DOCUMENT RESUME

ED 227 128 TM 830 115

AUTHOR Fuchs, Lynn S.; And Others
TITLE Use of Aggregation to Improve the Reliability of Simple Direct Measures of Academic Performance.
INSTITUTION Minnesota Univ., Minneapolis. Inst. for Research on Learning Disabilities.
SPONS AGENCY Special Education Programs (ED/OSERS), Washington, DC.
REPORT NO IRLD-RR-94
PUB DATE Oct 82
CONTRACT 300-80-0622
NOTE 33p.
AVAILABLE FROM Editor, IRLD, 350 Elliott Hall, 750 East River Road, University of Minneapolis, MN 55455 ($3.00)
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Elementary Education; *Evaluation Methods; Measurement Objectives; *Measures (Individuals); *Oral Reading; Research Needs; Special Education; Standardized Tests; Student Evaluation; *Test Reliability; Test Validity; *Writing Evaluation
IDENTIFIERS *Aggregation (Data); Ginn Reading 720 Series; Woodcock Reading Mastery Test

ABSTRACT

The effects of aggregation on the reliability of measures of academic performance were explored in two studies. In the first study, 30 elementary-age children were tested four times on the same forms of the Woodcock Reading Mastery Tests and the Ginn 720 Reading Passage measures. 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. On the standardized measure and on the oral passage reading correct rate score, aggregation had little impact; however, on the oral passage reading error rate score, aggregation substantially increased all reliability indices. In the second study, 78 children were tested 10 times on alternate forms of two reading measures and one written expression measure. Group stability coefficients were calculated on the basis of 2, 4, 6, 8, and 10 testings. For the oral words-in-isolation reading correct score, aggregation had little effect, whereas aggregating over occasions and test forms dramatically improved the stability of the oral words-in-isolation reading error score and the written expression score. The reliability and criterion validity of short, simple measures demonstrated their suitability as measures of academic performance. (Author/PN)

University of Minnesota

TM

Research Report No. 94

USE OF AGGREGATION TO IMPROVE THE RELIABILITY OF SIMPLE DIRECT MEASURES OF ACADEMIC PERFORMANCE

Lynn S. Fuchs, Stanley L. Deno, and Doug Marston

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 the Office of Special Education, Department of Education, through Title VI-G of Public Law 91-230. 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).

Research Report No. 94

USE OF AGGREGATION TO IMPROVE THE RELIABILITY OF SIMPLE DIRECT MEASURES OF ACADEMIC PERFORMANCE

Lynn S. Fuchs, Stanley L. Deno, and Doug Marston

Institute for Research on Learning Disabilities

University of Minnesota

October, 1982

Abstract

The effects of aggregation on the reliability of measures of academic performance were explored in two studies. In the first study, 30 elementary-age children were tested four times on the same forms of three reading measures; 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. On the standardized measure and on the oral passage reading correct rate score, aggregation had little impact; however, on the oral passage reading error rate score, aggregation substantially increased all reliablity indices. In the second study, 78 children were tested 10 times on alternate forms of two reading measures and one written expression measure; group stability coefficients were calculated on the basis of 2, 4, 6, 8, and 10 testings. For the oral words-in-isolation reading correct score, aggregation had little effect, whereas aggregating over occasions and test forms dramatically improved the stability of the oral words-in-isolation reading error score and the written expression score. Implications for the measurement of academic behavior are discussed.

Use of Aggregation to Improve the Reliability of Simple Direct Measures of Academic Performance

According to the Standards for Educational and Psychological Tests (APA, 1972), criterion validity is a broad class of test validity that assesses the usefulness of a measure as a predictor of other variables. Criterion validity questions typically address the suitability of substituting a test for a longer, more cumbersome, or more expensive criterion. Therefore, the concern is with verifying the existence and strength of useful relationships under applied conditions (Messick, 1980).

