Resource Consultant Training Program
Monograph No. 3

RCTP

Curriculum-Based Measurement Research: Reports from the Oregon Conference, 1989 -1991

Edited by
Jerry Marr
Gerald Tindal

University of Oregon, Division of Teacher Education, Special Education Area,
Eugene, Oregon, 97403-1215

Published by
Resource Consultant Training Program
Division of Teacher Education
College of Education
University of Oregon

Staff
Gerald Tindal, Program Director
Jerry Marr, Editor

Parker, Richard
Curriculum-Based Measurement Research: Reports from the Oregon Conference, 1989 -1991

Preparation of this document was supported in part by the U.S. Department of Education, grant number H029B80119-89, and by the Oregon Conference. Opinions expressed herein do not necessarily reflect the position or policy of the U.S. Department of Education, and no official endorsement by the Department should be inferred.

Cover design: George Beltran

Contents

Note: Page numbering reflects numbering in the original Oregon Conference Monographs *from which the articles are taken.*

Curriculum-Based Measurement Research: Reports from the Oregon Conference, 1989 -1991

Edited by
Jerry Marr
Gerald Tindal

Introduction

The Oregon Conference is a large regional colloquium for special educators, hosted annually by the University of Oregon College of Education. For the past three years, the Conference has published a monograph to document its proceedings. The monograph is essentially a compilation of articles to accompany many of the presentations. The area of assessment, which has grown in recent years, is well represented in the conference monograph. In fact, with the current attention provided by politicians and the media, this assessment has been highlighted in general education also. Therefore, we decided that a special monograph that includes only assessment articles from the Oregon Conference would be of value.

Because of the current research focus of the special education faculty at the University of Oregon, the major form of assessment presented in this monograph is curriculum-based measurement (CBM). While many varieties of curriculum-based assessment are available, this particular form began at the University of Minnesota's Institute for Research on Learning Disabilities (IRLD). With the rise in population of students labeled learning-disabled (LD), widespread concern existed about their identification and programming (as of the 12th Annual Report to Congress on Implementation of the Education of the Handicapped Act, 1990, the national count was almost 2 million, comprising 48% of all handicapping conditions). In the late 1970's, Congress funded five Institutes around the country, with a broad mission of investigating different aspects of learning disabilities. At IRLD, the focus was on assessment, with two major strands: (a) documentation of current practices (headed by Dr. James Ysseldyke) and development of alternative procedures (headed by Dr. Stan Deno). From 1978 through 1984, a number of investigations were conducted in an effort to validate curriculum-based measures of student performance. This line of research and the measures that were developed came to be known as curriculum-based measurement (CBM). In this monograph, you will see extensions of this work.

One area to which CBM has been applied is identification. In an article exploring the relationship between learning disabilities and regular education, CBM is used to highlight differences (or lack thereof) between students labeled LD and those in the general-education classroom. Also research is described for identifying special-needs students ranging from at-risk to talented-gifted. The issues raised in this article reflect a growing concern with the labeling process, given the considerable overlap that appears in student skill levels. In fact, this same theme appears in a later Oregon Conference monograph article, which has been extended into a model for reintegrating students back to the general education classroom. The general argument is that measures used in the labeling process could and should be applicable not only for identifying students (of a particular type), but as a vital sign for evaluating program outcomes and signalling changes in needed program options.

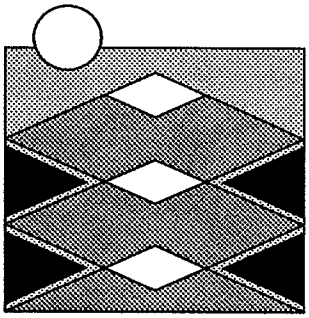
The other articles in this compilation offer a variety of classroom applications of CBM, which is where this line of research originally began and to which it should remain committed. This research agenda includes two dominant themes: (a) development of other measures that were not originally included in the initial validation research, and (b) use of CBM in teacher decision-making.

