DOCUMENT RESUME

ED 236 190 TM 830 699

AUTHOR Tindal, Gerald; And Others
TITLE Visual Analysis of Time Series Data: Factors of Influence and Level of Reliability.
INSTITUTION Minnesota Univ., Minneapolis. Inst. for Research on Learning Disabilities.
SPONS AGENCY Office of Special Education and Rehabilitative Services (ED), Washington, DC.
REPORT NO IRLD-RR-112
PUB DATE Mar 83
CONTRACT 300-80-0622
NOTE 49p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Data Analysis; Evaluation Methods; *Evaluation Utilization; Program Effectiveness; *Program Evaluation; *Reliability; Research Methodology; Visual Discrimination; *Visual Perception
IDENTIFIERS *Data Interpretation; *Time Series Analysis

ABSTRACT

The focus of this study was on the visual analysis of time series data for evaluating educational programs. Two characteristics of the data--changes in slope and variability--and two characteristics of evaluation--training in data utilization and the use of aimlines/decision rules--were manipulated. A total of 51 students and/or teachers in education evaluated a set of 28 graphs on two dimensions: (1) Was the program depicted on the graph an effective program? and (2) What about the data supported such a conclusion? Findings of the study indicated that visual analysis is not very reliable for evaluating educational programs, and is influenced considerably by the characteristics of the data array (specifically, slope and variability). Training in data utilization or the use of aimlines did not appear to be particularly powerful procedures for improving visual analysis. At the same time, the findings indicated evaluation consistent with established data analysis paradigms. Implications for training in visual analysis are discussed. (Author)

University of Minnesota

Research Report No. 112

VISUAL ANALYSIS OF TIME SERIES DATA: FACTORS OF INFLUENCE AND LEVEL OF RELIABILITY

Gerald Tindal, Stanley L. Deno, and James E. Ysseldyke

IRLD

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).

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 the Office of Special Education.

Research Report No. 112

VISUAL ANALYSIS OF TIME SERIES DATA: FACTORS OF INFLUENCE AND LEVEL OF RELIABILITY

Gerald Tindal, Stanley L. Deno, and James E. Ysseldyke
Institute for Research on Learning Disabilities
University of Minnesota

March, 1983

Abstract

The focus of this study was on the visual analysis of time series data for evaluating educational programs. Two characteristics of the data--changes in slope and variability--and two characteristics of evaluation--training in data utilization and the use of aimlines/decision rules--were manipulated. A total of 51 students and/or teachers in education evaluated a set of 28 graphs on two dimensions: (a) Was the program depicted on the graph an effective program? and (b) What about the data supported such a conclusion? Findings of the study indicated that visual analysis is not very reliable for evaluating educational programs, and is influenced considerably by the characteristics of the data array (specifically, slope and variability). Training in data utilization or the use of aimlines did not appear to be particularly powerful procedures for improving visual analysis. At the same time, the findings indicated evaluation consistent with established data analysis paradigms. Implications for training in visual analysis are discussed.

Visual Analysis of Time Series Data:
Factors of Influence and Level of Reliability

While behavioral interventions have been well documented and empirically supported over the past several decades, the appropriate analysis of behavioral data has not been explicated with the same results. Most behavioral research is based upon, indeed predicated upon, the use of time-series data, which typically are graphed on either equal-interval or semi-log graphs. Until recently, there have been ver few methods available for analyzing such data.

The historical roots of the experimental analysis of behavior (Sidman, 1960; Skinner, 1953) have held theoretical and practical sway against the use of statistics in data analysis. The visual analysis of graphed data has been the most accepted basis for judgments of the adequacy and meaningfulness of interventions:

> determination of change is dependent on the change being of sufficient magnitude to be apparent to the eye. Compared with the potential algebraic sophistication of statistical tests of significance, (not always realized in practice), the above procedure usually is relatively insensitive. (Parsonson & Baer, 1978, p. 111)

That is, it is contended that the reliance upon visual analysis, an admittedly less sensitive measurement technique, results in an inherent bias against the selection of weak and unstable variables (Baer, Wolf, & Risley, 1968). Minor effects are not seen as "change."

There is a very low probability of Type I errors and consequently a high probability of Type II errors. Type I errors result when a conclusion of "an effect" is made when in actuality no effect is present. Type II errors represent an error in the opposite direction: a conclusion of "no effect" is made when, in reality, an effect is present. Pechacek (1978) investigated the validity of N-1 designs (both reversal and multiple baseline designs) by means of a probabilistic model using visual analysis of effects as his criterion. Given three possible outcomes (increase, decrease, or no change in behavior), both the basic ABAB design and multiple baseline designs using four baselines were found to possess a probability estimate of Type I error well below the traditional .05 level. Although statistical analysis would often simply corroborate such findings, it is also true that effects would have been found for many less powerful and stable variables, serving "only to confound, complicate, and delay the development of a functional analysis of behavior" (Parsonson & Baer, 1978, p. 113). It is quite likely that if statistical analysis is needed to demonstrate certain effects, there will be problems in replicating those effects later (Kazdin, 1976).

Furthermore, as Michael (1974) notes, an emphasis on statistics and the elaboration of statistical control of unwanted sources of variation in the dependent variable, will likely result in a reduction in the necessity for developing experimental control. The harmful consequences engendered in devoting more time and effort to the use of statistics include the loss of a source of ideas for further experimentation, reliance upon less useful knowledge having limited applicability, the design of experiments having less generality and "replicability," excessive dependence upon statistical tests of significance, and experiments being designed in more complex and less flexible manners. In the final analysis, he believes that time spent in learning how to use and interpret statistical procedures will simply take time away from the primary subject of interest, a functional analysis of behavior.

