DOCUMENT RESUME

ED 268 169 TM 860 232

AUTHOR Fuchs, Lynn S.; Fuchs, Douglas
TITLE Effects of Long- and Short-Term Goal Assessment on Student Achievement.
PUB DATE Apr 86
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Analysis of Variance; *Effect Size; Elementary Secondary Education; Evaluation Methods; Evaluation Problems; *Interaction; Literature Reviews; *Meta Analysis; Special Education; *Student Educational Objectives

ABSTRACT

This meta-analysis explored how measuring progress toward long- versus short-term goals relates to contrasting outcome measures of student achievement. Twenty-one controlled studies, that provided sufficient data for the calculation of effect size, were coded in terms of measurement method (toward long- versus short-term goals) and type of achievement outcome (probe-like versus global achievement test). Analogues to analysis of variance conducted on weighted unbiased effect sizes (UES's) indicated an interaction: when progress was measured toward long-term goals, UES's on global measures were higher than on probe-like outcomes; when progress was measured toward series of short-term goals, the reverse was true. As demonstrated in this meta-analysis, short-term goal measurement may be misleading. Students may master a series of instructional objectives, even though progress may be limited on more global indices of achievement which better represent the true desired outcome performance. Consequently, for special education teachers who monitor mastery of short-term objectives, caution may be in order: curriculum-based assessment of long-term goals may represent a necessary supplementary strategy for validly assessing pupil progress. References, tables, figures, and a list of reports included in the meta-analysis are appended. (Author/PN)

Effects of Long- and Short-Term Goal Assessment on Student Achievement

Lynn S. Fuchs and Douglas Fuchs

Peabody College of Vanderbilt University

Address correspondence to:

Box 328, Department of Special Education

Peabody College

Vanderbilt University

Nashville, TN 37203

(615) 322-8165.

Requests for reprints should be sent to Lynn S. Fuchs, Box 328, Department of Special Education, George Peabody College, Vanderbilt University, Nashville, TN 37203.

Running Head: Curriculum-Based Assessment

Abstract

This meta-analysis explored how measuring student progress toward long- vs. short-term goals affects achievement outcomes. Twenty-one controlled studies were coded in terms of measurement method (toward long- vs. short-term goals) and type of achievement outcome (probe-like vs. global achievement test). Analogues to analysis of variance conducted on weighted unbiased effect sizes (UESs) indicated an interaction: When progress was measured toward long-term goals, UESs on global measures were higher than on probe-like outcomes; when progress was measured toward series of short-term goals, the reverse was true. Implications for special education practice are discussed.

Effects of Long- and Short-Term Goal Assessment on Student Achievement

In special education, commercial norm-referenced achievement tests are the traditional (Tindal et al., 1985) and continute to represent a prevalant (Goh, Teslow, & Fuller, 1981) measurement tool for generating individualized educational programs and for evaluating the effects of those programs. Nevertheless, it has been criticized increasingly (see Tindal et al., 1985; Ysseldyke & Thurlow, 1984). With respect to generating educational programs, critics contend that the abilities measured by these instruments frequently lack necessary conceptualization (Ysseldyke, 1979), and relatedly that the tests often fail to demonstrate adequate psychometric properties (Salvia & Ysseldyke, 1985). In terms of program evaluation, critics argue that these measures fail to: (a) indicate the extent to which specific educational objectives have been attained (Skager, 1971), (b) provide enough alternate forms to permit ongoing progress monitoring, (c) sample the domains of interest comprehensively (Zigmond & Silverman, 1984), and (d) relate to curricular materials (Armbruster, Stevens, & Rosenshine, 1977; Jenkins & Pany, 1978).

In response to these problems, ongoing criterion-referenced, curriculum-based assessment (CBA) strategies have been developed. With CBA, measurement procedures are designed to match students' program objectives. Alternate test forms are drawn directly from curricula specified in objectives and are administered at regular intervals during intervention; student progress data are evaluated regularly with reference to the performance criteria specified in objectives; and individualized programs are tested formatively and modified over time as required to insure effective instructional programs and attainment of objectives. Therefore, with CBA, instructional program evaluation is ongoing and based in the curriculum; program development is inductive, in response to the ongoing program evaluation data.

Research indicates that such ongoing CBA of students' attainment of goals and objectives represents an effective alternative approach to program development and evaluation. In a quantitative synthesis of available controlled studies, the average effect size was .70 (Fuchs & Fuchs, in press). This indicates that, in terms of the standard normal curve and an achievement test scale with a population mean of 100 and standard deviation of 15, the use of CBA to develop and evaluate instructional programs over time can be expected to raise the typical achievement outcome score from 100.00 to 110.50, or from the 50th to the 76th percentile.

Additionally, the requirements of federal legislation seem to indicate the importance of CBA: The IEP mandate of PL 94-142 requires special educators to specify long-term goals, short-term objectives, and evaluation procedures for monitoring students' attainment of objectives. Assuming that the intent of this legislation was to encourage compatibility and congruence between goals/objectives and pupils' curricula, then the IEP mandate requires a CBA approach to progress evaluation.

