AUTHOR	Ysseldyke, James E.; And Others
TITLE	Evaluation Research: An Integrative Summary of Findings.
INSTITUTION	Minnesota Univ., Minneapolis. Inst. for Research on Learning Disabilities.
SPONS AGENCY	Special Education Programs (ED/OSERS), Washington, DC.
REPORT NO	IRLD-RR-144
PUB DATE	Sep 83
CONTRACT	300-80-0622
NOTE	117p.
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; *Evaluation Methods; Individualized Education Programs; *Information Utilization; *Learning Disabilities; Mathematics; Oral Language; Reading Tests; Social Adjustment; Spelling; *Student Evaluation

ABSTRACT

The report summarizes findings on evaluation studies conducted over a 6 year period at the Institute for Research on Learning Disabilities. Major findings are highlighted in the first chapter. Findings are grouped under the following topics: current evaluation practices, reading evaluation, spelling evaluation, written expression evaluation, oral language evaluation, mathematics evaluation, social adjustment evaluation, and data utilization. In chapter 2, implications for practice are noted, specifically the need for identifying alternative evaluation measures in the areas of reading, spelling, and written expression. Each of the eight major topics is then explored, in chapters 3 through 10, in terms of data sources and specific evidence. Sample findings include that most teachers rarely use systematic evaluation procedures to assess mastery of Individualized Education Programs goals and that teachers should be trained to use data for judging intervention effectiveness and improving student performance. Recommendations are made regarding direct measurement of reading, spelling, written expression, oral language, mathematics, and social adjustment. A final chapter, chapter 11, summarizes data sources for the research findings presented throughout the report. (CL)

University of Minnesota

Research Report No. 144

EVALUATION RESEARCH: AN INTEGRATIVE SUMMARY OF FINDINGS

James E. Ysseldyke, Martha L. Thurlow, and Sandra Christenson

Institute for Research on Learning Disabilities

IRLD

Director: James E. Ysseldyke

The Institute for Research on Learning Disabilities is supported by a contract (300-80-0622) with Special Education Programs, Department of Education. Institute investigators are conducting research on the assessment/decision-making/intervention process as it relates to learning disabled students.

During 1980-1983, Institute research focuses on four major areas:

- Referral
- Identification/Classification
- Intervention Planning and Progress Evaluation
- Outcome Evaluation

Additional information on the Institute's research objectives and activities may be obtained by writing to the Editor at the Institute (see Publications list for address).

The research reported herein was conducted under government sponsorship. Contractors are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent the official position of Special Education Programs.

Research Report No. 144

EVALUATION RESEARCH: AN INTEGRATIVE SUMMARY OF FINDINGS

James E. Ysseldyke, Martha L. Thurlow, and Sandra Christenson
Institute for Research on Learning Disabilities
University of Minnesota

September, 1983

Acknowledgments

A collaborative endeavor by numerous individuals was responsible for the production of this integrative summary of IRLD research. Most of the research upon which this summary is based was conducted by Stanley Deno, Phyllis Mirkin, and their associates Lynn Fuchs, Doug Marston, Mark Shinn, Jerry Tindal, Caren Wesson, and others. Because this report was written after these primary researchers had moved on to new positions and responsibilities, the authors integrated and interpreted findings that had been reported in other publications. Thus, while most of the content in this report was derived from the work of others, the authors take responsibility for the interpretations presented. Special acknowledgment is due to Lynn Fuchs and Caren Wesson who reviewed initial drafts of this report.

Chapter 1

Overview of IRLD Evaluation Research

Over a six-year period, the Institute for Research on Learning Disabilities (IRLD) at the University of Minnesota conducted research on evaluation issues, especially as they relate to assessing educational progress of learning disabled students, identifying instructionally-relevant evaluation procedures, and using continuous evaluation in classrooms. Current evaluation practices, alternative measurement procedures, and the use of data to evaluate students' programs were studied by means of a systematic research program.

This report describes the results of IRLD studies that provide information on evaluation procedures, especially as they relate to students who are receiving special education services. Findings from separate studies have been integrated to address major issues and to produce recommendations for practice that are based on research results. The studies from which the findings and recommendations were derived used a variety of methodologies. Included among these were:

- Comparative studies
- Surveys and interviews
- Experimental studies
- Developmental studies
- Observations
- Single subject studies
- Analytical studies
- Implementation studies

Highlights of Major Findings

The major questions that we asked and the major findings are presented here in very brief form. Implications of the findings for practice are discussed in Chapter 2. Details of the evidence that supports the findings are presented in Chapters 3-10. Information on the data sources and specific research procedures are presented in Chapter 11.

Current Evaluation Practices

1. What do teachers report to be their typical evaluation practices?

   a. Most teachers evaluate student progress four times during the school year.

   b. Teachers primarily rely on informal observations or informal tests to assess student mastery of IEP goals; they rarely use systematic evaluation procedures.

   c. The confidence that teachers have regarding the accuracy of their judgments about student performance is unjustified.

   d. Regardless of the evaluation procedure used, the frequency of measurement varies greatly from one teacher to the next.

2. To what extent do teachers use direct and frequent measurement procedures for evaluation?

   a. Most special education teachers are familiar with direct and frequent measurement strategies, but few use them.

   b. Teachers believe that direct and frequent measurement is time consuming and takes away from instructional time.

   c. Teachers who do use direct and frequent measurement strategies, on the average, use only a small proportion of a student's instructional time.

3. To what extent do teachers use the information obtained from direct and frequent measurement to make instructional changes?

   a. Teachers primarily rely on personal observation and judgment to make changes in instructional programs. Few teachers use direct and frequent evaluation strategies to decide about changes in students' instructional plans or to decide when to reteach or review a skill.

   b. Teachers who are required to use direct and frequent measurement strategies make more instructional program changes for students than do teachers not required to use the strategies.

c. Changes made by teachers are variable; the most common characteristic of changes is the infrequency with which they are made.

d. Training in data evaluation procedures should include a focus on appropriate changes to make in instruction, motivation, and physical setting.

Reading Evaluation

4. What are the characteristics of a recommended direct measure of reading?

a. A direct measure of reading should focus on the behavior of reading aloud from text. Measures of this behavior are technically adequate (valid, reliable, and sensitive to student growth), have instructional utility, and are logistically feasible in the classroom. A second choice behavior to measure is reading aloud from word lists.

b. When assessing a student's level of performance, the difficulty level of the direct reading measure should be as close as possible to the age-grade appropriate level, without reaching a level so frustrating that the measure is insensitive to student growth.

c. When assessing a student's level of performance, reading test items (text passages or words) should be selected randomly from a mid-sized domain, such as stories or words within a basal reader.

d. When selecting passages from one basal reader, it is desirable to select several "parallel" forms.

e. When assessing a student's mastery within progress measurement, the reading mastery criterion should be an absolute raw score correct and incorrect criterion; a recommended criterion is 50-70 words correct per minute, with 7 or fewer errors.

5. How should the direct reading measure be administered and scored?

a. The duration of a direct reading measure should be from one to three minutes each time it is administered.

b. Reading performance or progress on a direct reading measure should be scored in terms of the number of words read correctly.

c. Within an evaluation system, the direct reading measure should be administered at least two to three times per week.

d. The determination of whether to measure performance or progress should be made in light of individual student and teacher needs. Both procedures produce technically adequate data.

6. To what extent are basal reader criterion-referenced tests technically adequate?

-- Despite the content and face validity of basal reader criterion-referenced tests, their technical adequacy is often questionable.

Spelling Evaluation

7. What are the characteristics of a recommended direct measure of spelling?

a. A direct measure of spelling should focus on the behavior of writing words dictated from lists. Measures of this behavior are technically adequate (valid, reliable, and sensitive to student growth), have instructional utility, and are logistically feasible in the classroom. A second choice behavior to measure is writing compositions.

b. When assessing a student's level of performance, the difficulty level of the direct spelling measure should be within one to two grades of the student's instructional level.

c. When assessing a student's level of performance, words included in a dictated spelling list should be selected randomly from the domain of words in the spelling text or basal reader.

8. How should the direct spelling measure be administered and scored?

a. The duration of a direct spelling measure should be from two to three minutes each time it is administered. Paced dictation at a rate of 15 seconds per word is an acceptable procedure.

b. Performance on a direct spelling measure should be scored in terms of either the number of words spelled correctly or the number of letters in correct sequence. Letters in correct sequence is preferred for low-functioning students.

c. Within an evaluation system, the direct spelling measure should be administered at least two times per week.

d. The determination of whether to measure performance or progress should be made on the basis of individual student and teacher needs. The two procedures produce similar results.

Written Expression Evaluation

9. What are the characteristics of a recommended direct measure of written expression?

-- A direct measure of written expression should focus on the behavior of writing compositions in response to a verbal stimulus. Certain measures of this behavior (total words written, total words spelled correctly, or letters in correct sequence) are technically adequate, have instructional utility, and are logistically feasible in the classroom.

10. How should the direct written expression measure be administered and scored?

a. The duration of a direct written expression measure should be three minutes each time it is administered.

b. Performance on a direct written expression measure should be scored in terms of either total number of words or number of correctly spelled words.

c. Within an evaluation system, two or three writing samples should be elicited on each measurement occasion.

Oral Language Evaluation

11. What are the characteristics of a recommended direct measure of oral language?

-- A direct measure of oral language should focus on the behavior of describing a picture stimulus.

12. How should the direct oral language measure be administered and scored?

a. Performance on a direct oral language measure should be scored in terms of the number of non-repetitive words spoken.

b. The oral language measure should be administered by a familiar examiner.

Mathematics Evaluation

13. What are the characteristics of a recommended direct measure of mathematics?

-- Preliminary data suggest that a direct measure of mathematics should focus on the calculation of math computation problems.

14. How should the direct mathematics measure be administered and scored?

a. The types of problems presented to a student may be determined by the grade level of the student or may sample from all types of math functions.

b. Performance on a direct mathematics measure should be scored in terms of the number of digits correct.

c. Within an evaluation system, several samples should be elicited on each measurement occasion.

Social Adjustment Evaluation

15. What are the characteristics of a recommended direct measure of social adjustment?

-- A direct measure of social adjustment should focus on general classroom conduct and social interaction. The specific behaviors should be identified within the specific setting of interest.

16. How should the direct social adjustment measure be administered and scored?

a. Administration of the direct social adjustment measure could involve observation of the target student and classmates on an interval-sampling schedule.

b. Performance could be scored by tallying occurrences of the target behaviors.

Data Utilization

17. What are recommended procedures for graphing data?

a. Correct performance should be graphed. Incorrect performance may also be graphed along with correct performance to provide information about accuracy of performance.

b. When graphing a student's level of performance, equal interval graph paper can be used rather than semi-logarithmic chart paper.

c. When graphing a student's reading or spelling progress through a curriculum, number of words spelled or pages read should be spaced along the ordinate axis according to the time of mastery expected of average students in the curriculum.

18. How should graphed data be used to evaluate students' programs?

a. Graphed data should be summarized and interpreted to determine whether the instructional program is effective or needs to be changed.

b. Goal-oriented analysis is preferred for monitoring progress toward IEP goals, obtaining information about when to change a student's instructional program, and explaining student progress to parents and other teachers.

c. Program-oriented analysis is preferred for obtaining information about what to change in a student's instructional program.

d. A combined goal-oriented and program-oriented procedure that is recommended involves drawing a trend line through 7 to 10 data points; if the trend is flatter than the goal line, a program modification should be introduced.

19. How should teachers be trained to use data for judging intervention effectiveness and improving student performance?

a. Direct inservice or workshop training, rather than self instruction, is recommended for training teachers to collect data frequently and to use the data to make instructional decisions.

b. Systematic procedural changes can increase teachers' efficiency in using direct and frequent measurement procedures.

c. Direct training of teachers in measurement activities is more likely to result in teacher use and efficiency than training through manuals alone.

d. Goal setting is integral to progress measurement activities; teachers should monitor student performance in relation to short-term objectives rather than long-term goals.

e. Direct and frequent measurement with curriculum-based

20. To what extent do measurement and data utilization by teachers affect students' learning?

a. Student performance increases more when teachers use specific data-utilization rules to monitor progress than when they rely on their own judgment about student progress.

b. The quality of instruction improves when teachers use direct and frequent measurement and evaluation.

c. Students' knowledge about their goals and progress is greater when teachers employ direct and frequent measurement and evaluation.

d. Measurement appears to be a necessary condition in producing student growth, but not a sufficient one; positive effects of measurement cannot be sustained unless data-utilization procedures also are used.

implications for educators. Some of the general implications are discussed in this chapter. More specific implications can be found in IRLD research reports and monographs.

At the most general level, the IRLD research indicates that there are viable alternatives to those current evaluation practices which lack technical adequacy and which frequently are unrelated to making instructional decisions. For the most part, evaluation of learning disabled students is characterized by pre and post testing on standardized measures and by informal teacher procedures during the course of instruction. The IRLD research findings suggest that procedures that de-emphasize standardized testing and that emphasize continuous monitoring of pupil performance represent a more efficient and effective approach to evaluation when providing special education services to students. Further, the alternative approaches we have developed require as little as one to three minutes of testing time in a specific area, can also be used to make identification and eligibility decisions, and broader decisions about program effectiveness and allocation of resources.

IRLD research focused mainly on identifying and analyzing alternative evaluation measures in the areas of reading, spelling, and written expression. Some initial work also was done in oral language and mathematics. The mathematics work is being continued by local school districts who participated in IRLD research. In addition, research on non-academic measures (social adjustment) also was [illegible] measured, and how the data obtained can be used to improve educational programs for students.

The specific nature of the alternative approach reflects the notion that students must be measured on instructionally relevant tasks that can be administered repeatedly, and that their performance must be monitored continuously to identify when instructional, motivational, or other types of changes are needed to maintain student performance growth. Furthermore, the information obtained must be used systematically to make changes for students. The need for this type of approach is inherent in federal law (P.L. 94-142) which requires that schools construct individual educational programs (IEPs) for special education students. The IEPs must specify curriculum-based goals, <u>and</u> procedures for measuring progress toward those goals. A critical component of these procedures is their usefulness in generating data that can verify the extent to which program changes lead to program goals.

The IRLD research verified that efficient measures could be developed for reading, spelling, written expression, oral language, and mathematics. Procedures also were identified for social adjustment, but these were more situation specific, thus limiting their usefulness. Extensive studies documented the technical adequacy of the developed measures. Numerous implementation studies examined the feasibility of using the developed measures and the alternative approach to evaluation within special education programs. Measures and procedures were revised on the basis of these studies.

adopted the recommended evaluation procedures in their special education programs. In some cases, the procedures were adopted only for monitoring progress of special education students. In other cases, the procedures were applied to the entire array of special education decisions, including eligibility and termination considerations. The types of programs adopting the procedures have been quite varied. For example, one school system is a rural educational cooperative comprised of six school districts. The school districts have a total of about 5,000 students, with approximately 250 served in special education. Another school system is a large urban district that has a total student population of over 37,000 students. The total minority population accounts for 34.8% of the school population. Special education services are provided to 5200 students in this district.

The adoption of the direct and frequent measurement procedures by school systems speaks for its usefulness and feasibility. An excellent case study of how such a measurement and evaluation system might be created and employed is provided in IRLD Monograph No. 20.

This chapter summarizes IRLD research findings related to the nature of evaluation procedures typically used by special education teachers. Three specific questions are addressed in this chapter:

- What do teachers report to be their typical evaluation practices?
- To what extent do teachers use direct and frequent measurement procedures for evaluation?
- To what extent do teachers use the information obtained from direct and frequent measurement to make instructional changes?

For each question, the major findings are summarized and the data sources from which the findings were obtained are listed (generally ordered in terms of recency). Specific evidence for the major findings then is presented.

1. <u>What Do Teachers Report to be Their Typical Evaluation Practices?</u>

Findings:

a. Most teachers evaluate student progress four times during the school year.

b. Teachers primarily rely on informal observations or informal tests to assess student mastery of IEP goals; they rarely use systematic evaluation procedures.

c. The confidence that teachers have regarding the accuracy of their judgments about student performance is unjustified.

d. Regardless of the evaluation procedure used, the frequency of measurement varies greatly from one teacher to the next.

Data Sources:

- Survey and observation of special education teachers (RR 81)
- Survey of LD teachers (RR 65, 80)

Evidence:

Surveys and observations revealed that special education teachers primarily use informal observation and teacher judgment to formulate [illegible] assess student progress on IEP goals (RR 61). Over half of a group of nearly 150 special education teachers (65.9%) indicated that they evaluate progress on IEP objectives quarterly, 20% indicated weekly evaluation or at periodic review, and less than 3% indicated only annual evaluation of student performance. The majority of teachers (65.3%) relied on informal observations compiled over each quarter to formulate their decisions as to whether IEP objectives had been met. Informal observation not only was used more often than norm-referenced tests, criterion-referenced tests, and consultation, but also was the only method of progress evaluation used by 20% of the teachers. The general pattern of choices of methods of evaluation was the same for elementary and secondary teachers. Assessment of a student's level of performance on material covered in daily lessons involved informal observation for 80% of the teachers. Almost all (over 90%) of the teachers were confident in their selected evaluation procedures for determining student mastery. In fact, these teachers indicated they were "sure" or "very sure" about the student's level of performance. However, observations revealed that these teachers failed to recognize when objectives were not met by their students; for students who actually had failed objectives, teachers frequently indicated that they had been met.