For example, in an article from the 1989 monograph, alternative writing measures are investigated, with both new objective measures (e.g., legible words written), and subjective measures (e.g., holistic evaluations). Other articles include mathematics assessment, which was never really addressed in the research at the IRLD, and curriculum-based assessment in the use of behavior ratings of students (1991). Presently, few, if any, studies have seriously incorporated this latter area within CBM, and little is known of the relationships between students' social and academic performance.

Probably the most critical area in which CBM research has moved and needs to continue is in teacher decision-making. A number of the articles in this compilation have such a focus, with a wide variety of decisions considered. For example, teachers need content-valid measures that reflect their instructional focus. Two interesting applications of CBM have been made in mathematics (using criterion-referenced diagnostics and dynamic assessments) and social studies (focusing on knowledge forms). Teachers also need measures for organizing their classrooms and grouping students, for which CBM appears particularly well suited. Given the small amount of time needed for measuring students, this decision should be made quickly and efficiently, with success of instruction used eventually as the outcome for adjusting such decisions. Another example of this area is the research in different metrics used for making decisions. Two articles exemplify such a focus. In one, slope of improvement is used to project reading growth, and, in the other, floating percentiles are used to ascertain performance within an instructional group. In both, the metric used to make decisions provides a context for understanding and interpreting performance. Such alternatives must replace our current dependence on group-administered, published achievement tests, which rely solely upon level of performance or relative standing within the norm group. As reflected in the article on their construct validity, a close look at the use of such tests reveals that "the emperor has no clothes." Finally, in probably the most unique and far-reaching application of classroom-based measures, the teacher certification board in Oregon has mandated use of student portfolios in judging student teacher effectiveness. While this is an exciting development with considerable potential, many issues need to be resolved, as reflected in the article on CBM portfolios.

In summary, this compilation of articles from the Oregon Conference represents a range of options to which CBM has been applied. Whether the interest is in comparing students with each other or measuring learning and growth, curriculum-based measurement has been expanded to address a large number of issues since its inception at the University of Minnesota. We hope such extensions encourage us all to consider seriously the reason for, and outcomes from, specific assessment systems.

The Oregon Conference



1989

Using Curriculum-Based Measures to Explore the Relationship Between Learning Disabilities and Regular Education

Mark R. Shinn
Kathleen Rodden-Nord
Victor Nolet
Nancy Knutson
University of Oregon

OVERVIEW

The relationship between special education for learning disabled students (LD) and regular education continues to be controversial. One theory for explaining the role that special education services for LD-labeled students play for regular education is offered by Gerber and Semmel (1984; 1985). Their theory was stimulated by the growth of services for LD students and by what they call the failure of the psychometric model (i.e., complex, expensive testing has failed to identify a constellation of cognitive or other characteristics that reliably identify learning disabilities).

The Gerber and Semmel (1984; 1985) theory is based on a micro-economic analysis of teacher resources. These researchers see tremendous pressures that have been placed on regular education teachers, who are expected to cope with increasing class sizes and the growing numbers of students unable to perform satisfactorily within the regular education curriculum. To cope with an increasing range of student abilities and decreasing educational resources (i.e., a single teacher per room), teachers must select instructional strategies that:(a) maximize mean outcomes, and/or (b) minimize variance in the classroom (Gerber & Semmel, 1984; 1985). Should teachers choose the former, more teacher effort and resources are committed to those students who are perceived to gain the most from what is provided (i.e., higher achievers). This approach is hypothesized to result in increased variability in student achievement, however. Should teachers choose the latter (i.e., minimizing classroom variance in academic achievement), their efforts are directed towards making student outcomes as similar as possible, regardless of ability. Consequently, the lowest-achieving students are the recipients of the greatest effort and the most resources. It has been demonstrated that teachers most frequently choose the first instructional strategy, directing their efforts and showing more positive affect toward "teachable" students and away from students who are more difficult to teach (Brophy & Good, 1974).

