DOCUMENT RESUME

ED 366 451 PS 022 073

AUTHOR Herman, David O.; And Others
TITLE Early Childhood Checklist. Pilot Study, Spring 1992. OREA Report.
INSTITUTION New York City Board of Education, Brooklyn, NY. Office of Research, Evaluation, and Assessment.
PUB DATE May 93
NOTE 34p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Check Lists; Early Intervention; *Elementary School Students; *Grade 1; *Informal Assessment; Mathematical Aptitude; Pilot Projects; Primary Education; Public Schools; *Reading Achievement; Remedial Programs; *Test Reliability; Test Selection; Test Validity
IDENTIFIERS New York City Board of Education

ABSTRACT

Each spring, the New York City Public Schools conduct evaluations of first-graders to determine which students are in need of remedial reading and mathematics programs. A new measure was designed to be more reliable than the multiple-choice, standardized tests used previously. A checklist of 20 items relating to language arts, reading, and mathematics was culled from various instruments and checklists. The format for each item required teachers to evaluate students' behaviors on a four-point scale. Expanded definitions were drafted for all items to help ensure that teachers would have a common understanding of the behaviors being rated. A pilot study was conducted to evaluate the new instrument, using nearly 1,700 first-graders from 14 schools; scores from the standardized Metropolitan Achievement Test (MAT) were available for about 1,400 of the students. A moderate correlation was found between checklist and MAT scores. (An appendix provides a copy of the directions given to teachers for completing the checklist with students.) (MDM)

OREA Report

EARLY CHILDHOOD CHECKLIST

Pilot Study, Spring 1992

EARLY CHILDHOOD CHECKLIST

Pilot Study, Spring 1992

New York City Public Schools
Office of Research, Evaluation, and Assessment
May 1993

NEW YORK CITY BOARD OF EDUCATION

7/6/93

PS

ACKNOWLEDGEMENTS

This report was written by David O. Herman of the Test Research, and Analysis Section, Betsy Taleporos, Administrator. The project was a joint effort of many individuals from all sections of the Office of Research, Evaluation, and Assessment, Robert Tobias, Director.

EARLY CHILDHOOD CHECKLIST FOR GRADE 1 STUDENTS

Spring 1992 Pilot Study

SUMMARY

Each spring, evaluations are needed for students completing first grade to identify those who could benefit from funded program remediation in the following school year. Up until now identification of such students has been based primarily on scores on a multiple-choice standardized test. This procedure, while providing objective information, has been criticized for a number of reasons relating to reliability, possible cultural bias, and serious concerns with validity.

In seeking an alternative, many instruments and checklists were examined to find one that would measure aspects of the first grade curriculum that would have relevance for the identification of students for Chapter I programs in second grade. No single checklist was found that satisfied a committee of early childhood and educational measurement specialists who reviewed them. Items culled from these various checklists formed an item pool, which was then reviewed by teachers and district staff to select those most relevant and important for identifying students for Chapter I second grade programs, and those that were easily observable in the daily events of most first grade classrooms.

The resulting checklist consisted of 15 items relating to language arts and reading, and five items relating to mathematics. The format of each item required teachers to evaluate students' behaviors on a four-point scale from "Has not done this yet" (scored as 1) to "Does this consistently" (scored as 4). Expanded definitions were drafted for all the items to help ensure that teachers completing the checklist would have a common understanding of the behaviors being rated.

Almost 1700 students from 14 elementary schools in five districts were rated by their teachers using the checklist. For about 1400 of these students, scores from the Metropolitan Achievement Test (sixth edition) in reading (MAT) were also available. Standard statistical tests showed that the checklist yielded reliable results, and that scores on it were related to MAT scores.

Moderate correlation was found between Checklist and MAT scores. While the relationship is not strong enough for application of a score-for-score equating of the two measures, the checklist did appear to be effective for separating those who scored at or below the 25th percentile on the MAT (the current second grade Chapter I cutoff) from those who scored above it. About 80 percent of the students who received a checklist score at or below 34 on the 15 reading items scored at or below the 25th percentile on the MAT.

We are encouraged by the findings of this pilot study. Most of the students who were identified as Chapter I eligible using the MAT were also identified by using a cutoff point on the checklist. The checklist is an unobtrusive measure. It can be used at any point during the school year by the teacher for instructional purposes. It can also be used to summarize a student's status at the end of first grade, and has been demonstrated to be valid for the purposes of Chapter I identification as well.

INTRODUCTION

Elementary schools in New York State need to be able to identify first grade students who could profit from Chapter I and PCEN remedial teaching in reading in the following school year. In the past, identifying such children has been based primarily on teacher judgment. Beginning in the spring of 1986, however, the Reading Survey Tests of the Metropolitan Achievement Tests, Sixth Edition (MAT Reading) were made available on an optional basis for administration to children in the New York City public schools in the spring of grade 1.[1] The MAT Reading results provided an objective measure that could be used to guide second grade reading instruction.

Despite the potential benefits of the MAT Reading test and similar tests, their use with first graders has been criticized. Reasons include allegations of inadequate coverage of what is taught in first grade reading instruction, possible cultural bias, and perhaps most important, a stressful or tiring effect on young children with consequent lack of validity. Thus some teachers and early childhood specialists would welcome a measure of reading skill other than a standardized test.

During the first year that the MAT Reading test was made available, the Office of Research, Evaluation, and Assessment, then called the Office of Educational Assessment, undertook a pilot study in which a specially selected sample of students who took the new reading test in grade 1 were also evaluated by their classroom teachers on a 30-item checklist of communication-arts skills. This experimental checklist was included in the study so that its validity and reliability could be explored as a possible replacement for the standardized reading test.

The 1986 pilot study found a correlation of .58 between the MAT Reading test results and a total score derived from the experimental checklist. This indicated that the domains measured by the test and the checklist had much in common, but were far from identical.

---

[1] The Primary 1 level of MAT Reading was used, which includes three tests: Vocabulary, Word Recognition Skills, and Reading Comprehension. Raw scores on the three tests are summed to yield the Total Reading score.

Teachers and administrators in schools that participated in the 1986 pilot study were surveyed for their opinions of the experimental checklist. Most felt that a teacher-completed checklist should be included as part of the citywide assessment process in grade 1, but that the checklist should not be the only measure of reading performance used in the assessment.