Criterion-relatedness is determined by correlational analyses and extensions of correlational analyses to multivariate analyses. The most elementary example is the correlation of an individual predictor test with an individual criterion (Nunnally, 1978), where the strength of that correlation specifies the degree of predictive efficiency between the measures. In most criterion-related or prediction problems, psychometric theorists agree that it is reasonable to expect only modest correlations between a criterion and predictor test (Nunnally, 1978; Terwilliger, 1980). One reason for these modest correlations is the imprecision or unreliability that attentuates observed correlations (Stanley, 1971).

In studies of criterion validity, one method commonly employed to reduce random error and simultaneously to improve the extent to which true relationships are observed is to increase the sample size. However, as Epstein (1980) makes clear, a fundamental but widely ignored alternative strategy is to aggregate observations over situations and/or occasions. The law of sampling distributions holds

that behavior aggregated over stimuli or occasions as well as over individuals should reduce measurement error and improve the basis for establishing reliable, generalizable relationships.

In a series of four studies, Epstein (1979) demonstrated that aggregating over occasions, in fact, did render more reliable correlations. He found that when a wide range of personality measures each were averaged over an increasing number of occasions, stability coefficients, indicative of a measure's reliability or precision, increased to high levels. In these studies, Epstein found that relations between variables observed on one occasion were lower than, and sometimes opposite from, relations between the same variables observed and averaged over several occasions. This pattern held not only for personality measures, but also for direct observations of behavior and even a physiological index of heart rate.

The two experiments reported here examined the hypothesis that this phenomenon may apply to the measurement of academic behaviors. These investigations are relevant for educational measurement, in general, because they provide information concerning how to measure more accurately students' academic performance. More specifically, they are relevant for frequent measurement and continuous time-series evaluation strategies, where the practice of aggregating performance across occasions and/or test forms is routine, but where the frequency with which measurement need occur is unclear. Results of these studies should provide practitioners, who measure student performance on goals frequently and who formatively evaluate student programs, with information concerning how many data points are necessary before

reliable and valid estimates of student performance are achieved.

The first experiment explored questions related to the measurement of reading behavior on the same test sampled over occasions. The second investigated issues concerning the measurement of reading and written expression performance when behaviors are sampled over occasions and over parallel test forms.

Study 1

Study 1 posed three questions. First, it asked: How does aggregating students' scores on a test administered on several occasions affect stability in the measurement of academic performance? The study compared stability coefficients for reading behavior measured and averaged over two occasions with coefficients for the same behavior measured and averaged over four occasions.

The second question addressed in Study 1 was: If aggregation improves the stability of academic measures as explored through correlational analyses, then to what extent does it allow one to predict more accurately an individual's true score? To explore this question, within-subject reliability coefficients were examined, with subjects' behavior first observed on two occasions, then observed on and averaged over four occasions.

Question 3 in this study explored: How does aggregation over testing occasions affect the strength of relations between measures of academic performance? Specifically, the study compared the strength of relations between two reading behaviors when the data were collected on a single occasion with the strength of relations when data were collected on and averaged within subjects over four

occasions.

Method

Subjects. Ninety English speaking students, distributed across the six elementary grade levels, were selected randomly from one midwestern metropolitan school for inclusion in a separate study. From this pool of 90 students to whom the dependent measures were administered as part of a larger battery of tests, 30 subjects (M=15, F=15) evenly distributed among grades 1-6 were selected randomly.

Measures. The measures were: (a) from the Woodcock Reading Mastery Tests (Woodcock, 1973), the Word Identification Test of Form A (WRMT); and (b) from the Ginn 720 reading series, a 200 word passage, representative of the average readability (3.6) from the last 25% of level 8. (See Fuchs & Deno, 1981 for passage selection procedure.)

Procedure. According to a standard format, the 30 students were tested individually four times by a trained examiner. On one of these occasions, the measures were administered within a larger battery of tests; this testing session was approximately 60 minutes. Each of the other three sessions lasted approximately 10 minutes. Each student was assigned randomly to one of four groups, each of which received the longer battery at a different point in the sequence of the four administrations. Additionally, the order in which the measures were administered within a test session was random.