An unfortunate side effect of this controversy is that far more effort has been invested in the development of statistical procedures than in explicating important variables in visual analysis. This is an important line of research in which more attention needs to be given to the technology of graphing and the development of major guidelines for use in "seeing change." Although visual analysis has been the most frequently used procedure for data analysis in applied behavioral research, there is little empirical evidence regarding its technical adequacy. Most of the studies that have been conducted have compared visual analysis with statistical analysis. This research indicates that inconsistent conclusions have occurred in judging N=1 data through visual analysis when compared to statistical inference criteria (DeProspero & Cohen, 1979; Glass, Willson, & Gottman, 1975; Jones, Vaught, & Weinrott, 1977; Jones, Weinrott, & Vaught, 1975).

These investigations clearly demonstrate that visual analysis of time series data may be suspect when compared to statistical analysis. However, it also is important to know what aspects threaten the statistical conclusion validity (Cook & Campbell, 1976) of this type of analysis. A better understanding of the components of visual analysis would provide a basis for improving its accuracy and reliability. Three studies have been conducted with this purpose in mind.

DeProspero and Cohen (1979) investigated the degree to which agreement in visual judgment could be attributed reliably to certain features of the graph. Using a set of simulated "ABAB reversal design" graphs, they systematically varied the pattern and degree of mean shift across phases, variability within phases, and trend. Their results indicated that the pattern of mean shift was a critical characteristic, with the average rating of effectiveness falling off very rapidly for any pattern other than the "ideal" one of change congruent with the hypothesized effect. They also found the degree of mean shift to have a reliable effect upon the average rating. The interrater agreement of the judges in this study was .61 overall; no data were reported on reliability within each of the graphic characteristics investigated. The evaluative criteria employed by the judges fell into four cluster statements. Most frequently mentioned was the topography of the scores - their trend, means, and stability. The format of presentation was mentioned next most frequently, followed by intra- and extra-experimental concerns. Although DeProspero and Cohen (1979) "attempted to assess the factors contributing to reliable or unreliable visual judgment," they concluded that "graphic characteristics appear to determine judgments in concert rather than singly" (p. 578).

An investigation to ascertain the extent to which serial dependency influenced the agreement between inferences based on visual or time-series analysis was conducted by Jones, Weinrott, and Vaught (1978). JABA graphs were presented to judges well versed in behavior charting and they were asked whether a meaningful change in level had been demonstrated from one phase to another. The authors selected graphs in which the effects were sufficiently "nonobvious" to warrant critical analysis, and serial dependency was apparent. The graphs were blocked further into three different levels of serial dependency by two levels of significance of differcnce in level between phases. Their results indicated that agreement between visual analysis and time-series analysis were inversely related to the magnitude of the serial dependency in the scores. That is, the more serial dependency present (with a significant difference in level), the less reliable visual analysis tended to be. Furthermore, they found that visual and time-series inferences agreed better when the statistical test showed non-significant changes in level than when significant changes in level were indicated. Finally, an interaction effect was present in which visual and time-series inferences agreed most when the data showed neither serial dependency nor significant differences in level. In effect, judges tended to agree with time-series analysis that no effect was present but disagreed most when an effect was present. Intercorrelations among the 11 judges ranged from .04 to .79, with a median of .39, suggesting fairly low consensus among judges and indicting the dependability of visual inferences. However, there was no relationship found between the reliability of the judges and the degree of agreement with time-series inferences.

Jones et al. (1978) consider their findings (low agreement with high serial dependency and statistically reliable changes in level, and high agreement with low serial dependency and unreliable changes in level) to be contrary to the unlikely and/or undesirable purpose of research using an operant paradigm. They conclude that "statistically reliable experimental effects may be more often overlooked by visual appraisals of data than nonmeaningful effects" (p. 280). Their suggestion to use time-series analysis to supplement visual analysis (Jones et al., 1977) would result in an increase in the number of meaningful changes inferred.

The final study (Wampold & Furlong, 1981) of visual inference focused on an explication based on schema theory. It was hypothesized that the process of visually analyzing time series data was primarily a classification problem controlled by previous training in visual inference through the use of model data - prototypes and exemplars. Furthermore, this training typically has been characterized by the presentation of prototypes and exemplars demonstrating large changes and little variability (small distance exemplars) in single subject designs.

The primary purpose of the study by Wampold and Furlong was to compare graph analyses of subjects trained in different analytic procedures. Specifically, it was hypothesized that subjects trained in behavior analysis (with a focus on prototypes and small distance exemplars) would analyze graphed data differently than subjects trained in advanced statistical procedures (having little or no contact with the prototype or exemplars typically found in the behavioral literature). Additionally, the ability to discriminate between different intervention effects was investigated by analyzing differential reactions to graphs that demonstrated either a change in level, a change in trend, or a change in both level and trend.

The stimulus materials to which all subjects responded included a series of three graphs, two of which were kept functionally equivalent (had the same size of intervention effect in relation to the variation), and the third depicting a smaller intervention effect (relative to the variability). In addition, each of these three types of graphs displayed a change from phase 1 to phase 2 in: (a) level, (b) trend, or (c) level and trend.

The results from this research provided support "for the hypothesis that subjects trained primarily in visual inference would be more prone to attend to large differences while ignoring variation in graphic data than would subjects primarily trained in statistics" (p. 89). Additionally, it was determined that the subjects trained in visual inference were less able to differentiate the intervention effects than were the subjects trained in classical statistical procedures. It must be noted, however, that neither group performed the sorting task exceptionally well, with only 36% of the N=1 subjects and 50% of the statistically trained group responding appropriately to the experimental stimuli.

In summary, it appears that visual analysis of time series data is a tenuous proposition at best and is influenced negatively by such characteristics of the data as: (a) nonconformity to an ideal and hypothesized pattern; (b) serial dependency; and (c) variability relative to certain changes in slope and trend. However, the studies have differed in major ways, including the stimulus materials used in the research and the population of subjects examined. The purpose of this research was to examine another variation in methodology and to focus on additional characteristics of time series data that influence visual analysis. The main reason for this pursuit is related to the shortcomings of previous research. The DeProspero and Cohen (1979) study provided few meaningful or specific findings having implications for improving the visual analysis of time series data. The Jones et al. (1978) study manipulated a statistical variable not readily amenable to manipulation in the field, although the findings are quite relevant. Finally, Wampold and Furlong (1981) looked at change between different time series rather than within various time series, limiting any interpretations that can be made.