Among the reasons for caution were that the items on the communication arts checklist covered only some of the skills taught in a first grade reading program; that the format of the checklist permitted teachers to rate whether or not a child demonstrated a skill but not how well the skill was performed; and that the checklist ratings are inherently subjective and not based on fixed standards. Nevertheless, the results of the 1986 pilot study were sufficiently encouraging that a modified version of the study was repeated in the spring of 1992. The major purposes of the new study were to refine the content of the checklist, and to explore the extent to which the checklist can be substituted for an objective test for assessing first grade reading performance.

PROCEDURES

Preparing the Item Pool

The 30 items of the 1986 experimental checklist were part of the pool of behavioral descriptions that were considered for inclusion on the 1992 instrument. Additional checklists, some already used informally in one or another school district in New York City, and some published early-childhood rating scales, were also considered. From these materials all items that appeared to relate to early reading or mathematics achievement were identified, and duplicates were eliminated. Because the items included relatively few that were mathematically oriented, additional item statements were drafted in this area.

The resulting set of 98 items was edited for clarity, grouped according to subtopic, and reproduced as a special rating scale for distribution to teachers and others concerned with reading and mathematics instruction in the early grades.

Rating the Experimental Items

Following each item on the rating form were blanks for entering two ratings. One was a rating of Importance - the relevance of the behavior in question to the development of reading skills. The other was a rating of Observability - the ease, convenience, and clarity with which the behavior can be observed in the classroom. The ratings of Importance and Observability were to be given on a three-point scale (1, 2, or 3). Raters were also asked to choose and mark the 20 items that they would most like to see included on the final Early Childhood Checklist.

The form was distributed in February of 1992 to groups of primary-grade teachers, early childhood specialists, staff developers, reading and mathematics specialists, and others whose training and background would give them understanding of factors that contribute to early learning.

After editing, a sample of 146 usable rating forms remained. This group represented 22 of the 32 school districts of the New York City public schools.

Preparing the Checklist for the Pilot Test

Three measures of "goodness" were derived for each item: the total number of respondents who gave the item a rating of "3" (high) for Importance, the number who rated the item "3" for Observability, and the number of respondents who marked the item as among the 20 they would most like to see retained for the final Early Childhood Checklist. The number of raters choosing the items as most desirable and the number who rated the items as highly observable were primary factors in selecting material for the version of the Checklist to be pilot-tested. (The number of raters marking an item as desirable for the final Checklist was so closely related to the ratings of Importance that ratings of Importance received little direct weight during item selection.)

A committee that included specialists in early child development, bilingual education, and educational measurement was convened to choose items for the Early Childhood Checklist. Together, the committee suggested additional factors to consider in the selection process. The items should be so chosen that the Checklist would be comprehensive and cover all significant areas. The behaviors reflected on the instrument should be those usually seen in children at the end of first grade. (This requirement follows from the Checklist's intended purpose of identifying children who may need remedial services in grade 2; therefore the Checklist should be most sensitive at the low end of the range of reading skills, and should focus largely on "easy" items.) Further, the instrument should contain items related to both reading and mathematics.

A total of 15 items related to reading achievement and five items related to mathematics were ultimately selected, and these were printed on a scannable answer document to facilitate processing the Checklist data. The format of each item required teachers to evaluate children's reading or mathematics behaviors on a four-point scale, from "Has not done this yet" (scored as 1) through "Does it consistently" (scored as 4).[2] Expanded definitions were drafted for all of the items to help ensure that teachers completing the Checklist would have a common understanding of the behaviors described. The 20 items chosen for the Checklist appear in Appendix A, together with their expanded definitions.

---

[2] On the final version of the Checklist, items are scored on a four-point scale ranging from 0 through 3, instead of 1 through 4.

Selecting the Sample for the Pilot Test

One of the goals of the present study was to examine the consequences of substituting the Checklist for a standardized, multiple choice test of reading achievement for Chapter I and PCEN program participation and evaluation. For this reason the sample of schools for the pilot study was drawn from school districts that chose to administer the MAT Reading test to their students in grade 1.

Five of the eleven districts that used the test in grade 1 in the spring of 1992 were chosen to take part in the pilot study. Schools within these five districts were selected with the help of recent information about the percentage of second grade students in each school who performed at or above grade level on a standardized reading test. The selection was carried out in such a way that the student populations of the participating schools varied considerably in reading ability; in addition, the chosen schools were thought to give good representation of reading ability within their districts.

Collecting the Pilot Test Data

In all, 14 elementary schools in the five school districts took part in the pilot study. In June of 1992, copies of the Early Childhood Checklist were distributed to these schools to be filled out by the regular teachers of all first grade classrooms in which the Primary 1 level of the MAT Reading test had been administered one month earlier. Bilingual and special education classes were excluded. Sets of the expanded item definitions were also sent to the schools so that all participating first grade teachers would have a copy of this clarifying material at hand while using the Checklist.

In three of the 14 cooperating schools it was possible to arrange for a second teacher to complete a Checklist for the children in each classroom, independently of the regular teacher. These double ratings were solicited in order to evaluate inter-rater agreement.

ANALYSIS OF DATA

Of the 2013 Checklists returned by the schools, some had one or more of the 20 items with omitted ratings or multiple ratings. After eliminating such cases, 1692 valid Checklists remained for analysis. This group of Checklists, completed by the children's regular classroom teachers, constituted the basic sample of data that were analyzed for this report.

First, frequency distributions for the basic sample were prepared for the sums of ratings on the 15 reading-related items, the five mathematics-related items, and the total of all 20 items of the Early Childhood Checklist. Although these distributions do not constitute formal norms, they provide a general picture of the average and variability of the summary "scores" that might be expected from future administrations of the Checklist.

The internal-consistency reliability of the Checklist was also studied for the basic sample. Inter-rater agreement, another aspect of reliability, was studied for the much smaller group of 103 children for whom a second valid Checklist had been independently prepared by another teacher.

Validity analyses were limited to the 15 reading-related items of the Early Childhood Checklist because the only available criterion measure was performance on the MAT Reading test. (No measure of mathematics performance for first grade students was available from the schools.) The validation sample consisted of those first grade students who had valid scores on MAT Reading as well as valid ratings on the Checklist. Scaled scores on MAT Reading were used for this analysis because more than one form of the test had been used in the study, and raw scores are not comparable from form to form. Some students had scaled scores of zero on one or more of the three parts of MAT Reading, and all such cases were omitted from the sample studied. (Scaled scores of zero on this test correspond to raw scores of zero, and were considered invalid for purposes of these analyses.)