Data analyses. The data were subjected to three analyses. The first analysis was to obtain group stability coefficients within variables. These coefficients were obtained for the following variables: (a) the WRMT raw score, (b) the words correct per minute

score on the Ginn 720 reading passage, and (c) the errors per minute score on the Ginn 720 reading passage. Odd-even stability coefficients (Epstein, 1980), averaged first across two days (correlation between behavior on Day 1 and behavior on Day 2) and then across four days (correlation between behavior averaged over Days 1 and 3 and behavior averaged over Days 2 and 4), were calculated and compared.

A second analysis was conducted to obtain within-subject reliability coefficients. For the Ginn 720 correct per minute score and the Ginn 720 error per minute score, a reliability coefficient (percentage of overlap) was calculated between (a) Day 1 and Day 2, and (b) the average of Days 1 and 3 and the average of Days 2 and 4. These coefficients were compared for each variable.

A third analysis examined group correlations between variables. Correlations were calculated between (a) the WRMT raw score and the words per minute correct score on the Ginn passages, and (b) the WRMT raw score and the error per minute score on the Ginn passages. First, these correlations were based on each subject's performance on the first occasion. Then, the correlations were recalculated on the basis of the average of each subject's performance on each variable across the four occasions. The strength of relations based on one occasion was compared to the strength of relations based on four occasions.

Results

Question 1: How does aggregating students' scores on a test administered on several occasions affect stability in the measurement of academic performance? As displayed in Table 1, 2-day and 4-day

group stability coefficients for the three dependent variables were statistically significant ($p < .001$). The correlations for the WRMT raw score and the Ginn words correct score were high and similar; correlations were low for the Ginn error rate score. This indicates greater precision or reliability for the WRMT and correct rate scores relative to the error rate score.

Insert Table 1 about here

Within each measure, stability coefficients increased from 2-day to 4-day aggregations. The 2-day error rate coefficient initially was .18 (22%) lower than both the 2-day WRMT and the 2-day correct rate coefficients. However, the error rate 4-day coefficient improved .15 (19%) over its 2-day coefficient while the WRMT and correct rate coefficients remained nearly the same. Therefore, the 4-day error rate was very similar to the 4-day WRMT and the 4-day correct rate coefficients. It appears, then, that aggregation positively affected the reliability of error rate scores; it had no impact on the WRMT or the correct rate scores.

Question 2: To what extent does aggregation allow one to predict more accurately an individual's true score? Table 2 displays, for each measure, the within-subject reliability coefficients: (a) the mean percentage of overlap between scores on Day 1 and Day 2, (b) the mean percentage of overlap between the average of scores on Days 1 and 3 with the average of scores on Days 2 and 4, and (c) the mean within-subject changes between 2-day and 4-day coefficients. As with the

stability coefficients, these mean reliability coefficients were highest for the WRMT and lowest for the Ginn error rate measures. Again, small differences were noted between the 2-day and 4-day WRMT coefficients; the difference was slightly larger for Ginn correct rate and largest for Ginn error rate. Mean within-subject changes were ordered in a similar manner. Therefore, whereas the Ginn 2-day error rate coefficient was .31 (47%) below the WRMT 2-day coefficient, the error rate 4-day coefficient was only .25 (34%) below the WRMT 4-day coefficient. It appears that aggregation allows one to predict an individual's score more accurately for error rate, but has little effect on WRMT or correct rate scores.