A group of LD teachers identified their evaluation procedures for learning disabled students in reading, math, written language, and spelling (RR 65, 80). No single procedure or general type of evaluation was favored in reading and math. In these areas, teachers most often mentioned criterion-referenced measures, teacher-made tests/oral quizzes, informal observations of student performance, direct and frequent measurement (i.e., precision teaching), and standardized achievement tests. Teachers also included workbook scoring as a frequently used procedure for evaluating math progress. Informal observation of student performance was the chief form of evaluation in written language, while teacher-made tests/oral quizzes were clearly the most relied on form of evaluation in spelling. Informal observation of student performance primarily was used to evaluate students in other academic areas.

Teachers' frequency of evaluation varied with the area in which evaluation was used. Weekly evaluations were most common for written language and spelling, while daily evaluation in reading and math was mentioned by one-third of the teachers.

Teachers noted a number of ways in which they use evaluation information. Among the most commonly noted were discussing progress with student and parent, changing instructional plans, reteaching skills, and monitoring progress on IEP goals and objectives. Few teachers indicated that evaluation information was used to assign grades or review progress with the child study team. Most of the teachers who used evaluation information to discuss progress with a student did this on a daily or weekly basis; teachers who used evaluation information when reviewing progress with the team did so much less frequently (i.e., semi-annually, annually).

Most teachers were satisfied with the amount of time spent in evaluation activities; one-fourth of the sample desired an increase in evaluation, while 12.8% desired a decrease. Three-fourths of the teachers indicated they spent 30% of their time in evaluation. The remaining teachers indicated that they spent more than 30% of their time in evaluation.

2. To What Extent Do Teachers Use Direct and Frequent Measurement Procedures for Evaluation?

Findings:

a. Most special education teachers are familiar with direct and frequent measurement strategies, but few use them.

b. Teachers believe that direct and frequent measurement is time consuming and takes away from instructional time.

c. Teachers who do use direct and frequent measurement strategies, on the average, use only a small proportion of a student's instructional time.

Data Sources:

- Surveys of experimental study participants (RR 124)
- Comparative study of formative evaluation effects (RR 97)
- Surveys of special educators (RR 67)
- Interviews of special educators (RR 41)

Evidence:

Most surveyed teachers indicated they were familiar with direct and frequent measurement strategies, but only from one-third to one-half used the procedures with their students, even though only a few believed such measurement was not useful (RR 67). Some teachers, who were interviewed following their participation in one research project in which they used direct and frequent measurement, indicated that the procedures took too much time (RR 41). However, only 26% of the participants in another direct and frequent measurement study (RR 97) and only 4% in another (RR 124) indicated on surveys that the procedures were very time consuming. Of those teachers who typically used direct and frequent measurement, most reported that 20% or less of their time was devoted to measurement activities (RR 67). However, variability in time was considerable; some teachers estimated that direct and frequent measurement activities took up to 30% of instructional time. Yet, comparison of teachers' estimated and actual measurement times indicated that teachers who used the techniques generally overestimated how much time was involved (RR 67).

3. <u>To What Extent Do Teachers Use the Information Obtained from Direct and Frequent Measurement to Make Instructional Changes?</u>

Findings:

a. Teachers primarily rely on personal observation and judgment to make changes in instructional programs. Few teachers use direct and frequent evaluation strategies to decide about changes in students' instructional plans or to decide when to reteach or review a skill.

b. Teachers who are required to use direct and frequent measurement strategies make more instructional program changes for students than do teachers not required to use the strategies.

c. Changes made by teachers are variable; the most common characteristic of changes is the infrequency with which they are made.

d. Training in data evaluation procedures should include a focus on appropriate changes to make in instruction, motivation, and physical setting.

Data Sources:

- Survey of LD teachers (RR 65, 80)
- Comparative study of data-utilization rules (RR 64)
- Comparative study of teacher goals (RR 61, 62)

Evidence:

The national survey of LD teachers revealed that subjective teacher judgments played a major role in influencing intervention decisions (RR 65, 80). Such factors were cited both in relation to initial decisions about a student's program, and in relation to program changes. Only 19% of the teachers said that changes in the student's program would be based on "objective performance data" such as direct and frequent measurement strategies and standardized tests.

In a comparative study, special education teachers were trained in and required to implement continuous evaluation procedures using two data-utilization rules (RR 64). The first rule involved comparing student performance to a prespecified goal; the second involved a general directive to improve continuously upon the student's current performance level. The results demonstrated that teachers who used either rule made more program changes and more often used student performance data to modify students' programs than teachers who did not use any data-utilization rule. Further, students' reading performance improved more when the data-utilization rules were implemented by their teachers than when such rules were not used.

In another study, the quality and quantity of teachers' changes were compared for teachers using long-term goals and introducing program changes at least every two weeks, and for teachers using short-term goals and introducing program changes only as frequently as necessary to ensure that their students would achieve goals. A greater percentage of teachers in the short-term goal group made no changes in students' reading programs (RR 62). When changes were made, all teachers made a greater percentage of changes that were characterized as instructional as opposed to either motivational or physical arrangement changes. Although teachers who set long-term goals made more changes overall, no differences in reading performance were revealed for the students in the two groups (RR 61). The finding that teachers rarely made changes in students' programs highlighted the need for more intensive training in data evaluation procedures.

Chapter 4

Reading Evaluation

This chapter summarizes IRLD research findings related to reading evaluation. Three specific questions are addressed in this chapter:

- What are the characteristics of a recommended direct measure of reading?
- How should the direct reading measure be administered and scored?
- To what extent are basal reader criterion-referenced tests technically adequate?

For each question, the major findings are summarized and the data sources from which the findings were obtained are listed (generally ordered in terms of recency). Specific evidence for the major findings then is presented.

4. What Are the Characteristics of a Recommended Direct Measure of Reading?

Findings:

a. A direct measure of reading should focus on the behavior of reading aloud from text. Measures of this behavior are technically adequate (valid, reliable, and sensitive to student growth), have instructional utility, and are logistically feasible in the classroom. A second choice behavior to measure is reading aloud from word lists.

b. When assessing a student's level of performance, the difficulty level of the direct reading measure should be as close as possible to the age-grade appropriate level, without reaching a level so frustrating that the measure is insensitive to student growth.

c. When assessing a student's level of performance, reading test items (text passages or words) should be selected randomly from a mid-sized domain, such as stories or words within one basal reader.

d. When selecting passages from one basal reader, it is desirable to select several "parallel" forms.

e. When assessing a student's mastery within progress measurement, the reading mastery criterion should be an absolute raw score correct and incorrect criterion; a recommended criterion is 50-70 words correct per minute, with 7 or fewer errors.

Data Sources:

- Norming study (RR 132)
- Analysis of readability formulas (RR 129)
- Comparative study of standardized and direct measures (RR 126)
- Direct measure reliability study (RR 109)
- Implementation study (RR 106)
- Aggregation study (RR 94)
- Study of curriculum differences (RR 93)
- Comparative study of formative evaluation effects (RR 88)
- Direct measures norm development (RR 87)
- Study of alternative reading performance criteria (RR 59)
- Comparative study of three reading placement procedures (RR 56, 57)
- Comparative study of reading domains (RR 55)
- Longitudinal study of learning trends on simple measures (RR 49)
- Comparative study of reading domains and durations (RR 48)
- Technical characteristics of direct reading measures (RR 20)

Evidence:

The issue of what specific behaviors to measure when evaluating reading was addressed by a series of studies on the technical characteristics of direct reading measures (RR 20). Correlations of five direct measures (reading aloud from text, reading aloud from word lists, reading isolated words presented in text; identifying deleted words in text, and giving word meanings) with standardized reading tests indicated that performance on three of the direct measures (reading aloud from text, reading aloud from word lists, and identifying deleted words in text) was correlated highly with performance on the standardized tests, with the validity coefficients ranging between .73 and .94. Significant correlations were replicated in several other studies (RR 56, 57, 88, 94). Both of the reading aloud measures consistently correlated higher with the standardized tests than did the cloze measure (identifying deleted words). Comparisons of correct performance on the three measures across grades (RR 20, 87) and across time within grades (RR 87) revealed that the cloze measure was much less sensitive to student growth than either of the reading aloud measures, and further that the reading aloud from text measure was somewhat more sensitive to student growth than the reading aloud from word lists measure. Separate analyses, however, confirmed that reading aloud from word lists was more sensitive to student growth than standardized tests (RR 126). The sensitivity of the reading aloud measure across and within grades, and its reliability, were confirmed in additional studies with different student samples (RR 49, 106, 109, 132).

The issue of how to select the difficulty level of a reading aloud measure was addressed in studies of validity (RR 57) and sensitivity to student growth (RR 20, 57). These investigations indicated that when correct performance scores were used, all difficulty levels were correlated significantly with achievement test scores (RR 57); however, reading aloud passages of mid-range difficulty maximized slope, indicating greater sensitivity to student growth (RR 20, 57). When error performance scores were used, difficulty level affected the size of the correlation of the direct measure with achievement test performance (RR 57).

The issue of the appropriate domain from which measurement items should be selected to assess a student's performance was addressed directly with respect to reading aloud from word lists (RR 48). In comparisons of measures derived from a limited (200 words) instructional level domain, an entire within-grade level domain, and an across grades (preprimer-grade 4) domain, it was found that as the size of the domain increased, sensitivity to student growth decreased. However, variability of slope was greatest for measures selected from the most limited domain size; minimal variability is desired. Analyses of the effect of domain size on the judged effects of instructional interventions did not produce clear results (RR 55). Given that it is easier to draw samples of items from a larger domain, and that a somewhat restricted domain results in greater sensitivity to student growth and reduced variability, a mid-size domain was recommended (RR 48). The widely-accepted procedure of random selection from the domain also was recommended (RR 20).

The issue of the appropriate procedure for selecting reading passages was highlighted by a study of the reliability and validity of alternative performance criteria (RR 59). In this study, reading passages were sampled randomly until the readability level of two passages coincided with the mean readability scores for the reading levels. The number of passages that had to be selected ranged from 5 to 14; over half of the 19 textbooks sampled required the selection of 10 or more passages before two representative passages could be identified. The problem is further complicated by the demonstrated inaccuracy of readability formulas (RR 129). First, there appears to be minimal agreement among several formulas. Second, the difficulty of a passage also seems to be influenced by the background of the student reading the passage. A suggested procedure for reducing error and increasing technical adequacy is both to create parallel forms of passages by selecting several alternative passages and to administer them on consecutive days so that pupils' scores can be aggregated or so that administrations can be repeated until results agree on at least two consecutive days (RR 59).

The issue of the appropriate criteria to apply to determine whether a student has achieved mastery of materials was addressed in a study that examined seven criteria recommended by various individuals. When the seven criteria were applied to reading aloud from text scores of students, four were found to be sensitive to student growth, to demonstrate good criterion validity with standardized tests, and to result in at least 50% agreement with teacher judgments (RR 57, 93). Given that criteria involving the calculation of percentages require extra teacher time, an absolute raw score criterion of 50-70 words correct per minute with 7 or fewer errors was recommended.

5. How Should the Direct Reading Measure be Administered and Scored?

Findings:

a. The duration of a direct reading measure should be from one to three minutes each time it is administered.

b. Reading performance or progress on a direct reading measure should be scored in terms of the number of words read correctly.

c. Within an evaluation system, the direct reading measure should be administered at least two to three times per week.

d. The determination of whether to measure performance or progress should be made in light of individual student and teacher needs. Both procedures produce technically adequate data.

Data Sources:

- Single subject study (RR 120)
- Direct measure reliability study (RR 109)
- Direct measures norm development (RR 87)
- Comparative study of data-utilization rules (RR 64)
- Comparative study of teacher goals (RR 61, 62)
- Comparative study of three reading placement procedures (RR 57)
- Teacher efficiency studies (RR 53)
- Comparative study of reading domains and durations (RR 48)
- Development of data-utilization systems (RR 23)
- Technical characteristics of direct reading measures (RR 20)

Evidence:

The issue of the duration of a direct reading measure was addressed in several studies. In studies of the technical characteristics of reading measures (RR 20) and in the development of data-utilization systems (RR 23), a one-minute assessment of reading was found to validly index reading proficiency. Although correlations between 30-second and 60-second reading aloud trials were as high as .92 (RR 20), the 30-second duration was less sensitive to student growth and was characterized by greater intra-individual variability (RR 48). Comparisons of 30-second and 3-minute durations indicated that the longer duration resulted in reduced intra-individual variability and increased reliability (RR 48). Given the logistical benefit of shorter tests weighed against the technical and instructional superiority of longer tests, the recommendation of a one to three minute duration was made.

Several studies provided evidence on the issue of how to score performance on direct reading measures; they consistently found that either correct rate or percentage correct is a more valid score than error rate. Studies of the technical adequacy of direct reading measures (RR 20) and a reliability study (RR 109) indicated that correct performance is a more valid measure of reading performance than is error performance. Correct performance scores were found to discriminate among reading proficiencies as well as scores reflecting a combination of correct and incorrect performance (RR 20). Further, correct rate stability coefficients, indicative of a measure's test-retest reliability, were higher than error rate stability coefficients (RR 87). In addition, validity correlations for error rate were unreliable (RR 20). Given that one additional step is required to calculate a percentage correct score, it was recommended that correct rate be scored. For instructional information, practitioners might want to monitor both correct rate and error rate.

The issue of the frequency with which the direct reading measure should be given in an evaluation system was addressed indirectly by data collected during the development of data-utilization systems (RR 23). Students who were measured on a daily basis showed greater progress than students who were measured on a weekly basis. Daily measurement is the ideal; however, teachers find daily measurement to be cumbersome and time consuming (RR 53). In light of this, a compromise solution of two to three times per week is recommended.

The issue of whether a reading evaluation system should use progress measurement (in which the measurement domain changes each time a student masters a segment of the curriculum) or performance measurement (in which the measurement domain remains the same) was examined in several studies. High correlations (concurrent validity) were found for both performance and progress measures of reading; both were highly predictive of scores on standardized achievement tests (RR 20, 57). The progress measures studied were based on mastery of books within a reading curriculum. When the effect of the measurement system (progress vs. performance) on student reading achievement was examined, no significant differences were found (RR 61); a similar finding for spelling performance was provided by a single-subject experiment (RR 120). However, in a study of goals and data utilization, teachers using progress measurement were more realistic and optimistic about their students' programs than were teachers who used performance measurement (RR 62); further, progress measurement teachers introduced fewer unnecessary program modifications. Also, in a school district where direct measurement procedures were adopted district-wide, teachers more often selected progress measurement for reading than they selected performance measurement (RR 64). Since there is no evidence of differences in the technical adequacy of the two approaches, the decision may be made appropriately on the basis of preferences and needs.

6. To What Extent Are Basal Reader Criterion-Referenced Tests Technically Adequate?

Findings:

-- Despite the content and face validity of basal reader criterion-referenced tests, their technical adequacy is often questionable.

Data Sources:

- Analyses of basal reader criterion-referenced tests (RR 113, 122, 128, 130)

Evidence:

Analyses of the technical characteristics of selected criterion referenced tests from Houghton-Mifflin (RR 113), Ginn 720 (RR 122), Scott-Foresman (RR 128), and Holt (RR 130) indicated considerable variability in technical adequacy. The reliability and validity of the Houghton-Mifflin end-of-level 11 basic reading test were found to be less than adequate. For the Ginn 720 end-of-level 11 mastery test, reliability and validity were acceptable for the composite test scores, but variable for the subtests. Reliability and validity of the Scott-Foresman end-of-book 9 criterion-referenced test appeared acceptable for the total test, but not for some of the scale scores. Analyses of the Holt management program level 13 test indicated that the criterion-related validity was acceptable, but that the test-retest reliability and the convergent and discriminant validity were questionable. It was concluded that test consumers must demand empirical validation before relying on criterion-referenced test data for making instructional decisions.

Chapter 5

Spelling Evaluation

This chapter summarizes IRLD research findings related to spelling evaluation. Two specific questions are addressed in this chapter:

- What are the characteristics of a recommended direct measure of spelling?
- How should the direct spelling measure be administered and scored?

For each question, the major findings are summarized and the data sources from which the findings were obtained are listed (generally ordered in terms of recency). Specific evidence for the major findings then is presented.

7. What Are the Characteristics of a Recommended Direct Measure of Spelling?

Findings:

a. A direct measure of spelling should focus on the behavior of writing words dictated from lists. Measures of this behavior are technically adequate (valid, reliable, and sensitive to student growth), have instructional utility, and are logistically feasible in the classroom. A second choice behavior to measure is writing compositions.

b. When assessing a student's level of performance, the difficulty level of the direct spelling measure should be within one to two grades of the student's instructional level.

c. When assessing a student's level of performance, words included in a dictated spelling list should be selected randomly from the domain of words in the spelling text or basal reader.

Data Sources:

- Norming study (RR 132)

- Direct measure reliability study (RR 109)
- Direct measures norm development (RR 87)
- Longitudinal study of learning trends on simple measures RR 49)
- Development of data-utilization systems (RR 23)
- Technical characteristics of direct spelling measures (RR 21)

Evidence:

The issue of what specific behaviors to measure when evaluating spelling was addressed by a series of studies on the technical characteristics of direct spelling measures (RR 21). Correlations of two direct measures (writing words dictated from lists and writing compositions) with standardized spelling tests indicated that performance on the writing words dictated from lists measure was correlated highly with standardized tests, with the validity coefficients ranging between .80 and .96. A moderately high correlation (.70) was obtained between spelling performance on the writing compositions measure and performance on a standardized spelling test. Test-retest, alternate-form, and interjudge reliability levels were high, at least when correct performance was scored (RR 109). Comparisons of correct performance on the writing words dictated from lists measure across grades and across time within grades revealed that this measure was sensitive to student growth (RR 21, 49, 87, 132).