Of the 1692 children in the basic sample analyzed in this report, 1378 had valid scaled scores on the MAT Reading test, and these constituted the special sample used in this portion of the study. The validity analyses focused on the degree of relationship between reading "scores" from the Checklist and Total Reading scores on MAT Reading, and on the consequences of substituting the Early Childhood Checklist for MAT Reading as a first grade screening device.

RESULTS

Distribution of Ratings

Table 1 presents frequency distributions of three scores derived from the Early Childhood Checklist: the sum of ratings on the 15 reading-related items, the sum of ratings on the five items related to mathematics skills, and the sum of ratings on all 20 items. It was noted earlier that each item on the Checklist was presented as a four-point rating scale. Therefore the sum of ratings on the 15 reading-related items could range from 15 (for a child rated "Has not done this yet" on each item) to 60 (for a child rated "Does it consistently" on each item).[3]

For the sum of ratings on the reading items, the mean was 45.5, or approximately the score a child would receive if he or she had been rated "Does it often" (scored as 3) on all 15 items. This is consistent with the design of the instrument, since the items were selected

---

[3] Because of differences in the numerical values assigned to ratings on the final version of the Checklist, the sum of ratings on the final version ranges from 0 through 45 for the 15 reading-related items.

so that average children would "Do it often" by the end of the first grade, and children who did not exhibit the behavior might be in need of remediation.

The distribution of the three sums of ratings shows a concentration of scores at and near the maximum. For example, 338 children, or one fifth of the sample, received total reading ratings of 59 or 60; and 440 children, or more than one quarter of the sample, received total reading ratings between 57 and 60. Likewise on the five-item portion of the Checklist that is related to mathematics learning, 674 children, or nearly 40 percent of the sample, had total ratings of 19 or 20 (the top of the scale). For the entire set of 20 items, 17 percent of the children had total ratings of 79 or 80. These results were expected; it will be recalled that by design the Checklist included only "easy" items, so that it would be maximally discriminating among children with relatively poor reading skills.

Data relating to the relative difficulty of the Checklist items may be found in Table 2, in which one column presents the mean rating on each item. These means vary from a low of 2.7 to a high of 3.5 (2.7 to 3.4 for the 15 reading-related items). The standard deviations of the ratings were all close to 1.0. Of the 15 reading-related items, five had low average ratings of 2.7 or 2.8; these were Items 7, 8, 10, 12, and 14, and dealt with such relatively advanced skills as sounding out unfamiliar words while reading, and being able to perceive the main idea of a story. Three items, numbers 4, 9, and 13, had high average ratings of 3.4, and dealt with early directional habits in reading and possessing a minimal sight vocabulary. The range of mean ratings suggests that even though the Checklist items generally describe behaviors that most children will have mastered by the end of the first grade, they still reflect an appreciable array of difficulty. Again, this is consistent with the design parameters of the instrument.

Reliability of Ratings

Internal Consistency

The internal consistency of ratings on the Early Childhood Checklist was explored in two ways. Coefficients of correlation were computed between individual items and the three summary ratings (reading, mathematics, and total) on the Checklist. The coefficients were corrected for statistical "contamination"; that is, the sums of ratings with which the individual items were correlated excluded the item being analyzed. These data, shown in Table 2, range from .71 to .88 and indicate strong relationships between the individual items and the scales to which they are assigned.

The impression that the items provide good measurement of their underlying constructs is reinforced by the alpha coefficients given at the bottom of Table 2: .97 for the set of 15 reading items, .93 for the five mathematics items, and .98 for the full set of 20 items. (It is interesting to note that for the total score derived from the experimental 30-item communication arts checklist tried out in 1986, coefficient alpha was also .98.) Coefficient

alpha provides an overall indication of the homogeneity of the Checklist ratings, or the extent to which they measure the same underlying construct. Together, this reliability information suggests that none of the Checklist items are extraneous to their purposes, and that the three summary scores derived from the teachers' ratings have high internal consistency.[4]

Inter-rater Agreement

Although high internal consistency of ratings given on a checklist is important, the consistency of ratings given by two independent raters is also of interest. A number of factors may affect the stability of ratings on the Early Childhood Checklist. Two teachers may interpret the items differently, or may have different standards of judging whether the behavior described occurs "often" or "consistently," for example. A child's behavior may be inconsistent from one occasion to another, so that two teachers' ratings may be based on different behavioral samples. Finally, teachers may vary in the care with which they use the Checklist, and relatively careless raters will introduce random error into their ratings. For all of these reasons it is essential to study the inter-rater reliability of the ratings.

A relatively small group of 103 first grade students in three of the 14 participating schools had complete sets of valid ratings from two separate teachers. The matched Checklists for these children were studied in two ways. For each of the 20 items in turn, the number of children to whom both teachers gave exactly the same rating was determined, and expressed as a percentage of the entire group. These percentages of agreement are presented in Table 3, and range from a low of 65.0% to a high of 90.3%. The number of children whose ratings were within a single point of each other was also determined and expressed as percentages of the full group. These percentages ranged from 91.3 to 98.1%.

Another aspect of inter-rater reliability is the correlation of the four-point ratings on each item that were given by two separate raters. This information is also presented in Table 3. The coefficients range from a low of .50 to a high of .82 for the full set of 20 items, and from .52 to .82 for the 15 reading-related items. It should be pointed out that these coefficients indicate the extent to which children rated high on an item by one rater are rated relatively high by the second rater, and children rated low by one rater are rated relatively low by the second rater. They do not indicate the degree to which the two raters give identical ratings, and for this reason may not correspond with percentages of agreement computed for the same data.

Suppose that, for a particular checklist item, all of the ratings given by the second rater are exactly one point higher than those given by the first rater. In this case the percentage of agreement between the two raters would be zero, but the coefficient of correlation of the two sets of ratings would be 1.00. In this hypothetical example the two

---

[4] Note that alpha coefficients of .98 are extremely high, so much so as to suggest that "Halo effect" or some other type of systematic bias influenced the ratings.

indicators of inter-rater reliability would appear to give contradictory information, whereas they actually give equally meaningful, but different, views of the consistency of the two sets of ratings. (Differences in the opposite direction may also occur, with fairly high percentages of agreement and relatively low inter-rater correlations.) Inspection of the agreement percentages and coefficients of correlation in Table 3 reveals that such discrepancies do in fact occur. Item 8, for example, has a relatively low percentage of agreement of 68.9, but a relatively high inter-rater correlation of .80; by contrast, Item 9 has a relatively high percentage of agreement (81.6) but a relatively low inter-rater correlation (.52). While being aware of these discrepancies, one must still keep in mind that the percentages of agreement and coefficients of correlation are, as a group, quite high, especially considering that these statistics refer to individual items and not to subscores derived from groups of several items.