According to Gerber and Semmel (1984), teachers have a third option that maximizes outcomes and reduces variability: Extremely low-performing students may be referred for, and subsequently placed in, special education. In current practice, the teacher's decision to refer a student for evaluation is a strong determinant of special education placement. It is estimated that nearly 75% to 90% of referred children eventually are found eligible for special education services (Algozzine, Christenson, & Ysseldyke, 1982). One possible outcome is that regular education classrooms are more manageable without these low-achieving students, who typically demand more teacher time and instructional resources than their higher-achieving classmates. Another result is that more educational resources (i.e., teacher time) are available for students who may gain the most. In addition, teachers may teach groups of students who are more homogeneous in their academic skills, thereby reducing the inherent difficulty in planning for and instructing students with a broad range of abilities (Kulik & Kulik, 1982; Winn & Wilson, 1983).

REQUIRED EVIDENCE FOR THE GERBER AND SEMMEL THEORY

Gerber and Semmel (1984; 1985) argue the need to understand learning disabilities as having a fundamental relationship to failure in the regular education classroom and believe that as currently operationalized, the LD phenomenon is explained best as regular education teachers' attempts to reduce classroom variance by placing instructional "outliers" and/or those students who exceed their "tolerance" for academic or social behaviors, in special education. Five data-based outcomes are necessary for empirical verification of this perspective. There must be evidence: (a) of a continuous influx of students into special education programs; (b) of a substantive number of students in the regular education classroom whose level of achievement may exhaust existing teaching resources; (c) that regular education teachers refer the most deviant, or lowest-performing, students for special education; (d) that those students who are placed in special education are extremely low achievers; and (e) that a substantive proportion of LD students have sufficient academic skills that suggest that they could be educated in the regular education classroom. This paper presents the results of four school-based studies that use Curriculum-Based Measures of reading (CBM) to explore the relationship between special and regular education.

An extensive body of literature supports the validity of CBM reading measures as tests of general reading skill (Marston, in press). Deno, Mirkin, and Chiang (1982), for example, documented concurrent validity coefficients ranging from .73 to .91 between CBM and published reading achievement tests, with most of these correlations above .80. Criterion-related validity also has been documented between reading fluency measures and four basal reading series' criterion-referenced mastery tests (Tindal et al., 1985). Fuchs, Fuchs, and Maxwell (1988) also provided data supporting CBM as a valid measure of reading comprehension. The primary advantage of CBM reading measures in studying the relationship between regular and special education is their relevance for assessing performance in the regular education curriculum.

Growth of Special Education

Extensive data support the first prediction. Nearly every source of data on referral and placement rates (Gerber, 1984; Kauffman, Gerber, & Semmel, 1988) suggests a constant number of students referred for special education services and a stable, if not growing, number of students placed in special education programs for the mildly handicapped each year.

Range of Skills in the Regular Education Classroom

Rodden-Nord and Shinn (1988) presented the results of an analysis of the range of reading skills in the regular education classroom. Subjects for this study were 2,815 regular education students in grades 1 through 6 from a large midwestern city. Although 16% of the sample received Chapter I services, special education students were systematically excluded from the sample. Testing materials for this study were derived from the students' regular education basal reading series. A reading passage of approximately 300 words was selected randomly from a third-grade level *Ginn* reader and a word list was generated by randomly selecting first- through fourth-grade words that appeared on the vocabulary list compiled by Harris and Jacobson (1972). Students were required to read from the Harris-Jacobson Word List for 1 minute and then read the third-grade passage for 1 minute. The number of words read correctly served as the dependent measures for this study.

The amount of overlap in reading skills between grades increased steadily from the first to sixth grades. For example, on the third-grade reading passage, 10% of second graders performed below the median of first graders; only 1% performed below the 25th percentile of first graders. A substantial number of students at the upper grades earned scores below lower grades' medians and first quartiles. Fully 36% of sixth graders performed below the fifth-grade median; almost one in five performed below the first quartile of fifth graders. About one in four and one in five sixth graders earned scores below the fourth-grade and third-grade medians, respectively. A significant proportion (2%) of regular education sixth graders earned scores *below the second-grade median*. When regular education fifth graders were asked to read the third-grade passage, 23% displayed reading fluency below the third-grade median; 1% were less fluent than median second-grade students, who should have found this material most difficult as it was above their grade placement! A similar pattern was found on the Harris-Jacobson word list. These findings support Gerber and Semmel's premise that there are many students in regular education classes whose reading skills are well below their grade-level peers and more similar to students in lower grades.