As important as these methodological considerations are, however, the populations of subjects used by these researchers provide another critical reason for conducting further research. The subjects in this previous research were graduate students and/or professionals with considerable experience in data analysis. Because of this, this previous research simply has been inadequate for answering the question of the effects of training on school teachers. The focus of the current research was on the interpretation of time series data for purposes of making educational decisions involved in program evaluation. Therefore, to provide external validity to this investigation, it was imperative that the population of subjects sampled was appropriate to the population of public school educators.

Method

Subjects

Subjects for this study were in-service and pre-service teachers from three different locations around a large midwestern city. Two of the sites were school districts, accounting for nine of the subjects, all of whom were currently teaching. Teachers in these two sites were randomly assigned to different treatment conditions, with the three subjects from one district assigned to the experimental group and the six from the other district assigned to the control group.

The remaining 42 subjects were students taking a required special education class at a large midwestern university. Most subjects were currently teaching or were former teachers. Subjects from this pool were randomly assigned to treatment groups in proportion to the number needed for bringing both groups to the same size. Twenty students were assigned to the control group and 28 assigned to the experimental group.

Training Procedures

The training of subjects involved both an in-service workshop and a 'take-home' training module. The teachers in the experimental group were given training in the analysis of graphed data for evaluating instructional programs. This entailed explanations and exercises in summarizing student performance and using it to make interpretations. Included in the summarization of time-series data were computations of step changes, medians, slopes (using the split-middle technique; White, 1971), variability (using total bounce; Pennypacker, Koenig, & Lindsley, 1972), and overlap (Parsonson & Baer, 1978). A portion of the workshop also was devoted to the use of this information for evaluating instruction.

The teachers in the control group were given training in the development of measurement techniques in the areas of reading, writing, and spelling. They were trained in assessing students to determine performance discrepancies, sampling curriculum materials to find an appropriate instructional level, and developing a measurement system to monitor student improvement.

Both workshops lasted approximately 2½ hours. Following the workshop, the experimental materials (graphs, response sheets, and directions) for 14 graphs were distributed. Following completion of these graphs (which ranged from one week for the subjects in the class to three weeks for subjects in the schools), a second set of 14 graphs were distributed. The completion and return of this material again took one week for the subjects in the class and three weeks for those in the schools.

Materials

A total of 28 different graphs was constructed in which slope and variability were systematically manipulated. Two phases were displayed in each graph - 11 data points in baseline and 15 data points in the intervention phase. A vertical line was drawn separating the two phases. The aimline represented a 30% improvement over the median of the last three days during baseline. To ensure comparability between the graphs with and without aimlines, the absolute level of this median value was nearly the same across both aimline conditions within each respective level of slope. Although the slope was manipulated only in the intervention phase, variability was manipulated in both baseline and during the intervention. A total of three levels of slope and four conditions of variability were included in the graphs. (Details of the procedures for constructing the graphs are presented in Appendix A.)

With variability manipulated in both baseline and intervention, two different combinations of variability were included: a bounce of 5 data points and one of 15 data points. For every combination of slope, variability increased (5-15), decreased (15-5), remained at the same low level (5-5) or remained at the same high level (15-15). This resulted in the following combinations of graphed data:

(a) Six graphs showed an increase in variability from baseline to intervention from 5 data points bounce to 15 data points bounce, with a concurrent increase in slope from 0 to 10 degrees for 2 graphs, an increase from 0 to 15 degrees in two graphs, and an increase from 0 to 20 degrees for the final two graphs. Of these six graphs, three had an aimline drawn in during the intervention phase, one for each combination of slope and variability.

(b) Six graphs showed a decrease in variability from 15 data points bounce in baseline to 5 data points bounce in the intervention phase. For two of these graphs, the change in slope from baseline to intervention involved an increase from 0 to 10 degrees, two graphs depicted an increase from 0 to 15 degrees, and two had an increase from 0 to 20 degrees. Again, an aimline was drawn in on half (three) of the above graphs, one from each combination.

(c) Six graphs showed steady (unchanging) variability at a low level (5 data points bounce) from baseline to intervention. Again, the slope changed from 0 to 10 degrees on two of the graphs, 0 to 15 degrees on two graphs, and 0 to 20 degrees on the final two graphs. For each pair of slope-variability, one had an aimline and one did not.

(d) Six graphs showed steady (unchanging) variability at a high level (15 data points bounce) from baseline to intervention. Each level of slope (10, 15, and 20 degrees) was represented; two graphs displayed a change of 0 to 10 degrees, two graphs displayed a change of 0 to 15 degrees, and two showed a change of 0 to 20 degrees. Again, aimlines were present on half of these - one in each combination of slope-variability.

The final four graphs had the following characteristics:

(e) Four graphs which were given at time 1 were again given at time 2, with exactly the same data array depicted. All of these graphs displayed a low slope change (0 to 10 degrees) and constant variability (either the same low or same high variability). For each of the two variability conditions, one had an aimline present and one had no aimline present.

Dependent Variables

As noted previously, each subject was given 14 of the graphs immediately following training. Each graph had a response sheet which included two primary questions (see Appendix B):

(1) Was the intervention depicted on the graph an effective one? Response to this question consisted of rating the effectiveness on a 1-4 scale, with 1 being definitely not effective and 4 being definitely effective.

(2) What about the data led them to the above conclusion? Response to this question was a short answer description of anything in the data array that they were particularly attentive to while making their judgment.

After the first set of 14 graphs and responses were collected, another set of 14 graphs was distributed. The order in which the graphs were organized (and completed) was determined randomly for both groups of subjects.

Results

What Influence Does Slope and Variability have on Ratings of Intervention Effectiveness?