One final observation about the inter-rater reliability study is worth adding. The mean ratings given to the 103 children in this study are everywhere higher than those given to the 1692 children in the principal sample described in Tables 1 and 2. This holds for each of the 20 items taken individually, and for the three summary scores (reading, mathematics, and total) derived from the Checklist results. Furthermore, the standard deviations of the three summary scores are lower for the twice-rated sample of 103 than for the principal sample of 1692 students. These statistics are 12.3, 4.0, and 15.9, respectively, for the three summary scores of the principal sample (see Table 2), and 9.0, 2.5, and 11.1 for the summary scores of "Rater 1" in the twice-rated sample. Because the checklist ratings are markedly restricted for the special twice-rated group, the statistics presented in Table 3 probably underestimate the true degree of agreement between the two independent sets of ratings.

Validity of Ratings

Correlation with MAT Reading Test

As noted in the "Analysis of Data" section of this report, 1378 first grade students were available for exploring the relationship between the Early Childhood Checklist and the MAT Reading test. Only the 15 reading-related items of the Checklist were involved in this portion of the study. Although it is not possible to compare the reading test scores of this group with the scores of all first graders in New York City, a comparison can be made with the standardization sample for the MAT Reading test. For the present validity sample, the mean and standard deviation of MAT Total Reading scaled scores were 515.0 and 50.2, respectively, as shown in Table 4. These statistics were 512.0 and 44.0 for the 8802 first grade students tested as part of the spring (April and May, 1984) standardization of the Primary 1 level of the MAT battery, Form L; this was one of the MAT forms used for the

May 1992 administration in New York City.[5] (The figures were nearly the same for the portion of the standardization sample that took Form M of the battery.) Thus the average MAT Reading score for the Checklist validity sample was insignificantly higher than that of the MAT standardization sample, and its variability about 15 percent greater. Although eight years separate the MAT standardization and the present Checklist study (and therefore the national MAT norms may be outdated), it is reasonable to conclude that the validation sample for the Early Childhood Checklist is similar in reading skills to a national sample of children tested at a comparable time of the school year.

The validity sample of 1378 children may also be compared with the basic Checklist sample of 1692 first grade children in terms of the Checklist ratings. For the sum of ratings on the 15 reading-related ratings on the Checklist, which will be referred to henceforth as the "Reading Rating," the mean and standard deviation were 45.9 and 12.1 for the validity sample, and 45.5 and 12.3 for the larger basic sample. In terms of the ratings, then, the validity sample is nearly identical to the larger sample from which it was drawn.

The MAT Reading test was chosen as the criterion measure of reading ability for this report in part because it has a history of use in the New York City schools (1986-1992) to help identify children in need of remedial reading instruction in the second grade. The test also has good psychometric characteristics: a carefully selected national standardization sample, and excellent internal consistency (the KR-20 reliability coefficient was .96 for the Total Reading score for Form L of the Primary 1 level administered in the spring of grade 1)[6].

The test has three separate subtests: Vocabulary, Word Recognition Skills, and Reading Comprehension. Raw scores on the subtests are summed to obtain the Total Reading score.

Coefficients of correlation of the Reading Rating and scores on MAT Reading are presented in Table 4. Several observations about the data are relevant to the present study.

For any given Checklist item, as well as for the Reading Rating, the ratings correlate almost identically with the three MAT Reading subtests.

In general the Checklist ratings correlate slightly better with the MAT Total Reading score than with the subtest scores.

---

[5] MAT standardization statistics from MAT6 Preliminary Technical Manual, The Psychological Corporation, 1985.

[6] MAT reliability data from MAT6 Preliminary Technical Manual, The Psychological Corporation, 1985.

The correlation of the individual item ratings and the Total Reading test score is generally moderate, ranging from .40 to .61.

The Reading Rating correlates .61 with the MAT Total Reading score. The size of this coefficient indicates that the ratings and test scores overlap considerably; put another way, there is a strong tendency for children with high Checklist ratings to have high MAT scores, and vice versa.

In view of the high internal consistency of both measures (.97 for the Reading Rating, as shown in Table 2, and .96 for the MAT Total Reading score, as mentioned earlier), the correlation of .61 between the two suggests that they are not measuring identical domains, and thus are not equivalent measures. Nevertheless it will be of interest to explore the consequences of substituting the Early Childhood Checklist for the MAT Reading test.

Figure 1 shows a scatter diagram, or bivariate distribution, of the Reading Rating (on the horizontal axis) and the scaled score on MAT Total Reading (on the vertical axis). All 1378 first grade students in the validity sample are represented on the diagram. Each number on the diagram indicates the number of children who obtained a particular combination of Reading Rating and MAT Reading scaled score. For example, the isolated "1" at the upper left corner of the diagram represents the single student who obtained a Reading Rating of 20 and a scaled score of about 600 on MAT Total Reading.

The shape of the plot in Figure 1 is roughly that of a right triangle, with the right angle at the lower right corner. The column of numbers at the far left (1, 4, 6, 1, and 1) shows the distribution of MAT Total Reading scores for the 13 children who obtained the lowest possible Checklist score, 15.[7] (It will be recalled that each Checklist item is rated on a scale that is scored from 1 through 4, so the lowest possible sum of ratings on 15 items is 15.) The column at the far right represents the distribution of MAT Total Reading scores for the 230 children in the validity sample who obtained the maximum Reading Rating of 60.[8]

The very large number of children with Reading Ratings of 60 is consistent with the shape of the distributions of ratings shown in Table 1. As pointed out earlier under "Distribution of Ratings," the Checklist was designed to include only "easy" items covering behaviors that most average children would have mastered by the spring of grade 1. Thus the Checklist does not differentiate among children with average or above-average reading skills, nor was it intended to.

---

[7] The lowest possible Reading Rating on the final version of the Checklist is 0; the highest possible Reading Rating is 45.

[8] All numbers in the diagram that are greater than 9 are represented by letters, as follows. The number 10 is printed as an "A," 11 as a "B," 12 as a "C," and so forth, through 29, which is represented as a "T."

Equivalence of Ratings and Test Scores

If the Early Childhood Checklist is to replace the MAT Reading Test, the two measures must be equated in some way so that each critical score (such as a cut score) on MAT Reading has an equivalent point on the Reading Rating derived from the Checklist. Equivalent scores on two measures are usually established throughout the entire range of both measures, and this was attempted in the present project. The equating was carried out independently by two statistical techniques, the equipercentile method and regression analysis.

