood education experience. Information on amount of this experience was available for about half of the pilot "matched" sample (n=2,153). Using this smaller sample, comparisons were made among the mean reading scores*

* For the purpose of statistical analyses, raw scores were converted to scaled scores, which provide an equal interval scale.

for three groups of first-grade children: those with pre-kindergarten and kindergarten, those with kindergarten only, and those with no early childhood experience at all.

Table 14 shows that children who had early childhood education demonstrated higher reading achievement than children without such experience. Children who had both pre-kindergarten and kindergarten scored significantly higher than those who had either kindergarten only or no formal experience at all. Even children with kindergarten experience only performed better than those with no early childhood education at all. It is interesting to note that most children for whom we had information had some early childhood education experience.

Table 14

Mean Scaled Scores: Reading Achievement
of Children with Differing Early Childhood Education Experiences
(n = 2,153)*

	Pre-K and K	K Only	No Pre-School
N of children	442	1,472	239
Mean scaled score	521.6**	500.5**	490.9**

F=57.82, p < .001

* Regular and Resource Room only.

The above results seem to support the importance of early educational experience in improving first-grade reading achievement. It is necessary to consider, however, that this analysis does not include information on other possible causes for higher test scores among these children. For example, research has shown that home environment is a critical factor in student achievement. It is possible that children who had the most early educational experience also had the most supportive home environment. In that case, attributing higher scores to educational experience per se is an inaccurate interpretation of the data.

To better understand how early educational experience affected reading test scores, a second analysis was conducted. The reading achievement score was correlated with the amount of experience (coded as "2" for two year's experience, "1" for one year of experience, and "0" for no experience). The results (Spearman $r = .20$, $p < .05$) reveal that previous school experience accounts for only a small portion of the variance (.04) in the test scores. This lends support to the hypothesis that significant differences in mean test scores among children with different educational experiences may be due to other factors which are related to pre-school participation, e.g., home environment. It is also possible that effects of early education experience would be confounded by a year's worth of first-grade instruction.

Influence of gender on reading achievement. Girls in the first-grade sample had higher average reading scores than boys. This finding is consistent with a large body of research which suggests that girls at this age are developmentally more mature and better able to read than boys.

Table 15

Comparison of MAT Scores for Boys and Girls in the Pilot Sample

	Number of Children	Mean Raw Score	Standard Deviation	Median Raw Score
Boys	2,078	50.9*	20.2	46
Girls	2,087	54.6*	21.2	50

* $t = 5.85$, $p < .001$

This difference is also reflected in the proportion of girls reading at or above grade level (40.9 percent) as compared to boys (32.7).

MAT Subtest Results

The MAT is made up of three subtests: Vocabulary, Word Recognition, and Reading Comprehension. The distribution of scores in these subtests does not always mirror the total test score distribution. Further, the content and level of the subtests are different. Thus, it makes sense to examine the subtest results separately.

Vocabulary subtest. The vocabulary subtest is made up of 22 items that "measure the meaning of words in context." All items are read by the child and require the student to fill in the missing word in a sentence. This subtest is essentially at the primer and grade 1 reading level.

The mean raw score for the Vocabulary Test for New York City first-grade students in regular and resource rooms is slightly lower than the national average. However, as was the case with the total raw score, the

median for the New York sample is clearly lower than the national sample score. The level of overall vocabulary achievement for New York City first graders is reflected in the finding that 37 percent of first graders have vocabulary skills at or above grade level, i.e., at or above the 50th percentile.

Table 16

Vocabulary Raw Scores: New York City* and National Samples

	Mean Raw Score	Standard Deviation	Median Raw Score
New York City sample	11.1	6.2	9
National sample	11.8	6.5	13

*This sample includes only general education children.

A graph of the distribution of Vocabulary subtest scores shows a very interesting pattern of scores. Although half of the sample received scores of nine or below, there is a group of about 500 students (or 12 percent of the sample) who scored perfect or almost perfect scores on the vocabulary subtest. The extremely high scores of this group of students raises the New York City mean score so it is close to the national mean. However, for the New York City sample, the subtest data show that in vocabulary achievement, half the students performed below the national sample's mean score but a smaller group strongly outperformed the national sample. The data also suggest that the test was not hard enough for these top students, i.e., it did not include enough difficult items to adequately measure their vocabulary level.

Word recognition subtest. The word recognition subtest contains 28 items that "measure phoneme/ grapheme; consonants, phoneme/grapheme; vowels, and word part clues." This subtest is a combination of teacher-dictated and printed items. For the first ten items, the child is given a picture and a list of four words and asked to choose the word that begins with (or ends with or includes, depending on the item) the same sound(s) as the picture. For each of these items, the teacher says aloud what the picture is. For the next ten items, a sound in a word is underlined and the child must choose, from a list of four words, the word that has the underlined sound. This section is not read by the teacher. The last eight items asks the child to read and complete sentences by choosing the correct word from a list of four words.

The mean raw score of New York City children on word recognition was exactly one point less than the national sample mean raw score. Although not quite as dramatic as the vocabulary findings, the distribution of scores does reveal that few students performed very poorly on the test, the majority perform at or somewhat below average and a subgroup of about 400 scored quite high. As above, this is reflected in a median score for New York City children which is lower than their mean score and lower than the national median score. The proportion of first graders reading at or above grade level on this subtest was 38.8 percent.

Table 17

Word Recognition Raw Score:
New York City* and National Samples

	Mean Raw Score	Standard Deviation	Median Raw Score
New York City sample	15.2	6.5	14
National sample	16.2	6.2	17

*This sample includes only general education children.

Reading comprehension subtest. The Reading Comprehension subtest contains 53 items measuring comprehension of rebus (4 items), sentences (4 items), and passages (45 items). The reading level for the nine passages begins at primer level and increases in difficulty to third grade level. The 45 passage-related items are designed to assess the child's ability to "recognize detail and sequence; infer meaning, cause and effect, main idea, and character analysis, and draw conclusions."