The average ratings of intervention effectiveness are summarized in Table 1. A significant difference was found between the three levels of slope, $F(2,98) = 116.4$, $p \leq .000$, and the four conditions of variability, $F(3,147) = 14.2$, $p \leq .000$, as well as the interaction between slope and variability, $F(6,294) = 22.8$, $p \leq .000$. The average ratings for the three levels of slope increased monotonically for 10, 15, and 20 degrees, respectively. For the four conditions of variability, the average ratings were higher for decreased variability

(2.81) and high constant variability (2.74), and lower for increased variability (2.47) and low constant variability (2.45).

Insert Table 1 about here

The interaction between slope and variability is depicted in Figure 1. When variability was constant, there was a linear increase in the ratings of intervention effectiveness across slope levels. When variability changed (either increased or decreased), similar ratings were given for both of the lower levels of slope (10 and 15 degrees), regardless of the direction of the change. However, with a 20 degree slope, there was a substantial increase in the ratings of effectiveness when variability decreased, with little change in the rating when variability increased.

Insert Figure 1 about here

Data on the reliability of ratings for the three levels of slope and four conditions of variability are summarized in Table 2. The relationship between slope and reliability appeared to be mediated by the influence of variability. Of the three levels of slope, the lowest reliability occurred with the intermediate slope level (15 degrees). While the greatest reliability occurred with the steepest slope (20 degrees), there was one exception. Under conditions of increased variability, the highest reliability occurred with the lowest slope (10 degrees). When variability decreased, the reliability was highest when the slope was steep (20 degrees). In these two conditions of variability, reliability deteriorated considerably with low increases in slope (from 10 to 15 degrees).

Inse:t Table 2 about here

The overall influence of variability on the average reliability of ratings was most pronounced when variability increased. Under that condition, the average reliability was the lowest. The difference between the other conditions of variability, however, was considerably less. The effect of variability on reliability also appeared to be mediated by the level of slope. With a low slope of 10 degrees, there was little change in reliability across the various conditions of variability. When the slope was higher (15 and 20 degrees), reliability changed with changes in variability. For a 15 degree slope, reliability was highest when variability was constant (either low or high). In contrast, with a slope of 20 degrees, the reliability was highest when variability decreased or remained low and constant.

There appeared to be little differential effect on the stability (reliability) of ratings from time 1 to time 2 under conditions of constant variability (see Table 3). Very similar findings appeared whether or not the variability had been low.

Insert Table 3 about here

What Influence Does the Use of Aimlines and Training in Data Utilization have on Ratings of Intervention Effectiveness?

The results of the rating of intervention effectiveness are summarized in Table 4. Although a significant effect was found for training, F(1,49) = 14.0, p $\leq$ .000, there was no effect found for the use of aimlines, F(1,49) = 0.36, p $\leq$ .552, or the interaction between the use of aimlines and training in data utilization, F(1,49) = 2.7, p $\leq$ .105. The average rating by trained subjects was less than the rating by untrained subjects. In contrast to this significant difference, nearly the same ratings were given when aimlines were present as when they were absent.

Insert Table 4 about here

What Influence Does the Use of Aimlines and Training in Data Utilization Have on the Reliability of Ratings of Intervention Effectiveness?

There was little difference in the average reliability (consensus) across training and aimline conditions (see Table 5), with the range from .51-.54. Trained subjects were slightly more reliable when aimlines were present (.54 vs .51), while untrained subjects showed no difference in reliability across this dimension (.52). The difference in reliability between trained and untrained subjects was very slight (.01 to .03).

Insert Table 5 about here

A greater difference was apparent in the reliability of ratings over time for the training and aimline condition (see Table 6). Trained subjects were considerably more reliable from time 1 to time 2 than untrained subjects, regardless of the presence (or lack) of aimlines. While trained subjects were more reliable when aimlines were present than when they were absent, untrained subjects were actually more reliable without aimlines from time 1 to time 2.

Insert Table 6 about here

What Type of Data Dimensions are Utilized by Trained and Untrained Subjects in their Ratings of Intervention Effectiveness?

For each graph, subjects were asked to describe any characteristic of the data array that influenced their judgments. Their responses were categorized into nine dimensions of time series data that summarize and describe change over time. These categories were structured around various statistical summarizations, each one providing unique information for evaluating change in performance. In most cases, there were many different descriptions of any particular characteristic of the data, though reference was obviously to the same dimension. Following is a list of the categories and a brief explanation/definition using the various terms listed by the subjects:

(a) Progress - nonspecific statements of changes in performance over time. Synonomous terms included slope, upward (downward) movement, rate increases (decreases), acceleration, improvement, gains.

(b) Variability - descriptions of day-to-day variation in performance. Synonomous terms included scatter, fluctuation, range, (in)consistency, (un)stable, steady, gradual, sporadic, (un)predictable.

(c) Jump - immediate change in performance from the last day of baseline to the first day of the intervention phase. Other terms included changes in step, or level, and immediate increase (decrease) in performance.

(d) Direction - comparison of slope from baseline to intervention or within the intervention phase from the beginning to the end. Also included in this category were statements describing a leveling off or a previously flat (downward) slope as now increasing.

(e) Number of Days of increases and decreases relative to any index: previous days, baseline, slope, aimline, overlap. Statements that implied counting also were included, allowing for descriptions of performance as being "consistently," "never," "always," "the majority of time" over (under) the above indices.

(f) Goal/Aim - use of goals or aimlines to qualify interpretations of performance, including any comparison of actual to expected performance.

(g) Average Performance - use of a composite summarizing index for measuring change between baseline and intervention or within the intervention phase, from beginning to end, including mean, average, median, or percent.

(h) Overlap - reference to the band within which scores fall across phases. Any statements taking note of simultaneous comparison of high and low points between phases were included in this category.

(i) Absolute Values - use of numbers from the graph representing single-point values, including high or low scores and/or the difference between them, or the last day of baseline, the last day of intervention and/or the difference between them.