Lowest Performers Are Referred

The second study was an investigation of the hypothesis that teachers refer their lowest-performing students for possible special education services. Shinn, Tindal, and Spira (1987) found that regular education teachers consistently referred their lowest-performing students for special education services for the mildly handicapped. Grade-level CBM reading procedures were used to assess 570 students in grades 2 through 6 who were referred because of reading difficulties. Shinn et al. (1987) determined that in four of the five grades, the referred group's median reading achievement approximated the 5th percentile of the reading achievement for same-grade regular education students in the district. Nearly three of every four referred students earned scores below the 10th percentile of their regular educaton peers. Thus, students whom teachers perceived as potentially "handicapped" were characterized by low reading performance relative to the norm group. The authors concluded that "consistent with Gerber and Semmel's (1984) speculation, it appears that the referral process may best be characterized as an index of teacher tolerance and an attempt to reduce the variance of students in their classrooms" (p. 38).

LD Students Are Extremely Low Performers

The third study researched the reading achievement performance of students typically on LD caseloads. Shinn, Tindal, Spira, and Marston (1987) provided data that students placed in special education are extremely low achievers relative to their regular education classroom peers. They compared the reading skills of 638 learning disabled (LD) students in grades 1 to 6 to a sample of 451 students who received Chapter I services and to 2,337 regular education students. Students were selected randomly from their respective groups within the district. Again, grade-level CBM reading measures were used. All of the LD students had an IEP objective in reading and had been placed in special education by means of an eligibility criterion of significant ability-achievement discrepancies. Across grades, significant differences in read-

ing scores were evidenced for the three groups. The typical student placed in LD programs via traditional procedures performed at about the 3rd percentile compared to their regular education peers. Three of four LD students scored at the 8th percentile or below. The typical Chapter I student performed around the 20th percentile; three of four Chapter I students scored at the 30th percentile or below. The authors concluded that lack of success in classroom reading materials, as measured by CBM procedures, may have been the basis for LD social policy. LD students were homogeneous in their low achievement.

Some LD Students May Exceed Teacher Tolerances

The fourth study was an examination of the degree to which there exists a pool of students who conceivably could be educated in the regular education classroom, given their reading skills approximate those of other students in the lowest regular education reading group. These students would be considered academically competent for integration into regular education classrooms. LD students were tested via CBM reading procedures and their performance compared to that of students in the lowest reading group in the regular education classroom. To make this determination, the reading performance of students who received reading instruction in resource room settings was compared with that of students who were instructed in the lowest reading group in the regular classroom.

Students from grades 3, 4 and 5 were selected from six rural and suburban schools in southern Oregon. Classroom teachers at each grade level identified 233 regular education students who received instruction in the lowest reading group in their room. Special education resource room teachers in each school identified 77 students at each grade level for whom they provided primary reading instruction.

All students read from the lowest reading level used in regular education at their grade level in their building. Each classroom teacher identified the specific curriculum material used by their lowest reading group and two passages were randomly selected from the lowest basal reader used at each grade level in each building. The mean number of words read correctly across the two passages represented the reading score for each student. The scores obtained by the special education students were compared with these distributions to see how many (academically competent) special education students obtained scores in the range of scores obtained by low reading group students. Academically competent special education students, those students whose scores were within the range of students in the lowest reading groups, were founds in all six school buildings. In only 2 of the 18 grades were no academically competent students found. The percentage of academically competent special education students among those tested ranged from 15% to 100% across grades and building. The median percentage of academically competent students in a building was 40%.

SUMMARY AND DISCUSSION

It seems logical that any discussion of special education reform for mildly handicapped students should be guided by a data-based understanding of the relationship between the two systems. To date, all five aspects of Gerber and Semmel's model have garnered empirical support: (a) special education services for mildly handicapped students continue to grow, (b) there is a considerable range of reading skills in the regular education classroom, (c) teachers refer the lowest-performing students academically, (d) the lowest-achieving students are placed in special education, and (e) a substantive proportion of these low-achieving students have academic skills that conceivably could allow them to be educated in the lowest reading group in the regular education classroom.