The principal definition underlying the equipercentile method is that scores on two different, but parallel, measures are considered equivalent if they lie at the same percentile rank on a given sample of people. For example, a Reading Rating of 37 falls at about the 25th percentile for the group of 1378 children that had valid ratings on the Checklist as well as valid scores on MAT Reading, and a scaled score of 474 on MAT Total Reading falls at the same percentile level for the same student sample. According to the equipercentile procedure, then, a Reading Rating of 37 is considered functionally equivalent to an MAT Total Reading scaled score of 474, providing that the two scores provide parallel measures of the same skill.

The equipercentile method of equating is intuitively appealing and easy to apply. In addition, the results are considered reversible; to use the example above, the equivalence relationship is the same whether one asks for the Reading Rating that is equivalent to an MAT Total Reading score of 474, or for the MAT Total Reading score that is equivalent to a Reading Rating of 37. One of its disadvantages, however, is the need to "smooth" score distributions, usually by smoothing curves on a graph. The smoothing process introduces an element of subjectivity, and therefore an unknown amount of error into the results, especially at extreme score levels. For this reason the equivalence of the Reading Rating and MAT Total Reading scores was also explored through regression analysis, an analytic technique that yields unique results free of smoothing error and other subjectivity. A condition of equating through regression procedures is that the results are not reversible, unlike the results of equipercentile equating; different results are obtained depending on which variable is considered the "given," and which variable its estimated equivalent.

In the present situation it is proposed to substitute the Early Childhood Checklist for MAT Reading. Should this be put into practice, teachers would have Reading Ratings available from the Checklist, and need to know a given child's most likely scaled score on MAT Reading. The regression analysis was set up accordingly. The results will be summarized here briefly rather than presented in detail.

An important product of the analysis was a set of regression equations that expressed the estimated MAT Reading score as a function of the Reading Rating on the Checklist. Such equations can then be used to produce a table of Reading Ratings and their estimated equivalents in terms of MAT Reading scores. Comparison of the equations showed that the

cube of the Reading Rating actually gave slightly better prediction of the MAT test score than did the unmodified Reading Rating. The advantage of nonlinear prediction was small in absolute terms even though it was statistically significant.

A by-product of the analysis was the discovery of an unacceptably large number of students with very large discrepancies between their actual and predicted MAT test scores. This was true of both linear and nonlinear prediction of MAT results. Furthermore, this was found to be particularly true of students with medium or high Reading Ratings. Such a finding is consistent with the triangular shape of the bivariate distribution shown in Figure 1, which was discussed earlier. Figure 1 clearly shows that the distribution of MAT Total Reading scores of children with low Reading Ratings (see the column at the far left) is quite narrow, while the distribution of MAT Total Reading scores of those with high Reading Ratings (the column at the far right) is broader, with greater variability. Thus predictions of MAT test scores were more accurate for students with low Reading Ratings. In relation to the goals of the present study, this indicates that equivalent MAT Total Reading scores may be given with confidence only for children who have relatively low Reading Ratings on the Early Childhood Checklist. For this reason it would be inappropriate and misleading to present a table of MAT Reading equivalents of the full range of Reading Ratings on the Checklist.[9]

Expectancy Table

Because the regression analysis described above found that predictions of MAT Total Reading results were most accurate for the students with low Reading Ratings on the Checklist, an expectancy table was prepared to summarize the probability of different MAT Reading test results for children with various Reading Ratings. This approach should help readers see at a glance not only the most likely test score obtained by children at different levels of the Reading Rating, but also the extent of error associated with this prediction.

In Table 5, scaled scores on MAT Total Reading have been grouped into four ranges corresponding to the four quartiles based of the MAT national norms. For example, students obtaining scaled scores of 485 or lower fell into the bottom 25 percent of the national normative sample. The Reading Rating is also grouped into four ranges. This grouping was based on the equivalence data developed as part of the equivalence study.

---

[9] The correlation between the Reading Rating and the MAT Total Reading score was recomputed after excluding children with high Reading Ratings. The purpose was to explore whether a table of equivalent scores on the two measures could be justified for a narrower range of Reading Ratings. However, the level of correlation dropped markedly, from .61 to .42, when Reading Ratings of 0 through 48 were considered (N = 713). This approach was therefore considered unfruitful.

The Reading Ratings were grouped in other ways as well, to see whether using different cut points would change the results described below. Although particular percentages did shift, of course, conclusions drawn from the variants of Table 5 were materially the same.

The table is read as follows. The data in the top row show that of the 439 first grade students whose Reading Ratings were between 55 and 60, 7.1 percent obtained Total Reading scaled scores of 485 and below, 15.3 percent obtained scaled scores between 486 and 510, and so forth. It is the bottom row of the table, however, that is of most interest in the context of this report. Here it may be seen that the MAT Reading scores of students with Reading Ratings between 15 and 34 were much more narrowly distributed than in the other three Reading Rating groups. Of the 270 students in this group, 81.1 percent obtained MAT Reading scores in the lowest quartile on the national norms, 19.7 percent received scores in the next higher quartile, 3.3 percent scored in the third quartile, and only 1.9 percent scored in the highest quartile on the test.

Table 6 presents the data in Table 5 reworked by compressing the top three rows of Table 5 into a single row, and combining the last three columns of the table into one. The result simplifies examining the consequences of using a cutting score on the Early Childhood Checklist to forecast the likelihood that children will score in the lowest quartile on MAT Reading. If we make the assumption that these results would be obtained in similar studies in the future, the data suggest that if we used a Reading Rating of 34 or below[10] to identify first grade students who would score in the lowest quartile on the test (and thus be eligible for Chapter I programs in the second grade), we would correctly forecast a reading test score in the bottom quartile 81.1 percent of the time for students in this range of Reading Ratings, but we would be wrong 18.9 percent of the time. Likewise, for students with Reading Ratings of 35 and above[11] we would correctly forecast a test score above the bottom quartile 73.6 percent of the time, but would be wrong 26.4 percent of the time.

Whether or not such results are acceptable will depend on the seriousness of false predictions, in both directions. Is it satisfactory to single out about 19 percent of low-rated children as needing remedial reading instruction, when their MAT Reading test scores would not have identified them thus? Likewise, among children with Reading Ratings above 34, would it be acceptable to identify about 26 percent as not needing remedial reading instruction when in fact their MAT Reading results would have qualified them for this instruction? It is important to note that these error rates of about 19 and 26 percent for the two groups apply to base groups of quite different size. The 18.9 percent of the 270 low-rated students who would be mistakenly assigned to the at-risk group amounts to only 51

---

[10] 19 or below on the final version of the Checklist.