Table 7 contains the means and standard deviations of the number of references made to each of the characteristics. There was no difference between trained and untrained subjects on only two dimensions: progress and the number of days improved. For the remaining dimensions there were significant differences between the two groups. Trained subjects referred more often to every dimension except absolute values. Untrained subjects referred to this characteristic significantly more often than trained subjects. In addition, the range of frequencies across the various dimensions was quite great. Reference was made most often to progress and variability for both groups. The only dimension not used very frequently by trained subjects was absolute values. In contrast, untrained subjects rarely referred to jump, direction, and overlap.

Insert Table 7 about here

Another analysis of this same variable - frequency of reference to data characteristics - was conducted on the number of different dimensions mentioned for each graph. The results indicated a significant main effect for changes in slope, F(2,98) = 11.2, p $\leq$ .000. The difference between the three levels of slope revealed an interesting relationship (see Table 8). More dimensions were referred to when the slope was 15 degrees. In contrast, when the slope was 10 or 20 degrees, this number dropped. No significant effects were found for variability, F(3,147) = .49, p $\leq$ .690, or the interaction between slope and variability, F(6,294) = 1.1, p $\leq$ .348.

Insert Table 8 about here

Table 9 is a summary of the frequency of reference to data dimensions as a function of aimline condition and training condition. All three sources of variance were found to be significant - both main effects - aimlines, F(1,49) = 49.3, $p \leq .000$, and training, F(1,49) = 39.5, $p \leq .000$, as well as the interaction between them, F(1,49) = 82.0, $p \leq .000$. Trained subjects used more dimensions than untrained subjects. Fewer dimensions were referenced when aimlines were present than when no aimlines were present. The interaction between training in data utilization and the use of aimlines appears in Figure 2. While there was no difference between trained and untrained subjects when aimlines were present, there was a great difference when no aimlines were present. In this latter condition, trained subjects referred to a far greater number of dimensions than the untrained subjects.

Insert Table 9 and Figure 2 about here

Discussion

In general, the findings from this research are consistent and logical within the framework of data utilization. For instance, successively higher levels of slope were rated higher in intervention effectiveness and, for the four conditions of variability, the lowest ratings were given when variability either increased or was low and constant while the highest ratings were given when variability either decreased or was high and constant. Both of these interpretations would be consistent with established data utilization practice: steeper slopes mean higher (faster) rates of improvement and increased variability signifies lack (loss) of control of those variables relevant to performance (Parsonson & Baer, 1978). Yet, the ratings of interventions followed by high constant variability were higher than those followed by low constant variability. The interpretation apparently is one of considering erratic performance as at least including some high scores, which was viewed as a more positive aspect than consistent control of performance.

While the above finding was true in general, the presence of an interaction between slope and variability necessitates a qualification of that result. The effect of increased variability, relative to the other three conditions, reveals the highest rating of intervention effectiveness to occur when the slope is 10 degrees, the lowest by only a small margin to occur when the slope is 15 degrees, and the lowest by a significant margin to occur when the slope is 20 degrees. That is, when there is minimal improvement over time (a low slope), increased variability is not viewed as a negative component of performance. As the rate of improvement increases, increases in variability result in lower ratings of effectiveness, relative to the other conditions. At the same time, if variability does not change, but remains high, ratings of effectiveness also remain high, nearly the same as if variability had decreased. Again, some degree of variability actually is found acceptable and there is attention to large changes.

This finding is somewhat in keeping with that reported by Wampold and Furlong (1981). In that study, subjects failed to appreciate the functional equivalence of two different time series in which the change in intervention effects were the same relative to the variation present. That is, a steeper slope or trend (with proportionately greater variability) should be rated the same as a modest slope or trend, in which the variability is proportionately smaller. In this study, subjects rated graphs with low slope and variability as reflections of no intervention effect and graphs with a high slope and high variability as reflecting a very strong intervention effect. However, because variability was manipulated in both phases in this study, it was possible to ascertain subjects' responses to this factor <u>within</u> a time series (between phases), as well as between different time series, a condition lacking in the Wampold and Furlong (1981) study. In analyzing this factor, it is apparent that subjects reacted differentially to various changes in variability between phases.

The findings for the two evaluation variables, the use of aimlines and training in data utilization, revealed less of an effect and less consistency in the effects. The use of aimlines did not appear to have any significant effect on the ratings of intervention effectiveness. Subjects rated intervention effectiveness the same regardless of the presence (or lack) of aimlines. The apparent effect of training was to create a more cautious perspective in evaluating programs, with untrained subjects rating intervention effectiveness significantly higher than trained subjects. This may be, in part, a function of the number of data dimensions that trained subjects attended to during their evaluations. It is possible that trained subjects were attending to different elements of the data array in concert and not simply responding to any one element.

This characteristic of time-series data - the capacity of generating several summary statistics - is both an advantage and a disadvantage. There is flexibility in summarizing performance in many different ways, allowing change to be reflected in a sensitive and appropriate manner. At the same time, the use of such data becomes more problematic, because not all of the indices are changing in concord with each other. That is, when the data array depicts both an increase in slope and variability, judgments of effects may be tempered. Because the trained subjects had at their disposal a more complete and detailed procedure for evaluating effects, it is possible that the net result was one of moderating conclusions of effectiveness.

The lack of a significant interaction between the use of aimlines and training in data utilization represents an interesting finding. It is possible that the training session was not effective and/or the skills developed as a result of training were not sufficient to differentiate that group of subjects from the untrained group. That is, trained subjects evaluated graphs with aimlines the same as they evaluated graphs with no aimlines, failing to apply the decision rule criteria of three days above or below the aimline. On the other hand, it is possible that the critical factor in the training-aimline interaction is the aimline, not the training. A group of untrained subjects may be evaluating program effectiveness in the same differential manner on graphs with and without aimlines as subjects trained in the use of aimline decision rules.