Insert Table 2 about here

Question 3: How does aggregation over testing occasions affect the relation between measures of academic performance? Two sets of correlations were computed between (a) WRMT raw score and Ginn words correct rate score, and (b) WRMT raw score and Ginn error rate score. The first set was based on scores on one day; the second set was based on the average score across the four occasions. The correlations and their p-values are displayed in Table 3. All correlations were statistically significant. For the stable measures, WRMT and Ginn words correct rate scores, the 1-day coefficient was high (.91) and remained at approximately the same level when calculated on the basis of four days (.89). However, the correlation between WRMT scores and the least stable measure of error rate based on one day (-.46)

increased 18% when calculated on the basis of four days (-.53). As with the other analyses, then, it appears that aggregation affects the strength of relation when error rate is involved, but does not affect the strength of relation when correct rate is involved.

Insert Table 3 about here

Discussion

The WRMT and the Ginn correct scores initially were precise, reliable meaures, as evidenced by all three statistics, the 2-day group stability coefficients, and 2-day within-subject reliability coefficients, and the 1-day correlation between the WRMT and Ginn correct rate scores. For these initially reliable measures, aggregating on the same test over occasions made no important contribution to the measures' stability or to the strength of the relations between measures.

However, aggregating on the same test over occasions appeared to have an important effect on the least stable measure, the Ginn error rate. Aggregating over four days substantially enhanced the error rate group stability coefficients, the within-subject reliability coefficients, and the strength of relation between variables.

Additionally, the finding that error rate, the least reliable measure, manifested an initially weak relation with other measures corroborates other studies of criterion-relatedness between simple measures and achievement tests (Deno, Mirkin, Chiang, & Lowry, 1980; Fuchs & Deno, 1981). However, this study suggests that when

performance is sampled and aggregated across time, as is routinely done in frequent measurement and continuous evaluation, error rate becomes a more stable, reliable, precise measure and its criterion validity with other measures improves.

Study 2

While the effects of sampling on the same test form over occasions were explored in Study 1, the impact of sampling on parallel test forms over occasions was examined in Study 2. By aggregating performance across stimuli (test forms) in addition to aggregating performance over occasions, two types of error in pupils' scores potentially are reduced. First, with respect to aggregation across stimuli, the unique effects associated with particular stimuli are cancelled relative to their contribution to the test concept/skill on which all items converge. Second, aggregating over occasions cancels incidental effects associated with specific sessions. Both types of aggregation should enhance the reliability of a measure and increase the replicability of findings (Epstein, 1980). Therefore, the purpose of the second study was to examine the effect of aggregation across both test forms and occasions on group stability coefficients for academic measures.

Method

Subjects. Subjects were 78 children (M=48, F=30) selected from three public schools in a midwest metropolitan area. Each child, selected as "high-risk" for receiving special education services, scored at or below the 15th percentile on a short duration measure of written expression within his/her grade level (see measurement

procedures in Deno, Marston, & Mirkin, 1982). The numbers of children in grades 3-6, respectively, were 26, 17, 19, and 21.

Procedure. Once per week over a 10-week period, an alternate form of an oral word reading measure was administered individually to each child (Deno et al., 1980). Each alternate form was generated by randomly selecting words from the third grade level of the Harris-Jacobson Word List (Harris & Jacobson, 1972). The children's task was to read aloud words for one minute while the examiner recorded errors. Words read correctly per minute and errors per minute were scored.

During each testing session, a writing sample also was obtained. For this measure of written expression, each student was presented with an alternate form of a story starter each week and required to write on the story topic for three minutes. Number of correctly spelled words was scored.

Data analysis. Group stability coefficients were calculated for the reading word correct rate score, the error rate score, and the written expression measure score. The odd-even stability coefficients first were averaged across two observations (correlation between behavior on Week 1 and behavior on Week 2), then across four observations (correlation between behavior averaged over Weeks 1 and 3 and behavior averaged over Weeks 2 and 4), then across six observations (the average behavior over Weeks 1, 3, and 5 correlated with the average behavior over Weeks 2, 4, and 6), then across eight observations, and finally across 10 observations. Within variables, these correlations were compared.