Out of 53 items on this section of the MAT, the mean scores and standard deviation for New York City children and for the national sample were very similar. As with the other two subtests, the median score for reading comprehension was lower than the mean score and than the national sample median. Unlike the other two subtests, the Reading comprehension scores were more normally distributed. This likely occurred because there were enough difficult items (this subtest included items up to third grade difficulty) to spread out the distribution of scores. The percentage of children reading at or above grade level in this subtest was 38.3 percent.

Table 18

Reading Comprehension Raw Scores:
New York City* and National Samples

	Mean Raw Score	Standard Deviation	Median Raw Score
New York City sample	26.5	9.8	24
National sample	27.0	10.3	28

*This sample includes only general education children.

Relationship Between the MAT and Checklist

The MAT and the checklist are two different kinds of measures, each assessing different aspects of communication arts abilities and each using different assessment approaches. Thus, a significant correlation between scores on these two measures implies that knowledge about a child's score on one helps to predict the other but does not necessarily mean that the two are measuring the same skills. A strong and positive relationship (Pearson correlation = .58) between children's reading achievement on the standardized test (total score on MAT) and teacher observations of communication arts skills recorded on the checklist (total checklist score) was found. In other words, children who do well on the MAT are also likely to be rated highly by their classroom teacher.

Although the total test and total checklist scores are based on different kinds of items, there are selected individual items within each measure which seemed to be assessing the same concept. For example, on the MAT word recognition subtest, there were five items in which the teacher said the name of a pictured object and the child chose one of four

given words (printed next to a picture of that object) that began with the same sound(s). On the checklist, item 21 asks the teacher to judge whether the child "recognizes initial sounds and letters". Although the MAT is a direct measure of child performance and the checklist depends on teacher assessment, the concept being measured is similar. Thus, a series of correlations were computed to determine the relationship between individual items on the MAT and on the checklist which were thought to be measuring similar behaviors. The items chosen for this analysis were initially selected by O.E.A. and subsequently reviewed by and chosen in conjunction with the Early Childhood Unit of C. & I.

Correlations between individual items ranged from .11 to .36; all were statistically significant. It is not surprising that item correlations were lower than total score correlations due to the nature of the correlation statistic. What these findings suggest is that there is indeed a relationship between the two ways of assessing students' performance on similar communication arts tasks, but the relationship is far from perfect.

MAT Results for Special Education Children

Based on individual needs, modifications to testing procedures were made for children in special education classes. Modifications, which are permitted when they appear on a student's IEP, included: time limit extended or waived; examination administered in special location; questions read aloud; and answers recorded in any manner. However, the test (MAT6, Form L) given to special education children was the same as that administered to the rest of the first-grade children.

There were 45 readiness-delayed students for whom both MAT and checklist information was available. Table 16 shows that both the mean and median raw scores for these children were about 15 points lower than those for the general-education children in the pilot sample. The distribution of test scores shows that most children obtained scores between 26 and 35, out of a possible raw score total of 103. The highest score obtained in this sample was 61. It is interesting that in this group of 45 children classified as developmentally delayed, there were four children who scored above the national norm and ten who scored above the New York City median. Thus, by the time the MAT was administered in the spring, close to one-fourth of these children were performing at a a level comparable with general-education first graders in the New York City pilot sample. However, considering the whole group of readiness-delayed learners in the sample, only 8.9 percent of the MIS IV first graders were reading at or above grade level.

Table 19

MAT Test Scores: Special Education Children Compared with General Education Children*

	Mean Raw Score	Standard Deviation	Median Raw Score
Special Education (n = 45)	38.0	11.0	35
General Education (n = 4,189)	52.7	20.9	48

*New York City pilot sample

There was less variation among the scores of children in the MIS IV special education classes when compared to the general education or the national norm sample. This makes sense since children were purposefully grouped to make homogeneous instructional groups. The standard deviation of 11 for this group of children means that about two-thirds of this pilot sample had total MAT scores between 27 and 49 (as compared to a range of between 32 and 74 for two-thirds of the general education sample.)

As with the general education sample, results for each of the three subtests were separately examined to judge whether students' performance varied from one category of reading achievement to another. Out of a possible 22 points on the vocabulary subtest, the mean score for the special education children in the sample was 7.3 and the median was 6. These average scores are about three points lower than the general education students in the New York City sample. It is particularly interesting to note that three MIS IV children had vocabulary scores that were at the level of the national median score (raw score = 13) and three others had very high scores. However, the majority did perform poorly compared to the national norm group: only 13 percent were reading at or above grade level.

There were three children in this sample of special education children who scored at or above the national mean raw score of 16 out of a possible 28 points in word recognition. However the average score of readiness-delayed learners on this subtest was 9.8, well below the national mean and the New York City mean (15.2). The percent of children in this special education sample whose word recognition skills were at or above grade level was 4.4 percent.

The Reading Comprehension subtest included the most difficult items on the test, some of which were at the third-grade level. The mean raw score for the national and the New York City pilot sample was about 27 of the 53 items correct. Interestingly, there were eight children in the readiness-delayed pilot group who scored 27 or higher on this subsection. The mean score of 20.9 for the entire group of special education children was six points lower than the national and New York City sample average. The proportion of the special education sample who had a reading comprehension performance at or above grade level was 15.6 percent.

Although it is inappropriate to make generalizations based on one sample of 45 students, it is interesting to observe that reading performance of this sample of students was best in the area of comprehension and poorest in word recognition.

Opinions on MAT

Responding to the same questionnaire which asked for opinions on the checklist, 323 teachers and 154 administrators gave their reactions to the MAT. Some of the questions were closed-option, such as "Was the difficulty level of the test 1) too easy, 2) just right, or 3) too difficult?" Other questions, such as "What do you see as the strengths of a test such as this?", allowed for open-ended responses. As was done for comments on the checklist, each comment was systematically categorized using a content analysis approach which classified statements with similar meaning into one category. Then, the number of statements in each category was added up to understand the degree of consensus on each of the strengths and weaknesses of the MAT identified by teachers and administrators.