Despite the apparent effectiveness of and seeming necessity for CBA, it remains unclear how practitioners should design CBA procedures to monitor students' attainment of goals and objectives and how alternative practices relate to student achievement outcomes. Currently, practitioners can select between two types of CBA, one focusing on the attainment of long-term goals (CBA-goal) and the other of short-term objectives (CBA-objective).

With the CBA-goal approach, an annual curriculum-based goal is specified and a large pool of related measurement items is created. From this measurement pool, subsets of items, or monitoring probes, are drawn randomly (see Fuchs, Deno, & Mirkin, 1984). The difficulty level of the monitoring probe remains constant over a long time. Contrastingly, with the CBA-objective approach, a series of objectives corresponding to steps within a hierarchical curriculum is

specified, and a series of relatively circumscribed, small pools of items are created, each of which corresponds to a specific objective (see Lindsley, 1971; White & Haring, 1980). The difficulty level of material on which students are measured increases as students master the sequentially-related objectives.

Both types of CBA are ongoing, criterion-referenced, curriculum-based, and enjoy strong curricular validity or correspondence between tests and programmatic goals and objectives (McClung cited in Yalow & Popham, 1983). However, these systems do differ conceptually. CBA-objective appears to have stronger instructional validity or correspondence between tests and instruction (Yalow & Popham, 1983). The monitoring probes for short-term measurement are related directly to current instructional material; so, for example, if an instructional intervention is introduction of the r-controlled phonics rule, the monitoring measure is reading r-controlled words. Alternately, with CBA-goal, the monitoring probes are not related to the instructional material. The instructional intervention may be introduction of the r-controlled phonics rule, whereas the monitoring measure may involve oral reading fluency, accuracy, and/or comprehension on second grade passages.

Although CBA-objective may enjoy stronger instructional validity, CBA-goal is advantageous in other respects. It possesses better content validity or representation of the ultimate desired performance, i.e., reading fluency/comprehension (Yalow & Popham, 1983). Additionally, its concurrent validity or correlation with other measures of achievement appears to be stronger than that of CBA-objective (Fuchs, 1982).

The emergent question, and the focus of the current meta-analysis, is how well these types of ongoing criterion-referenced, curriculum-based assessment strategies relate to outcome measures of student achievement. The investigation of this question should help practitioners assess the relative merits of the two types of CBA and select CBA monitoring procedures.

Method

Search Procedure

The search for pertinent studies to include in the meta-analysis comprised four steps. First, employing the Thesaurus of Psychological Index Terms (APA, 1982), multiple descriptors were generated for key terms. For example, student achievement alternately was represented by "student progress," "goal attainment," and "educational effects." Second, these terms facilitated a computer search of three on-line data bases: (a) ERIC, a data base of educational materials from the Educational Resources Information Center consisting of abstracts from Research in Education and Current Index to Journals in Education; (b) Comprehensive Dissertation Abstracts; and (c) Psychological Abstracts. Third, employing similar key descriptors, a manual search was conducted of five educational journals for the years 1973 through 1983. These journals were: American Educational Research Journal, Journal of Learning Disabilities, Journal of Precision Teaching, Journal of Special Education, and Learning Disability Quarterly. Fourth, the reference sections of relevant papers along with identified bibliographies were explored for additional studies.

Criteria for Relevant Studies

A study was considered for inclusion if it employed a control group to evaluate the effects of curriculum-based monitoring on academic achievement. Such monitoring was defined as curriculum-based data collection that occurred at least twice weekly, with decisions concerning the adequacy of programs formulated on an individual, not group, basis. Studies were excluded that (a) monitored social behaviors, (b) primarily focused on the use of behavior modification, while employing time series to test experimental effects, (c) provided test feedback

only to students, and/or (d) employed college age subjects. (Other factors, such as instrumentation, methodological rigor, and adequacy of decisionmaking were coded as variables, and results related to these variables are reported elsewhere [e.g., Fuchs & Fuchs, in press].)

The search yielded 29 studies that met the criteria established for inclusion. From these studies, 11 were eliminated because of insufficient data for calculating effect size.

Data Extracted Studies

Calculation of, related assumptions about, and interpretation of effect size. Results of the studies were transformed to estimates of effect size, typically calculated by subtracting the treatment means and dividing by the control group standard deviation. For studies reporting relevant means and standard deviations for both groups, effect sizes were calculated from these measurements. For studies not reporting means and standard deviations, effect sizes were calculated from other statistics, such as F or p values (see Glass, McGaw, & Smith, 1981). When pretest differences or analysis of covariance were reported, alternative procedures for calculating effect size were used, as possible, to control for those differences (see Glass et al., 1981). For purposes of analysis, an effect was given a positive sign if subjects achieved greater scores in the systematic monitoring treatment.

Since this estimator of effect size is biased positively, especially for small N, each effect size was converted to an unbiased effect size (UES) by multiplying the estimated effect size by a correction factor (see Hedges, 1981). This procedure corrects for inconsistency in estimating true from observed effect sizes (Hedges, 1981). The difference between the observed and unbiased effect sizes was neglible ($\bar{X}$ = .019, SD = .025) as has been demonstrated elsewhere (Bangert-Drowns, Kulik, & Kulik, 1983). Nevertheless, UESs were employed to

insure the mathematical tractability of the data.