The issue of the appropriate difficulty level of a measure of writing words dictated from lists to assess a student's performance level was addressed by a study of the technical characteristics of direct spelling measures (RR 21) and a study on the development of norms for direct measures (RR 87). When correct performance scores were used, list difficulty had little effect on the validity of the dictated word list measure (RR 21). When materials were selected from material around the student's instructional grade level, the measure was sensitive to student growth (RR 87). Given that it is generally easier to select words from a grade-level spelling text or from the lists of words in a basal reader, the recommendation was made that words included in a dictated spelling list measure be within one to two grades of the student's instructional level.

The issue of the appropriate domain from which words should be selected to assess a student's performance was addressed by comparing student progress when teachers made program changes on the basis of student performance on words from a small domain (a within-grade-level list of words) and when teachers made changes on the basis of student performance on words from a large domain (a list of words selected from across several grade levels) (RR 23). Both domains produced measures that were sensitive to student growth over time. Examinations of the validity of curriculum-based spelling measures when words were selected in three ways (randomly, arbitrarily, and ordered from easy to difficult) indicated that both randomly selected words and arbitrarily selected words had high correlations with achievement tests, but ordered words had low concurrent validity (RR 21). Given the lack of additional research, the widely-accepted procedure of random selection from the domain was recommended.

8. How Should the Direct Spelling Measure be Administered and Scored?

Findings:

a. The duration of a direct spelling measure should be from two to three minutes each time it is administered. Paced dictation at a rate of 15 seconds per word is an acceptable procedure.

b. Performance on a direct spelling measure should be scored in terms of either the number of words spelled correctly or the number of letters in correct sequence. Letters in correct sequence is preferred for low-functioning students.

c. Within an evaluation system, the direct spelling measure should be administered at least two times per week.

d. The determination of whether to measure performance or progress should be made on the basis of individual student and teacher needs. The two procedures produce similar results.

Data Sources:

- Single subject study (RR 120)
- Direct measure reliability study (RR 109)
- Direct measures norm development (RR 87)
- Comparative study of data-utilization rules (RR 64)
- Teacher efficiency studies (RR 53)
- Development of data-utilization systems (RR 23)
- Technical characteristics of direct spelling measures (RR 21)

Evidence:

Intercorrelations among scores from three test durations (1, 2, and 3 minutes) were all high; further, all test durations demonstrated acceptable concurrent validity with standardized achievement tests (RR 21). Given that limited behavior samples reduce a measure's sensitivity to student growth and that low-functioning students will write few words during a short duration test, it was recommended that the duration of the test be from two to three minutes. Paced dictation at a rate of 15 seconds per word was used in seven different studies with demonstrated validity, reliability, and sensitivity to student growth (RR 21, 23, 87, 109, 120). Given that the behavior sample from low-functioning students probably would be low without pacing, paced dictation was recommended; the 15-second pacing appeared appropriate on the basis of its demonstrated technical adequacy.

Studies on the issue of how to score performance on direct spelling measures consistently found that correct performance scores were more valid and reliable than error scores (RR 21, 109). Both the number of words spelled correctly and the number of correct letter sequences showed high correlations with standardized achievement tests (RR 21). In addition, interscorer reliability was very high for both types of scores (RR 87, 109). However, correct letter sequence scores were found to be more sensitive to student growth than correct word scores (RR 87).

The issue of the frequency with which the direct spelling measure should be given in an evaluation system was addressed by data collected during the development of data-utilization systems (RR 23). Students who were measured in spelling on a daily basis showed greater progress than students who were measured on a weekly basis. Daily measurement is the ideal since seven data points are needed to make program decisions; however, teachers find daily measurement to be cumbersome and time consuming (RR 53). Thus, a compromise of at least two times per week is recommended.

The issue of whether a spelling evaluation system should use progress measurement (in which the measurement domain changes each time a student masters a segment of the curriculum) or performance measurement (in which the measurement domain remains the same) was addressed in a study that compared the effect of the two systems on one student's spelling performance (RR 120). No differences were found in spelling performance as a function of the system. Given that teachers sometimes prefer one system over the other (RR 64), it was recommended that the decision be made on the basis of teacher and student preferences and needs.

Chapter 6

Written Expression Evaluation

This chapter summarizes IRLD research findings related to written expression evaluation. Two specific questions are addressed in this chapter:

- What are the characteristics of a recommended direct measure of written expression?
- How should the direct written expression measure be administered and scored?

For each question, the major findings are summarized and the data sources from which the findings were obtained are listed (generally ordered in terms of recency). Specific evidence for the major findings then is presented.

9. What Are the Characteristics of a Recommended Direct Measure of Written Expression?

Findings:

-- A direct measure of written expression should focus on the behavior of writing compositions in response to a verbal stimulus. Certain measures of this behavior (total words written, total words spelled correctly, or letters in correct sequence) are technically adequate (valid, reliable, and sensitive to student growth), have instructional utility, and are logistically feasible in the classroom.

Data Sources:

- Norming study (RR 132)
- Comparative study of standardized and direct measures (RR 126)
- Direct measure reliability study (RR 109)
- Longitudinal study of learning trends on simple measures RR 49)
- Technical characteristics of direct written expression measures (RR 22)

Evidence:

A series of studies demonstrated that story starters, topic sentences, and picture stimuli could be used to collect written compositions from students (RR 22). When compositions obtained from these approaches were scored in terms of total words, words spelled correctly, and correct letter sequences, correlations between scores on the direct measures and standardized achievement tests were high. Internal consistency reliability also was high for all three. However, since pictorial stimuli generally are more expensive to produce and are less easily incorporated into a response form, verbal stimuli are preferred. Both story starters and topic sentences may be printed at the top of lined paper to allow students to look at the stimulus as well as listen to it.

Comparisons of story starter performance in terms of words written and correct letter sequences, across grades and within grades, indicated that both measures demonstrated adequate sensitivity to student growth (RR 49, 132). Further, the direct measures of written expression were found to be much more sensitive to pupil progress over 10 weeks than a standardized test, on which virtually no growth was evident (RR 126). Test-retest, alternate-form, and interjudge reliabilities generally were quite high when correct performance was scored (RR 109), although in some studies reliability has been below .70 (RR 132).

10. How Should the Direct Written Expression Measure be Administered and Scored?

Findings:

a. The duration of a direct written expression measure should be three minutes each time it is administered.

b. Performance on a direct written expression measure should be scored in terms of either total number of words or number of correctly spelled words.

c. Within an evaluation system, two or three writing samples should be elicited on each measurement occasion.

Data Sources:

- Direct measure reliability study (RR 109)
- Aggregation study (RR 94).
- Direct measures norm development (RR 87)
- Comparative study of written expression scoring procedures (RR 84)
- Reliability of written expression measures (RR 50)
- Longitudinal study of learning trends on simple measures RR 49)
- Technical characteristics of direct written expression measures (RR 22)

Evidence:

Correlations between performance on the direct written expression measure and a developmental sentence score at the end of three, four, and five minutes were all high (RR 22). The three-minute samples of writing produced the widest range of scores. Use of a three-minute duration in other studies produced data that were very sensitive to student growth across and within grade levels (RR 22, 87).

Comparisons of six scoring procedures (mean T-unit length, mature words, total words written, large words, words spelled correctly, and corret letter sequences) in terms of validity, reliability, and sensitivity to student growth indicated that three (total words written, words spelled correctly, and correct letter sequences) had the greatest technical adequacy (RR 22, 50). Scores of mature words, total words written, words spelled correctly, and correct letter sequences correlated significantly with standardized written expression measures (RR 22) and evidenced good test-retest and parallel-form reliability. Discriminative validity with respect to grade levels also was demonstrated (RR 49, 87). However, since mature words is more difficult to score, and correct letter sequences is quite time consuming, scoring either total words written or number of correctly spelled words was recommended. Scoring of correct performance is recommended since the reliability of incorrect performance is too low for it to be used in educational decision making. (RR 109). Inter-judge agreement in scoring total words written, words spelled correctly, and correct letter sequences was very high (RR 84).

Low test-retest and parallel-form reliability coefficients were found for single written expression samples (RR 50). Aggregating three writing samples and using the mean score resulted in acceptable reliability (RR 94). On this basis, it was recommended that at least two, and preferably three, writing samples should be elicited on each measurement occasion.

Chapter 7

Oral Language Evaluation

This chapter summarizes IRLD research findings related to oral language evaluation. Two specific questions are addressed in this chapter:

- What are the characteristics of a recommended direct measure of oral language?
- How should the direct oral language measure be administered and scored?

For each question, the major findings are summarized and the data sources from which the findings were obtained are listed (generally ordered in terms of recency). Specific evidence for the major findings then is presented.

11. What Are the Characteristics of a Recommended Direct Measure of Oral Langauge?

Findings:

-- A direct measure of oral language should focus on the behavior of describing a picture stimulus.

Data Sources:

- Study of expressive language (RR 83)

Evidence:

An initial investigation of the relationship between a direct measure of oral language and more elaborate, psychometrically adequate measures of the quality of language (semantic/syntactic complexity and descriptive accuracy scores) indicated that certain measures of children's picture descriptions (number of non-repetitive words) were highly correlated with the more elaborate methods of analyzing language samples (RR 83). Concurrent validity of the direct oral language measure was supported by correlations between .89 and .97 with the semantic/syntactic complexity and descriptive accuracy scores. Additional research is needed on other technical characteristics (reliability, sensitivity to student growth) of a direct oral language measure.

12. How Should the Direct Oral Language Measure be Administered and Scored?

Findings:

a. Performance on a direct oral language measure should be scored in terms of the number of non-repetitive words spoken.

b. The oral language measure should be administered by a familiar examiner.

Data Sources:

- Study of expressive language (RR 83)

Evidence:

When children's oral language samples were scored in terms of the number of non-repetitive words spoken, high correlations (.89 to .97) were found with psychometrically adequate and more complicated measures of semantic/syntactic complexity and descriptive accuracy scores (RR 83). In addition, both the quality and quantity of spoken language was greater when the tester was familiar rather than unfamiliar, suggesting that optimal performance will be obtained by a familiar examiner.

Chapter 8

Mathematics Evaluation

This chapter summarizes IRLD research findings related to mathematics evaluation. Two specific questions are addressed in this chapter:

- What are the characteristics of a recommended direct measure of mathematics?
- How should the direct mathematics measure be administered and scored?

For each question, the major findings are summarized and the data sources from which the findings were obtained are listed (generally ordered in terms of recency). Specific evidence for the major findings then is presented.

13. What Are the Characteristics of a Recommended Direct Measure of Mathematics?

Findings:

-- Preliminary data suggest that a direct measure of mathematics should focus on the calculation of math computation problems.

Data Sources:

- Norming study (RR 132)
- Direct measure reliability study (RR 109)

Evidence:

A study of the test-retest reliability, alternate-form reliability, and interjudge reliability indicated that correct performance scores on most computation problems was good (RR 109). Interjudge reliability was very high across all types of problems (.90 to .99) and test-retest reliability was good (.78 to .93), but alternate-form reliability was only moderate on addition, subtraction, and multiplication (.61 to .72) and low on division (.48). In a local norming study, the alternate-form correlation was low for both multiplication (.61) and division (.48) (RR 132). Although math measures used in the local norming study showed grade level differences, they did not always reflect higher performance by older students. However, the measurement task in that study did vary for different grades in some cases. Additional research is needed on sensitivity to student growth and other technical characteristics (e.g., validity) of a direct mathematics measure. Such research may lead to refinement of the recommended direct measure of mathematics.

14. How Should the Direct Mathematics Measure be Administered and Scored?

Findings:

a. The types of problems presented to a student may be determined by the grade level of the student or may sample from all types of math functions.

b. Performance on a direct mathematics measure should be scored in terms of the number of digits correct.

c. Within an evaluation system, several samples should be elicited on each measurement occasion.

Data Sources:

- Direct measure reliability study (RR 109)

Evidence:

When students in grades 4 and 5 were tested on math problems limited according to their grade level, most reliability coefficients were in an acceptable range (RR 109). Only interjudge reliability (.93) and test-retest reliability (.93) were calculated for a single measure that included all math functions. Additional data are needed before a specific recommendation can be made as to the scope of problems included in a direct measure of mathematics.

Reliability data clearly indicated that correct performance on math problems should be scored (RR 109). While correct performance scores produced good to high reliability coefficients, incorrect performance scores often produced very low reliability coefficients (e.g., .09). The correct performance scores were calculated by counting the number of digits correct; a digit was considered correct if it appeared in the correct place within the answer.

Alternate-form reliability coefficients for direct mathematics measures sometimes were lower than desirable (e.g., division = .48), suggesting that several alternate forms should be administered on each testing occasion (RR 109). The student's score would be an average of the scores on the repeated administrations.

Chapter 9

Social Adjustment Evaluation

This chapter summarizes IRLD research findings related to mathematics evaluation. Two specific questions are addressed in this chapter:

- What are the characteristics of a recommended direct measure of social adjustment?
- How should the direct social adjustment measure be administered and scored?

For each question, the major findings are summarized and the data sources from which the findings were obtained are listed (generally ordered in terms of recency). Specific evidence for the major findings then is presented.

15. What Are the Characteristics of a Recommended Direct Measure of Social Adjustment?

Findings:

-- A direct measure of social adjustment should focus on general classroom conduct and social interaction. The specific behaviors should be identified within the specific setting of interest.

Data Sources:

- Study of variables influencing direct social adjustment measures (RR 82)
- Technical characteristics of direct social adjustment measures (RR 24)
- Measuring classroom behavior (RR 6)

Evidence:

Observational studies of behaviors that index social adjustment indicated that the specific behaviors associated with social functioning variables vary with the specific setting, and to some extent with the sex of the student (RR 24, 82). An initial study revealed that the degree of discrepancy between the rate of a target student and his or her peers on several specific measures (noise, out of place, physical contact or destruction, off task) agreed with teachers' identifications of problem students (RR 6). Another study suggested that either the frequency of occurrence of peers talking with the target child or the number of different peers talking with the target child was a valid indicator of social status (RR 24). In a third study, only the frequency of occurrence of peers talking with the target child reliably correlated with social status (RR 24).

An extensive study of behaviors that correlated with social functioning (both social status and teacher-perceived behavior problems) suggested that peer behavior toward the target student correlated with social status and the target student's behavior (e.g., aggression) correlated with teacher ratings (RR 82). However, in this study, differences in the patterns of correlations existed between boys and girls across settings. Peer approaches related consistently to social status for boys in both structured academic and unstructured non-academic settings, but only in an academic setting for girls. Problem behaviors clearly related to teacher ratings of girls in academic settings, but aggression was the consistent predictor of teacher ratings of boys in the same settings. Because of the pervasive influence of setting, it was recommended that data be collected on the target student's general classroom conduct and social interaction and on a classmate's general classroom conduct and social interaction.

At this point, it appears that the rate of student initiations is a prime behavior to monitor. A discrepancy between the two students would provide a basis for monitoring the target student's social adjustment.

16. How Should the Direct Social Adjustment Measure be Administered and Scored?

Findings:

a. Administration of the direct social adjustment measure could involve observation of the target student and classmates on an interval-sampling schedule.

b. Performance could be scored by tallying occurrences of the target behaviors.

Data Sources:

- Study of variables influencing direct social adjustment measures (RR 82)
- Technical characteristics of direct social adjustment measures (RR 24)

Evidence:

During investigations of the technical adequacy of various measures of social adjustment, a 60-second observation interval was used to collect data on the occurrence of specific behaviors (RR 24, 82). This schedule could be applied to a situation where the target student and one classmate would be observed during alternate intervals to obtain a measure of target student discrepancy from a peer. During each observation interval, behavior was coded simultaneously on different categories; two behaviors within a category were coded during one interval if a 5-second break clearly separated the behaviors. Additional research is needed to establish the logistical feasibility of this procedure and its utility for interventions.

Chapter 10

Data Utilization

This chapter summarizes IRLD research findings related to the use of data collected on students to make decisions regarding pupil progress and program success. Four specific questions are addressed in this chapter:

- What are recommended procedures for graphing data?
- How should graphed data be used to evaluate students' programs?
- How should teachers be trained to use data for judging intervention effectiveness and improving student performance?
- To what extent does measurement and data utilization by teachers affect students' learning?

For each question, the major findings are summarized and the data sources from which the findings were obtained are listed (generally ordered in terms of recency). Specific evidence for the major findings then is presented.

17. What are Recommended Procedures for Graphing Data?

Findings:

a. Correct performance should be graphed. Incorrect performance may also be graphed along with correct performance to provide information about accuracy of performance.

b. When graphing a student's level of performance, equal interval graph paper should be used rather than semi-logarithmic chart paper.

c. When graphing a student's reading or spelling progress through a curriculum, number of words spelled or pages read should be spaced along the ordinate axis according to the time of mastery expected of average students in the curriculum.

d. Students may be taught to chart their own performance to increase teacher efficiency and facilitate student satisfaction.

Data Sources:

- Comparison of student self-management techniques (RR 115)
- Direct measure reliability study (RR 109)
- Comparative study of graph papers (RR 101)
- Aggregation study (RR 94)
- Direct measures norm development (RR 87)
- Comparative study of written expression scoring procedures (RR 84)
- Comparative study of three reading placement procedures (RR 57)
- Reliability of written expression measures (RR 50)
- Technical characteristics of direct written expression measures (RR 22)
- Technical characteristics of direct spelling measures (RR 21)
- Technical characteristics of direct reading measures (RR 20)

Evidence:

Studies in the area of reading, spelling, and written expression consistently have indicated that correct performance has greater technical adequacy than does incorrect performance (RR 20, 21, 22, 50, 57, 84, 87, 94, 109). Thus, graphing of data should focus on correct performance.