Survey responses provided important information from administrators and teachers who were testing first-grade children with the MAT for the first time. Their attitudes toward testing first graders with the MAT have implications for future testing of children in this age group. Quotes from teachers and administrators are included below to more clearly illustrate the reactions of participants in the pilot sample.

Appropriateness for New York City curriculum and students. Almost half of the teachers (47 percent) and more than half of the administrators (57 percent) responded "yes" to the question "Does the test adequately reflect the New York City communication arts curriculum?" Open-ended comments reflect this almost even division of opinion as to how well the MAT represents the curriculum. Some teachers thought the MAT "covers what is taught throughout the school year" and that "each of the reading skills are adequately covered." An administrator agreed that "the MAT contains and tests all reading skills for which first graders should be held responsible." For those who feel the MAT is not an adequate test of what was taught, concerns range from specific comments -- "The test given is not valid in view of our phonics oriented program (Lippincott)" -- to more generalized comments -- "It didn't evaluate many of the things which were taught in the first grade."

Other feelings about how appropriate the test is for New York City children were obtained in response to the open-ended questions about the MAT's weaknesses. The responses reveal concerns about cultural bias in the test items. Administrators (n=10) and teachers (n=25) thought "the subject matter and much of the vocabulary...were inappropriate fc inner-

city children." An even stronger issue was the appropriateness of testing first-grade children. "The average first-grade student is not developmentally mature enough to attend to the same task (such as reading stories and comprehension questions) for 35 minutes," said one of the 38 teachers and 27 administrators who commented that children this age should not be given standardized tests. Some teachers (n=25) specifically questioned the appropriateness of the MAT for low-ability children. Their perspective is illustrated by the following comments: "It did not accurately test the abilities of children in the lower third of the grade" or "did not adequately test the abilities of a child who is just beginning to read".

One of the survey questions designed to examine the appropriateness of the MAT for first graders asked teachers and administrators to classify the difficulty of the MAT: virtually no one selected "too easy"; about ten percent said "just right", and most (close to 90 percent) chose "too difficult." Indeed, the overall difficulty of the test was mentioned as a weakness of the MAT by both administrators (n=38) and teachers (n=81). Typical statements were that the test was "too difficult" or "much too difficult" or "too difficult fo the average first grader." There were additional comments that identified subtests as being especially difficult. The comprehension section was criticized the most: "There were too many stories. The children were restless and didn't attempt to do their best" or "The comprehension difficulty threatens the children" or "The reading comprehension passages are high above first-grade reading ability."

The latter comment is accurate to some extent. The MAT Directions for Administering state that the comprehension section includes passages that

are of third-grade difficulty and tells teachers to say to children before they take the comprehension section "you may not be able to read all of them, just do your best." While the Directions also ask teachers to encourage pupils even though "Some of the pupils may become discouraged", the teachers' comments suggest that some children moved beyond discouragement to frustration.

Administrators (n=33) and teachers (n=58) also expressed concern that the length of the MAT was inappropriate for New York City first-grade students. The most frequent comment simply was that "the test was too long" for children to sit through. One teacher said "the test was so long that I believe it was more of a test of endurance than reading".

The length and difficulty of the test, given the age of the students, contributed to the most common criticism of the MAT, that it was a stressful and frustrating experience for the children. Many teachers (n=89) and administrators (n=38) made strong statements about how stressful the experience of taking the MAT was. Comments included "It made most of the students who took it highly anxious and frustrated," "many students became frustrated and cried", "children became frustrated just looking at the passage and did not take time to read carefully", "The frustration level surfaced very quickly in my class of slow learners...they either put their heads down to cry or just filled in any circle." and "The comprehension part of the test frustrated the children who are still struggling with decoding words and who lost the aim of this portion of the test." The frequency and strength of these comments considered in conjuntion with other statements made about the test's length, difficulty, and relevance for this population raise concerns.

On a more positive note, some administrators (n=6) and teachers (n=19) suggested that the MAT may be an appropriate test for identifying or assessing gifted children.

MAT format. The majority of teachers* felt that MAT test items were clear (75 percent) and that directions were understandable (77 percent). Almost all teachers (96 percent) agreed that directions to the teacher were understandable. Responses to the open-ended questions reveal why one-fourth of teachers felt directions to the children were not understandable: directions were too wordy, there were too many examples and too many changes of directions.

Other criticisms of the test format arose (n=34) in response to questions about weaknesses of the MAT. One concern was that "there was no progression in complexity", i.e., the test should have started with easier items and gradually increased in difficulty. The layout of test items was also perceived to be a problem. For example, some sections ended in the middle of the page and one teacher thought the column layout was confusing. Another format problem identified by a number of teachers was the small print size, including the STOP signs designed to signal the end of each subtest section. Finally, a number of teachers did not feel it was appropriate or necessary to give first graders a timed test.

Use of results. When asked to rate how useful MAT results would be in instructional planning, only eight percent of teachers indicated it would be "very useful" and 30 percent said it would be "moderately useful." The

---

* Administrators were not asked these questions.

majority thought the MAT results would be "minimally useful" (39 percent) or "not at all useful" (23 percent). Administrators viewed the test results somewhat more favorably: ten percent thought the MAT results would be "very useful" for instructional planning, 43 percent thought results would be "moderately useful" and only 12 percent indicated "not at all useful". Some teachers (n=22) and administrators (n=25) did specifically comment that the MAT results would be useful, e.g., "to guide the teacher in planning an instructional program if used correctly". There was clearly a range of opinion on the instructional value of the MAT results though relatively few were very enthusiastic about this use of test results.

Responses to the specific questions on the usefulness of test results for overall assessment of first graders' communication arts skills were very similar to those discussed above. Comparatively few respondents felt MAT results would be "very useful" for this purpose and most felt results would be moderately or minimally useful. However, responses to the question asking for strengths of the MAT suggested some interest by administrators (n=51) and teachers (n=71) for evaluating students' achievement and, in particular, strengths, weaknesses, and needs. Teachers' comments included: "It could be an objective measure of the skills the children have been taught", "The only strength would be to aid teachers in determining in what areas the children need the most help," and "The test indicates how well the first grader reads and knows his skills." Administrators also believed the results could serve a diagnostic purpose as well as provide an objective evaluation of a child's ability. Some respondents felt the MAT might serve as a tool to identify or assess gifted children.