In the former case, the implication is that training should be more extensive than that implemented in this research. Although the procedures of analysis and the paradigm of evaluation were fully described and modeled, the subjects were given very little practice and no feedback prior to their evaluation of the graphs. In the latter argument, the implication is that training in data utilization is unnecessary, as long as the graphs being evaluated contain aimlines. Explanation of decision rule criteria need not be included either. A simple depiction of performance relative to an aimline is all that is necessary.

Further support for the lack of training hypothesis comes from an analysis of reliability. Not only was there no differential use of the data by trained and untrained subjects on graphs with and without aimlines, but little effect was found for either of the two factors on the reliability of ratings of intervention effectiveness. Untrained subjects were nearly as reliable as trained subjects, and little difference existed in the use of aimlines. Although trained subjects were slightly more reliable on graphs with aimlines, untrained subjects were not. Therefore, the use of aimlines without training, does not appear to be a critical factor.

In contrast to the lack of training effects on the use of aimlines and reliability of ratings, there was an effect on the stability (reliability over time) of ratings: Trained subjects were more reliable than untrained subjects; ratings of effectiveness were more reliable when aimlines were present; and there was an interaction between the use of aimlines and training in data utilization, with trained subjects more reliable on graphs with aimlines and untrained subjects more reliable on graphs without aimlines. In general, the range and absolute values of reliability coefficients are in keeping with previous investigations (DeProspero & Cohen, 1978; Jones et al., 1977). Visual analysis of time series data nas modest reliability at best.

Two factors that appear to influence reliability include both slope and variability. The effect of variability was most pronounced when it increased (resulting in low reliability), with little difference among reliabilities in the other three conditions. The effect of slope was most noticeable when it was steep (resulting in the highest reliability). Generally, the differences between the reliability coefficients for the various conditions of variability increased as the slopes increased. When the slope was 10 degrees, the range was from .51-.53; the range was .46-.64 for a slope of 20 degrees. This finding again indicates that not all data indices are equivalent stimulus dimensions for rating intervention effectiveness. When the slope is low, there is little differentiation and the absolute level of reliability quite low (.52). When the slope is steep, the reliability of ratings of effectiveness is very low (.46) when variability has increased, and modest (.64) when variability was low and constant. Nevertheless, the range is greater with a steeper slope.

A descriptive analysis of the data dimensions utilized for evaluating effectiveness provides a partial explanation for the low levels of reliability and problems with training. Subjects' responses reflected the influence of many characteristics of the data, rather than any single dimension. The three most frequently cited dimensions, however, were those that were manipulated in this study

-slope, variability, and aimlines. In addition, several other characteristics appeared influential, greatly expanding the type and frequency of interactions possible. The dimensions attended to by the trained subjects were both more varied and cited with greater frequency than those attended to by the untrained subjects. Thus, for any given graph, the subject's response was under the control of eight different characteristics (for trained subjects) or six different characteristics (for untrained subjects), excluding those that rarely were considered.

There was also a difference in the kind of data characteristics used by trained versus untrained subjects. The only dimension consistently referred to more frequently by untrained subjects was absolute values. This particular characteristic is probably the most static, least informative, and most potentially biasing of any of the possible dimensions. Given a time-series data array, the use of a single score to summarize change in performance has many problems, not the least of which is the failure to take advantage of that characteristic unique to time-series data - changes in scores over time. In contrast, the remaining characteristics reflect changes over time and consistently were referred to more frequently by trained subjects. Furthermore, there was an indication that the data array itself influenced the number of data characteristics mentioned. It appears that, in general, when the changes were more obvious, there was a reliance on fewer characteristics. For instance, the use of aimlines provided a clear indication of relative improvement, resulting in less reliance on other data characteristics. Or when growth was either minimal (10 degrees) or maximal (20 degrees), fewer dimensions were referred to in the evaluation process. Finally, it was only when the data became unpredictable (high and constant or increased variability) that reference to other dimensions was increased.

In conclusion, before an adequate and valid analysis of time series data using visual inspection can be established, some consistent data utilization needs to occur. As this study has demonstrated, there are several factors that influence this process, including training in data analysis, and the data array itself. The fact that these influences all occur together simply makes the task at hand more difficult. The simple use of aimlines did not appear to result inherently in a better analysis. Rather a decision-making system needs to be empirically established that takes into account both the fact that judgment is based on several dimensions at the same time and that such factors often conflict with each other.

References

Baer, D. M., Wolf, M. M., & Risley, T. R. Some current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis, 1968, 1, 91-97.

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette & J. P. Campbell (Eds.), Handbook of industrial and organizational research. Chicago: Rand McNally, 1976.

DeProspero, A., & Cohen, S. Inconsistent visual analysis of intrasubject data. Journal of Applied Behavior Analysis, 1979, 12, 573-579.

Glass, G. V., Willson, V. L., & Gottman, J. M. Design and analysis of time-series experiments. Boulder, Colo: University of Colorado Press, 1975.

Jones, R. R., Vaught, R. S., & Weinrott, M. Time-series analysis in operant research. Journal of Applied Behavior Analysis, 1977, 10, 151-166.

Jones, R. R., Weinrott, M., & Vaught, R. S. Visual vs. statistical inference in operant research. In A. E. Kazdin (Ed.), The use of statistics in N=1 research. Symposium presented at the annual convention of the American Psychological Association, Chicago, September, 1975.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. Effects of serial dependency on the agreement between visual and statistical inference. Journal of Applied Behavior Analysis, 1978, 11, 272-283.

Kazdin, A. E. Statistical analyses of single-case experimental designs. In M. Hersen & D. Barlow (Eds.), Single case experimental designs: Strategies for studying behavior change. New York: Pergamon Press, 1976.

Michael, J. Statistical inference for individual organism research: Mixed blessing or curse? Journal of Applied Behavior Analysis, 1974, 7, 647-653.

Parsonson, B. S., & Baer, D. M. The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), Single subject research: Strategies for evaluating change. New York: Academic Press, 1978.

Pechacek, T. F. A probabilistic model of intensive designs. Journal of Applied Behavior Analysis, 1978, 11, 357-362.