Studies of graphing procedures within performance measurement have examined the relative merits of equal interval and semi-logarithmic graph paper (RR 101). Analyses of deviations between actual scores and scores predicted from graphs on each type of paper indicated that predictions were more accurate when data had been graphed on equal interval paper.

Within progress measurement, a critical problem is the lack of equal intervals from one curriculum unit to the next. It appears that the technical adequacy of progress measurement might be improved if the system were conceptualized as progress through pages read or words spelled of a curriculum, with the number of pages or words spaced along the ordinate axis according to the time of mastery expected of average students in the curriculum. Research should be conducted to examine the validity of this assumption.

Although having students graph their own performance data does not necessarily result in increased student achievement (RR 115), it may facilitate increased student satisfaction and reduce teacher time for evaluation activities. Evidence suggests that by increasing student responsibility in charting tasks, increased student achievement also can be attained.

18. How Should Graphed Data be Used to Evaluate Students' Programs?

Findings:

a. Graphed data should be summarized and interpreted to determine whether the instructional program is effective or needs to be changed.

b. Goal-oriented analysis is preferred for monitoring progress toward IEP goals, obtaining information about when to change a student's instructional program, and explaining student progress to parents and other teachers.

c. Program-oriented analysis is preferred for obtaining information about what to change in a student's instructional program.

d. A combined goal-oriented and program-oriented procedure that is recommended involves drawing a trend line through 7 to 10 data points; if the trend is flatter than the goal line, a program modification should be introduced.

e. Data obtained from several students can be used to make decisions regarding general program components.

Data Sources:

Analysis of statistical properties of data (RR 125, 138)

- Norming study (RR 132)
- Evaluation of program effectiveness (RR 123)
- Assessment of alternative data summary procedures (RR 112, 118)
- Comparative study of data utilization rules (RR 64)
- Comparative study of teacher goals (RR 61, 62)
- Analysis of program components (RR 12)
- Demonstration study of data utilization (RR 10)

Evidence:

When teachers summarize student data and implement data-utilization rules, student performance increases more than when data-utilization does not occur (RR 10) or when constant efforts are made to improve the student's performance without data-utilization procedures (RR 64).

Two basic procedures may be used to summarize student data: visual analysis or statistical analysis. One investigation of visual analysis (RR 112) revealed that it is not very reliable for evaluating educational programs, and that it is influenced considerably by the characteristics of the data array (e.g., slope and variability). Another study (RR 118) indicated that the relationship between results of visual analysis and statistical analysis procedures was modest at best. In other words, many interventions were judged visually to be significant in their effects when they were not statistically significant, and vice versa. Further research has suggested that specific factors can affect the accuracy of visual analysis (RR 125). Given that training in statistical analysis and access to statistical programs is limited for most teachers, it is likely that graphed data generally will be analyzed visually. Thus, training becomes

especially important. Initial research has indicated that training can increase the accuracy of visual analysis judgments (RR 118). Additional research is needed to identify specific training components that can increase even further the accuracy of visual analysis for summarizing student data. For example, an analysis of statistical properties of data has suggested that some training should focus on the interactions among time-series characteristics when making judgments based on visual inference (RR 138).

In analyzing graphed data, teachers who used both goal-oriented procedures (set the goal and a data on which it is to be reached, draw a goal line, and then compare student performance trends to the goal line) and program-oriented procedures (test student performance frequently and change program when it appears needed or after a specified number of tests, usually 7-10) reported that they preferred the goal-oriented approach for (a) monitoring student progress toward IEP goals, (b) obtaining information about when to change a student's instructional program, and (c) explaining student progress to parents and teachers (RR 64). They also indicated that the goal-oriented approach was easier to use, more efficient, and a more accurate representation of student performance. The program-oriented approach was preferred only as a guide for what to change in a student's instructional program.

Teachers also were more accurate in summarizing data when using goal-oriented procedures (47% correct summarizations) than when using program-oriented procedures (12% correct summarizations) (RR 64). Further, the timing of changes in students' programs was more accurate in goal-oriented analysis (70%) than in program-oriented analysis (33%). In another study, teachers who used goal-oriented analysis made more correct decisions about whether to change a student's program (79%) than teachers who used program-oriented analysis (68%) (RR 61). Teachers also judged effective interventions more accurately when they applied goal-oriented procedures (100%) than when they applied program-oriented procedures (80%) (RR 61). Further, teachers believed they were more effective when using a goal-oriented approach than when using a program-oriented approach, even though there actually were no student performance differences (RR 61, 62).

Program component research is a viable outcome of data collected through direct and frequent measurement procedures. An illustration of this approach in a Child Service Demonstration Center for Children with Learning Disabilities (RR 12) indicated that it could provide immediate payoff for decision makers and could be used to identify effective intervention variables within a program. System-wide data collection also has revealed that the data-based assessment approach offers not only a measurement alternative for the student, but a comprehensive reviewing procedure that is sensitive to the needs and problems of the school system as a whole (RR 123). Local normative data can be obtained by sampling as few as 20 students per grade, with the result providing a median for normative comparisons that is very close to that of the entire class or to an estimated true median (RR 132).

19. How Should Teachers Be Trained to Use Data for Judging Intervention Effectiveness and Improving Student Performance?

Findings:

a. Direct inservice or workshop training, rather than self instruction, is recommended for training teachers to collect data frequently and to use the data to make instructional decisions.

b. Systematic procedural changes can increase teachers' efficiency in using direct and frequent measurement procedures.

c. Direct training of teachers in measurement activities is more likely to result in teacher use and efficiency than training through manuals alone.

d. Goal setting is integral to progress measurement activities; teachers should monitor student performance in relation to short-term objectives rather than long-term goals.

e. Direct and frequent measurement with curriculum-based tests can increase the reliability of scores and may provide the best measure for determining reading placement.

Data Sources:

- Experimental study of formative evaluation effects (RR 88, 96, 97, 111, 116)
- Comparative study of data-utilization rules (RR 64)
- Study of self-instructional training (RR 63)
- Comparative study of teacher goals (RR 61, 62)
- Comparative study of three reading placement procedures (RR 56)
- Teacher efficiency studies (RR 53)
- Interviews of special educators (RR 41)
- Development of data utilization systems (RR 23)

Evidence:

An early study of data utilization provided participating teachers only with minimal training (1½ hrs) in data analysis for making program decisions (RR 23). Results indicated that, in general, teacher use of decision rules was more effective than teacher judgment in improving student performance. These tenuous findings suggested that a fruitful area of research, with respect to the development of an effective evaluation system, would be to test alternative data utilization approaches that involve more intensive teacher training in monitoring and evaluating student progress. In fact, when interviewed one year after the study, several teachers indicated the need for more intensive and relevant training, including modeling of the procedures (RR 41).

Teachers have been trained via (a) a week-long workshop prior to start-up in the fall and semi-weekly workshops throughout the school year (RR 63), (b) a self-instructional manual plus four workshops (RR 64), and (c) training of district personnel who in turn directly train teachers using the self-instructional manual (RR 88, 96, 97). Regardless of training procedure, teachers have had difficulty using the data systematically. However, direct training of teachers was more effective in promoting efficiency in measurement tasks (RR 53). For this reason, direct training is recommended. Increased attention to data utilization during training is needed. Imprecise teacher implementation of measurement and decision rules has been observed even in controlled studies (RR 61, 88, 111, 116), thus emphasizing the need for on-going training in measurement, graph interpretation, and data utilization procedures.

A series of studies on teacher efficiency in employing direct and frequent measurement strategies indicated that teachers initially required 13½ minutes to prepare, administer, score, and graph measurement tasks on four academic behaviors for one student; the number of students on teachers' caseloads made the time commitment burdensome (RR 53). In addition, this time commitment did not include the time needed to read and analyze graphs, important tasks if the data are to be employed meaningfully to improve student progress. Procedural changes, such as administration of reading and spelling tasks prior to the written expression task, and measurement at the beginning of the period, resulted in greater teacher satisfaction. Other factors suggested as effective in increasing teacher efficiency were precounting the words in oral reading passages, group administration of tasks, the use of mechanical devices to administer the measures, and student graphing of measurement results. Direct training of teachers in efficient procedures was more effective than training via a self-instructional manual and periodic inservices. Teachers appeared to need prompting to improve their efficiency with direct and frequent measurement strategies.

In a comparative study, special education teachers who had set long-term goals for students, graphed students' word recognition performance, and made a program adjustment every two weeks, predicted that students would master a greater number of words than did teachers who set short-term objectives and compared graphed student performance with a short-term aimline (RR 62). The predictions of the teachers in the short-term goals group were more accurate. However, no actual differences were found in student progress (RR 61). Imprecise teacher implementation of measurement and designated decision rules was observed, suggesting that on-going training of teachers to measure, interpret graphs correctly, and use student data consistently are critical. For example, teachers who were to measure students daily actually measured only three times per week. And, teachers in the long-term goal setting group who were to measure daily only correctly employed their decision rules 56% of the time; teachers who measured weekly did so 78% of the time. Teachers in the short-term goal setting group appropriately moved to new reading lists only two-thirds of measurement days. Obviously, both the use of data-utilization rules and specific training on the rules is an essential dimension of a measurement system effective in improving student achievement.

In a comparative study of three procedures for placing students in reading curricula (teacher judgment, standardized testing, curriculum-based assessment), correlations among the three placement procedures were high but the agreement among scores from the three measures was not (RR 56). Placement from curriculum-based measures agreed better with students' actual reading placements than did norm-referenced test scores. Achievement test scores and curriculum-based placement scores agreed for only about one-half of the students. It is proposed that direct and frequent measurement strategies provide a resolution to this problem. Since curriculum-based measures can be used with any curriculum, and a student's score is calculated as the median of scores on repeated samplings, measurement error may be reduced, resulting in improved accuracy of curriculum-based tests for reading placement.

20. To What Extent Do Measurement and Data Utilization by Teachers Affect Students' Learning?

Findings:

a. Student performance increases more when teachers use specific data-utilization rules to monitor progress than when they rely on their own judgment about student progress.

b. The quality of instruction improves when teachers use direct and frequent measurement and evaluation.

c. Students' knowledge about their goals and progress is greater when teachers employ direct and frequent measurement and evaluation.

d. Measurement appears to be a necessary condition in producing student growth, but not a sufficient one; positive effects of measurement cannot be sustained unless data-utilization procedures also are used.

Data Sources:

- Comparison of student self-management techniques (RR 115)
- Surveys of experimental study participants (114, 124)
- Experimental study of formative evaluation effects (RR 88, 96, 97, 111, 116)
- Instructional rating scale validation (RR 107)
- Implementation study (RR 106)
- Causal model analysis (RR 105)
- Comparative study of data utilization rules (RR 64)
- Analysis of program components (RR 12)
- Demonstration study of data utilization (RR 10)

Evidence:

In a comparative study of data-utilization rules, student reading performance increased more when specific data-utilization strategies were used than when teachers were making a constant effort to improve upon the students' current performance levels without data-utilization procedures (RR 64). Further analyses indicated that time or maturation alone did not explain the increase in student performance from the no data-utilization phase to the first data-utilization phase.

In a demonstration study of the implementation of data-utilization techniques, students exhibited greater reading achievement when rules for the utilization of measurement data were included as part of the formative evaluation system (RR 10). When teachers measured student reading performance daily in relation to daily goals and altered both goals and consequences contingent upon measured student performance relative to goals, superior achievement occurred. In another study, a significantly higher proportion of elementary students attained mastery more rapidly when daily performance was graphed than when it was not graphed (RR 12). Variations in the implementation of direct and frequent evaluation procedures also appear to influence student achievement. For example, in experimental conditions where students selected their instructional activities and then charted their own performance, significant increases in student achievement occurred on both direct and standardized measures (RR 115).

A series of implementation studies indicated that the extent to which a formative evaluation system is implemented may determine the extent to which effects are seen in terms of instructional structure or student achievement (RR 88, 111, 116). The observational scale used to assess instructional structure in these studies was determined to be a useful research tool from the standpoint of technical adequacy and heuristics (RR 107). In the studies, it also was found that the lack of communication among special educators, regular educators, administrators, and parents might be reduced through the use of formative evaluation procedures (RR 114). There was some indication that, even with minimal implementation of the formative evaluation system, students were more aware of working toward a goal and were more optimistic about their progress; teachers also seemed better able to realistically judge their students' progress (RR 124).

A causal model analysis was conducted on the relationships among the degree of implementation of the formative evaluation system, the amount of structure in the students' reading instructional program, and the students' rate of academic progress over one year (RR 105). Causal modeling techniques allow inferences to be made about the logic of directional hypotheses for obtained correlations. Teacher implementation of measurement procedures, student achievement, and degree of teaching structure were found to be stable over time (e.g., if a teacher designed a highly structured program for a student, that student continued to receive highly structured instruction throughout the school year). While measurement had a strong effect on structure and achievement, these effects were short-lived and not evident at the end of the study. Specifically, silent reading practice related to reading achievement gains and the routine of measuring student progress influenced structure; however, the hypothesis that measurement would result in increased structure and student achievement was unsupported. It appeared that measurement activities were important initially in the implementation of data-based modification, but that student achievement gains could be sustained only if evaluation of data occurred.

Another study of the effects of direct and frequent measurement and evaluation on students' reading achievement, teachers' quality of instruction, and students' knowledge about their own goals and progress, strongly supported the use of direct and frequent measurement and evaluation (RR 96, 97). In comparison to 21 teachers who used typical special education evaluation procedures, 18 New York City special education teachers who employed frequent curriculum-based measurement and evaluation procedures (a) affected greater student reading achievement, (b) delivered more structured reading lessons, and (c) were more successful in communicating accurate information to their pupils concerning student goals and progress.

An implementation study also confirmed that teachers generally found that the data obtained from a data monitoring system in reading were useful for tracking student progress (RR 106). Some of the teachers in this study reported that the system was helpful in communicating with parents and teachers.

7

Table 1

Evaluation Research Data Sources

Data Source	Research Reports	Questions
Direct measure reliability study	109	4, 5, 7, 8, 9, 10, 13, 14, 17
Comparative study of data utilization rules	64	3, 5, 8, 18, 19, 20
Direct measures norm development	87	4, 5, 7, 8, 10, 17
Norming study	132	4, 7, 9, 13, 14, 18
Comparative study of teacher goals	61, 62	3, 5, 18, 19
Comparative study of three reading placement procedures	56, 57	4, 5, 17, 19
Longitudinal study of learning trends on simple measures	49	4, 7, 9, 10
Development of data utilization systems	23	5, 7, 8, 19
Experimental study of formative evaluation effects	88, 96, 97, 111, 116	2, 19, 20
Technical characteristics of direct reading measures	20	4, 5, 17
Aggregation study	94	4, 10, 17
Teacher efficiency studies	53	5, 8, 19
Technical characteristics of direct spelling measures	21	7, 8, 17
Technical characteristics of direct written expression measures	22	9, 10, 17
Survey of LD teachers	65, 80	1, 3
Interviews of special educators	41	2, 19
Surveys of experimental study participants	114, 124	2, 20
Comparative study of reading domains and durations	48	4, 5
Comparative study of standardized and direct measures	126	4, 9
Implementation study	106	4, 20
Single subject study	120	5, 8
Reliability of written expression measures	50	10, 17
Comparative study of written expression scoring procedures	84	10, 17
Study of expressive language	83	11, 12
Technical characteristics of direct social adjustment measures	24	15, 16
Study of variables influencing social adjustment measures	82	15, 16
Comparison of student self-management techniques	115	17, 20
Demonstration study of data utilization	10	18, 20
Analysis of program components	12	18, 20
Survey and observation of special ed teachers	81	1
Surveys of special educators	67	2
Comparative study of reading domains	55	4
Study of alternative reading performance criteria	59	4
Study of curriculum differences	93	4
Analysis of readability formulas	129	4
Analyses of basal reader criterion-referenced tests	113, 122, 128, 130	6
Measuring classroom behavior	6	15
Comparative study of graph papers	101	17
Assessment of alternative data summary procedures	112, 118	18
Evaluation of program effectiveness	123	18
Analysis of statistical properties of data	125, 138	18
Study of self-instructional training	63	19
Causal model analysis	105	20
Instructional rating scale validation	107	20

Chapter 11

Data Sources

This chapter provides a summary of the data sources and research procedures used to obtain the research findings presented in the previous chapters. An overview of the data sources is provided in Table 1. The IRLD research reports in which more detailed explanations may be found are listed in the table, as are the numbers of the corresponding research questions. The data sources are ordered within this chapter (and the table) according to the frequency with which they are cited as evidence for various research questions.

Direct Measure Reliability Study (RR 109)

Two separate investigations were conducted to examine the test-retest reliability, alternate-form reliability, and interjudge reliability for direct and repeated measures in the areas of reading, spelling, written expression, and math.

In study I (1979-80), a sample of 566 students (275 males) enrolled in grades 1-6 from three states was administered direct measures of reading (6 one-minute tests) spelling (2 three-minute tests), and written expression (2 three-minute tests). All students were selected randomly from the school districts that volunteered to participate in the study. The students were approximately equally distributed among grades 1-6. Each student was administered the measures during late fall and again during early spring on an individual basis by a trained examiner.

In study II (1981-82), 76 students randomly sampled from grades 4 and 5 were subjects in a math reliability investigation. Thirty students in grade 5 were involved in the test-retest reliability investigation; the 46 students in grade 4 were involved in the alternate-form reliability investigation. Measurement materials in math included computation problems printed on forms. All testing was group administered, with 10 students tested at a time and a one-week interval separating the two testing periods. All testing and scoring was done by trained educational aides. The number of digits correct and incorrect was computed for each math function.