The statistical properties of effect size depend on the model for the observations in the experiment. In this meta-analysis, it was assumed that observations are distributed independently normally within groups of the experiment. The related interpretation of the population effect size is that it represents the mean difference one would obtain if the dependent variable were scaled to have unit variance within groups of the experiment. Thus, the effect size is the mean differen e reexpressed in unit scaled so that $\sigma=1$ to remove the dependence of the arbitrary scale factor σ. When observations in the experimental and control groups are distributed normally, effect siz. can be used to quantify the degree of overlap betwen the distribtions of observations in the experimental and control groups. Because this effect size is the standardized score of the experimental group mean in the control group distribution, it represents the proportion of control group scores that are less than the average score in the experimental group (Hedges & Olkin, 1985).

This parametric point estimate for effect size was selected over nonparametric estimators because nonparametric estimators can be computed only when raw data of each study are available and because such estimators probably are less efficient than parametric counterparts when the assumptions of parametric procedures are satisfied. Thus, Hedges and Olkin (1985) recommend that nonparametric estimators be used only when it is suspected that parametric assumptions are violated seriously.

The parametric point estimate for effect size also was selected over estimation of an effect magnitude based on the idea of variance accounted for due to the introduction of an explanatory variable, such as correlation coefficients and ratios, intra-class correlation coefficients, and the omega-squared index. Although such indices are intuitively appealing, they are not suited for combination across studies (Hedges & Olkin, 1985): They are nondirectional and

depend on functions of arbitrary design decisions, such as the particular definition of groups or patterns of X values selected, as well as on the underlying relation between theoretical constructs (Hedges & Olkin, 1985).

Effect size aggregation. Guidelines were established to ensure that each relevant effect was counted only once in analyses. When an effect was measured by tests that failed to represent dimensions relevant to the meta-analysis (i.e., Reading Comprehension and Structural Analysis Subtests of the Stanford Diagnostic Reading Test), these results were pooled. For example, if achievement within a study were measured with three global tests and two probe-like measures, the three effect sizes for the global tests would be aggregated as would be done for the two probe-like tests. So two, rather than five, effect sizes would be included for such a study.

There were 96 effect sizes, with between 1 and 12 effect sizes per study. Analyses indicated no statistical dependency between effect size magnitude and number of comparisons per study ($r = .12$). Therefore, UESs were aggregated at the individual effect size level.

In combining UESs, a weighting procedure was employed to account for the fact that the variance of the estimator depends on sample size, in which estimates from studies with larger sample sizes are more precise than those from studies with smaller sample sizes (Hedges & Olkin, 1985). Hence, the weighting procedure gives weight inversely proportional to the variance within each study: With a larger N / smaller variance, a larger weight is assigned.

To combine UESs, a direct weighted linear combination of estimators procedure was employed because Hedges and Olkin (1985) have demonstrated that such a method is comparable to, but simpler and more intuitively appealing than, alternative procedures. In such an aggregation, large sample statistical theory for estimating effect size from a series of studies is employed, and Hedges and Olkin (1985) demonstrated this theoretical orientation is reasonably accurate

when effect sizes are less than 1.5 in absolute magnitude and sample sizes are at least 10. These conditions were met in the current meta-analysis. Nevertheless, such an aggregation also assumes perfect linear equatability between dependent measures and relatedly similar operationalizations of the constructs measured. These assumptions may be violated through noncomparability, measurement error, and presence of unique factors or invalidity (Hedges & Olkin, 1985). Therefore, results must be considered within the confines of these potential statistical problems.[1]

Study features. To describe study features pertinent to the current investigation, two major substantive variables were identified and coded for each study. The first study feature was type of goal. This variable had two levels that differentiated studies in which progress toward long-term goals (CBA-goal) was moritored from studies in which progress toward a short-term objective or a series of short-term objectives (CBA-objective) was monitored.

Studies in which progress toward long-term goals was monitored involved the specification of a level of material on which a student was expected to be proficient within the next 15 or more weeks. For example, for a student currently reading proficiently on primer material, a student's goal might specify that, in 25 weeks, a student would read 75 words per minute correct with 90% accuracy on second grade reading passages. Then, for the next 25 weeks, measurement probes would be sampled randomly from second grade reading passages, representing approximately equivalent samples of measurement material.

Studies in which progress toward short-term goals was monitored required the identification of a sequence of small segments in a hierarchical curriculum to be mastered by the student. For example, the series of objectives might specify that the student would read, with 90% accuracy, flashcards first with consonant-vowel-consonant words, second with final e words, and third with double vowel words. Proceeding in a fashion parallel to the specification of

objectives, measurement probes first would be drawn from flashcards with consonant-vowel-consonant words until the mastery criterion was achieved by the student on that domain. Then, the measurement domain would change so that probes were flashcards with final e words, and so on.

The second study feature was outcome measure. This variable also had two levels: dependent measures similar to the monitoring probes and more global achievement tests. Employing the examples provided above, probe-like outcome indices were oral reading rate on second grade passages or percentage read correctly from flashcards with final e words; global achievement tests were the Structural Analysis and Reading Comprehension Subtests of the Stanford Diagnostic Reading Test.