Pennypacker, H. S., Koenig, C. H., & Lindsley, O. R. The handbook of the standard behavior chart (prelim. ed.). Kansas City, KS: Precision Media, 1972.

Sidman, M. Tactics of scientific research. New York: Basic Books, 1960.

Skinner, B. F. Science and human behavior. New York: MacMillan, 1953.

Wampold, B. E., & Furlong, M. J. The heuristics of visual inference. Behavioral Assessment, 1981, 3, 79-92.

White, O. R. A pragmatic approach to the description of progress in the single case. University of Oregon: Unpublished Doctoral Dissertation, 1971.

Table 1

Average Rating of Intervention Effectiveness
for All Levels of the Slope and Variability Factors

	Slope			
Variability	10^0	15^0	20^0	Average
Increase	2.6	2.2	2.6	2.5
Decrease	2.6	2.3	3.4	2.8
Low	1.9	2.5	3.0	2.5
High	2.3	2.6	3.3	2.7
Average	2.3	2.4	3.1	2.6

Table 2

Comparison of Trained and Untrained Subjects on the Reliability of Ratings (Agreement/Agreement + Disagreement) for Each Combination Slope and Variability

Slope	Inc. Var.	Dec. Var.	Low Var.	High Var.	Average
10°	.51	.52	.51	.53	.52
15°	.43	.48	.51	.52	.49
20°	.46	.62	.64	.55	.57
Average	.47	.54	.56	.54	.53

Table 3

Reliability of Ratings (Agreement/Agreement + Disagreement) from Time 1 to Time 2 for Graphs with Variability Manipulated

Low	High
.52	.54

Table 4

The Average Rating of Intervention Effectiveness for Both Levels of the Aimline and Training Factors

	Trained	Untrained	Average
Aimline	2.4	2.8	2.6
No Aimline	2.5	2.7	2.6
Average	2.5	2.8	2.6

Table 5

Reliability of Ratings (Agreement/Agreement + Disagreement) for Both Levels of the Aimline and Training Factor

Trained		Untrained	
Aimline	No Aimline	Aimline	No Aimline
.54	.51	.52	.52

Table 6

Reliability of Ratings (Agreement/Agreement + Disagreement) from Time 1 to Time 2 by Trained and Untrained Subjects for Graphs with Aimline Manipulated

Independent Variable	Trained	Untrained	Average
Aimline	.66	.40	.53
Without Aimline	.53	.48	.51
Average	.62	.44	.52

Table 7

Average Number of References Made to Various Characteristics of the Data by Trained and Untrained Subjects

Data Characteristic	Trained				Untrained			
	$\overline{X}$	S.D.	Min.	Max.	$\overline{X}$	S.D.	Min.	Max.
Progress (slope)	18.5	3.2	14	26	19.7	5.5	7	28
Variability*	21.8	3.6	12	27	15.7	5.9	5	29
Jump*	9.4	4.8	1	18	.4	1.4	0	7
Direction*	6.8	4.6	0	19	2.7	4.2	0	18
No. Days Improved	7.9	3.3	1	15	6.0	6.2	0	25
Goal/Aim*	12.0	2.1	4	15	8.7	4.4	0	15
Average Performance*	13.5	4.2	7	23	6.2	7.0	0	24
Overlap*	11.4	4.6	2	20	1.0	2.1	0	7
Absolute Values*	3.5	3.5	0	15	7.8	6.2	0	26

*Significant at $p \leq .05$.

Table 8

Number of Data Dimensions Referenced for All Levels of the Slope and Variability Factors

	Slope			
Variability	10°	15°	20°	Average
Increase	2.8	3.2	2.7	2.9
Decrease	2.8	3.1	2.9	2.9
Low, constant	2.8	3.0	2.7	2.9
High, constant	2.7	3.1	2.9	2.9
Average	2.8	3.1	2.8	2.9

Table 9

Number of Data Dimensions Referenced for Both Levels of the Aimline and Training Factors

	Trained	Untrained	Average
Aimline	2.5	2.4	2.5
No Aimline	4.5	2.2	3.4
Average	3.5	2.3	2.9