Comparative Study of Data Utilization Rules (RR 64)

Ten special education teachers in a midwestern rural educational cooperative implemented direct and frequent measures and data utilization procedures with at least two students each over the 1980-81 school year. Teaching experience ranged from 0 to 10 years, 8 teachers were female. Students in the study were functioning dramatically below their peers in academic, language, and/or social areas.

The teachers were trained to implement frequent measurement systems during one week of full-day workshops prior to the school year, and in half-day sessions periodically throughout the school year. By February 1981, each teacher was measuring and graphing the students' reading performance at least three times per week. At this time, two data utilization systems, experimental and therapeutic analysis, were introduced to the teachers. In therapeutic data analysis, the teacher's objective was to insure that a student's performance reached a prespecified goal by a certain date. In experimental data analysis, no student performance level and attainment date were specified; rather, the teacher's objective was to improve continuously upon a student's current performance level by introducing and evaluating a series of unending program changes. One half of the teachers implemented experimental teaching and the other half implemented therapeutic teaching; after nine weeks of data collection the teachers switched systems.

Three data utilization strategies (no data utilization, therapeutic, experimental) were compared in terms of their effects on the number of modifications teachers made. Every two weeks, IRLD staff inspected each student's graph and counted the number of instructional changes made. To assess the effect of the data utilization strategies on student performance, every two weeks teachers measured the students' oral reading rate correct on a random list of K-3 words. At the end of the school year, teachers completed surveys regarding their preferences for different measurement strategies.

Direct Measures Norm Development (RR 87)

During 1979-80, direct measures of reading, spelling, and written expression were administered to 566 elementary students from three states in order to (a) investigate the feasibility of using a standard task to measure the reading, spelling, and writing proficiency of elementary children, and (b) describe procedures for establishing local norms on the standard tasks. The grade 1-6 students from Minnesota, Pennsylvania, and Washington were selected randomly from school districts that volunteered to participate in the study. There were 275 males and 291 females in the total sample, which included 92 first graders, 85 second graders, 96 third graders, 99 fourth graders, 101 fifth graders, and 93 sixth graders.

The Minnesota sample consisted of 134 of the 566 students, 63 boys and 71 girls. Most of these subjects (73%) were selected from two urban areas with populations of 50,000 and 100,000 people. These elementary students were approximately equally distributed among grades 1 to 6. The Pennsylvania sample of students included 157 boys and 169 girls, equally distributed across the six grade levels. These elementary students were randomly selected from two areas (rural and urban) in Central Pennsylvania. The remaining 106 elementary students tested were from the Seattle, Washington area; 55 were male and 51 were female.

Each child was administered direct measures of reading, spelling, and written expression during the fall and the spring on an individual basis by an examiner trained in the administration of the measures. Data were examined in terms of grade level differences, annual growth, stability over time, and state, demographic, and sex differences.

Norming Study (RR 132)

During 1982-83, fall, winter, and spring local norms for student performance on direct measures of reading, spelling, math, and written expression were developed. Samples of regular education students from six school districts were asked to (a) read aloud from two basal reading passages, (b) spell words from a dictated word list taken from either a spelling series or a reading series, (c) complete math problems in addition, subtraction, multiplication, and division, and (d) complete a written composition in response to a story starter.

A total of almost 1800 students participated in this local norming, with approximately equal numbers from each grade (1-6). Data were summarized on the effect of using different measurement sampling plans, the reliability of the measures, and the distribution of scores within a grade level. Also, the effects of different population sampling plans were analyzed. The local norms also were compared to national norms and to the effects of the norms on the percentages of students served.

Comparative Study of Teacher Goals (RR 61, 62)

During 1979-80, 20 special education resource teachers from a midwestern metropolitan area participated in a 12-week study to examine the effects on student reading achievement of (a) goal size and data-utilization rule, and (b) measurement frequency. The majority of teachers were female; they had an average of 9.6 years teaching experience. Each teacher selected four to six students from his/her caseload, resulting in a student sample of 88 boys and 20 girls. The students' mean age was 10.3 years; their mean grade level was 3.9.

Teachers were assigned randomly to one of two experimental treatment groups for the purpose of measuring student progress: Long-Term Goal Measurement (LTGM) or Short-Term Goal Measurement (STGM). In LTGM, teachers tested students' oral reading performance by administering a 30-second word recognition test comprised of 25 words randomly selected from the large set of words to be introduced within the 12-week study. Teachers in this condition were required to make an instructional intervention every 10 days. In the STGM group, teachers tested a student's reading performance by administering a 30-second word recognition test comprised of 25 words that included vocabulary words introduced in the current instructional period plus words sampled from preceding stories. Teachers compared the student's performance against a short-term aimline related to the current short-term goal and made program adjustments accordingly. Both groups of teachers randomly assigned their students to one of three frequency of measurement conditions: daily, weekly, or pre-post measurement. During the first, seventh, and twelfth weeks of the study, teachers administered curriculum-based measures (both word recognition and oral reading passages) to all students in the study.

Teacher decision-making information was assessed weekly through the use of an interview checklist. Specific questions related to how, why, and when program adjustments were made and teacher re-estimates of long-term and short-term goals. Teachers also rank ordered the five most effective student program changes for each student from among eight instructional, eight motivational, and eight administrative and physical arrangement alternatives. These rankings occurred after the 3rd, 6th, 9th, and 12th weeks of the study.

Comparative Study of Three Reading Placement Procedures (RR 56, 57)

Two comparative studies involving the accuracy of reading placements were conducted during 1980-81 with 91 randomly selected students, distributed across grades 1-6 in one midwestern metropolitan public elementary school. All students were English speaking, 15 students received special education resource service, and 23 were enrolled in Title I programs for children who were "seriously behind" in reading.

In the first study, the correlations and agreements among scores on curriculum-based measures, scores on technically adequate achievement tests, and teacher judgments (actual placements) were investigated. Five trained examiners administered two standardized subtests and 10 reading passages during an hour session. The reading passages were administered for one minute each in a random order, following systematic procedures. Seven instructional criteria (e.g., 70 wpm with 10 or fewer errors) were applied to the scores from these passages; the students' placement score was the highest level at which a criterion was met before unsatisfactory performance on two consecutive levels.

In the second study, the concurrent validity of curriculum-based reading measures was examined for two basal reading programs. The measures and procedures employed were identical to the first study with one exception. In this study, two reading series were involved, resulting in a total of 19 reading passages. For each passage, the seven different instructional criteria were applied to the students' scores. An instructional level was identified as the highest level at which the criterion was met before an unsatisfactory performance was demonstrated on two consecutive levels.

Longitudinal Study of Learning Trends on Simple Measures (RR 49)

During 1979-80, 58 children randomly selected from the elementary schools of a small midwestern city were tested on direct measures of reading, spelling, and written expression. The grade 1-6 students ranged in age from 6.3 years to 12.2 years. None of the students was receiving special education services. The direct measures used were those described in RR 20, 21, and 22. All measures were administered in the fall, winter, and spring of the school year by researchers.

Development of Data-Utilization Systems (RR 23)

During the 1978-79 school year, the effectiveness of direct and frequent spelling measurement procedures were investigated in LD resource programs. Twenty-two volunteer special education resource teachers and 80 grade 2-6 students receiving spelling instruction participated in the study. The students, who were of low to middle income SES within a large metropolitan area, were at least two years below age/grade placement in spelling achievement. The number of students per teacher ranged from two to seven. Three-quarters of the students were male; most teachers were female. All teachers had taught special education for a minimum of three years.

Prior to the first experimental period, each student's spelling performance was assessed on grade specific tests. Instructional placements were determined by the rate of letter sequences spelled correctly. Teachers organized a 15-minute daily spelling period using 100 words from a level-appropriate grade specific test; they were responsible for all instructional decisions, such as the number of words to introduce daily, and were encouraged to change the instructional program as needed. To facilitate program change, a list of 12 spelling interventions was distributed to each teacher in a checklist format. Teachers were instructed to check the strategies they used for each student daily.

Three different formative evaluation systems were designed and implemented as treatments. Teachers were assigned randomly to use one system for a three-week period; they were trained during a 1½ hour workshop.

In the first system, *daily measurement and data-based rules* (DMDB), teachers taught for 10 minutes and used the remaining five minutes for testing. A weekly spelling goal was established and teachers used an aimline to indicate the need for an instructional intervention. If the student's performance fell below the aimline for three consecutive days, the teacher drew a new aimline and implemented a different teaching strategy. If performance was above the line, no new teaching strategy was implemented.

In the second treatment, *daily measurement and teacher judgment* (DMTJ); the same measurement procedures were used as in the DMDB treatment, however, rules were not specified regarding when to change teaching strategies. Teachers graphed student performance and were asked to judge whether the students' progress was sufficient to continue using the same teaching methods, or whether a new teaching strategy would increase performance.

In the third treatment, *weekly measurement and teacher judgment* (WMTJ), measurement of spelling performance occurred only once during the week and the students' scores were recorded in a grade book. The teacher judged the need for an instructional program change consistent with the guidelines from the DMTJ condition.

For a second three week period, half of the teachers in each treatment were randomly reassigned to one of the other two treatments and were again trained in the procedures. Thus, at the conclusion of the study, each teacher had participated in two of the three treatments. Students were tested by researchers before the study and after each experimental period on three grade specific tests and a grand master test that included words from all elementary grade levels.

Experimental Study of Formative Evaluation Effects (RR 88, 96, 97, 111, 116)

An experimental-control comparison was conducted during 1981-82 to determine the effects of training teachers in the use of continuous direct measures in reading on student achievement and the structure of the learning environment. The subjects included three different samples; these are described below. After extensive training in the use of direct measurement procedures, teachers were directed to measure experimental students daily using one-minute timed samples of reading from the student's curriculum, to develop IEP long-range goals and short-term objectives, and to use the data to evaluate the instructional program, over the entire school year. Visits by observers and frequent phone contacts provided feedback to the teachers on the accuracy of their implementation of the measures.

Both experimental and control subjects were administered two achievement measures (timed samples and subtests from a standardized test) and the Structure of Instruction Rating Scale. In addition, the Accuracy of Implementation Rating Scale was completed for experimental subjects. The Structure of Implementation Rating Scale (SIRS) was designed to measure the degree of structure of the instructional lesson that a student received. The observers rated 12 factors on a scale of 1 (low) to 5 (high). Inter-rater agreement was high (.92); in addition, the reliability of the SIRS as indicated by measures of homogeneity was .86. The Accuracy of Implementation Rating Scale

(AIRS) was designed to assess the degree of implementation of the continuous direct measures. The AIRS consisted of 12 items rated on a 1 (low) to 5 (high) scale. Parts of the scale require direct observation whereas other items on the checklist are completed by inspection of student reading graphs and reading IEP forms. The reliability of the AIRS as indexed by internal consistency of items was .62, which is adequate for research purposes.

Sample 1 (RR 88, 116). The subjects were 40 grade 1-8 students in a rural educational cooperative, representing 20 experimental-control matched pairs. Three fourths of the students were boys and the mean grade level of the students was 3.8. All subjects were functioning dramatically below their peers in reading. The students were studied in the resource room setting; their teachers were seven special education resource teachers whose experience ranged from two to six years.

Sample 2 (RR 96, 97). A total of 39 special education teachers and their students, from a large urban school district in the eastern part of the U.S., participated in the study. Most of the teachers were female; students selected from their caseloads read about three years below grade level (fifth grade). Students were in programs for the emotionally handicapped, or the brain-injured, or were placed in resource rooms.

Sample 3 (RR 111, 116). The subjects were 38 elementary grade 1-6 students in a suburban school district. Most of the students (84%) were male.

Technical Characteristics of Direct Reading Measures (RR 20)

Three concurrent validity studies of direct reading measures were conducted during 1978-79 in order to examine (a) relationships between the direct measures and standardized achievement measures, (b) resource vs regular program differences in student performance, and (c) grade level differences in student performance.

In the first study, 18 regular class students and 15 LD program students in grades 1-5 from a suburban public school were tested on five direct measures of reading (words in isolation, words in context, oral reading, cloze comprehension, and word meaning) and two standardized measures (Stanford Diagnostic Reading Test, Woodcock Reading Mastery Tests). In the second study, 27 regular students and 18 LD program students in grades 1-6 from two urban public schools were tested on the same five direct measures as used in Study I, but with some minor modifications made in them. No standardized tests were used in Study II. In the third study, 43 regular students and 23 LD program students in grades 1-6 from three urban schools were tested on four direct measures of reading (third-grade word list, sixth-grade word list, third-grade oral reading passage, sixth-grade cloze passage) and three standardized measures (Phonetic Analysis and Reading Comprehension subtests of Stanford Achievement Test and Reading Comprehension subtest of Peabody Individual Achievement Test).

Aggregation Study (RR 94)

The effects of aggregation on the reliability of measures of academic performance were explored in two studies during 1980-81. In the first study, subjects were 30 elementary-age students randomly selected from a pool of 90 students involved in another study. The students were all English speaking and attended a midwestern metropolitan school. The students were tested four times on the same forms of a reading passage measures and a standardized achievement test. Group stability coefficients, within-subject reliability coefficients, and group correlations between variables each were calculated on the basis of one or two testings and then on the basis of aggregations over four testings.

In the second study, 78 children in grades 3-6 who were described as "high-risk" for receiving special education services, were tested 10 times on alternate forms of two direct reading measures and one written expression measure. Once per week over a 10-week period, students read aloud words for one minute; two measures of reading, words read correctly per minute and errors per minute, were scored. During each testing session, a writing sample was obtained. Each student was presented with an alternate form of a story starter and required to write on the story topic for three minutes. The number of correctly spelled words was scored. Group stability coefficients were calculated on the basis of 2, 4, 6, 8, and 10 testings.

Teacher Efficiency Studies (RR 53)

A series of studies examined teacher efficiency in employing repeated curriculum-based measurement. The studies involved a group of 10 special education teachers in a midwest rural educational cooperative (see p. 1). In addition, five female teachers in a suburban school district served as a contrast group.

Dependent measures included teacher efficiency (teacher time and student transition to task time) and teacher satisfaction. Teacher time data were obtained through observations; student transition to task time was estimated by teachers on a self-report questionnaire. Teacher satisfaction was measured using two self-report surveys: the first measured teacher satisfaction with the efficiency modifications immediately following the experimental phases and the second obtained information on actual teacher practices several weeks following experimental phases.

After training teachers to organize, administer, score, and graph academic measures, teachers' efficiency in using procedures and the reliability of self-observation was measured. Teachers administered the measurement tasks in any order they preferred. During the following week, teachers administered the tasks to the same student in a prescribed order (reading, spelling, then written expression). The prescribed order was designed to allow teachers to use the students' response time for the written expression task to score and graph previously administered tasks. Efficiency also was assessed as a function of when measurement occured. In week one, teachers administered the three measurement tasks at the middle or end of the instructional period. During the next week, the teachers administered the tasks as soon as the student entered the room. In addition to recording the amount of time taken, the teachers completed a teacher satisfaction survey.

After obtaining the results from these comparisons, teachers selected ways in which they would try to increase their efficiency. These were studied in 8 single case studies using an ABA reversal design. Each phase lasted about two weeks during which time approximately six data points were collected.

One year after their original training, the efficiency of the 10 teachers was compared to that of five suburban teachers who had been trained via a self-instructional manual and periodic inservices, and who had not been systematically prompted to improve their efficiency. Both groups monitored their own measurement activities to arrive at a time representative of their end-of-the-year efficiency.

Technical Characteristics of Direct Spelling Measures (RR 21)

Three concurrent validity studies of direct measures of spelling were conducted during 1978-79 in order to examine (a) relationships between the direct measures and standardized achievement measures, (b) resource vs regular program differences in student performance, (c) grade level differences in student performance, and (d) various scoring procedures, time limits, and word lists.

In the first study, 27 regular class students and 15 LD program students in grades 2-6 from two urban public schools were tested on two direct spelling measures (3 dictated word lists and a picture stimulus written sample) and one standardized measure (Test of Written Spelling). In the second study, 35 regular students and 10 LD program students in grades 2-6 from two different urban public schools were tested on four word lists (selected from various grade levels) and the spelling section of the Peabody Individual Achievement Test. In the third study, 32 regular students and 29 LD program students in grades 2-6 from two urban public schools and four urban parochial schools were tested on four word lists (3 of which had been used in Study II, plus one developed by selecting randomly from a basal reading series) and the spelling section of the Stanford Achievement Test.

Technical Characteristics of Direct Written Expression Measures (RR 22)

Three concurrent validity studies of direct measures of written expression were conducted during 1978-79 in order to examine (a) relationships between the direct measures and standardized achievement measures, (b) resource vs. regular program differences in student performance, (c) grade level differences in student performance, and (d) various scoring procedures.

In the first study, 16 regular class students and 12 LD program students in grades 3-6 from two urban schools were given two direct measures of written expression (story starter and picture stimulus) and one standardized measure (Test of Written Language). Six scoring procedures were applied to the written samples (T-unit length, mature words, large words, words spelled correctly, total words written, and rates of words written). In the second study, 24 regular class students and 28 LD program students in grades 3-6 in one urban public school were tested on three direct measures (story starter, picture stimulus, and topic sentence) and two standardized measures (Test of Written Language, and Language section of Stanford Achievement Test). Scoring procedures used were identical to those of Study I. In the third study, 51 regular class students and 31 LD program students in grades 3-6 from five urban elementary schools were tested with the same direct measures and standardized measures as in Study II. In addition, the Developmental Sentence Scoring System was employed as an additional validation measure.

Survey of LD Teachers (R[illegible]65, 80)

During 1980-1981, 128 teachers of learning disabled students completed a survey on instructional program planning and implementation practices. The survey was sent to teachers randomly selected from the national membership list of the Council for Learning Disabilities (CLD) of the Council for Exceptional Children; a follow-up reminder was sent. The responding teachers were from 42 states distributed fairly evenly among rural, suburban, and urban school districts. The majority of teachers were female, held graduate degrees, taught in elementary schools, and provided direct service instruction to learning disabled students. The average number of years of experience teaching special education students was 6.3 years.