In addition to these two substantive features, a third, methodological variable was coded for each study, duration of the treatment. This variable had three levels: treatments implemented for less than 3 weeks (coded "1"); treatments lasting between 3 and 10 weeks (coded "2"); and treatments continued for more than 10 weeks (coded "3").

Two raters independently coded 10 of the 18 studies (56%). Percentage of agreement[2] for the raters on type of goal was 77 and 83 when progress toward long- and short-term goals was the respective level of the variable, with a mean intercoder agreement of 80. For outcome measure, the percentages of agreement for probes and global achievement tests were 94 and 86, respectively, with a mean percentage of 90. Percentage of agreement for duration of treatment was 100 for all levels of the variable.

A previous investigation (Fuchs & Fuchs, in press) explored methodological quality of the studies and identified no relation between effect size magnitude and study quality. Additionally, this previous study reported an overall effect size as well as the related fail-safe N. Thus, these results are not repeated here.

Characteristics of the Sample

Of the 20 references listed in the Appendix, which represent 18 separate investigations,[3] there are 4 dissertations, 11 unpublished studies, and 5 journal articles. Among the published papers, 2 appeared in Exceptional Children, 2 in American Educational Research Journal, and 1 in American Journal of Mental Deficiency. A total of 3665 subjects participated in these studies, with 83% of the investigations employing handicapped subjects. Of these handicapped pupils, 93% were mildly to moderately handicapped and 7% were severely handicapped. The grade level of these subjects ranged from preschool through high school, with a median grade level of 3.8. Among the 18 investigations, 8 (44%) focused solely on the academic area of reading, 3 (17%) on reading and math, 2 (11%) only on math, and 1 (6%) each on (a) high school content areas, (b) preschool skills, (c) spelling, (d) reading and spelling, and (e) reading, math, and spelling.

Results

Of the 96 effect sizes, 27 related to long-term goal measurement and 69 to short-term goal measurement. Of the 27 long-term goal effect sizes, 14 were associated with probe-like and 13 with global outcome measures. Of the 69 short-term goal effect sizes, 37 were related to probe-like and 32 to global outcome measures.

Relation between treatment duration and other effect size features. A pair of t tests was run to determine whether measurement goal or outcome measure was related to the duration of treatment. These tests indicated no statistically significant associations. For the long-term goal effect sizes, the mean coded level of treatment duration (see above) was 2.92 (SD = .27); for the short-term goal effect sizes, 2.75 (SD = .56), t (95) = 1.52, ns. The average level of

treatment duration for effect sizes associated with probe-like and global outcome measures, respectively, were 2.78 (SD = .51) and 2.76 (SD = .23), t (95) = .24, ns.

Relation between magnitude and features of effect sizes. Table 1 displays the weighted UESs by (a) the type of goal factor (long-term goal vs. short-term objective) and (b) the outcome measure factor (probe-like vs. global achievement test). To examine the relation between these variables and effect size magnitude, Hedges's (1984) chi-square analogue to analysis of variance was employed. When conventional analysis of variance is conducted on effect sizes, problems exist because of the possibility that systematic variance will be pooled into the estimate of error variance. Moreover, violation of the homoscedasticity assumption is severe in research synthesis, and there is little reason to believe that the usual robustness of the F test will prevail (see Hedges, 1984). Thus, Hedges's chi-square analogue was employed to avoid these conceptual and statistical problems.

As indicated in Table 1, type of goal was not related to UES, but outcome measure produced a statistically significant difference, with the mean effect size of probe-like measures .11 of a standard deviation higher than that of global measures. Nevertheless, tests for the homogeneity of effect size (Hedges, 1984) indicated that none of the four pools of UESs represented a homogeneous set: Statistical values for long- and short-term goals and for probe-like and global measures, respectively, were $\chi^2(26) = 208.37$, $\chi^2(68) = 1029.66$, $\chi^2(44) = 859.20$, and $\chi^2(50) = 357.54$. Therefore, additional analyses were conducted.

Insert Table 1 about here

These additional analyses addressed the effect of type of outcome

measure within each of the type of goal conditions, and suggested the presence of an interaction. As described in Table 2 and illustrated in Figure 1, within the type of goal conditions, there were statistically significant differences between UESs associated with the probe-like and the global outcome measures. With CBA-objective, UESs associated with probe-like outcome measures were higher than those of global measures. For CBA-goal, the reverse was true: UESs associated with global measures were higher than those related to probe-like outcome measures. Specifically, within short-term goal measurement, the mean effect size for probe-like measures was .40 higher than that of global measures; within long-term goal measurement, the average effect size for probe-like measures was .51 lower than that of global measures.

Insert Table 2 and Figure 1 about here

Discussion

The purpose of this meta-analysis was to investigate how well measuring progress toward long- vs. short-term goals relates to contrasting outcome measures of student achievement. Toward this end, a literature search was conducted, resulting in the identification of 18 relevant studies that provided sufficient information for the calculation of effect size. These studies were coded for long- vs. short-term goal measurement and for probe-like vs. global outcome achievement measures. To investigate the possibility that short- and long-term goal measurement or probe-like and global achievement measures might be related to the duration of the experimental treatment, study duration also was coded. Analyses indicated no reliable association between either substantive

variable and treatment duration.