Teachers (n=36) and administrators (n=19) suggested that administering the MAT in grade one was helpful in introducing formal testing procedures to children and gives them "practice with test-taking techniques." A few teachers emphasized that "test-taking skills are a necessary tool in our society." and suggested the MAT's "greatest strength is future preparation for standardized testing in the second grade."

Another strength of the MAT testing program was that results could be used for peer comparisons, e.g., "Because it is a standardized test you are able to compare scores of children throughout the district." As one of the administrators remarked, "As a supervisor, I am interested in how well my students perform on a national basis."

Although a number of uses for the MAT results were suggested, there were 60 teachers who emphatically believed there were no values in testing with the MAT. Reactions range from "I don't see any strength in a test such as this as it does not recognize the developmental levels of a first grade child" to a succinct "The test had no strengths." Fewer administrators (n=15) had a negative view of the value of the MAT.

Others expressed concern that results would be of limited value because they were not a true reflection of a child' ability. For example, one teacher remarked "Results can be deceiving. A few of my best readers didn't finish because they worked too slowly. Their scores will surely be deceptive." Others were sure that children were guessing and thus test scores would present an inaccurate and inflated picture of reading achievement. There was also the concern that the MAT "does not truly measure the 'real' progress many first graders have made. We have so many youngsters

coming to school severely lacking in skills. Over the years these youngsters have made great strides in reading. The MAT does not measure this."

Administration of the MAT. Test administration procedures ran smoothly and no major problems were identified. Most concerns expressed had to do with the length and difficulty of the MAT and resultant student stress and frustration. A few administrators noted there were a large number of absences from the test due either to illness or giving the test on a Monday. Other than that, the administration process itself seemed relatively uneventful.

ADMINISTRATOR AND TEACHER OPINIONS ON OVERALL FIRST-GRADE ASSESSMENT PROGRAM

Most administrators and teachers felt the checklist should be a part of an overall first-grade assessment program, either in combination with a standardized test or alone. Almost no one recommended assessment of first graders using a standardized reading test alone. A greater proportion of administrators than teachers thought there was a role for standardized testing in a first-grade assessment program, albeit not as the sole measure. Some teachers (13 percent) did not feel either assessment approach was appropriate.

Each administrator and teacher was given an opportunity at the end of the survey questionnaire to offer comments or recommendations regarding a citywide first-grade assessment program and what it should include. The most frequent comment was a criticism of the MAT. Many of the suggestions elaborated upon choices reported above, i.e., use the checklist either in combination with a standardized test or by itself. One teacher endorsed

"a short checklist two to three times a year regarding what skills should have been completed by a certain time. These checklists should be citywide and be used as a standard for all first-grade teachers." Others agreed that "The communication arts checklist seems like the right tool for assessing first graders. Teachers and parents would have an excellent basis upon which to assess each child's needs." Those who supported the combined use of the checklist and a standardized test typically recommended a test other than the MAT. Examples of alternative tests included: "test from basal reading programs, e.g., Houghton Mifflin"; or a shorter standardized test given under more relaxed conditions. Some administrators and teachers did not discuss the possible use of the checklist but did comment on the need for an alternative to the MAT. A number of teachers did not feel that first graders should be tested on a citywide basis and that reading series tests provide a better assessment of what is learned.

Table 20

Administrator and Teacher Opinions on Approach to First-Grade Assessment

	Administrators %	Teachers %
Checklist alone	33	40
Test alone	4	3
Checklist and test	60	44
Neither checklist nor test	3	13

IV. CONCLUSIONS

The pilot test provides important information to be used in deciding the most appropriate means of assessing first-grade students' reading achievement. Student performance data, as measured by a standardized test and a teacher observation checklist, revealed the range of abilities in New York City students and suggested strengths and weaknesses of children's reading performance and also of the measures of their performance. Teacher and administrator reactions to the pilot test and their ideas for future assessment approaches suggested ways to assess first graders that they felt would be both fair and informative.

All findings on student performance must be tempered by the fact the pilot sample probably under-represents the achievement of all first graders in New York City. If all first graders were to be tested, it is likely test and checklist scores would be higher. Nevertheless, the results are still valuable and led to the following conclusions.

STUDENT PERFORMANCE

- It is likely that the level of reading achievement of first graders is slightly less than but comparable to that of second graders in New York City, i.e., 42.3 percent of second graders in the spring of 1986 were reading at or above grade level.
- There is wide variation in reading achievement test scores among children in the first grade. For example, although there is a larger group of children in the pilot sample who performed below the national average, there was also a larger than expected group of children whose performance was much higher than the national average.

- Children performed at roughly the same level for each of the three reading subtests: vocabulary, word recognition, and reading comprehension.
- According to checklist findings, over two-thirds of the children could perform "most of the time" basic reading skills such as recognizing initial and final sounds and letters, identification of sight words and associating letter with sound. Fewer than half the children (40 percent) could routinely perform the more complex skills, such as using contextual clues when coming upon unknown words, or making inferences from materials read.
- There is a strong relationship between students' performance on the standardized achievement test and teachers' ratings on the observation checklist, i.e., students who perform well on the test are likely to be rated highly by teachers.
- Children who participate in early childhood education programs perform better on both the standardized test and the teacher checklist than children without such experience. However, it is likely that other factors not measured in this study are contributing to these performance differences.
- The special education sample in the pilot study included readiness-delayed learners only. A few of these children performed at levels comparable to children in regular classrooms though most scored lower than the New York City average.