After interviewing 25 learning disabilities teachers, a comprehensive eight-section survey was designed. Each responding teacher randomly selected one student (according to specific guidelines) from his/her caseload and provided information about this student's program, including school and teacher information, student information, selection of IEP goals and objectives, program description, determinants of the program, changes in the original instructional plan, evaluation of progress, and other topics (e.g., teacher satisfaction, general comments). Teachers were provided with a repertoire of responses for some questions; however, the list was not viewed as exhaustive and teachers were encouraged to use "other" as a response.

Interviews of Special Educators (RR 41)

During 1980, 18 elementary teachers were interviewed about their participation in a 1979 study investigating the impact of different measurement systems on instructional decision making with LD students. The major purpose of the 16-question structured interview was to determine the teachers' perceptions of the strengths and weaknesses of the original study and to further ascertain whether the research had any effects on individual teaching styles. The interviewers were not staff members of the Institute and had no prior involvement with the original study. Each interview lasted about one half hour.

Surveys of Experimental Study Participants (RR 114, 124)

Students, parents, teachers, and administrators in four rural and suburban Minnesota school districts provided survey information related to an experimental study on formative evaluation effects (see RR 88, 111, 116). One survey focused on the communication of IEP goals and student progress. This survey was completed by 12 parents of experimental students, 25 regular classroom teachers (16 teachers of experimental students and 9 teachers of control students), and 11 administrators from three school districts. The surveys and stamped return envelopes were sent to these individuals at the end of the school year. The surveys differed slightly as a function of the role of the respondent. Parents completed a 10-item survey designed to assess their confidence in the placement committee's decision on the delivery of special education service in the area of reading, their knowledge of and satisfaction with the child's year-end reading goal and progress toward it, and their knowledge of the child's academic status compared to other students of the same age. Teachers completed an 11-item survey on students they had referred and who received part-time special education services. The survey focused on (a) participation in the IEP or periodic review conference, (b) satisfaction with and usefulness of assessment information, (c) clarity of and satisfaction with the student's reading program and progress, and (d) student performance relative to other children in the classroom. Administrators completed a 9-item survey focusing on their participation in the students' conferences, satisfaction with assessment information, clarity of student's reading goal and system for monitoring progress, and their views of parents' understanding of special education services provided to the student.

A second basic survey focused on the effects of the experimental study on instruction, teacher estimates of student progress, and student knowledge of performance. This survey was completed by 31 special education teachers and (through an interview procedure) by 135 elementary-age resource room students. Teachers completed three surveys over the course of the school year. Two surveys completed during the year focused on student progress, goals, and level of functioning in reading. A 12-item survey was completed by these teachers at the end of the year. It asked teachers to rate and describe how the experimental procedures were different from their normal evaluation procedures and to indicate whether, and if so, how they would use the procedures during the subsequent year. A four-item interview survey was given to students to assess their knowledge about (a) their reading progress, (b) their reading goals, and (c) the likelihood that they would attain their reading goals. Two additional items required interviewers to assess the accuracy of student responses against the student's reading graphs and records.

Comparative Study of Reading Domains and Durations (RR 48)

Three studies were conducted during 1979-80 to examine the effects of variations in procedures used for curriculum-based assessment of reading proficiency. The first study addressed the question of the influence of sample duration on the concurrent validity and variability of the measures. Two groups of students served as subjects. The first group included 27 students randomly selected from grades 1-6 in two public urban elementary schools in a large metropolitan area. The second group included 18 students from LD resource programs in these two schools. The five curriculum-based measures (Words in Isolation, Words in Context, Oral Reading, Cloze Comprehension, Word Meaning) were administered individually in one session to each student. The students completed two 30-second and two 60-second parallel forms for each of the word recognition measures. For the Cloze measure, each test was two minutes.

The second study addressed the question of the influence of sample duration on the level, slope, and variability of performance over repeated measurements. Two second grade, eight year old girls in the same classroom were selected as subjects because of their consistent school attendance, similarity to each other, and seriously delayed reading performance. Both students received Title I programming daily. They read from the same reader, worked on the same phonics categories, and, over a five-week interval, both consistently scored within five words of each other on weekly, one-minute samples of the number of correct C-V-C words read from flashcards. A multiple baseline across subjects and reversal design was used and consisted of four experimental phases: Phase A, a daily 30-second measurement sample; Phase B, a daily three-minute measurement sample; Phase C, return to a daily 30-second measurement sample; and Phase D, return to a daily three-minute sample. Data were collected on the number of correctly and incorrectly read C-V-C words per minute. The Title I reading teacher individually collected the data at the end of the students' standard 20-minute instructional session.

The third study was designed to examine the effect that varying the size of the pool from which items are drawn has on slope and variability of performance on the measure. Subjects were 20 students in a metropolitan school district, reading at grade 2-4 instructional levels. Teachers instructed the students using the grade specific word lists representing their instructional level. Instruction occurred for 10 minutes daily, followed by teacher administration of three 30-second lists: one from the appropriate grade level, one from the appropriate instructional level, and one from the across-grade domain.

Comparative Study of Standardized and Direct Measures (RR 126)

The effectiveness of direct measurement techniques and standardized achievement tests for assessing within-individual change were compared over a 10-week period. A total of 83 grade 3-6 low-achieving students (ones who performed below the 15th percentile on a measure of written expression) from a rural midwestern area were administered the Reading Comprehension and Language subtests from the Stanford Achievement Tests and a direct measure of reading (see RR 20) in October and again in December.

Implementation Study (RR 106)

During 1981-82, educational personnel provided information on the feasibility and cost effectiveness of a continuous pupil progress monitoring system that was implemented in two elementary schools and involved 552 students. A total of 38 educational personnel participated in the weekly measurement of the students. Included were teachers, tutors, aides, a school psychologist, and a principal. Twenty-five of these individuals completed a survey at the end of the year that focused on their use of information, the time required by the system, and their reactions to specific aspects of it.

Single Subject Study (RR 120)

During 1980-81, the effects of two data-utilization rules on spelling achievement were compared for an 11 year old male fifth-grader who had been diagnosed as learning disabled in an upper middle-class midwestern school. He spent one hour daily in the school's resource room, receiving small group instruction in reading, language arts, and math.

At the beginning of the study, the student received five minutes of daily direct instruction on a random selection of words from each of two word packs. Two lists of spelling demons and difficult words were randomly divided into two word packs, which were assumed to be equivalent in difficulty. During instructional sessions, the boy was taught and measured on sets of difficult spelling words. Graphed data were analyzed using a concurrent schedule design whereby equivalent behaviors are treated simultaneously with different approaches to determine relative treatment effects. One treatment approach involved the following data-utilization rule: If the students' performance fell below the expected level on three consecutive days, the teacher introduced a program change. In the second treatment, the teacher made changes in the student's program every 5 to 10 days. Throughout the study, the measurement task was an analogous one-minute timing of the subjects' writing randomly selected words from a word pack. Dependent data were words corret and errors per minute.

Reliability of Written Expression Measures (RR 50)

During 1981, the reliability of four measures of written expression (Total Words Written, Mature Words, Words Spelled Correctly, and Letters in Correct Sequence) was examined. The subjects varied for the four types of reliability examined. Twenty-eight learning disabled students attending a summer program in a metropolitan midwestern elementary school were used to examine test-retest reliability. Parallel form reliability was examined with 161 elementary students selected randomly from two urban midwestern cities. To determine test-retest and parallel form reliabilities each student was administered two identical story starters and was given five minutes to write a composition for each. The administration of the story starters was three weeks apart for the test-retest reliability investigation. To determine split-half reliability (a measure of internal consistency), the written compositions of 105 elementary students in grades 1 through 6, randomly selected from six schools in a large midwestern city, were examined to determine how far each student had written at the end of minutes 1, 2, 3, 4, and 5. Inter-scorer reliability was examined for 20 students enrolled in grades 1-6 from a school in a large city in the eastern region of the U.S.

Comparative Study of Written Expression Scoring Procedures (RR 84)

During 1982, written expression samples from 50 students in grades 3-6 were scored in terms of correct word sequences to investigate (a) the consistency among scorers using the procedures, (b) the typical performance levels of students in grades 3-6 on this measure, and (c) the validity of this measure relative to criterion measures of written expression. The students were selected randomly from a set of students who had participated in a previous study; their average age was 10 years and their average grade level was 4.7.

Three trained graduate research assistants tested [illegible]s on an individual basis. Students were asked to write for [illegible]nutes in response to a story starter or topic sentence, and were given the Test of Written Language. Each composition was scored using several criterion measures (Developmental Sentence Scoring, Hunt's mean t-unit length, P[redacted] checklist of written expression, holistic rating scale, word spelled correctly, and total words written). The written samples also were scored by one teacher and one non-teacher for the number of correct word sequences, which was defined as two adjacent, correctly spelled words that are acceptable within the context of the phrase to a native speaker of the English language. In addition, six teachers, many of whom had master's degrees and were certified in special education, rated the 30 written samples according to two procedures.

Study of Expressive Language (RR 83)

Post-hoc analyses were conducted during 1982 to determine (a) whether subjects' expressive language was semantically and syntactically more complex when tested by a familiar examiner than when tested by an unfamiliar examiner, and (b) whether the quality of spoken language was related to fluency. Subjects were 34 preschool children whose speech and/or language functioning represented a moderate to profound handicap. The students were enrolled in a special education preschool program within a large urban midwestern metropolitan school district. The mean age of the students was 4-9 years; there were almost twice as many boys as girls, and minorities represented 38% of the sample. All but two subjects performed within the normal range on individually administered intelligence tests. A multi-categorical scale consisting of salient syntactic characteristics and semantic relations was used to score records of the subjects' expressive language performance.

Two experienced speech clinicians each scored 68 protocols (responses of 34 subjects to familiar and unfamiliar examiners). The subjects' spoken language was separated into utterances and the raters assessed the protocols for semantic/syntactic complexity. Subjects' responses also were judged as correct or incorrect with respect to the content of illustrations being described. Inter-rater agreement was .88.

Technical Characteristics of Direct Social Adjustment Measures (RR 24)

During the 1980-81 school year, two studies were conducted to identify simple and efficient measures of children's social adjustment and to determine their relationship to other measures of a student's classroom social status.

In the first study, subjects were 67 third and fourth graders from three different classrooms in a large metropolitan school. Slightly over half of the subjects were boys. Both sociometric status inventories (roster ratings and peer nominations) and teacher rating scales were used to estimate the social status of the students. Using an interval recording system, trained observers recorded five behaviors (initiations by peers to target, one-way and two-way verbal interactions between peers and target, aversive behavior, ignoring behavior, inappropriate behavior) in a variety of situations (academic, recess, transition time) in two of the classrooms. Each child was observed for five consecutive six-second intervals; observers rotated through the class list as many times as possible within the observation period. Ten days of data were collected over a three-week period. Observer agreement ranged from a mean of .66 to .87, depending on which of three reliability formulas was used.

In the third classroom, observers recorded behaviors of students as they functioned in cooperative groups of four students each; group membership was rotated systematically over five 30-minute observation sessions. The groups were presented with a different cooperative task during each session. Data were collected on three behaviors (verbal interaction, aversive behavior, ignoring behavior) using the interval recording system. Data were collected on five separate occasions over a two-week period.

In the second study, the subjects were 58 students from two third-grade classrooms in a suburban elementary public school within a large metropolitan area; 34 students were boys. The sociometric status instruments and teacher rating scale were similar to those used in the first study. Observations were made by two trained observers for two hours per day in each classroom over a three-week period. Data were collected on only two events: (a) frequency of peer talks to target, and (b) number of different peers with whom interaction occurred. The observation interval was increased from six to 30 seconds; an event rather than interval recording system was used. Observer reliability ranged from a mean of .73 to .92, depending upon which of three formulas was applied.

Study of Variables Influencing Social Adjustment Measures (RR 82)

During 1980-81, observations were conducted to identify student behaviors that relate to students' social functioning, defined first as social status within the group and second as behavior problems perceived by the teacher. Fifty-four fifth grade boys and girls from seven classrooms that were organized into two units were observed over a 10-week period during both informal and formal school periods. The students attended a midwestern urban public school. Measures included a roster and rating sociometric instrument, a peer nomination procedure and a school behavior profile. Behaviors observed in structured academic settings included noisy, out of place, target aggression, peer aggression, peer initiation, and off-task alone.

Behaviors observed in unstructured settings differed in that out of place was not observed and off-task alone was simply alone. Observer reliabilities ranged from .80 to 1.00 across observational categories.

An event recording system was used by the observers who moved through the list of names, observing each student for 60 seconds, with a 5-second break between students. Both boys and girls were observed over 10 weeks during a structured academic period; only boys were observed during an unstructured lunch period and free time prior to school. The teachers completed the School Behavior Profile prior to the behavioral observations. After nine weeks of data collection, the two sociometric measures were administered individually.

Comparison of Student Self-Management Techniques (RR 115)

During 1981-82, the effects of student charting and student selection of instructional activities were examined. In addition, the nature of student-selected activities was compared to the nature of teacher-selected activities. Forty-two elementary resource room students from a rural special education cooperative participated in the study. They were selected from the caseloads of 8 resource teachers who had agreed to participate in the study.

Demonstration Study of Data Utilization (RR 10)

During 1978, 52 children in grades 2-6 who had been previously classified as learning disabled or educable mentally retarded participated in a study on two components of formative evaluation (frequency of measurement and data utilization rules). The students were enrolled in regular class programs and were receiving daily reading instruction from 13 special education resource teachers in four metropolitan school districts in Minnesota.

Four students were selected randomly from each resource teacher's existing caseload and randomly assigned to either an untreated control group or one of three experimental treatment groups: (a) pre-post measurement, non-data-based change, (b) daily measurement, non-data-based change, or (c) daily measurement, data-based change. Each group contained 13 students.

Measures of oral reading rate correct, oral reading rate incorrect, vocabulary meaning, and comprehension were obtained for all students both prior to and following treatment. Baseline performance was established for each student and a 30% increase in oral reading rate correct was established arbitrarily as the 18-day objective for students in the experimental conditions.

Reading instruction was similar for all the treatment groups; it involved 20 minutes of reading instruction daily from the resource teacher. Prior to and following treatment, students read aloud for three minutes at each of three placement levels, were asked to define five words from each story, and were given standardized reading comprehension measures. The treatment groups differed only in the frequency of measurement and specific data utilization rule employed.

Analysis of Program Components (RR 12)

During its final year of funding, the Child Service Demonstration Center for Children with Learning Disabilities in the Minneapolis Public Schools served as a setting for a series of studies on several program components, including the daily data collection procedures used to monitor students' progress, the data utilization techniques, and the instructional techniques. A total of 32 students (18 elementary and 14 secondary) participated in a within-subject design. The research also served as an early test of the feasibility of integrating experimental research within existing service programs in a way that directs and benefits both research and service.

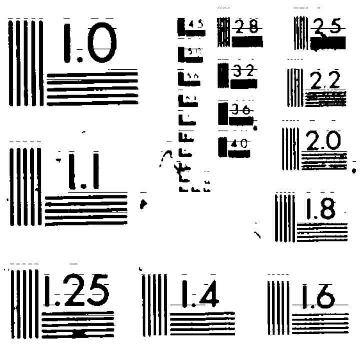
Survey and Observation of Special Education Teachers (RR 81)

During 1982, surveys of 147 special education teachers and observations of 20 practicing teachers and 20 cooperating teachers were used to (a) determine the procedures used most often by special education teachers in their evaluation of student progress, and (b) assess the adequacy of those procedures. A one-page survey was developed to investigate how special educators assess students' mastery of both IEP objectives and instructional material presented in daily lessons, their confidence in their estimates of students' daily performance on instructional objectives, and the frequency of evaluation of student progress toward IEP goals. The surveys were mailed to members of the Massachusetts Federation of the Council for Exceptional Children and were to be completed by teachers only. The responding teachers were predominantly female, had taught an average of 8 years, with half conducting resource programs. More than half of the teachers held graduate degrees.

Data also were collected from 20 practicing (student) teachers and 20 cooperating teachers through observation of the practicing teacher with the target student. A lesson plan and behavioral objective with criterion performance were provided to observers. While the practicing teacher instructed, the observer recorded the student's actual performance on the behavioral objective and the methods employed by the practicing teacher to assess the student's performance. Following the lesson, the practicing and cooperating teachers independently rated the success of the lesson, provided a rationale for their rating, indicated whether the student mastered the behavioral objective, and estimated the actual level of performance on the objective if the student failed to master the objective. The accuracy of practicing and cooperating teachers' estimates of child performance on the behavioral objective were compared. All trainees and cooperating teachers were female. The trainees were completing their final practicum for a special education degree. The cooperating teachers had taught for an average of 7 years; two-thirds had advanced degrees. Only two teachers were in a private school setting; the teachers were in either resource or special self-contained classrooms.

Surveys of Special Educators (RR 67)

During 1981, three separate groups of teachers were surveyed to document their familiarity with and use of direct and frequent measurement of student behavior. Teachers indicating use of the procedures were asked to specify the amount of time allotted to measurement of student behavior in their classrooms, while teachers indicating they did not use the procedures were asked to specify factors that inhibit their use of the procedures. The specific questions asked and procedures varied for the three groups of teachers. The first group included 136 LD teachers who responded to a postcard survey sent to randomly selected members of the Council for Learning Disabilities. The overall response rate for this sample was 45.3%. The teachers were from all regions of the U.S. No follow-up


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

contacts were made. The second survey group involved a national sample of 128 LD teachers who responded to an in-depth survey (see p. 23). The final sample included 10 special education elementary resource teachers (2 male, 8 female) in a rural educational cooperative in the midwest who were required by their special education directors to employ direct and frequent measures in conjunction with a research project (see p. 1).