If there is "truth" to the Gerber and Semmel (1984) theory, a significant contribution can be made toward understanding the pros and cons of REI. From the special education perspective, without reform of regular education, special education will continue to assume responsibility for difficult-to-teach children (Gerber & Semmel, 1984) and will continue to grow unchecked because there is a seemingly unending supply of extremely low performers in regular education classrooms. Special educators are concerned that their resources will become stretched as they too then face the problem of increasing numbers of students and greater heterogeneity of academic need. It could be argued that the REI is, in fact, a micro-economic response to just this problem! From the regular education perspective, it would not be desirable to place large numbers of students who are outside the range of modal instruction (i.e., those students currently in programs for the mildly handicapped) back into classrooms that, by definition, already lack the resources to address the problems of these students.

Further investigation of the Gerber and Semmel (1984; 1985) model may contribute a viable approach to regular (and special) education reform. In particular, studies are needed on how to best allocate regular and special education resources to (a) maximize the increase in mean achievement and (b) reduce the variability in regular and special education classrooms.

REFERENCES

Algozzine, B., Christenson, S., & Ysseldyke, J. (1982). Probabilities associated with the referral to placement process. *Teacher Education and Special Education*, 5, 19-23.

Brophy, J., & Good, T. (1974). *Teacher-student relationships: Causes and consequences*. New York: Holt, Rinehart & Winston.

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

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, 9, 20-28.

Gerber, M. (1984). The Department of Education's sixth annual report to Congress of PL 94-142: Is Congress getting the full story? *Exceptional Children, 51*, 209-224.

Gerber, M., & Semmel, M. (1984). Teacher as imperfect test: Reconceptualizing the referral process. *Educational Psychologist, 19*, 137-148.

Gerber, M., & Semmel, M. (1985). Micro-economics of referral and reintegration: a paradigm for evaluation of special education. *Studies in Educational, 11*, 13-29.

Harris, A. P., & Jacobson, M. D. (1972). *Basic elementary reading vocabularies*. New York: Macmillan.

Kauffman, J. M., Gerber, M. M., & Semmel, M. I. (1988). Arguable assumptions underlying the Regular Education Initiative. *Journal of Learning Disabilities, 21*, 6-11.

Kulik, C. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Education Research Journal, 19*, 415-428.

Marston, D. (in press). A curriculum-based measurement approach to assessing performance: What is it and why do it? In M. Shinn (Ed.), *Curriculum-based measurement and special services to children*. School Psychology Practitioner's Series. New York: Guilford.

Rodden-Nord, K., & Shinn, M. R. (1988). *The range of reading skills with and across regular education classrooms: Contributions to understanding special education for the mildly handicapped*. Unpublished manuscript. Eugene, OR: University of Oregon.

Shinn, M. R., Tindal, G. A., & Spira, D. A. (1987). Special education referrals as an index of teacher tolerance: Are teachers imperfect tests? *Exceptional Children, 54*, 32-40.

Shinn, M. R., Tindal, G. A., Spira, D. A., & Marston, D. (1987). Practice of learning disabilities as social policy. *Learning Disabilities Quarterly, 10*, 17-28.

Tindal, G. A., Fuchs, L., Fuchs, D., Shinn, M., Deno, S., & Germann, G. (1983). *The technical adequacy of a basal reading series mastery test: The Scott Foresman reading program* (Research Report No. 128). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.

Winn, W., & Wilson, A. P. (1983). The affect and effect of ability grouping. *Contemporary Education, 54*, 11-125.

Classroom Uses of Curriculum-Based Reading and Written Expression Measures

Jan Hasbrouck
Richard Parker
University of Oregon

To make meaningful decisions about instruction, teachers need an adequate measure of student performance. This article covers both of these aspects of assessment: making decisions and measuring performance. In the first section, the focus is on placement of students into reading groups; in the second section, the focus is on measuring performance in written expression.

PART 1: PLACING STUDENTS IN READING GROUPS

When teachers are making decisions about placing students in reading groups within or across classrooms, what objective measures can give them the most useful information in the most efficient manner? Traditionally, students have been administered informal reading inventories (IRIs) that are (a) provided by publishers with their basal readers, (b) developed by teachers from their classroom reading programs, or (c) produced as "generic" publications not associated with any particular reading program (Harris & Lalik, 1987).