USEFULNESS OF THE MAT AND CHECKLIST

- The usefulness of the MAT results was seriously weakened by administrators' and teachers' concerns regarding the difficulty and length of the test as well as its inappropriateness for first graders.
- The checklist results were thought to be useful in guiding teachers as part of their effort to assess individual student's strengths, weaknesses, and progress during the school year. A number of important considerations for future use were suggested:
  - In deciding how frequently to use the checklist during the school year, it is imperative to recall that it adds approximately three hours to a teacher's workload each time a class is evaluated.
  - The three-point scale should be carefully re-evaluated to judge whether <u>frequency</u> of skill performed (i.e., "not yet" to "most of the time") is an important measure or should <u>quality</u> of skill performance be judged instead (or, in addition). Also, does the three-point scale provide adequate differentiation or might a five-point scale be better.
  - In light of the pilot test findings are skills listed on the checklist providing teachers and others with new or useful information? Should other skills, suggested by teachers in this pilot study, such as "blending" or "word families" be added?

- Teachers and administrators would like to see a teacher checklist included as part of a citywide first-grade assessment process. Most feel, however, it was not comprehensive enough to be the only measure of reading performance. Although many would also like to see a standardized test as part of the assessment program, there were strong concerns expressed about using the MAT.

In sum, the pilot assessment program yielded valuable information which has both practical and theoretical implications. Based on these findings, New York City did not mandate citywide testing of first graders for the 1986-1987 school year and will consider test results and teacher and administrator opinion in planning future testing of first graders. They further point to the instructional value of a revised checklist but indicate caution in its use as a citywide assessment measure.

RECOMMENDATIONS FOR FIRST-GRADE ASSESSMENT

The following recommendations are made as a result of the pilot study:

- A citywide assessment program for the first graders should reflect the need for diverse types of assessment instruments to suit diverse purposes, such as identifying students' strengths and weaknesses, comparing of New York City children with national norms, and evaluating early childhood programs
- Because of the generally negative reaction to using standardized citywide tests at this grade level, a

sampling approach to citywide testing should be considered. A carefully chosen sample would give the data needed for citywide program evaluation without requiring that every student be tested, or that results be reported for every school and district.

- Since teachers and administrators reacted more negatively to the MAT than to the concept of standardized testing of first graders in general, an alternative might be to seek a more acceptable standardized test to administer to the selected sample. The benefits of any alternative test would have to be weighed against the benefits of a uniform citywide testing program from grade to grade, since the MAT is used citywide at grade 2.
- Those districts that continue to use the MAT for their own evaluation purposes should offer appropriate staff development in the use and interpretation of the test results, particularly in light of the pilot respondents' strong concerns about the test at this grade level.
- If the checklist continues to be used either as an option or as part of a citywide assessment program, aspects of the checklist such as which skills it measures and what kind of rating scale it should have need to be re-examined.
- The MAT test results should be shared with the Division of Curriculum and Instruction so they are aware of the large number of first graders performing below the national average on the skills assessed by this test.

In sum, these results have implications for both test developers and public education decision-makers, particularly regarding attitudes of teachers toward standardized testing of young children and appropriate ways to gather standardized reading achievement data for the first graders.

New York City Board of Education
Spring, 1986 First-Grade Pilot Study
Communication Arts First-Grade Checklist
(Monolingual Classes)

Please complete the following information.

STUDENT INFORMATION

I.D. Number 1 2 3 4 5 6 7 8 9

Name 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
(First) (Last)

Birthdate 27 28 (Month) 29 30 (Day) 31 32 (Year)

Limited-English Proficient 1) ___ Yes 2) ___ No (33)

OTHER INFORMATION

District 34 35 School 36 37 38 Classroom 39 40 41

For each statement, circle the number which indicates the degree to which the child has developed the communication arts skill in English. (The skills are defined in the attached Teacher's Guide.)

	Not Yet (1)	Some-Times (2)	Most of the Time (3)	DO NOT WRITE IN THIS COLUMN
A) LISTENING SKILLS				
1. Listens to others reading aloud with interest and pleasure.	1	2	3	(42)
2. Retells a simple story in sequence.	1	2	3	(43)
3. Perceives the main idea of a story.	1	2	3	(44)
4. Follows oral directions.	1	2	3	(45)
5. Recognizes rhyming words aurally.	1	2	3	(46)
B) SPEAKING SKILLS				
6. Looks at pictures and demonstrates understanding of content.	1	2	3	(47)
7. Relates own experiences, ideas, and feelings.	1	2	3	(48)
8. Asks questions.	1	2	3	(49)
9. Reveals understanding through replies and reactions to questions.	1	2	3	(50)
10. Expresses thoughts clearly enough to be understood.	1	2	3	(51)
11. Predicts next probable event in a sequence.	1	2	3	(52)
C) WRITING SKILLS				
12. Writes upper and lower case letters.	1	2	3	(53)
13. Uses invented spelling	1	2	3	(54)
14. Writes simple stories with minimal assistance from adult.	1	2	3	(55)

D) READING SKILLS	Not Yet (1)	Some-Times (2)	Most of the Time (3)	DO NOT WRITE IN THIS COLUMN
15. Distinguishes between realism and fantasy.	1	2	3	(56)
16. Establishes left to right and top to bottom directionality on a printed page.	1	2	3	(57)
17. Identifies sight words, print in the environment, and signs and labels.	1	2	3	(58)
18. Reads experience charts.	1	2	3	(59)
19. Reads and understands a variety of mathematical symbols, e.g., numerals, clocks, calendars.	1	2	3	(60)
20. Follows written directions.	1	2	3	(61)
21. Recognizes initial sounds and letters.	1	2	3	(62)
22. Recognizes final sounds and letters.	1	2	3	(63)
23. Associates letters of the alphabet with their sounds.	1	2	3	(64)
24. Reads aloud to and with others from books and own stories.	1	2	3	(65)
25. Sounds out words.	1	2	3	(66)
26. Uses contextual clues when coming upon unknown words.	1	2	3	(67)
27. Reads high-frequency words easily in any format or context.	1	2	3	(68)
28. Uses texts to find answers to questions posed by adults.	1	2	3	(69)
29. Makes inferences from materials read.	1	2	3	(70)
30. Recognizes the sound of different consonant clusters (e.g., bl, tr).	1	2	3	(71)

New York City Board of Education
Spring, 1986 First-Grade Pilot Study
Communication Arts First-Grade Checklist
(Bilingual Classes)

Please complete the following information.