Analogues to analysis of variance indicated that the magnitude of effect size was not related to type of goal on which monitoring occurred, but was associated with the type of outcome measure employed, with a mean difference in effect size of .11. This indicates that in terms of a standard normal curve and achievement test scale with a population mean of 100 and standard deviation of 15, assessing outcome with a probe-type measure can be expected to raise the typical achievement score associated with global measures from 100.00 to 101.65. Such an effect also indicates that the upper 50% of the distribution of effect sizes associated with outcomes assessed via probe-like measures exceeds approximately 54% of the distribution for which effect size was assessed on global outcomes measures. Therefore, this statistically significant difference appears to represent one of minor practical effect.

Additional analyses suggested a more important effect, one of interaction. When progress was measured toward long-term goals, effect sizes calculated on global outcome measures were higher than on probe-like outcomes. On the other hand, when progress was measured toward a series of short-term goals, effect sizes were lower on global than on probe-like outcome measures.

This finding may be explained in terms of the relative strengths associated with the different goal measurement strategies. Long-term goal measurement corresponds poorly with instructional activities, but comparatively well with global measures of reading skills, including tests of decoding, word recognition, and comprehension (Deno, Mirkin, & Chiang, 1982; Fuchs, 1981). On the other hand, with short-term goal monitoring, correspondence between instruction and measurement is one-to-one; however, as Quilling and Otto (1971) demonstrated, mastery of a hierarchy of decoding skills relates inconsistently to global achievement indices.

Of course in interpreting findings of this synthesis, as with any

quantitative integregation, one must limit generalizations to situations similar to the experimental/control treatment and dependent measures in the primary research. Within such confines, alternative explanations for findings exist. For example, some may interpret results to suggest that, in order to demonstrate special education effectiveness, practitioners should select outcome measures that reflect the type of goal monitoring they have conducted. However, an alternative and perhaps more productive interpretation suggests the reverse: In order to promote the type of outcome special educators desire (i.e., global growth vs. mastery of discrete curriculum units), goal monitoring methods need to be selected carefully. Specifically, as practitioners develop their programmatic or IEP goals and objectives and related curriculum-based assessment procedures for monitoring pupil progress toward those goals and objectives, both the curricular and content validity of their measurement procedures must be addressed. Curricular validity refers to the match between testing and IEP goals and objectives; content validity, the correspondence between testing and the true domain in which proficiency is desired (Yalow & Popham, 1983). Curricular and content validity are addressed simultaneously only when practitioners write "significant rather than trivial" IEP goals and objectives, which relate well to the true desired outcome performance (Popham et al., 1985). Attention to this dual criterion allows "measurement-driven instruction" (Popham et al., 1985), or ongoing assessment of pupil progress, to assume an important effect on achievement. It implies that practitioners monitor progress toward long-term goals, an approach that appears to promote a global effect on achievement. Practitioners may wish to use this strategy to complement analyses of short-term objective mastery, a system that, on the other hand, can guide instructional programming decisions more directly.

The finding that long-term goal monitoring relates better to global achievement outcome measures may be especially important in the education of

handicapped students, who typically have poorly developed strategies for maintaining and transfering skills (Anderson-Inman, Walker, & Purcell, 1984; White, 1984). Short-term goal measurement focuses on instructionally related, relatively restricted domains of material for a period of time and then, upon mastery of that material, the measurement and instructional focus simultaneously changes. Such a paradigm may be problematic for at least two reasons. First, it may discourage teachers from reviewing material sufficiently to allow for long-term skill maintenance. Second, a close connection between instruction and measurement may encourage teachers to present new skills to students within the framework of the measurement task. For example, if the measurement procedure requires the pupil to read consonant-vowel-consonant words from a list, the teacher may focus instruction on reading consonant-vowel-consonant words from a list. As noted by Goodstein (1982), there may be danger in tying the instructional format too closely to the assessment device or of narrowly defining content-x-format domains of criterion-referenced assessment. Such a restricted instructional format may limit the transfer of skills. A more global, long-term goal approach to measurement may encourage teachers to incorporate instructional procedures that better allow for skill maintenance and generalization.

Teachers may prefer short-term goal measurement because it is easier to understand and it guides instruction more directly by providing information about when to progress from one skill to another (Fuchs, Wesson, Tindal, Mirkin, & Deno, 1982). In fact, evidence suggests the predominant monitoring strategy is assessment of short-term objective mastery through the periodic use of commercial criterion-referenced measures such as basal series mastery tests and the Brigance (1978) Diagnostic Inventory (Fuchs, Fuchs, & Warren, 1982). Nevertheless, as demonstrated in this meta-analysis, short-term goal measurement may be misleading. Students may master a series of instructional objectives, despite that progress may be limited on more global indices of achievement, which better

represent the true desired outcome performance. Consequently, for teachers who monitor mastery of short-term objectives, caution may be in order: Curriculum-based assessent of long-term goals may represent a necessary supplementary strategy for validly assessing pupil progress.

References

American Psychological Association. (1982). Thesaurus of psychological index terms (3rd ed.). Washington, D.C.: Author.