Results

Table 4 displays 2, 4, 6, 8, and 10-day group stability coefficients for the three dependent variables. All correlations were statistically significant, and were consistently higher for the reading words correct score than for the reading error score or the written expression score.

Insert Table 4 about here

Within each measure, stability coefficients increased as the number of observations increased. The 2-day reading error rate coefficient initially was .69 (280%) lower than the 2-day correct rate coefficient; yet, the difference between the correct and error rate coefficients decreased as the number of observations increased so that, when coefficients were based on 10 observations, the error rate correlation was only .12 (13%) lower than the correct rate correlation. Consequently, the stability coefficient for the error rate score improved dramatically .62 (254%) over the increasing number of observations.

Similarly, the 2-day written expression coefficient was .39 (70%) lower than the 2-day reading words correct coefficient. Again, the difference between the reading words correct and written expression coefficients decreased as the number of observations increased. When coefficients were based on 10 observations, the written expression correlation was only .10 (11%) lower. It appears, then, that aggregation over test forms and occasions dramatically affects oral

reading error and written expression stability coefficients, but does not affect correct oral reading stability.

Discussion

The correct rate oral reading score again was an initially precise measure as evidenced by the group stability coefficients. For this initially reliable measure, aggregating over alternate forms of a test and over occasions made no real contribution to the measure's stability. However, as in Study 1, oral reading error rate was initially quite imprecise. Additionally, the written expression score initially was unreliable. Aggregating over alternate forms of a test and over occasions had a dramatic effect on these unstable measures, enhancing their stability to well within an acceptable level of alternate-form/test-retest reliability when the stability coefficients were based on aggregations over 10 observations.

Implications

The results of these two studies have several implications for the measurement of academic behavior. First, it appears that some academic behaviors initially are measured precisely. The WRMT, by all indices, rendered reliable student scores even when measurement was based on one observation. Given the documented strong psychometric adequacy of the WRMT, this may not be surprising. However, an interesting finding of these studies is that the simple, short duration measures of either oral correct word reading or oral correct passage reading were very precise, just as precise as the WRMT, when measurement was based on one occasion and/or on one test form. For these behaviors, aggregating on the same test over occasions had

little or no effect on group stability coefficients, on within-subject reliability, or on the strength of relations with other reliably observed behaviors. Similarly, for these initially precise measures, aggregating over alternate forms of the same test and over occasions did not affect group stability coefficients.

A second implication of these studies, nevertheless, is that other academic behaviors, such as the error Ginn passage reading measure, the error word reading measure, and the written expression measure, are not measured reliably on the same test form on one occasion. For those behaviors, aggregating over occasions had a positive impact on group stability coefficients, on within-subject reliability, and on the strength of relations between variables; similarly, aggregating over alternate test forms and over occasions dramatically affected group stability coefficients. Therefore, for certain academic behaviors, sampling on the same test form across time or on alternate test forms across time provides more precise information. This suggests the importance of aggregating a student's academic test performance across observations and/or test forms for certain behaviors, in order to ensure accurate information for decision making. These studies indicate a minimum of 5 to 10 data points are required for reliable estimation of children's performance on relatively imprecise measures such as oral reading errors or a written expression measure. As teachers increasingly use curriculum-based measurement to formulate decisions about students' progress toward goals, they might well consider aggregation as a means of improving the accuracy of their estimates of student performance and

the decisions they make.

Nevertheless, results of this study suggest that certain very simple, short duration academic measures, such as a one-minute correct oral word reading task and a one-minute correct oral passage reading test, are very stable and correlate highly with more elaborate, global, norm-referenced standardized tests such as the WRMT. Results of these studies demonstrate the reliability and criterion validity of such short, simple measures, and suggest the suitability of substituting them for more elaborate and time-consuming measures of academic performance.

References.

APA. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1972.

Deno, S. L., Marston, D., & Mirkin, P. K. Valid measurement procedures for continuous evaluation of written expression. Exceptional Children, 1982, 48, 368-370.

Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. Relationships among simple measures of reading and performance on standardized achievement tests (Research Report No. 21). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1980. (ERIC Document Reproduction Service No. ED 197 508)

Epstein, S. The stability of behavior: I. On predicting most of the people much of the time. Journal of Personality and Social Psychology, 1979, 37(7), 1097-1126.

Epstein, S. The stability of behavior: II. Implications for psychological research. American Psychologist, 1980, 35(9), 790-806.

Fuchs, L. S., & Deno, S. L. The relationship between curriculum-based mastery measures and standardized achievement tests in reading (Research Report No. 57). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981. (ERIC Document Reproduction Service No. ED 212 662)

Harris, A. J., & Jacobson, M. D. Basic elementary reading vocabularies. New York: MacMillan, 1972.

Messick, S. Test validity and the ethics of assessment. American Psychologist, 1980, 35(11), 1012-1027.

Nunnally, J. C. Psychometric theory. New York: McGraw Hill, 1978.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Terwilliger, J. Personal communication. October, 1980.

Woodcock, R. Woodcock Reading Mastery Tests. Circle Pines, MN: American Guidance Service, 1973.

Table 1

Group Stability Coefficients[a] (N=30)

Measures	Stability Coefficients 2-day	4-day
Woodcock Word Identification Test - raw score	.96	.98
Ginn 720, 3rd grade reading passage - words correct per minute	.96	.96
Ginn 720, 3rd grade reading passage - errors per minute	.78	.93

[a]All correlations are statistically significant ($p < .001$).

Table 2

Within Subject Reliability Coefficients (N=30)

Measure	2-day coefficient	4-day coefficient	within subject change from 2-day to 4-day coefficient
WRMT	.96	.97	.012
Ginn Correct Rate	.85	.88	.036
Ginn Correct Rate	.65	.72	.080

Table 3

Correlations Between Variables Calculated on One-Day Scores and on the Means of Four-Day Scores (N=30)

Between	Correlations and p-values 1-day coefficient	p-value	4-day coefficient	p-value
WRMT and Ginn Correct Rate	.91	.001	.89	.001
WRMT and Ginn Error Rate	-.46	.011	-.54	.003

Table 4

Two, Four, Six, Eight, and Ten-Day Observation Stability Coefficients[a] (N=78)

	Stability Coefficients				
	2-Observation	4-Observation	6-Observation	8-Observation	10-Observation
Reading Words Correct Rate	.94	.96	.98	.98	.99
Reading Error Rate	.25*	.58	.76	.83	.87
Writing Words	.55	.72	.85	.88	.89

[a]All correlations are statistically significant (p=.001, except *p=.015).

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 $3.00 per document, 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.

Ysseldyke, J. E. Assessing the learning disabled youngster: The state of the art (Research Report No. 1). November, 1977.

Ysseldyke, J. E., & Regan, R. R. Nondiscriminatory assessment and decision making (Monograph No. 7). February, 1979.

Foster, G., Algozzine, B., & Ysseldyke, J. Susceptibility to stereotypic bias (Research Report No. 3). March, 1979.

Algozzine, B. An analysis of the disturbingness and acceptability of behaviors as a function of diagnostic label (Research Report No. 4). March, 1979.

Algozzine, B., & McGraw, K. Diagnostic testing in mathematics: An extension of the PIAT? (Research Report No. 5). March, 1979.

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

Ysseldyke, J. E., & Mirkin, P. K. Proceedings of the Minnesota round-table conference on assessment of learning disabled children (Monograph No. 8). April, 1979.

Somwaru, J. P. A new approach to the assessment of learning disabilities (Monograph No. 9). April, 1979.

Algozzine, B., Forgnone, C., Mercer, C. D., & Trifiletti, J. J. Toward defining discrepancies for specific learning disabilities: An analysis and alternatives (Research Report No. 7). June, 1979.

Algozzine, B. The disturbing child: A validation report (Research Report No. 8). June, 1979.