Comparative Study of Reading Domains (RR 55)

During 1979-80, five special education resource teachers in a large metropolitan school volunteered to participate in a study examining the effects of varying the size of the population of words from which test items for daily testing were sampled. For each teacher, four students were selected randomly from among those reading at or between the second and fourth grade instructional levels; the 20 students served as subjects in the study.

Three populations of reading vocabulary words were created using the Harris-Jacobson Word List. The largest population, called Across-Grade list (AG), consisted of the entire pool of words from preprimer through grade 4. The second population, called the Grade-Level list (GL) consisted only of those words within the students' grade level. The third, Instructional-Level list (IL) was a subset of 200 words drawn at random from the GL population. Daily word lists for testing were created by drawing 60 words at random from each of the three populations; 20 different word lists for each domain were created by random sampling with replacement.

The appropriate grade level for instruction was determined for each student. Students were instructed individually for 10 minutes daily on 200 words from this instructional level. Following each instructional period the student took a 30-second word reading test on each of the three populations of words using the daily tests and lists that had been created. Teachers recorded the number of words read correctly and incorrectly on each type of word list. Throughout the study, the students' performance graphs were evaluated weekly to determine the need for an instructional modification. After 15 days, an instructional change was required.

Study of Alternative Reading Performance Criteria (RR 59)

During 1980-81, analyses of the technical adequacy of informal reading inventories were conducted using data from 91 randomly selected students, distributed across grades 1-6 in a midwestern metropolitan elementary school. All students were English speaking, 15 received special education resource service, and 23 were enrolled in Title I programs for children who were "seriously behind" in reading. Correlational and congruency analyses were conducted to determine the technical adequacy of (a) choosing a criterion of 95% accuracy for word recognition to determine an instructional level, (b) arbitrarily selecting a passage to represent the difficulty level of a basal reader, and (c) employing one-level floors and ceilings to demarcate levels beyond which behavior is not sampled.

Study of Curriculum Differences (RR 93)

The performances of 660 elementary students in six school districts on two curriculum-based reading aloud tasks and one non-curriculum-based measure were examined. Four different basal reading series were compared. The two basal reading measurement tasks consisted of a reading passage and a vocabulary word list; the non-curriculum measure was a word list. The 660 students, who were selected randomly, attended schools in a rural midwestern education cooperative. No attempt was made to obtain equal representation of males and females.

The testing of the 660 students was conducted within the first month of school by 10 trained educational aides. All testing was completed on an individual basis and involved the administration of two one-minute oral reading passages, one basal word list, and the non-curriculum word list. The order of administration of the three measures was counterbalanced.

Analysis of Readability Formulas (RR 129)

During 1981-82, 285 special education students in grades 1-9 were tested twice on three passages of a Passage Reading test. The students were from either rural and suburban Minnesota (n=117) or from New York City (NYC). Six readability formulas were applied to the three passages to examine the agreement among the formulas. In addition, difficulty rankings by the formulas were compared to rankings produced by students' actual performance. Student performance in the two settings also was compared to explore the contribution of pupil background to text difficulty.

Analyses of Basal Reader Criterion-Referenced Tests (RR 113, 122, 128, 130)

Four studies were conducted during 1982-83 on the technical adequacy of the criterion-referenced tests associated with basal readers commonly used in public schools. In each study, students' performance on basal reader tests was compared to their performance on a standardized test and a direct measure word reading test. Various analyses were then conducted on the data to examine the technical adequacy of the basal reader tests. The specific subject samples and tests included in each study are detailed below.

Houghton-Mifflin Basic Reading Test (RR 113). Subjects were 47 sixth graders who were tested on the SRA Reading Achievement Test, the Houghton-Mifflin End-of-Level 11 Test, and the Word Reading Test. A subgroup of 20 students was tested a second time on the Basal Reading Test. All students were from a school district in a rural midwestern cooperative.

Ginn 720 Series Mastery Test (RR 122). Subjects were 47 fifth graders who were tested on the SRA Reading Achievement Test, the Ginn 720 End-of-Level 11 Mastery Test, and the Word Reading Test. A subgroup of 22 students was tested a second time on the Mastery Test. All students were from a school district in a rural midwestern cooperative.

Scott-Foresman Criterion-Referenced Test (RR 128). Subjects were 25 fourth graders in a rural educational cooperative who were tested on the SRA Reading Achievement Test, the Scott-Foresman End-of-Book 9 Criterion-Referenced Test, and the Word Reading Test. All students were tested a second time on the Criterion-Referenced Test.

Holt Basic Reading Series Management Program Level 13 Test (RR 130). Subjects were 21 fourth graders in a rural educational cooperative who were tested on the SRA Reading Achievement Test, the Holt Basic Reading Series Management Program Level 13 Test, and the

Word Reading Test. All students were tested a second time on the Management Program Test.

Measuring Classroom Behavior (RR 6)

Observations were conducted on 11 students enrolled in a midwestern inner city elementary school. These students had been identified by their teachers as ones having the most difficulty adjusting socially. Students were observed during periods of structured academic work. A sample of 10 peers was observed during the same observation period, producing 10 minutes of data on each target student and 10 minutes of data on each target student's peers. Observations focused on five categories of behavior: (1) noise, (2) out of place, (3) physical contact or destruction, (4) off task, and (5) other.

Comparative Study of Graph Papers (RR 101)

During 1980-81, student performance on direct, repeated measures of reading and written expression were collected over a 2½ month period for 83 low-achieving elementary students identified during the screening of all 785 elementary students from grades 3-6 enrolled in three rural elementary schools. The students had no history of special education services, but scored at or below the 15th percentile on a short duration measure of written expression that significantly discriminated LD and non-LD students. The students (32 were females) were fairly evenly distributed across grades 3-6.

The students were administered two tasks on a weekly basis for 10 weeks. First, students were asked to read aloud for one minute from a third-grade list of words. The number of words read correctly and incorrectly were scored and graphed. Students in grades 4-6 also read a list of words selected from their grade levels. Second, story starters were used weekly to obtain writing samples from the students. These were scored for Total Words Written, Words Written Correctly, Words Written Incorrectly, and Correct Letter Sequences Written.

A computer program was used to simulate charting on both interval and semi-logarithmic graphs. Each students' data were entered into the computer at the end of the seventh week; the slope of each student's performance on the two types of graphs was used to predict student performance at weeks 8, 9, and 10 of the data collection period. The estimates of student performance at three times was contrasted with the actual data collected at weeks 8, 9, and 10 by determining the absolute deviation between the scores. The graphing approach with the smaller average deviation score was considered to be the one making better predictions of student performance.

Assessment of Alternative Data Summary Procedures (RR 112, 118)

A study of two basic procedures for analyzing time series data (visual analysis and statistical analysis) was conducted during 1981-82. Student performance represented on 28 hypothetical graphs was evaluated by 52 in-service and pre-service teachers from three locations around a large midwestern city. The slope and variability of data presented in the graphs were varied systematically. In addition, two other conditions (training in data utilization and use of aimlines) were varied in the study. Subjects evaluated each graph in terms of the effectiveness of the program depicted on the graph, and also indicated what about the data supported each judgment.

Statistical analyses also were conducted on the data presented in each graph:

Evaluation of Program Effectiveness (RR 123)

During 1982-83, a system-level analysis of the effectiveness of special education was conducted in an educational cooperative comprised of six school districts. A total of 96 special education students in grades 1-6 were assessed three times during the year on direct, curriculum-based measures of achievement in reading, math, and spelling. Analyses of student performance data were conducted across all six districts, for each district, by teachers, and by student classification (LD or EMR), grade, and sex. All measurement materials were developed from the curricula in use in the school districts.

Analysis of Statistical Properties of Data (RR 125, 138)

During 1981-82, reading performance data were collected on 68 resource room students over a period of six months. The grade 1-7 students were from four Minnesota school districts. All were participating in research on the effects of teachers using frequent curriculum-based measures of student performance when the data were collected. The students' data were subjected to further analysis in this investigation. First, the slope, standard error of estimate, mean level of performance, and number of data points were calculated for each graph to document the characteristics of the time-series data collected through frequent curriculum-based measurement. Second, a principal components factor analysis was performed to summarize relationships among the time-series properties and properties of the measurement system. In addition, multiple regression analyses were used to identify the relationship of such variables to achievement.

Study of Self-Instructional Training (RR 63)

Eight special education resource teachers pilot tested a manual designed to train teachers to use direct and frequent measurement techniques to monitor students' progress toward individualized goals and to evaluate the effectiveness of the students' instructional programs. All teachers were certified in special education and two held graduate degrees. The teachers, whose teaching experience ranged from 1 to 35 years, taught in a suburban school district.

Causal Model Analysis (RR 105)

Causal modeling techniques were used to examine the relationships among implementation of a formative evaluation system, structure of instructional programs, and reading achievment for 117 students in grades 1-7. Most of the students were boys in grades 2-5; their average age was 9.5 years. For the most part, they received special education services in resource rooms. The 31 teachers were predominantly female and had an average of 8.8 years teaching special education. The greatest percentage of teachers had no experience teaching regular education.

Three major types of measures were employed. The measure of the degree of implementation of the monitoring system (Accuracy of Implementation Rating Scale - AIRS) and the measure of the degree of structure of the students' instructional programs (Structure of Instruction Rating Scale - SIRS) were used to determine how the evaluation system influences teaching practices. Both scales involve observation and the rating of multiple items on a 1 (lowest) to 5 (highest) scale. The third set of measures were student achievement indices. At three different points in time during the study (separated by approximately two months each and synchronized with AIRS and SIRS observations), three one-minute oral reading measures were administered to the student. Posttest measures included two subtests from a standardized reading test.

Three formats were used to train the teachers to carry out a specific set of procedures that included establishing an appropriate reading measurement level, writing long-range goals and short-term objectives, administering direct reading measures, graphing, and data utilization in making decisions about the effectiveness of students' reading instructional programs. The training formats included: (a) three half-day workshops at the beginning of the school year supplemented by a training manual and research feedback, (b) training by district personnel with the aid of the same manual, supplemented by phone contact with the researchers, and (c) one week of full-day workshops and periodic ongoing inservice.

Instructional Rating Scale Validation (RR 107)

During 1981-82, a bi-polar rating scale was developed for use in an experimental study on the effects of teachers using direct and frequent measurement of special education students' reading performance. The scale was developed as a measure to monitor the structure of instruction provided to target students; it included variables identified in educational literature as important in predicting classroom achievement. Data collected from 158 elementary school children in four school districts were analyzed to examine the technical characteristics of the scale. The data were examined in terms of reliability and evidence of a consistent factor structure.

Research Report References

No. 6 Deno, S. L. A direct observation approach to measuring classroom behavior: Procedures and application. April, 1979.

No. 10 Mirkin, P. K., & Deno, S. L. Formative evaluation in the classroom: An approach to improving instruction. August, 1979.

No. 12 Deno, S. L., Chiang, B., Tindal, G., & Blackburn, M. Experimental analysis of program components: An approach to research in CSDC's. August, 1979.

No. 20 Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. Relationships among simple measures of reading and performance on standardized achievement tests. January, 1980.

No. 21 Deno, S. L., Mirkin, P. K., Lowry, L., & Kuehnle, K. Relationships among simple measures of spelling and performance on standardized achievement tests. January, 1980.

No. 22 Deno, S. L., Mirkin, P. K., & Marston, D. Relationships among simple measures of written expression and performance on standardized achievement tests. January, 1980.

No. 23 Mirkin, P. K., Deno, S. L., Tindal, G., & Kuehnle, K. Formative evaluation: Continued development of data utilization systems. January, 1980.

No. 24 Deno, S. L., Mirkin, P. K., Robinson, S., & Evans, P. Relationships among classroom observations of social adjustment and sociometric rating scales. January, 1980.

No. 41 Meyers, B., Meyers, J., & Deno, S. Formative evaluation and teacher decision making: A follow-up investigation. September, 1980.

No. 48 Fuchs, L., Tindal, J., & Deno, S. Effects of varying item domain and sample duration on technical characteristics of daily measures in reading. January, 1981.

No. 49 Marston, D., Lowry, L., Deno, S., & Mirkin, P. An analysis of learning trends in simple measures of reading, spelling, and written expression: A longitudinal study. January, 1981.

No. 50 Marston, D., & Deno, S. The reliability of simple, direct measures of written expression. January, 1981.

No. 53. Fuchs, L., Wesson, C., Tindal, G., & Mirkin, P. Teacher efficiency in continuous evaluation of IEP goals. June, 1981.

No. 55 Tindal, G., & Deno, S. L. Daily measurement of reading: Effects of varying the size of the item pool. July, 1981.

No. 56 Fuchs, L. S., & Deno, S. L. A comparison of teacher judgment, standardized tests, and curriculum-based approaches to reading placement. August, 1981.

No. 57 Fuchs, L., & Deno, S. The relationship between curriculum-based mastery measures and standardized achievement tests in reading. August, 1981.

No. 59 Fuchs, L., Fuchs, D., & Deno, S. Reliability and validity of curriculum-based informal reading inventories. October, 1981.

No. 61 Tindal, G., Fuchs, L., Christenson, S., Mirkin, P., & Deno, S. The relationship between student achievement and teacher assessment of short- or long-term goals. November, 1981.

No. 62 Mirkin, P., Fuchs, L., Tindal, G., Christenson, S., & Deno, S. The effect of IEP monitoring strategies on teacher behavior. December, 1981.

No. 63 Wesson, C., Mirkin, P., & Deno, S. Teachers' use of self instructional materials for learning procedures for developing and monitoring progress on IEP goals. January, 1982.

No. 64 Fuchs, L., Wesson, C., Tindal, G., Mirkin, P., & Deno, S. Instructional changes, student performance, and teacher preferences: The effects of specific measurement and evaluation procedures. January, 1982.

No. 65 Potter, M., & Mirkin, P. Instructional planning and implementation practices of elementary and secondary resource room teachers: Is there a difference? January, 1982.

No. 67 King, R., Wesson, C., & Deno, S. Direct and frequent measurement of student performance: Does it take too much time? February, 1982.

No. 80 Mirkin, P. K., & Potter, M. L. A survey of program planning and implementation practices of LD teachers. July, 1982.

No. 81 Fuchs, L. S., Fuchs, D., & Warren, L. M. *Special education practice in evaluating student progress toward goals.* July, 1982.

No. 82 Kuehnle, K., Deno, S. L., & Mirkin, P. K. *Behavioral measurement of social adjustment: What behaviors? What setting?* July, 1982.

No. 83 Fuchs, D., Dailey, Ann Madsen, & Fuchs, L. S. *Examiner familiarity and the relation between qualitative and quantitative indices of expressive language.* July, 1982.

No. 84 Videen, J., Deno, S., & Marston, D. *Correct word sequences: A valid indicator of proficiency in written expression.* July, 1982.

No. 87 Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. *The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study.* August, 1982.

No. 88 Skiba, R., Wesson, C., & Deno, S. L. *The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparison.* September, 1982.

No. 93 Tindal, G., Marston, D., Deno, S. L., & Germann, G. *Curriculum differences in direct repeated measures of reading.* October, 1982.

No. 94 Fuchs, L. S., Deno, S. L., & Marston, D. *Use of aggregation to improve the reliability of simple direct measures of academic performance.* October, 1982.

No. 96 Fuchs, L. S., Deno, S. L., & Mirkin, P. K. *Effects of frequent curriculum-based measurement and evaluation on student achievement and knowledge of performance: An experimental study.* November, 1982.

No. 97 Fuchs, L. S., Deno, S. L., & Mirkin, P. K. *Direct and frequent measurement and evaluation: Effects on instruction and estimates of student progress.* November, 1982.

No. 101 Marston, D., & Deno, S. L. *Measuring academic progress of students with learning difficulties: A comparison of the semi-logarithmic chart and equal interval graph paper.* November, 1982.

No. 105 Wesson, C., Deno, S., Mirkin, P., Sevcik, B., Skiba, R., King, R., Tindal, G., & Maruyama, G. Teaching structure and student achievement effects of curriculum-based measurement: A causal (structural) analysis. December, 1982.

No. 106 Marston, D., & Deno, S. L. Implementation of direct and repeated measurement in the school setting. December, 1982.

No. 107 Deno, S. L., King, R., Skiba, R., Sevcik, B., & Wesson, C. The structure of instruction rating scale (SIRS): Development and technical characteristics. January, 1983.

No. 109 Tindal, G., Marston, D., & Deno, S. L. The reliability of direct and repeated measurement. February, 1983.

No. 111 King, R. P., Deno, S., Mirkin, P. & Wesson, C. The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparison. February, 1983.

No. 112 Tindal, G., Deno, S. L., & Ysseldyke, J. E. Visual analysis of time series data: Factors of influence and level of reliability. March, 1983.

No. 113 Tindal, G., Shinn, M., Fuchs, L., Fuchs, D., Deno, S., & Germann, G. The technical adequacy of a basal reading series mastery test. April, 1983.

No. 114 Sevcik, B., Skiba, R., Tindal, G., King, R., Wesson, C., Mirkin, P., & Deno, S. Communication of IEP goals and student progress among parents, regular classroom teachers, and administrators using systematic formative evaluation. April, 1983.

No. 115 Wesson, C. Two student self-management techniques applied to data-based program modification. April, 1983.

No. 116 Wesson, C., Skiba, R., Sevcik, B., King, R., Tindal, G., Mirkin, P., & Deno, S. The impact of the structure of instruction and the use of technically adequate instructional data on reading improvement. May, 1983.