Anderson-Inman, L., Walker, H., & Purcell, J. (1984). Promoting the transfer of skills across settings: Transenvironmental programming for handicapped students in the mainstream. In W.L. Heward, T.E. Heron, D.S. Hill, & J. Trap-Porter (Eds.), Focus on behavior analysis in education (pp. 17-37). Columbus, OH: Merrill.

Armbruster, B.B., Stevens, R.J., & Rosenshine, B. (1977). Analyzing content coverage and emphasis: A study of three curricula and two tests (Technical Report No. 26). Urbana-Champaign: Center for the Study of Reading, Univerity of Illinois.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.C. (1983). Effects of coaching programs on achievement test performance. Review of Educational Research, 53, 571-585.

Brigance, A.H. (1978). Brigance Diagnostic Inventory of Early Development. Woburn, MA: Curriculum Associates.

Deno, S.L., Mirkin, P.K., & Chiang, B. (1982). Identifying valid measures of reading. Exceptional Children, 49, 36-45.

Fuchs, L.S. (1981). The concurrent validity of progress measures of basal reading material. Unpublished doctoral dissertation, University of Minnesota.

Fuchs, L.S. (1982). Reading. In P.K. Mirkin, L.S. Fuchs, & S.L. Deno (Eds.), Considerations for designing a continuous evaluation system: An integrative review (Monograph No. 10) (pp. 29-74). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Fuchs, L.S., Deno, S.L., & Mirkin, P.K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedgagogy, student achievement, and student awareness of learning. American Educational Research Journal, 21, 449-460.

Fuchs, L.S., & Fuchs, D. (in press). Effects of systematic formative evaluation: A meta-analysis. Exceptional Children.

Fuchs, L.S., Fuchs, D., & Warren, L.M. (1982). Special education practice in evaluating student progress toward goals (Research Report No. 82). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. ED 224 198)

Fuchs, L.S., Wesson, C., Tindal, G., Mirkin, P.K., & Deno, S.L. (1982). Instructional changes, student performance, and teacher preferences: The effects of specific measurement and evaluation procedures (Research Report No. 64). Minneapolis: Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. 218 849)

Glass, G., McGaw, B., & Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills: Sage.

Goh, D.S., Teslow, C.J., & Fuller, G.B. (1981). The practices of psychological assessment among school psychologists. Professional Psychology, 12, 699-706.

Goodstein, H.A. (1982). The reliability of criterion-referenced tests and special education: Assumed versus demonstrated. Journal of Special Education, 16, 37-48.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 359-361.

Hedges, L.V. (1984). Advances in statistical methods for meta-analysis. In W.H. Yeaton & P.M. Wortman (Eds.), Issues in data synthesis (pp. 25-42). New Directions for Program Evaluation, 24. San Francisco: Jossey-Bass.

Hedges, L.V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando: Academic Press.

Jenkins, J.R., & Pany, D. (1978). Standardized achievement tests: How useful for special education? Exceptional Children, 44, 448-453.

Lindsley, O. (1971). Precision teaching in perspective: An interview with Ogden R. Lindsley. Teaching Exceptional Children, 3 (3), 114-119.

Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.D., & Williams, P.L. (1985). Measurement-driven instruction: It's on the road. Phi Delta Kappan, 66, 628-634.

Quilling, M., & Otto, W. (1971). Evaluation of an objective based curriculum in reading. Journal of Educational Research, 65, 15-18.

Salvia, J., & Ysseldyke, J. (1985). Assessment in special and remedial education (3rd ed.). Boston: Houghton-Mifflin.

Skager, R. (1971). The system for objectives-based evaluation -- reading. Evaluation Comment, [illegible].

Thompson, R.H., White, K.R., & Morgan, D.P. (1982). Teacher-student interaction patterns in classrooms with mainstreamed mildly handicapped students. American Educational Research Journal, 19, 220-236.

Tindal, G., Fuchs, L.S., Fuchs, D., Shinn, M.R., Deno, S.L., & Germann, G. (1985). Empirical validation of criterion-referenced tests. Journal of Educational Research, 78, 203-209.

White, O.R. (1984). Descriptive analysis of extant research literature concerning skill generalization and the severely/profoundly handicapped. In M. Boer (Ed.), Investigating the problem of skill generalization: Literature review (pp. 1-19). Seattle: University of Washington, Washington Research Organization.

White, O.R., & Haring, N.G. (1980). Exceptional teaching (2nd ed.). Columbus, OH: Merrill.

Yalow, E.S., & Popham, W.J. (1983). Content validity at the crossroads. Educational Researcher, 12 (8), 10-14, 21.

Ysseldyke, J.E. (1979). Psychoeducational assessment and decision making. In J.E. Ysseldyke & P.K. Mirkin (Eds.), Proceedings of the Minnesota Roundtable Conference on Assessment of Learning Disabled Children (Monograph No. 8) (pp. 15-36). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.

Ysseldyke, J.E., & Thurlow, M.L. (1984). Assessment practices in special education: Adequacy and appropriateness. Educational Psychologist, 19, 123-136.

Zigmond, N., & Silverman, R. (1984). Informal assessment for program planning and

evaluation in special education. Educational Psychologist, 19, 163-171.