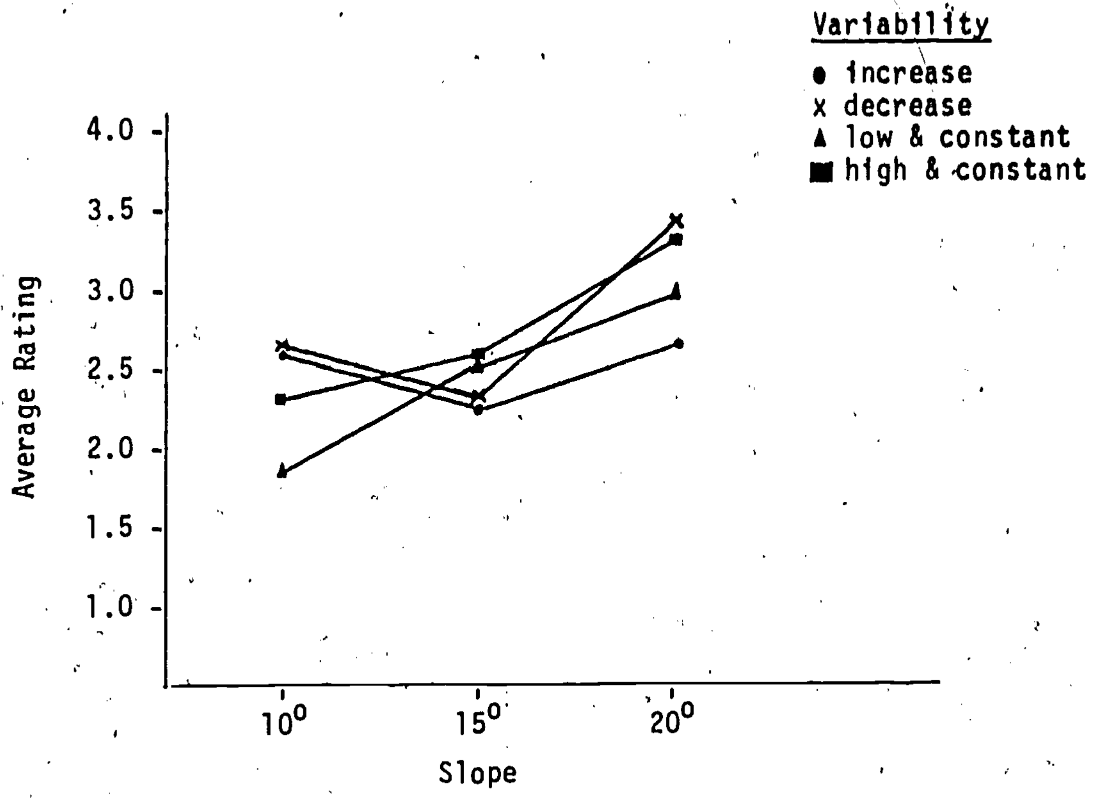

Figure 1. Interaction between changes in slope and variability on the average rating of intervention effectiveness.

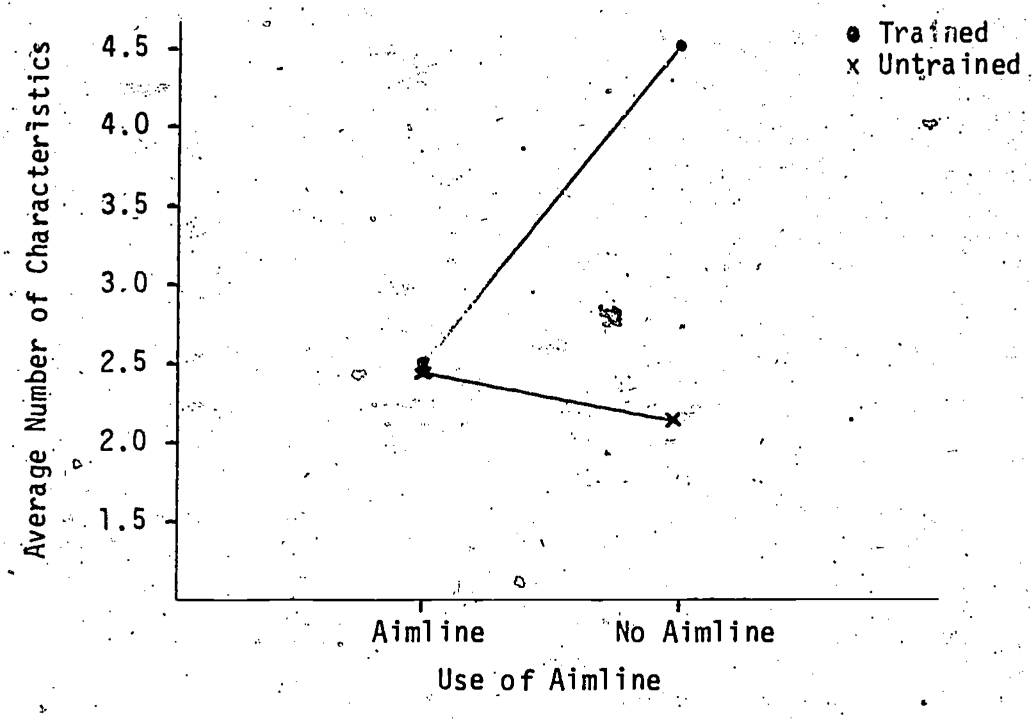

Figure 2. Interaction between training in data utilization and the use of aimlines in the number of data characteristics mentioned.

Appendix A

Procedures for Constructing Graphs

The first step in constructing the graphs involved drawing in the slope line: a slope of 0 degrees was drawn in during baseline and either 10, 15, or 20 degrees drawn in during the intervention. The lines that defined total bounce were then drawn in. These lines were parallel to the slope line, with one passing through the data point farthest above the slope and one passing through the data point farthest below the slope. Bounce around the slope line was kept nearly equidistant above and below the line. That is, if the total bounce involved five data points, two lines were drawn parallel to the slope line: one that was two data points above the slope line and one that was three data points below the slope line. If the total bounce was 15 data points, the envelope included data points 7-8 units above and 7-8 units below the slope line. The graph at this point had a defined slope and variability that was used as a guideline in plotting the actual data points.

Data points then were plotted onto the graphs using the quarter-intersect method (Pennypacker, Koenig, & Lindsley, 1972). All data points had to fall within the range of the total bounce. To use this procedure for systematically varying the slope, a data point had to be determined that intersected the median of each half on the middle day of that half. That is, using only the first half of the graph, the median was determined and plotted, at any point (on any day) during the first half. Then an equal number of data points above and below this point were plotted on the remaining days. The median of this half, when plotted on the middle day of the half, defined a point through which the slope line would pass.

The same procedure was used for the second half of the graph. The median level necessary for the slope line to pass through on the middle day of the second half was determined and plotted, on any day for that half. Following this, the remaining data points were plotted such that half of the data points fell above and half fell below this value. When this median value was plotted on the middle day, the slope line would pass through it. This entire process resulted in a data pattern having a given slope and variability.

In generating a data array during baseline, the slope line was kept horizontal (with a slope of zero). However, for the data during the intervention, the slope was predetermined at some fixed value (10, 15, or 20 degrees). In order to provide an adequate test of the influence of slope alone in determining judgments, the change in step (jump or level) from baseline to intervention was kept minimal. The difference between the last data day of baseline and the first data day of the intervention phase was kept to a maximum of two data points. Given this one constraint on the actual value of the data points plotted, all others were plotted in a random manner, given the particular levels of slope and variability.

Appendix B

Evaluation Response Form

1. Rate whether the instructional program was an effective one for increasing the student's reading rate.

1	2	3	4
Definitely Not Effective	Possibly Effective	Moderately Effective	Very Effective

2. What about the student's performance makes you think so?

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. 108). 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.