IRIs are flexible, allow timely feedback, and often are directly related to the reading curriculum. Although their forms vary, shared characteristics are that they: (a) are individually administered, (b) require students to orally read passages, (c) require students to answer comprehension questions orally after reading a passage, and (d) are used to determine students' independent, instructional, and frustration levels of reading. In addition, some IRIs use word lists to select reading passages, include silent reading, allow students to look at the text while answering questions, or include timing the students' oral reading performance.

Teachers like and trust IRIs for making placement decisions. They are flexible, subjective, multi-faceted, and holistic, and they measure reading in a "natural" mode (Burry, Catterall, Choppin, & Dorr-Bremme, 1982; Haller & Waterman, 1985; Salmon-Cox, 1981; Strike, 1983). But they are far from perfect (Powell, 1968; Schwartz, 1984). Potential problems that can reduce effective communication between teachers about students' reading include: (a) They require subjective teacher judgments in their construction, administration, scoring, and interpretation; (b) they can be very time consuming for teachers to administer; and (c) there is a lack of agreement among teachers on what constitutes independent, instructional, and frustration levels.

In addition to IRIs, teachers use a wide variety of other data to make grouping and placement decisions, including direct observation of students' behaviors, study and work habits, daily performance on classroom work, and recommendations of previous teachers. Since much of this information is subjective in nature, it may be argued that some other reliable, objective assessment also is required. This objective assessment also should be efficient and related to the classroom curriculum. It also should compare favorably with IRIs in the accuracy of its group placement recommendations. Such an assessment tool should be able to be administered and interpreted consistently from one time to the next, and should allow accurate communication among teachers for cross-classroom grouping.

In a study conducted in a fourth- through sixth-grade elementary school with 280 pupils, three reading measures—IRI, Maze (multiple choice Cloze), and oral reading fluency—were examined as to how well they matched teachers' cross-classroom reading group placement decisions. Six months before the study began, teachers had established five reading groups at each grade level on the basis of information that included cumulative folder reports, recommendations of the previous year's teacher, and informal classroom observations.

Approximately 40 students in each grade were selected for the study from across the five reading groups (special education/remedial, low, low-medium, medium-high, and high). These 120 students were tested using (a) passages from a commercial IRI, the Analytic Reading Inventory (ARI) (Woods & Moe, 1985), (b) 1-minute timed oral readings of three 250-word passages from the students' basal reading program, and (c) three 250-word Maze tests, also produced from the basal reading program.

The ARI was adapted for this study: Word list reading and miscue error analyses were omitted, and consistent decision rules were established for how many and which passages each student was to read. The three Maze and oral reading fluency passages had reading levels at grade level, 1 year below, and 1 year above, according to averaged Spache, Fry, and Raygor readability formulas. The ARI and oral reading fluency assessments were individually administered, while the Maze tests were group-administered during students' reading periods.

On both the ARI and the oral reading tests, omissions, additions, substitutions, and 3-second hesitations were coded as errors; self-corrects and repetitions were not considered errors. If students skipped a line in the text, they were interrupted and the correct line pointed out to them. Their oral reading score was the percent of words read correctly per minute, averaged over the three passages.

Although each of the three measures was a reasonably good predictor of group placement, the Maze most consistently matched the students' existing group placements. Combining the Maze and oral reading fluency results offered the same predictive accuracy as the more time-consuming IRI.

Table 1: Classification Accuracy (Percent Correct) in Reading Group Placement Using Combined Maze and Oral Reading Fluency (OR) Results and Analytical Reading Inventory (ARI) Results

	Within One Level			Exact Match		
Grade	4	5	6	4	5	6
Maze + OR	94	100	84	56	59	48
ARI	87	98	82	46	55	58

As a group test, the Maze was the most efficient to administer. Teachers reported that most students completed the three Maze passages within 20 minutes. The oral reading fluency measure required an average of 1.5 minutes per passage to administer, or 5 minutes for the three passages for each student. In contrast, administration of the adapted version of the ARI, *an individualized test*, required 20 minutes. The complete *student* version of the ARI would have required about an hour to administer to each student.