Footnotes

[1] In addition to these problems associated with noncomparability, measurement error, and sources of invalidity, research design features and multiple behavioral causes can be among limiting factors in interpreting quantitative integrations.

[2] Percentage of agreement was calculated using the following formula (Coulter cited in Thompson, White, & Morgan, 1982): Percentage of agreement = agreements between observer A & observer B/(agreements between A & B + disagreements between A & B + omissions by A + omissions by B).

[3] One paper authored by Haring (1971) and two additional reports by Haring & Krug (1975a, 1975b) described aspects of the same investigation. Therefore, although it is reported that 18 studies were employed in the meta-analysis, 20 appear in the Appendix due to the separate listing of the Haring (1971) and the Haring and Krug (1975a, 1975b) papers.

Table 1

Weighted Mean UESs, z Values, and Chi-Square Statistics as Analogues to Analysis of Variance by Type of Goal and Outcome Measure Factors

Factor	Weighted $\bar{X}$	z Value[a]	N[b]	χ^2	df
Type of Goal			96	.69	1
Long-term	.63	16.58	27		
Short-term	.67	24.82	69		
Outcome Measure			96	6.63[c]	1
Probe-like	.72	23.23	45		
Global	.61	19.06	51		

[a] A signficant z value indicates that the weighted mean is reliably different from zero. All z values are significant beyond the .001 level.

[b] N represents number of UESs not number of studies.

[c] $p < .05$.

Table 2

Weighted Mean UESs, z Values, and Chi-Square Statistics as Analogues to Analysis of Variance for Probe-Like and Global Outcome Measures within Type of Goal Conditions

Type of Goal/ Outcome Measure	Weighted $\bar{X}$	z Value[a]	N[b]	χ^2	df
Short-Term Goal					
Outcome Measure			69	56.78[c]	1
Probe-Like	.85	22.97	37		
Global	.45	11.54	32		
Long-Term Goal					
Outcome Measure			27	41.59[c]	1
Probe-Like	.41	7.32	14		
Global	.92	16.73	13		

[a] A significant z value indicates that the weighted mean is reliably different from zero. All z values are significant beyond the .001 probability level.

[b] N represents number of UESs not number of studies.

[c] $p < .001$.

Figure Caption

Figure 1. Unbiased mean effect sizes (UESs) for CBA-objective (----) and CBA-goal (——) on probe-like and global outcome measures.

Appendix

Reports Included in the Meta-Analysis

Beck, R (1976). Report for the Office of Education dissemination review panel. (Unpublished manuscript available at Precision Teaching Project, 3300 Third St. N.E., Great Falls, MT 59404.)

Beck, R. (1979). Report for the Office of Education dissemination review panel. (Unpublished manuscript available at Precision Teaching Project, 3300 Third St. N.E., Great Falls, MT 59404.)

Beck, R. (1981). Curriculum management through a data base. (Unpublished manuscript available at Precision Teaching Project, 3300 Third St. N.E., Great Falls, MT 59404.)

Beck, R. (1981). High school basic skills improvement project. (Unpublished manuscript available at Precision Teaching Project, 3300 Third St. N.E., Great Falls, MT 59404.)

Bohannon, R.M. (1975). Direct and daily measurement procedures in the identification and treatment of reading behaviors in children in special education. Unpublished doctoral disseration, University of Washington.

Brandstetter, G., & Merz, C. (1978). Charting scores in precision teaching for skill acquisition. Exceptional Children, 45, 42-48.

Bruening, S.E. (1978). Precision teaching in the high school classroom: A necessary step towards maximizing teacher effectiveness and student performance. American Educational Research Journal, 15, 125-140.

Dubrule, M.N. (1984). The study of precision teaching as a remedial method. Unpublished doctoral disseration, Clark University.

Fuchs, L.S., Deno, S.L., & Mirkin, P.K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. American Educational

Research Journal, 21, 449-460.

Frumess, S.C. (1973). A comparison of management groups involving the use of the standard behavior chart and setting performance aims. Unpublished doctoral dissertation, University of Washington.

Haring, N.G. (1971). Investigation of systematic instructional procedures to facilitate academic achievement in mentally retarded disadvantaged children. Final report. (ERIC Document Reproduction Service No. ED 071 248)

Haring, N.G., & Krug, D.A. (1975a). Evaluation of a program of systematic instructional procedures for extremely poor retarded children. American Journal of Mental Deficiency, 79, 627-631.

Haring, N.G., & Krug, D.A. (1975b). Placement in regular programs: Procedures and results. Exceptional Children, 41, 413-417.

King, R., Deno, S.L., Mirkin, P.K., & Wesson, C. (1983). The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparision (Research Report No. 111). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Mirkin, P.K. (1978). A comparison of the effects of three formative evaluation strategies and contingent consequences on reading performance. Unpublished doctoral disseration, University of Minnesota.