Note: Monographs No. 1 - 6 and Research Report No. 2 are not available for distribution. These documents were part of the Institute's 1979-1980 continuation proposal, and/or are out of print.

Ysseldyke, J. E., Algozzine, B., Regan, R., & Potter, M. Technical adequacy of tests used by professionals in simulated decision making (Research Report No. 9). July, 1979.

Jenkins, J. R., Deno, S. L., & Mirkin, P. K. Measuring pupil progress toward the least restrictive environment (Monograph No. 10). August, 1979.

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

Thurlow, M. L., & Ysseldyke, J. E. Current assessment and decision-making practices in model programs for the learning disabled (Research Report No. 11). August, 1979.

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

Ysseldyke, J. E., Algozzine, B., Shinn, M., & McGue, M. Similarities and differences between underachievers and students labeled learning disabled: Identical twins with different mothers (Research Report No. 13). September, 1979.

Ysseldyke, J., & Algozzine, R. Perspectives on assessment of learning disabled students (Monograph No. 11). October, 1979.

Poland, S. F., Ysseldyke, J. E., Thurlow, M. L., & Mirkin, P. K. Current assessment and decision-making practices in school settings as reported by directors of special education (Research Report No. 14). November, 1979.

McGue, M., Shinn, M., & Ysseldyke, J. Validity of the Woodcock-Johnson psycho-educational battery with learning disabled students (Research Report No. 15). November, 1979.

Deno, S., Mirkin, P., & Shinn, M. Behavioral perspectives on the assessment of learning disabled children (Monograph No. 12). November, 1979.

Sutherland, J. H., Algozzine, B., Ysseldyke, J. E., & Young, S. What can I say after I say LD? (Research Report No. 16). December, 1979.

Deno, S. L., & Mirkin, P. K. Data-based IEP development: An approach to substantive compliance (Monograph No. 13). December, 1979.

Ysseldyke, J., Algozzine, B., Regan, R., & McGue, M. The influence of test scores and naturally-occurring pupil characteristics on psycho-educational decision making with children (Research Report No. 17). December, 1979.

Algozzine, B., & Ysseldyke, J. E. Decision makers' prediction of students' academic difficulties as a function of referral information (Research Report No. 18). December, 1979.

Ysseldyke, J. E., & Algozzine, B. Diagnostic classification decisions as a function of referral information (Research Report No. 19). January, 1980.

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

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

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

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

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

Thurlow, M. L., & Ysseldyke, J. E. Factors influential on the psychoeducational decisions reached by teams of educators (Research Report No. 25). February, 1980.

Ysseldyke, J. E., & Algozzine, B. Diagnostic decision making in individuals susceptible to biasing information presented in the referral case folder (Research Report No. 26). March, 1980.

Thurlow, M. L., & Greener, J. W. Preliminary evidence on information considered useful in instructional planning (Research Report No. 27). March, 1980.

Ysseldyke, J. E., Regan, R. R., & Schwartz, S. Z. The use of technically adequate tests in psychoeducational decision making (Research Report No. 28). April, 1980.

Richey, L., Potter, M., & Ysseldyke, J. Teachers' expectations for the siblings of learning disabled and non-learning disabled students: A pilot study (Research Report No. 29). May, 1980.

Thurlow, M. L., & Ysseldyke, J. E. Instructional planning: Information collected by school psychologists vs. information considered useful by teachers (Research Report No. 30). June, 1980.

Algozzine, B., Webber, J., Campbell, M., Moore, S., & Gilliam, J. Classroom decision making as a function of diagnostic labels and perceived competence (Research Report No. 31). June, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R. R., Potter, M., Richey, L., & Thurlow, M. L. Psychoeducational assessment and decision making: A computer-simulated investigation (Research Report No. 32). July, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R. R., Potter, M., & Richey, L. Psychoeducational assessment and decision-making: Individual case studies (Research Report No. 33). July, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R., Potter, M., & Richey, L. Technical supplement for computer-simulated investigations of the psychoeducational assessment and decision-making process (Research Report No. 34). July, 1980.