[11] 20 and above on the final version of the Checklist.

students out of the sample studied, while the 26.4 percent of the 1108 high-rated students who would mistakenly be assigned to the not-at-risk group amounts to 293 students.

Whether such assignment "errors" are acceptable may also depend on how accurately the MAT Total Reading score is perceived as identifying children in need of remedial work in reading, and how appropriate the test is seen for young children. Many primary grade teachers and early childhood experts feel, for example, that the test is too long and tiring for children at this age, and results in underestimates of their true reading skills.

Other factors should be considered, too. How do first grade teachers feel about completing a reading-related rating scale for each child in their classes? All persons involved in the present pilot study understood that the Checklist results would have no direct effects on the students described. Should the Early Childhood Checklist become an official screening instrument, would the accuracy and fairness of teachers' ratings be subtly affected? To what extent might teachers consciously bias their ratings to achieve a particular screening outcome with particular children? Though beyond the scope of this report, such questions reflect legitimate issues in deciding whether the MAT Reading test should be abandoned in favor of this Checklist for early screening purposes.

CONCLUSION

A special-purpose Early Childhood Checklist has been prepared as an aid in identifying students in the spring of the first grade who would profit from remedial reading instruction in the second grade. First grade students in a sample of New York City school districts were rated by their regular teachers on the Checklist, and the ratings on 15 reading-related items were summed to yield a "Reading Rating."

In recent years the 25th percentile on the MAT Reading test (Primary 1 level) has been used in the New York City schools as a cut score for identifying first grade students in need of remedial work in reading. The present study has shown that a score of 34 or below[12] on the Reading Rating is an effective predictor of scoring at or below the 25th percentile on MAT Reading. Accordingly, the Reading Rating obtained from the Early Childhood Checklist is an acceptable proxy for the MAT Reading test for the limited purpose of Chapter I screening. The Checklist may also be used as a teacher's guide for the systematic observation of pupils in ways that are considered relevant to early reading learning. Records of such observations may be useful in planning instruction at any time of the school year, and as a summary of student status at the end of grade 1.

---

[12] 19 or below on the final version of the Checklist.

On the other hand there are several uses that are not appropriate for the Checklist:

It is not a general substitute for the MAT Reading test. Particularly in the case of medium and high scores on the Reading Rating, prediction of MAT Reading scores is inaccurate.

It is not a substitute for any reading achievement tests given above grade 1.

It is not suitable for use in Chapter 53 screening in any grade.

Frequency Distributions of Sums of Ratings on Reading and Mathematics Items, and on Total of All Items, for Students in Grade 1

N = 1692

Sum of Ratings	Reading (15 Items) N	Reading (15 Items) Cum. %	Mathematics (5 Items) N	Mathematics (5 Items) Cum. %	Total (20 Items) N	Total (20 Items) Cum. %
79-80					291	100.0
77-78					112	82.8
75-76					75	76.2
73-74					72	71.7
71-72					84	67.5
69-70					75	62.5
67-68					73	58.1
65-66					68	53.8
63-64					55	49.8
61-62					59	46.5
59-60	338	100.0			80	43.0
57-58	102	80.0			44	38.3
55-56	89	74.0			56	35.7
53-54	84	68.7			66	32.4
51-52	95	63.8			56	28.5
49-50	92	58.2			42	25.2
47-48	89	52.7			45	22.7
45-46	97	47.5			42	20.0
43-44	73	41.7			42	17.6
41-42	75	37.4			32	15.1
39-40	71	33.0			41	13.2
37-38	69	28.8			29	10.8
35-36	59	24.7			39	9.0
33-34	62	21.2			16	6.7
31-32	53	17.6			34	5.8
29-30	43	14.4			14	3.8
27-28	42	11.9			10	3.0
25-26	42	9.4			11	2.4
23-24	39	6.9			10	1.7
21-22	27	4.6			11	1.1
19-20	18	3.0	674	100.0	8	0.5
17-18	11	2.0	241	60.2		
15-16	22	1.3	214	45.9		
13-14			202	33.3		
11-12			127	21.3		
9-10			140	13.8		
7-8			59	5.6		
5-6			35	2.1		
Mean	45.5		15.9		61.5	
SD	12.3		4.0		15.9	

Table 2

Item Means and Coefficients of Correlation with Summary Ratings, with Alpha Coefficients for Summary Ratings, for Students in Grade 1

N = 1692

Item No.	Rating		Correlation with Summary Ratings[a]		
	Mean	SD	Reading	Mathematics	Total
1	3.2	0.9	.74		.73
2	3.1	0.9	.84		.83
3	3.0	1.0	.79		.79
4	3.4	0.8	.80		.81
5	3.2	0.9	.85		.85
6	3.0	1.0	.88		.87
7	2.8	1.1	.87		.86
8	2.7	1.1	.86		.85
9	3.4	0.9	.80		.80
10	2.7	1.1	.85		.85
11	3.0	0.9	.87		.87
12	2.8	1.0	.85		.84
13	3.4	0.8	.79		.80
14	2.8	1.1	.77		.76
15	3.0	1.1	.82		.83
16	3.3	0.9		.85	.82
17	3.5	0.8		.84	.80
18	3.3	0.9		.88	.84
19	3.1	1.0		.88	.84
20	2.7	1.1		.71	.72
Mean			45.5	15.9	61.5
SD			12.3	4.0	15.9
Coefficient Alpha			.97	.93	.98

[a] Each coefficient relates the rating on an item with the sum of the ratings on the remaining items on the appropriate group of items. For example, .74 is the correlation of the rating on Item 1 and the sum of ratings on Items 2 through 15.

Table 3

Percentages of Agreement for Items, and Inter-Rater Reliability Coefficients for Items and Summary Ratings, for Students in Grade 1 Rated Twice

N = 103

Item No.	Mean Rating		Percentage of Agreement	r_{12}[a]
	Rater 1	Rater 2		
1	3.7	3.6	73.8	.52
2	3.5	3.3	65.0	.62
3	3.6	3.4	76.7	.66
4	3.9	3.8	90.3	.78
5	3.5	3.5	74.8	.7[illegible]
6	3.5	3.4	81.6	.82
7	3.1	3.0	69.9	.79
8	3.1	3.0	68.9	.80
9	3.8	3.8	81.6	.53
10	3.1	3.0	68.9	.73
11	3.7	3.4	76.7	.67
12	3.5	3.3	77.7	.71
13	3.9	3.8	85.4	.77
14	3.0	3.0	66.0	.68
15	3.4	3.3	82.5	.82
Reading Rating	52.3	50.7		.86
16	3.8	3.6	73.8	.60
17	3.9	3.7	82.5	.50
18	3.8	3.6	84.5	.70
19	3.7	3.6	83.5	.69
20	3.1	2.9	69.9	.80
Mathematics Rating	18.2	17.4		.69
Total Rating	70.5	68.1		.86

[a] r_{12} is the correlation of the ratings given by Rater 1 with those given by Rater 2.