STUDENT INFORMATION

I.D. Number 1 2 3 4 5 6 7 8 9

Name 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 25
(First) (Last)

Birthdate 27 28 (Month) 29 30 (Day) 31 32 (Year)

Limited-English Proficient 1) Yes 2) No (33)

OTHER INFORMATION

District 34 35 School 36 37 38 Classroom 39 40 41

For each statement, circle the number which indicates the degree to which the child has developed the communication arts skill in his or her native language. (The skills are defined in the attached Teacher's Guide.)

	Not Yet (1)	Some-Times (2)	Most of the Time (3)	DO NOT WRITE IN THIS COLUMN
A) LISTENING SKILLS				
1. Listens to others reading aloud with interest and pleasure.	1	2	3	(42)
2. Retells a simple story in sequence.	1	2	3	(43)
3. Perceives the main idea of a story.	1	2	3	(44)
4. Follows oral directions.	1	2	3	(45)
5. Recognizes rhyming words aurally.	1	2	3	(45)
B) SPEAKING SKILLS				
6. Looks at pictures and demonstrates understanding of content.	1	2	3	(47)
7. Relates own experiences, ideas, and feelings.	1	2	3	(48)
8. Asks questions.	1	2	3	(49)
9. Reveals understanding through replies and reactions to questions.	1	2	3	(50)
10. Expresses thoughts clearly enough to be understood.	1	2	3	(51)
11. Predicts next probable event in a sequence.	1	2	3	(52)
C) WRITING SKILLS				
12. Writes upper and lower case letters.	1	2	3	(53)
13. Uses invented spelling.	1	2	3	(54)
14. Writes simple stories with minimal assistance from adult.	1	2	3	(55)

BILINGUAL CLASSES

D) READING SKILLS	Not Yet (1)	Some-Times (2)	Most of the Time (3)	DO NOT WRITE IN THIS COLUMN
15. Distinguishes between realism and fantasy.	1	2	3	(56)
16. Establishes left to right and top to bottom directionality on a printed page. (Not applicable in all languages.)	1	2	3	(57)
17. Identifies sight words, print in the environment, and signs and labels.	1	2	3	(58)
18. Reads experience charts.	1	2	3	(59)
19. Reads and understands a variety of mathematical symbols, e.g., numerals, clocks, calendars.	1	2	3	(60)
20. Follows written directions.	1	2	3	(61)
21. Recognizes initial sounds and letters.	1	2	3	(62)
22. Recognizes final sounds and letters.	1	2	3	(63)
23. Associates letters of the alphabet with their sounds.	1	2	3	(64)
24. Reads aloud to and with others from books and own stories.	1	2	3	(65)
25. Sounds out words.	1	2	3	(66)
26. Uses contextual clues when coming upon unknown words.	1	2	3	(67)
27. Reads high-frequency words easily in any format or context.	1	2	3	(68)
28. Uses texts to find answers to questions posed by adults.	1	2	3	(69)
29. Makes inferences from materials read.	1	2	3	(70)
30. Recognizes the sound of different consonant clusters (e.g., bl, tr).	1	2	3	(71)

NEW YORK CITY BOARD OF EDUCATION
SPRING, 1986 FIRST-GRADE PILOT STUDY
COMMUNICATION ARTS FIRST-GRADE CHECKLIST

TEACHER'S GUIDE

The purpose of this assessment form is to identify some salient characteristics of each first-grade child in a natural setting, i.e., the classroom. The teacher, who is in daily contact with the child, is able to provide an ongoing evaluation and to give a comprehensive picture of the child at a particular time. Observing the child at work during the independent and small group work/play time provides opportunities to fill in the observation checklist on an ongoing basis. Completion of the items may take place over a period of days.

To help define checklist items more clearly and to establish a uniform observation guide, illustrations of the items which will help focus on the child's behaviors in a more detailed manner are included. These items may be manifested in different ways.

(A) LISTENING SKILLS

1. Listens to others reading aloud with interest and pleasure
   Listens attentively and identifies aspects of the story. Is interested in listening even when not being addressed specifically.
   Example: responds appropriately to humorous parts of a story either by facial expression and/or verbally.

2. Retells a simple story in sequence.
   Is able to recall or reconstruct verbally, or in picture form, a story in proper sequence.
   Example: uses puppet or felt board for a retelling of Three Little Pigs, or other stories. Draws pictures illustrating different parts of a story.

3. Perceives the main idea of a story.
   Is able to understand the most important idea of a story told, and tell, dramatize, write or draw about it.

4. Follows oral directions.
   Is able to follow oral directions that have two or three different commands.
   Example: cooperates with transition routines such as putting away materials and choosing books for quiet reading.

5. Recognizes rhyming words aurally.
   Uses rhyming words as a way for enjoying language;
   uses rhyming words in informal classroom situations;
   is able to find a rhyme for a given word, e.g., my, pie, eye;
   understands and uses rhyming language to evoke emotional responses, e.g., laughter. When asked to rhyme an unfamiliar word the child will substitute letters until a rhyme is formed (fling, swing).

(B) SPEAKING SKILLS (continued)

6. Looks at pictures and demonstrates understanding of content.
   Makes personal associations with pictures presented;
   makes up a story about a picture;
   tells sequential story from a book using only pictures
   Example: engages in pretend reading of familiar books to friend.

7. Relates own experiences, ideas, feelings.
   Can give verbal explanation of a picture or story based on personal experience; relates story to own experiences; gives evidence of own fears, preferences and values in discussion or circle time.
   Example: fears of animals, witches, giants, getting lost; preferences for foods, activities, books.