Algozzine, B., Stevens, L., Costello, C., Beattie, J., & Schmid, R. Classroom perspectives of LD and other special education teachers (Research Report No. 35). July, 1980.

Algozzine, B., Siders, J., Siders, J., & Beattie, J. Using assessment information to plan reading instructional programs: Error analysis and word attack skills (Monograph No. 14). July, 1980.

Ysseldyke, J., Shinn, M., & Epps, S. A comparison of the WISC-R and the Woodcock-Johnson Tests of Cognitive Ability (Research Report No. 36). July, 1980.

Algozzine, B., & Ysseldyke, J. E. An analysis of difference score reliabilities on three measures with a sample of low achieving youngsters (Research Report No. 37). August, 1980.

Shinn, M., Algozzine, B., Marston, D., & Ysseldyke, J. A theoretical analysis of the performance of learning disabled students on the Woodcock-Johnson Psycho-Educational Battery (Research Report No. 38). August, 1980.

Richey, L. S., Ysseldyke, J., Potter, M., Regan, R. R., & Greener, J. Teachers' attitudes and expectations for siblings of learning disabled children (Research Report No. 39). August, 1980.

Ysseldyke, J. E., Algozzine, B., & Thurlow, M. L. (Eds.). A naturalistic investigation of special education team meetings (Research Report No. 40). August, 1980.

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

Fuchs, D., Garwick, D. R., Featherstone, N., & Fuchs, L. S. On the determinants and prediction of handicapped children's differential test performance with familiar and unfamiliar examiners (Research Report No. 42). September, 1980.

Algozzine, B., & Stoller, L. Effects of labels and competence on teachers' attributions for a student (Research Report No. 43). September, 1980.

Ysseldyke, J. E., & Thurlow, M. L. (Eds.). The special education assessment and decision-making process: Seven case studies (Research Report No. 44). September, 1980.

Ysseldyke, J. E., Algozzine, B., Potter, M., & Regan, R. A descriptive study of students enrolled in a program for the severely learning disabled (Research Report No. 45). September, 1980.

Marston, D. Analysis of subtest scatter on the tests of cognitive ability from the Woodcock-Johnson Psycho-Educational Battery (Research Report No. 46). October, 1980.

Algozzine, B., Ysseldyke, J. E., & Shinn, M. Identifying children with learning disabilities: When is a discrepancy severe? (Research Report No. 47). November, 1980.

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

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 (Research Report No. 49). January, 1981.

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

Epps, S., McGue, M., & Ysseldyke, J. E. Inter-judge agreement in classifying students as learning disabled (Research Report No. 51). February, 1981.

Epps, S., Ysseldyke, J. E., & McGue, M. Differentiating LD and non-LD students: "I know one when I see one" (Research Report No. 52). March, 1981.

Evans, P. R., & Peham, M. A. S. Testing and measurement in occupational therapy. A review of current practice with special emphasis on the Southern California Sensory Integration Tests (Monograph No. 15). April, 1981.

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

Fuchs, D., Featherstone, N., Garwick, D. R., & Fuchs, L. S. The importance of situational factors and task demands to handicapped children's test performance (Research Report No. 54). June, 1981.

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

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

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

Christenson, S., Graden, J., Potter, M., & Ysseldyke, J. Current research on psychoeducational assessment and decision making: Implications for training and practice (Monograph No. 16). September, 1981.

Christenson, S., Ysseldyke, J., & Algozzine, B. Institutional constraints and external pressures influencing referral decisions (Research Report No. 58). October, 1981.

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

Algozzine, B., Christenson, S., & Ysseldyke, J. Probabilities associated with the referral-to-placement process (Research Report No. 60). November, 1981.

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

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

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.

Martson, 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.