Correlation of Item Ratings and Reading Rating with Scaled Scores on MAT Reading Test, for Students in Grade 1[a]

N = 1378

Item No.	Coefficients of Correlation with Scores on MAT Reading Test				Rating	
	Vocabulary	Word Recognition Skills	Reading Comprehension	Total Reading	Mean	SD
1	.40	.38	.37	.40	3.3	0.9
2	.46	.45	.46	.48	3.1	0.9
3	.47	.45	.45	.48	3.0	1.0
4	.47	.48	.45	.49	3.4	0.8
5	.50	.51	.48	.52	3.2	0.9
6	.50	.50	.48	.52	3.0	1.0
7	.58	.56	.55	.60	2.8	1.0
8	.59	.56	.57	.61	2.7	1.1
9	.46	.46	.45	.48	3.4	0.9
10	.60	.56	.57	.61	2.8	1.1
11	.49	.48	.48	.51	3.0	0.9
12	.47	.45	.46	.49	2.9	1.0
13	.44	.44	.42	.45	3.4	0.8
14	.44	.44	.44	.47	2.8	1.0
15	.52	.51	.53	.55	3.0	1.0
Reading Rating	.59	.57	.57	.61	45.9	12.1
Mean	492.1	528.2	518.5	515.0		
SD	67.8	57.6	47.7	50.2		

[a] Checklist ratings given in June 1992. MAT Reading Test administered in May 1992.

Table 5

Expectancy Table Showing Relationship Between Reading Rating and Scaled Score on MAT Total Reading, for Students in Grade 1

N = 1378

Reading Rating	Number of Students	Percentage in Each Score Band on MAT Total Reading			
		485 and Below	486-510	511-537	538 and Above
55-60	439	7.1	15.3	15.5	62.2
45-54	375	28.3	24.8	26.1	20.8
35-44	294	53.1	25.5	10.5	10.9
15-34	270	81.1	13.7	3.3	1.9

Table 6

Expectancy Table Showing Probability of Scoring In or Above Lowest National Quartile on MAT Total Reading, for Students in Grade 1 in Two Categories of Reading Rating

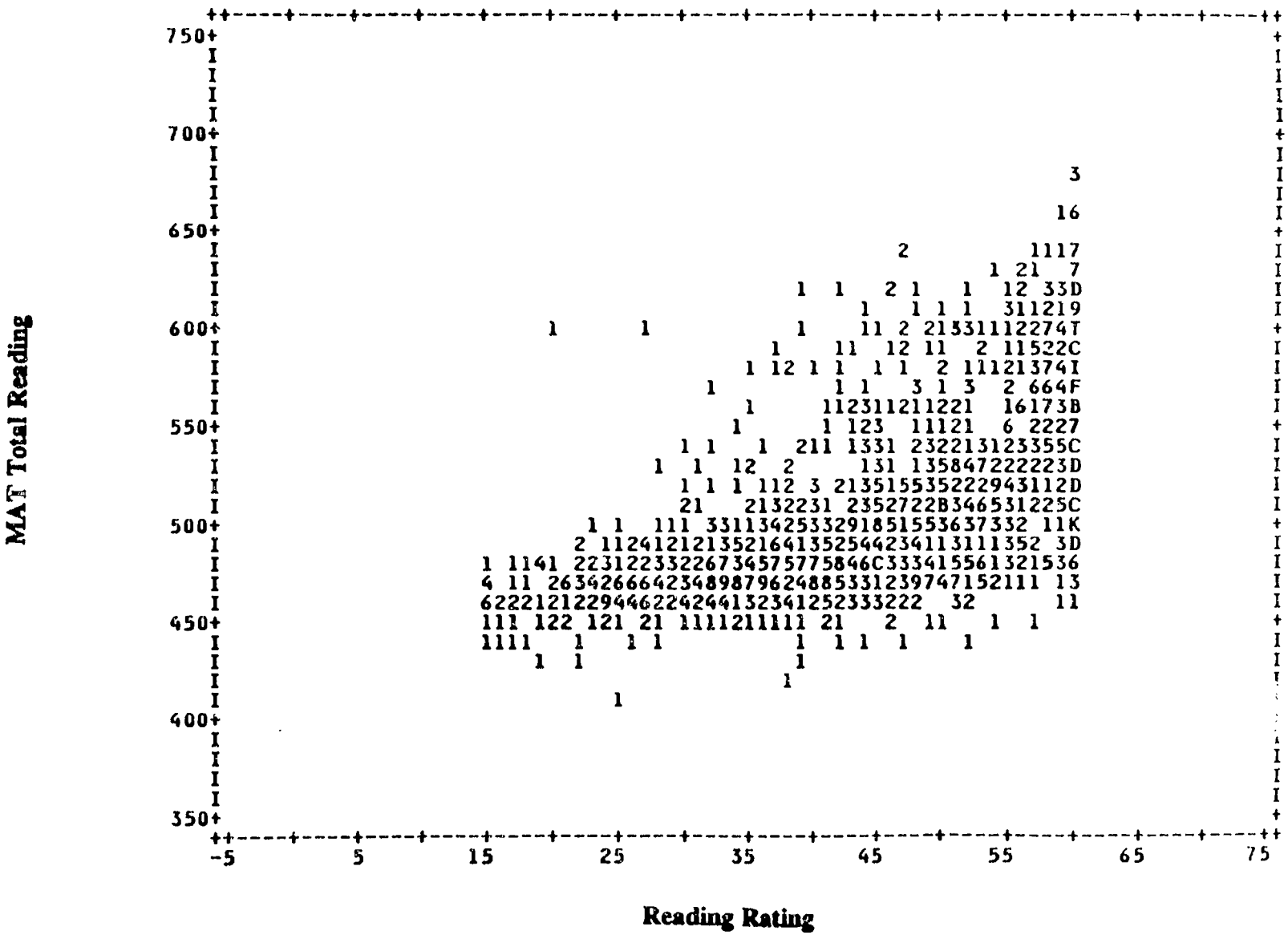
N = 1378

Reading Rating	Number of Students	Percentage in Each Score Band on MAT Total Reading 485 and Below (Lowest Quartile)	486 and Above (Highest 3 Quartiles)
35-60	1108	26.4	73.6
15-34	270	81.1	18.9