8. Asks questions.
   Is able to use language appropriately to ask questions in a variety of settings and experiences.
   Example: asks questions about a classroom pet: Where does it sleep? What does it eat?

9. Reveals understanding through replies and reactions to questions.
   Is able to react to questions with appropriate responses either verbally or through non-verbal expression.
   Example: able to select or choose appropriate dress for various weather situations presented; can express appropriate emotional reactions to a given situation.

10. Expresses thoughts clearly enough to be understood.
    Is able to use language appropriately in formal and informal settings;
    relates incidents in simple terms even with few details;
    uses sentences averaging 3 - 5 words;
    uses sentences with grammatical structure appropriate to age and developmental stages;
    uses many parts of speech.
    Example: children share personal experiences in a small or large group; children will sometimes respond differently in a small group from their response in a large group.

11. Predicts the next probable event in a sequence.
    Is able to report past events and predict future events either verbally or in picture form;
    is able to give verbal responses to questions based on predicting a story ending.
    Example: engages in scientific activities like planting seeds and can record in pictures the sequence of the experience.
    Example: responds to questions such as, "What do you think will happen next?"

(C) WRITING SKILLS

12. Writes upper and lower case letters.
    Can copy and write independently most of the upper and lower case letters.

13. Uses invented spelling.
    Uses invented spelling through experiential and language contexts, such as verbal cues, rhyming words and knowledge of sound; uses invented spelling to enrich independent writing projects.
    Example: Sistr, eyscrim, toi for sister, ice cream, toy.

14. Writes simple stories with minimal assistance from adult.
    Writes stories independently based on simple, personal and common experiences. Stories may consist of two or more sentences.

(D) READING SKILLS

15. Distinguishes between realism and fantasy.
    Is able to understand real and imaginary representation of ideas and show evidence of this understanding through group discussions, play activities and drawings.
    Examples: questions about whether a story is true/real or pretend will elicit responses from the children such as:
    "that's a make-believe story";
    "let's pretend we're doing this";
    "that's not a real story"; or
    "I'm only fooling".

16. Establishes left to right and top to bottom directionality on a printed page (not applicable to all languages), as evidenced by teacher observation of the child's interaction with printed materials, e.g., experience charts, big books.
    Example: runs finger under story (sentence) written under picture.

17. Identifies sight words, print in the environment, and signs and labels.
    Reads aloud or matches as evidenced by child's performance using these materials.
    Example: puts materials away appropriately as indicated by signs and labels at clean-up time; indicates understanding of signs such as: Stop, Go, Up, Down.

18. Reads experience charts.
    Reads experience charts to complete a recipe, recall events of a trip, follow a sequence of class rules at clean-up time, etc.
    Example: enjoys re-reading a chart or story when a discussion is recorded about classroom activities such as making play dough, spring time, etc.

19. Reads and understands a variety of mathematical symbols, e.g., numerals, clocks, calendars.
    Is able to respond verbally and in written form to questions by using mathematical terms appropriately, e.g., identifying class room number, finding a date on the calendar.

(D) READING SKILLS (continued)

20. Follows written directions.
    Is able to understand and respond to sequentially ordered instructions of two to three items.
    Example: Can understand directions to color, cut, write, circle, and/or underline.

21. Recognizes initial sounds and letters.
    Identifies some of the initial sounds and letters (more than ten).

22. Recognizes final sounds and letters.
    Identifies some of the final sounds and letters (more than ten).

23. Associates letters of the alphabet with their sounds.
    Identifies most of the letters of the alphabet and associates them with their sounds.
    Example: Demonstrates this skill in individual conference or group activity.

24. Reads aloud to and with others from books and own stories.
    Example: Reads original stories and/or trade books to the teacher or other children.

25. Sounds out words.
    Is independently able to sound out words while reading aloud, as evidenced by reading experience charts, classroom signs.
    Is able to read independently by using word attack skills; uses familiar sounds, rhyming words, similar words as clues.

26. Uses contextual clues when coming upon unknown words.
    Is able to understand unfamiliar word meanings through experiential and language clues, such as pictures, intra-sentence clues and in relation to meanings in surrounding sentences.
    Example: reads ahead to look for context clues for meanings of unknown words, reads sentence and then goes back to fill in unknown word.

27. Reads high frequency words easily in any format or context.
    Reads fluently the words used in the classroom.
    Example: signs and labels, experience charts, recipes, work charts, learning center directions, as well as common words used outside of the classroom.

OFFICE OF EDUCATIONAL ASSESSMENT

SPRING, 1986 FIRST-GRADE PILOT
ADMINISTRATOR SURVEY

Please fill in the following information:

DO NOT WRITE IN THIS COLUMN

District ___ ___ (1-2)

School ___ ___ ___ (3-5)

Please respond to the questions that follow. Your comments will help us as we study different methods of assessing first-graders' reading achievement.

Communication Arts First-Grade Checklist

1. Does the checklist adequately cover the skills included in the New York City first-grade communication arts curriculum?

   1) _____ Yes 2) _____ No (6)

2. Do the difficulty levels of the skills on the checklist adequately reflect the difficulty level of the first-grade curriculum?

   1) _____ Yes 2) _____ No (7)

3. Are the items clearly defined on the Teacher's Guide?

   1) _____ Yes 2) _____ No (8)

4. Are directions for completing the checklist understandable?

   1) _____ Yes 2) _____ No (9)

5. How useful do you think results from this checklist would be for instructional planning?

   1) _____ Very useful 3) _____ Minimally useful
   2) _____ Moderately useful 4) _____ Not at all useful (10)

DO NOT WRITE IN THIS COLUMN

6. How useful do you think results from this checklist would be for overall assessment of first-graders' communication arts skills?

1) _____ Very useful
2) _____ Moderately useful
3) _____ Minimally useful
4) _____ Not at all useful

(11)

7. At what time of year would administration of this checklist be most helpful to you?

1) _____ Fall
2) _____ Mid-year
3) _____ Spring

(12)

8. How, if at all, would you like to see a checklist such as this used in an overall first-grade assessment program?

1) _____ The checklist alone would be most useful.