No. 118 Tindal, G., & Deno, S. Factors influencing the agreement between visual and statistical analyses of time series data. June, 1983.

No. 120 Fuchs, L. S., Deno, S. L., & Roettger, A. The effect of alternative data-utilization rules on spelling achievement: An n of 1 study. June, 1983.

No. 122 Fuchs, L., Tindal, G., Shinn, M., Fuchs, D., Deno, S., & Germann, G. Technical adequacy of basal readers' mastery tests: The Ginn 720 series. June, 1983.

No. 123 Tindal, G., Germann, G., Marston, D., & Deno, S. The effectiveness of special education: A direct measurement approach. June, 1983.

No. 124 Sevcik, B., Skiba, R., Tindal, G., King, R., Wesson, C., Mirkin, P., & Deno, S. Curriculum-based measurement: Effects on instruction, teacher estimates of student progress, and student knowledge of performance. July, 1983.

No. 125 Skiba, R., Marston, D., Wesson, C., Sevcik, B., & Deno, S. L. Characteristics of the time-series data collected through curriculum-based reading measurement. July, 1983.

No. 126 Marston, D., Deno, S., & Tindal, G. A comparison of standardized achievement tests and direct measurement techniques in measuring pupil progress. July, 1983.

No. 128 Tindal, G., Fuchs, L., Fuchs, D., Shinn, M., Deno, S., & Germann, G. The technical adequacy of a basal series mastery test: The Scott-Foresman reading program. July, 1983.

No. 129 Fuchs, L. S., Fuchs, D., & Deno, S. L. The nature of inaccuracy among readability formulas. July, 1983.

No. 130 Fuchs, L., Tindal, G., Fuchs, D., Shinn, M., Deno, S., & Germann, G. The technical adequacy of a basal reading mastery test: The Holt basic reading series. July, 1983.

No. 132 Tindal, G., Germann, G., Deno, S. Descriptive research on the Pine County norms: A compilation of findings. July, 1983.

No. 138 Skiba, R., & Deno, S. A correlational analysis of the statistical properties of time-series data and their relationship to student achievement in resource classrooms. September, 1983.

Monograph

No. 20 Mirkin, P. K., Fuchs, L. S., & Deno, S. L. (Eds.). Considerations for designing a continuous evaluation system: An integrative review. December, 1982.

PUBLICATIONS

Institute for Research on Learning Disabilities
University of Minnesota

The Institute is not funded for the distribution of its publications. Publications may be obtained for $4.00 each, a fee designed to cover printing and postage costs. Only checks and money orders payable to the University of Minnesota can be accepted. All orders must be prepaid. Requests should be directed to: Editor, IRLD, 350 Elliott Hall; 75 East River Road, University of Minnesota, Minneapolis, MN 55455.

The publications listed here are only those that have been prepared since 1982. For a complete, annotated list of all IRLD publications, write to the Editor.

Wesson, C., Mirkin, P., & Deno, S. Teachers' use of self instructional materials for learning procedures for developing and monitoring progress on IEP goals (Research Report No. 63). January, 1982.

Fuchs, L., Wesson, C., Tindal, G., Mirkin, P., & Deno, S. Instructional changes, student performance, and teacher preferences: The effects of specific measurement and evaluation procedures (Research Report No. 64). January, 1982.

Potter, M., & Mirkin, P. Instructional planning and implementation practices of elementary and secondary resource room teachers: Is there a difference? (Research Report No. 65). January, 1982.

Thurlow, M. L., & Ysseldyke, J. E. Teachers' beliefs about LD students (Research Report No. 66). January, 1982.

Graden, J., Thurlow, M. L., & Ysseldyke, J. E. Academic engaged time and its relationship to learning: A review of the literature (Monograph No. 17). January, 1982.

King, R., Wesson, C., & Deno, S. Direct and frequent measurement of student performance: Does it take too much time? (Research Report No. 67). February, 1982.

Greener, J. W., & Thurlow, M. L. Teacher opinions about professional education training programs (Research Report No. 68). March, 1982.

Algozzine, B., & Ysseldyke, J. Learning disabilities as a subset of school failure: The oversophistication of a concept (Research Report No. 69). March, 1982.

Fuchs, D., Zern, D. S., & Fuchs, L. S. A microanalysis of participant behavior in familiar and unfamiliar test conditions (Research Report No. 70). March, 1982.

Shinn, M. R., Ysseldyke, J., Deno, S., & Tindal, G. A comparison of psychometric and functional differences between students labeled learning disabled and low achieving (Research Report No. 71). March, 1982.

Thurlow, M. L. Graden, J., Greener, J. W., & Ysseldyke, J. E. Academic responding time for LD and non-LD students (Research Report No. 72). April, 1982.

Graden, J., Thurlow, M., & Ysseldyke, J. Instructional ecology and academic responding time for students at three levels of teacher-perceived behavioral competence (Research Report No. 73). April, 1982.

Algozzine, B., Ysseldyke, J., & Christenson, S. The influence of teachers' tolerances for specific kinds of behaviors on their ratings of a third grade student (Research Report No. 74). April, 1982.

Wesson, C., Deno, S., & Mirkin, P. Research on developing and monitoring progress on IEP goals: Current findings and implications for practice (Monograph No. 18). April, 1982.

Mirkin, P., Marston, D., & Deno, S. L. Direct and repeated measurement of academic skills: An alternative to traditional screening, referral, and identification of learning disabled students (Research Report No. 75). May, 1982.

Algozzine, B., Ysseldyke, J., Christenson, S., & Thurlow, M. Teachers' intervention choices for children exhibiting different behaviors in school (Research Report No. 76). June, 1982.

Tucker, J., Stevens, L. J., & Ysseldyke, J. E. Learning disabilities: The experts speak out (Research Report No. 77). June, 1982.

Thurlow, M. L., Ysseldyke, J. E., Graden, J., Greener, J. W., & Mecklenberg, C. Academic responding time for LD students receiving different levels of special education services (Research Report No. 78). June, 1982.

Graden, J. L., Thurlow, M. L., Ysseldyke, J. E., & Algozzine, B. Instructional ecology and academic responding time for students in different reading groups (Research Report No. 79). July, 1982.

Mirkin, P. K., & Potter, M. L. A survey of program planning and implementation practices of LD teachers (Research Report No. 80). July, 1982.

Fuchs, L. S., Fuchs, D., & Warren, L. M. Special education practice in evaluating student progress toward goals (Research Report No. 81). July, 1982.

Kuehnle, K., Deno, S. L., & Mirkin, P. K. Behavioral measurement of social adjustment: What behaviors? What setting? (Research Report No. 82). July, 1982.

Fuchs, D., Dailey, Ann Madsen, & Fuchs, L. S. *Examiner familiarity and the relation between qualitative and quantitative indices of expressive language* (Research Report No. 83). July, 1982.

Videen, J., Deno, S., & Marston, D. *Correct word sequences: A valid indicator of proficiency in written expression* (Research Report No. 84). July, 1982.

Potter, M. L. *Application of a decision theory model to eligibility and classification decisions in special education* (Research Report No. 85). July, 1982.

Greener, J. E., Thurlow, M. L., Graden, J. L., & Ysseldyke, J. E. *The educational environment and students' responding times as a function of students' teacher-perceived academic competence* (Research Report No. 86). August, 1982.

Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. *The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study* (Research Report No. 87). August, 1982.

Skiba, R., Wesson, C., & Deno, S. L. *The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparison* (Research Report No. 88). September, 1982.

Marston, D., Tindal, G., & Deno, S. L. *Eligibility for learning disability services: A direct and repeated measurement approach* (Research Report No. 89). September, 1982.

Thurlow, M. L., Ysseldyke, J. E., & Graden, J. L. *LD students' active academic responding in regular and resource classrooms* (Research Report No. 90). September, 1982.

Ysseldyke, J. E., Christenson, S., Pianta, R., Thurlow, M. L., & Algozzine, B. *An analysis of current practice in referring students for psychoeducational evaluation: Implications for change* (Research Report No. 91). October, 1982.

Ysseldyke, J. E., Algozzine, B., & Epps, S. *A logical and empirical analysis of current practices in classifying students as handicapped* (Research Report No. 92). October, 1982.

Tindal, G., Marston, D., Deno, S. L., & Germann, G. *Curriculum differences in direct repeated measures of reading* (Research Report No. 93). October, 1982.

Fuchs, L.S., Deno, S. L., & Marston, D. *Use of aggregation to improve the reliability of simple direct measures of academic performance* (Research Report No. 94). October, 1982.

Ysseldyke, J. E., Thurlow, M. L., Mecklenburg, C., & Graden, J. *Observed changes in instruction and student responding as a function of referral and special education placement* (Research Report No. 95). October, 1982.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. Effects of frequent curriculum-based measurement and evaluation on student achievement and knowledge of performance: An experimental study (Research Report No. 96). November, 1982.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. Direct and frequent measurement and evaluation: Effects on instruction and estimates of student progress (Research Report No. 97). November, 1982.

Tindal, G., Wesson, C., Germann, G., Deno, S. L., & Mirkin, P. K. The Pine County model for special education delivery: A data-based system (Monograph No. 19). November, 1982.

Epps, S., Ysseldyke, J. E., & Algozzine, B. An analysis of the conceptual framework underlying definitions of learning disabilities (Research Report No. 98). November, 1982.

Epps, S., Ysseldyke, J. E., & Algozzine, B. Public-policy implications of different definitions of learning disabilities (Research Report No. 99). November, 1982.

Ysseldyke, J. E., Thurlow, M. L., Graden, J. L., Wesson, C., Deno, S. L., & Algozzine, B. Generalizations from five years of research on assessment and decision making (Research Report No. 100). November, 1982.

Marston, D., & Deno, S. L. Measuring academic progress of students with learning difficulties: A comparison of the semi-logarithmic chart and equal interval graph paper (Research Report No. 101). November, 1982.

Beattie, S., Grise, P., & Algozzine, B. Effects of test modifications on minimum competency test performance of third grade learning disabled students (Research Report No. 102). December, 1982

Algozzine, B., Ysseldyke, J. E., & Christenson, S. An analysis of the incidence of special class placement: The masses are burgeoning (Research Report No. 103). December, 1982.

Marston, D., Tindal, G., & Deno, S. L. Predictive efficiency of direct, repeated measurement: An analysis of cost and accuracy in classification (Research Report No. 104). December, 1982.

Wesson, C., Deno, S., Mirkin, P., Sevcik, B., Skiba, R., King, R., Tindal, G., & Maruyama, G. Teaching structure and student achievement effects of curriculum-based measurement: A causal (structural) analysis (Research Report No. 105). December, 1982.

Mirkin, P. K., Fuchs, L. S., & Deno, S. L. (Eds.). Considerations for designing a continuous evaluation system: An integrative review (Monograph No. 20). December, 1982.

Marston, D., & Deno, S. L. Implementation of direct and repeated measurement in the school setting (Research Report No. 106). December, 1982.

Deno, S. L., King, R., Skiba, R., Sevcik, B., & Wesson, C. The structure of instruction rating scale (SIRS): Development and technical characteristics (Research Report No. 107). January, 1983.

Thurlow, M. L., Ysseldyke, J. E., & Casey, A. Criteria for identifying LD students: Definitional problems exemplified (Research Report No. 108). January, 1983.

Tindal, G., Marston, D., & Deno, S. L. The reliability of direct and repeated measurement (Research Report No. 109). February, 1983.

Fuchs, D., Fuchs, L. S., Dailey, A. M., & Power, M. H. Effects of pretest contact with experienced and inexperienced examiners on handicapped children's performance (Research Report No. 110). February, 1983

King, R. P., Deno, S., Mirkin, P., & Wesson, C. The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparison (Research Report No. 111). February, 1983.

Tindal, G., Deno, S. L., & Ysseldyke, J. E. Visual analysis of time series data: Factors of influence and level of reliability (Research Report No. 112). March, 1983.

Tindal, G, Shinn, M., Fuchs, L., Fuchs, D., Deno, S., & Germann, G. The technical adequacy of a basal reading series mastery test (Research Report No. 113). April, 1983.

Sevcik, B., Skiba, R., Tindal, G., King, R., Wesson, C., Mirkin, P., & Deno, S. Communication of IEP goals and student progress among parents, regular classroom teachers, and administrators using systematic formative evaluation (Research Report No. 114). April, 1983.

Wesson, C. Two student self-management techniques applied to data-based program modification (Research Report No. 115). April, 1983.

Wesson, C., Skiba, R., Sevcik, B., King, R., Tindal, G., Mirkin, P., & Deno, S. The impact of the structure of instruction and the use of technically adequate instructional data on reading improvement (Research Report No. 116). May, 1983.

Wesson, C. Teacher vs student selection of instructional activities (Research Report No. 117). May, 1983.

Tindal, G., & Deno, S. Factors influencing the agreement between visual and statistical analyses of time series data (Research Report No. 118). June, 1983.

Skiba, R. S. Classroom behavior management: A review of the literature (Monograph No. 21), June, 1983.

Graden, J. L., Thurlow, M. L., & Ysseldyke, J. E. When are students most academically engaged? Academic responding time in different instructional ecologies (Research Report No. 119). June, 1983.

Fuchs, L. S., Deno, S. L., & Roettger, A. The effect of alternative data-utilization rules on spelling achievement: An n of 1 study (Research Report No. 120). June, 1983.

Skiba, R., Sevcik, B., Wesson, C., King, R., & Deno, S. The non-effect of process-product variables in resource classrooms (Research Report No. 121). June, 1983.

Fuchs, L. Tindal, G., Shinn, M., Fuchs, D., Deno, S., & Germann, G. Technical adequacy of basal readers' mastery tests: The Ginn 720 series (Research Report No. 122). June, 1983.

Tindal, G., Germann, G., Marston, D., & Deno, S. The effectiveness of special education: A direct measurement approach (Research Report No. 123). June, 1983.

Sevcik, B., Skiba, R., Tindal, G., King, R., Wesson, C., Mirkin, P., & Deno, S. Curriculum-based measurement: Effects on instruction, teacher estimates of student progress, and student knowledge of performance (Research Report No. 124). July, 1983.

Skiba, R., Marston, D., Wesson, C., Sevcik, B., & Deno, S. L. Characteristics of the time-series data collected through curriculum-based reading measurement (Research Report No. 125). July, 1983.

Ysseldyke, J., Christenson, S., Graden, J., & Hill, D. Practical implications of research on referral and opportunity to learn (Monograph No. 22). July, 1983.

Marston, D., Deno, S., & Tindal, G. A comparison of standardized achievement tests and direct measurement techniques in measuring pupil progress (Research Report No. 126). July, 1983.

Fuchs, D., Fuchs, L. S., Tindal, G., & Deno, S. L. Variability of performance: A "signature" characteristic of learning disabled children? (Research Report No. 127). July, 1983.

Tindal, G., Fuchs, L., Fuchs, D., Shinn, M., Deno, S., & Germann, G. The technical adequacy of a basal series mastery test: The Scott-Foresman reading program (Research Report No. 128). July, 1983.

Fuchs, L. S., Fuchs, D., & Deno, S. L. The nature of inaccuracy among readability formulas (Research Report No. 129). July, 1983.

Fuchs, L., Tindal, G., Fuchs, D., Shinn, M., Deno, S., & Germann, G. The technical adequacy of a basal reading mastery test: The Holt basic reading series (Research Report No. 130). July, 1983.

Ysseldyke, J. E., Christenson, S., Algozzine, B., & Thurlow, M. L. Classroom teachers' attributions for students exhibiting different behaviors (Research Report No. 131). July, 1983.

Tindal, G., Germann, G., & Deno, S. *Descriptive research on the Pine County norms: A compilation of findings* (Research Report No. 132). July, 1983.

Skiba, R. J. *The relationship between classroom management strategies and student misbehaviors* (Research Report No. 133). July, 1983.

Fuchs, D., Fuchs, L. S., Power, M. H., & Dailey, A. M. *Systematic bias in the assessment of handicapped children* (Research Report No. 134). July, 1983.

Fuchs, D., & Fuchs, L. S. *The importance of scorer bias to handicapped preschoolers' stronger performance with familiar examiners* (Research Report No. 135). July, 1983.

Foster, G. G., Ysseldyke, J. E., Casey, A., & Thurlow, M. L. *The congruence between reason for referral and placement outcome* (Research Report No. 136). August, 1983.

Potter, M. L. *Instructional decision-making practices of teachers of learning disabled students* (Research Report No. 137). September, 1983.

Skiba, R., & Deno, S. *A correlational analysis of the statistical properties of time-series data and their relationship to student achievement in resource classrooms* (Research Report No. 138). September, 1983.

Potter, M. L. *Decision research and its application to educational settings: A literature review* (Monograph No. 23). September, 1983.

Graden, J., Berquist, B., & Burnside, C. W. *Helping teachers increase the time their students spend in learning* (Research Report No. 139). September, 1983.

Graden, J., Casey, A., & Bonstrom, O. *Prereferral interventions: Effects on referral rates and teacher attitudes* (Research Report No. 140). September, 1983.

Thurlow, M. L., Christenson, S., & Ysseldyke, J. E. *Referral research: An integrative summary of findings* (Research Report No. 141). September, 1983.

Ysseldyke, J. E., & Thurlow, M. L. *Identification/classification research: An integrative summary of findings* (Research Report No. 142). September 1983.

Thurlow, M. L., & Ysseldyke, J. E. *Instructional intervention research: An integrative summary of findings* (Research Report No. 143). September, 1983.

Ysseldyke, J. E., Thurlow, M. L., & Christenson, S. *Evaluation research: An integrative summary of findings* (Research Report No. 144). September 1983.