Figure 1

Scatter Diagram of Relationship between Reading Rating and Scaled Score on MAT Total Reading, for Students in Grade 1

N = 1378



APPENDIX A

EARLY CHILDHOOD CHECKLIST
SPRING 1992 PILOT STUDY

Directions for Completing the Checklist

1. **What is the Early Childhood Checklist?**

Each item on the checklist describes a behavior considered characteristic of young children near the end of first grade. The checklist includes 15 items related to children's reading, and five items related to mathematics achievement. It is being pilot-tested now to gather evidence on its usefulness in helping to identify children who could benefit from Chapter I or PCEN programs in second grade.

2. **Who should be rated?**

First graders in general education and resource room programs, who took the MAT Reading test this spring, are the only children who should be described on the checklist. Children in bilingual classes or self-contained special education programs should not be included. In order to ensure valid data, only children well known to the classroom teacher should be rated. For example, teachers should not complete checklists for children who entered the class after February 1, 1992. See **OREA Test Memorandum No. 20** for further details.

3. **Who should complete the checklists?**

Each child's own classroom teacher should fill out the child's checklist. In certain selected school buildings, the eligible children are to be rated a second time as well, when possible. Whenever two checklists are called for, the second rater is to be be another teacher such as a cluster teacher. **Paraprofessionals should not complete the checklists.** NOTE: If the second rater cannot confidently complete the five Mathematics items for a child, these items may be left blank. However, the 15 Reading items must be completed on all checklists.

4. **How are the checklists to be completed?**

No pupil preparation is involved. The evaluation is completed by reflecting on the child's performance in various settings.

Each of the 20 behaviors on the checklist is followed by four choices:

Has not done this yet
Does it occasionally
Does it often
Does it consistently

Use "Has not done this yet" to indicate you have never seen the child engaging in the behavior; and use "Does it consistently" to indicate that, in appropriate circumstances, the child always or nearly always responds as described. Use "Does it occasionally" and "Does it often" as intermediate ratings.

Responses to the checklist items will be electronically scanned. Therefore the use of a soft-lead (no. 2) pencil to complete the items is a necessity. Do not use a pen because the scanning equipment will not pick up ink marks.

After rating the 20 items of the checklist, indicate in the bubbles provided in the lower right corner whether you are the child's classroom teacher, a cluster teacher familiar with the child, or another kind of teacher. Remember, for this pilot study, paraprofessionals should not submit ratings.

The directions on the following pages give expanded definitions of the checklist items and examples of relevant critical behaviors that might be seen in the classroom. Before using the checklist to describe each of your students, you should read through the items and their definitions carefully. Appropriate personnel involved in administering the checklist should then meet to discuss the items and choices. This will help ensure that all raters have the same understanding of the items, and help resolve any questions that may arise.

For specifics regarding dates, delivery, administration, and packaging, consult **OREA Test Memorandum No. 20.**

Reading Items

1. **Listens with interest and pleasure to others reading aloud.**

Listens attentively to a story. Is interested in listening even when not being addressed specifically.
Example: Responds appropriately to humorous parts of a story verbally or by facial expression.

2. **Responds to stories by writing, drawing, or role-playing.**

Shows understanding of a story or part of a story in a sketch, dramatic presentation, or other artistic production.

3. **Relates or answers questions about own experiences, ideas, and feelings.**

Can give a verbal explanation of a picture or story based on personal experience; relates the story to own experiences; gives evidence of own fears, preferences and values in discussion or circle time.

4. **Follows a line of print from left to right.**

Example: Runs finger in the correct direction under the caption of a picture, or when asked to "point as you read."

5. **Associates most letters of the alphabet with their sounds.**

Identifies at least 20 letters of the alphabet and associates them with their sounds.
Example: Demonstrates this skill in an individual conference or group activity. In the case of vowels, associates each with at least one major sound.

6. **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 a new story; draws pictures illustrating sequential parts of a story.

7. **Sounds out unfamiliar words.**

Sounds out words independently while reading aloud, as evidenced by reading experience charts, classroom signs, or other printed material. Reads independently by using word attack skills; uses familiar sounds, rhyming words, or similar words as clues.

8. **Guesses the meaning of unfamiliar words using contextual cues and/or illustrations.**

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.

9. **Has a sight vocabulary of at least 10 simple words.**

Identifies sight words in signs, labels, and other print in the environment. Reads aloud or otherwise responds appropriately to such printed words as: Stop, Down, On, With, Tall.

10. **Uses self-correction strategy when reading aloud.**

Recognizes errors of pronunciation or sense while reading aloud, and attempts correction, whether successful or not.

11. **Recalls details from a story.**

Retains surface or nonessential details of a story and includes them in a discussion or retelling of the story.

12. **Perceives the main idea of a story.**

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

13. **Prints words or sentences from left to right in own writing.**

This is a behavior that teachers will have observed in the classroom. It does not refer to reversal of individual letters.

14. **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 for sister, toi for toy, izskrim to ice cream.

15. **Writes simple sentences.**

Writes sentences independently. Sentences must have two or more words, one of which is a verb.

Mathematics Items

1. **Recognizes equality of two groups (e.g., of cubes or pennies) by one-to-one matching.**

Matches elements of two groups of up to 15 objects, and tells whether one group has the same number as (more than, less than) the other.

2. **Identifies common shapes (circle, square, triangle, rectangle).**

Can point to and name all 4 shapes in the environment, or pick a named shape from a box of mixed shapes.

3. **Can add and subtract by manipulating objects (e.g., blocks, chips).**

Example: Can add 6 blocks to a group of 12, and give the total number of blocks; can show subtraction by removing 4 blocks from a group of 15, and give the final total.

4. **Reads and understands numerals through 100.**

Can recite the numbers in correct order from 1 through 100; understands the value of the numbers.
Example: Can point to 6 objects. Knows that 82 is greater than 58; knows that 33 comes just after 32.

5. **Knows values of common coins (dime, nickel, quarter).**

Recognizes and identifies all 3 coins and understands their equivalence.
Example: Knows that a dime is equivalent to 2 nickels, and that a quarter is equivalent to 5 nickels or to 2 dimes plus 1 nickel.