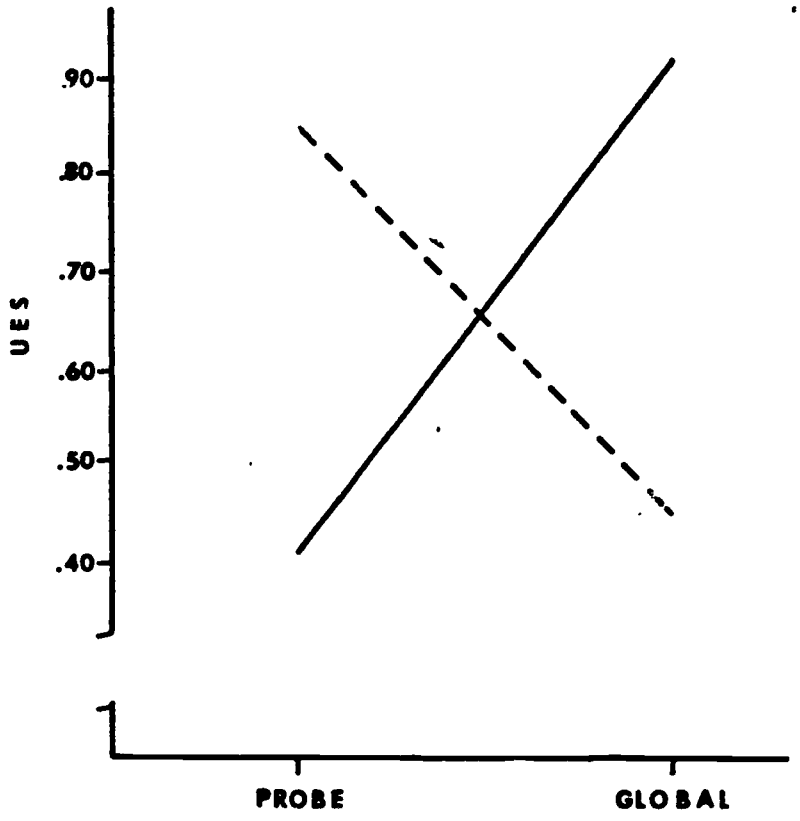
Mirkin, P.K., Deno, S.L., Tindal, G., & Kuehnle, K. (1980). Formative evaluation: Continued development of data utilization systems (Research Report No. 23). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. ED 197 510)

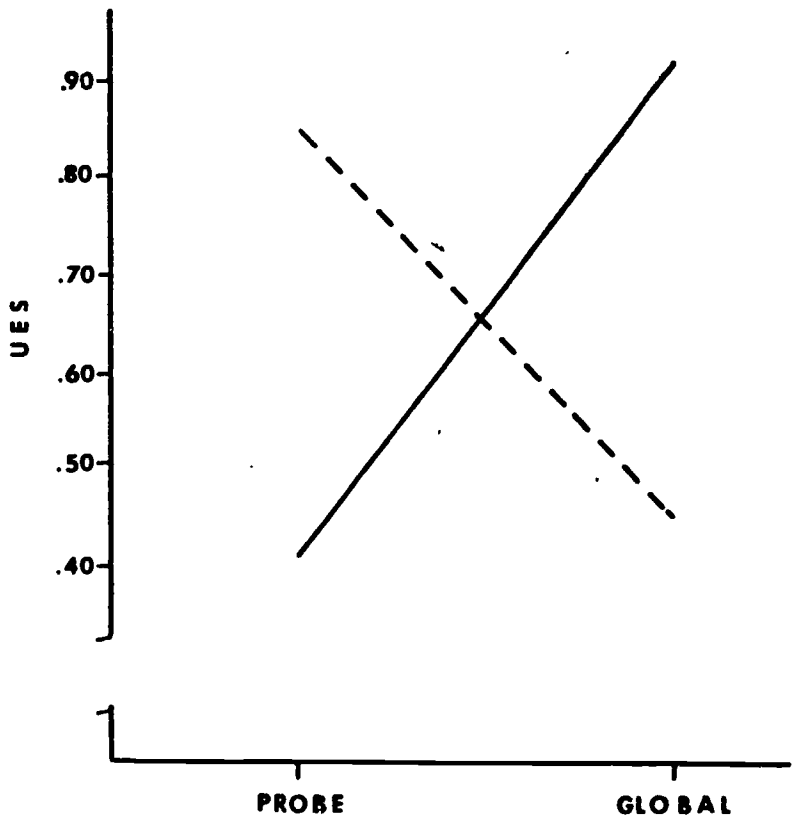
Peniston, E.G. (1975). An evaluation of the Portage Project: A comparison of a home visit program for mulitple handicapped preschoolers and Headstart program. (ERIC Document Reproduction Service No. ED 112 570)

Sevcik, B., Skiba, R., Tindal, G., King, R., Wesson, C., Mirkin, P.K., & Deno, S.L. (1983). Curriculum-based measurement: Effects of instruction, teaching estimates of student progress, and student performance (Research Report No. 124). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Skiba, R., Wesson, C., & Deno, S.L. (1982). The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparison (Research Report No. 88). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Tindal, G., Fuchs, L.S., Christenson, S., Mirkin, P.K., & Deno, S.L. (1981). The relationship between student achievement and teacher assessment of short- or long-term goals (Research Report No. 61). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. ED 218 846)


.90
.80
.70
.60
.50
.40
UES
PROBE
GLOBAL


.90
.80
.70
.60
.50
.40
UES
PROBE
GLOBAL