2) _____ A standardized first-grade reading test alone would be most useful.

3) _____ A combination of the checklist and a standardized test would be most useful.

4) _____ Neither the checklist nor the test would be useful.

(13)

9. What do you see as the strengths of a checklist such as this?

10. What do you see as the weaknesses of a checklist such as this?

11. Please describe any problems that occurred with the administration of this checklist in your school.

(Please go on to next page and respond to questions about the standardized test administered as part of the pilot.)

DO NOT WRITE IN THIS COLUMN

METROPOLITAN ACHIEVEMENT TEST

12. Did the test adequately reflect the New York City communication arts curriculum?

1) _____ Yes 2) _____ No (I4)

13. Was the difficulty level of the test

1) _____ Too easy

2) _____ Just right

3) _____ Too difficult (I5)

14. How useful do you think results from this test would be for instructional planning?

1) _____ Very useful 3) _____ Minimally useful
2) _____ Moderately useful 4) _____ Not at all useful (I6)

15. How useful do you think results from this test would be for overall assessment of first-graders' communication arts skills?

1) _____ Very useful 3) _____ Minimally useful
2) _____ Moderately useful 4) _____ Not at all useful (I7)

16. What do you see as the strengths of a test such as this?

17. What do you see as the weaknesses of a test such as this?

18. Please describe any problems that occurred with the administration of this test in your school.

SUMMARY

19. Please indicate any comments or recommendations you have regarding a citywide first-grade assessment program and what it should include.

Thank you for completing this questionnaire. Please return it in the envelope provided to your district test liaison.

OFFICE OF EDUCATIONAL ASSESSMENT

SPRING, 1986 FIRST-GRADE PILOT
TEACHER SURVEY

Please fill in the following information:

DO NOT WRITE IN THIS COLUMN

District ___ ___ (1-2)

School ___ ___ ___ (3-5)

Class ___ ___ ___ (6-8)

Type of Class: 1) ___ Monolingual 2) ___ Bilingual (9)

1) ___ Special Education 2) ___ General Education (10)

Please respond to the questions that follow. Your comments will help us as we study different methods of assessing first-graders' reading achievement.

Communication Arts First-Grade Checklist

1. Does the checklist adequately cover the skills included in the New York City first-grade communication arts curriculum?

   1) _____ Yes 2) _____ No (11)

2. Do the difficulty levels of the skills on the checklist adequately reflect the difficulty level of the first-grade curriculum?

   1) _____ Yes 2) _____ No (12)

3. Are the items clearly defined on the Teacher's Guide?

   1) _____ Yes 2) _____ No (13)

4. Are directions for completing the checklist understandable?

   1) _____ Yes 2) _____ No (14)

5. How useful do you think results from this checklist would be for instructional planning?

   1) _____ Very useful 3) _____ Minimally useful
   2) _____ Moderately useful 4) _____ Not at all useful (15)

DO NOT WRITE IN THIS COLUMN

6. In which of the following ways would you use the checklist? (Check all that apply.)

1) _____ For grouping (16)
2) _____ For curriculum planning (17)
3) _____ For instructional purposes (18)
4) _____ For planning individualized activities (19)
5) _____ For assessing children's progress (20)

7. How useful do you think results from this checklist would be for overall assessment of first-graders' communication arts skills?

1) _____ Very useful
2) _____ Moderately useful
3) _____ Minimally useful
4) _____ Not at all useful (21)

8. How many checklists did you complete?

1) _____ 1-5
2) _____ 6-10
3) _____ 11-15
4) _____ 16-20
5) _____ 21-25
6) _____ more than 25 (22)

9. How much time did it take you to complete <u>all</u> the checklists?

1) _____ Less than one hour
2) _____ At least one hour but less than two
3) _____ At least two hours but less than three
4) _____ Three hours or more (23)

10. At what time of year would administration of this checklist be most helpful to you?

1) _____ Fall
2) _____ Mid-year
3) _____ Spring (24)

11. In the future, if you were asked to complete these checklists once a year for every child in your class, would you need additional time in order to do this?

    1) _____ Yes 2) _____ No (25)

12. How, if at all, would you like to see a checklist such as this used in an overall first-grade assessment program?

    1) _____ The checklist alone would be most useful.

    2) _____ A standardized first-grade reading test alone would be most useful.

    3) _____ A combination of the checklist and a standardized test would be most useful.

    4) _____ Neither the checklist nor the test would be useful. (26)

13. What do you see as the strengths of a checklist such as this?

_____

_____

_____

_____

_____

14. What do you see as the weaknesses of a checklist such as this?

_____

_____

_____

_____

_____

(Please go on to next page and respond to questions about the standardized test administered as part of the pilot.)

DO NOT WRITE IN THIS COLUMN

METROPOLITAN ACHIEVEMENT TEST

15. Did the test adequately reflect the New York City communication arts curriculum?

1) _____ Yes 2) _____ No (27)

16. Was the difficulty level of the test

1) _____ Too easy

2) _____ Just right

3) _____ Too difficult (28)

17. Were test items, in general, clear?

1) _____ Yes 2) _____ No (29)

18. Were directions to the children understandable?

1) _____ Yes 2) _____ No (30)

19. Were directions to the teacher understandable?

1) _____ Yes 2) _____ No (31)

20. How useful do you think results from this test would be for instructional planning?

1) _____ Very usef ! 3) _____ Minimally useful
2) _____ Moderately useful 4) _____ Not at all useful (32)

21. How useful do you think results from this test would be for overall assessment of first-graders' communication arts skills?

1) _____ Very useful 3) _____ Minimally useful
2) _____ Moderately useful 4) _____ Not at all useful (33)

22. What do you see as the strengths of a test such as this?

23. What do you see as the weaknesses of a test such as this?

SUMMARY

23. Please indicate any comments or recommendations you have regarding a citywide first-grade assessment program and what it should include.

Thank you for completing this questionnaire. Please return it in the envelope provided to your district test liaison.