The combined administration of Maze and oral reading fluency assessment was efficient and accurate in predicting teachers' placement of students into reading groups. These two measures contain most of the popular features of an IRI: (a) individual administration of the oral reading fluency passages, which gives teachers the chance to evaluate each student's reading ability holistically, (b) an assessment of students' comprehension, and (c) the use of students' actual reading materials for assessing their skills. Their advantage is that they can be developed, administered, and scored more objectively than an IRI.

PART 2: SCORING STUDENTS' WRITING

Before setting objectives for writing instruction in special education, teachers must identify their students' skill deficiencies. Unfortunately, most published writing tests are time consuming to administer and score, difficult to interpret, and of questionable relevance to low-performing students. Such tests are of little use to teachers in setting goals and planning instruction. In contrast, a number of features of actual student writing samples (e.g., T-units, correct word sequences, percent of correctly spelled words) can be measured directly and objectively (Isaacson, 1985). Simple counts of the features can help highlight areas of needed remedial writing instruction for both groups and individuals.

Research conducted at the University of Oregon over the past 3 years has found moderate-to-strong correlations between several measures and teachers' holistic judgments of the quality of these same writing samples. Not all direct, objective measures are appropriate for assessing the writing of all students, however. Research on writing samples from elementary and secondary students in regular, remedial, and special education programs has indicated that any one of these measures is best applied to a limited range of grade and/or ability levels (Tindal & Parker, in press).

Assessing Communication Properties of Writing

When scoring students' writing samples, teachers must decide on the aspects of writing to assess. A valid assessment procedure should measure properties of writing clearly related to communication (Wallace, Cohen, & Polloway, 1987). Researchers including Isaacson (1985) and Stewart and Grobe (1979) have identified components of writing related to communication, including legibility, spelling, syntax, use of mechanics, vocabulary usage, and content.

This study assessed the relationship between several direct, objective writing indices and the holistic judgments of the communicative quality of the writing by experienced special education and remedial teachers. Students in remedial (N=142) and special education (N=30) programs contributed writing samples written in response to a common story starter. Four of the objective indices—percent of correctly sequenced words, percent of correctly spelled words, percent of legible words, and the average length of continuous correct word sequences—were moderate to strong predictors of teachers' holistic ratings (r = .6-.8). The weakest predictor was the total number of words written (Tindal & Parker, in press). The remainder of this paper details the scoring procedures used in the study.

Writing Sample Collection Procedures

Standardized procedures were used to collect the writing samples. Students wrote for 3 minutes from the following story starter: "I went up to the old, deserted house. The door was open so I walked in. Suddenly . . ." The students were given 1 minute to think about the story they would write and then directed to begin writing. After 3 minutes, the teacher cued the students to mark a star on their papers, after which they were given a few more minutes to finish their story. The samples were scored up to the star only.

Holistic Scoring Procedures

The holistic judgments of the writing samples were completed first because they must be made on clean, un-

marked samples. Holistic scoring relies on global impressions of quality, not line-by-line scrutiny. The scorer does not focus on single aspects of a paper such as organization, mechanics, and persuasiveness, even though these traits undoubtedly influence the final judgment (Rafoth & Rubin, 1984; Spandel, 1981). With holistic scoring, students are rated according to overall writing proficiency, typically on a 5-point scale. Model papers, or "range finders," help guide the scoring decisions and improve consistency among different scorers. In this study, the following definition of good writing also helped to ensure consistency in scoring:

Good writing clearly communicates to the reader the ideas/story of the writer. Good writing requires legible handwriting or printing, as well as distinguishable words, phrases, and sentences. Coherent linking of ideas from one sentence to the next also contributes to good writing.

Objective Scoring Procedures

Percent of legible words. In order for writing to communicate, it must first be "readable" or legible. To determine legibility in an objective manner, teachers first counted the number of *illegible* words, which were defined as groups of letters not recognizable as single, particular words *outside of the context of the phrase or sentence*. Teachers judged each word separately from the others by working backward from the last word written to the first word of the story. The number of legible words was divided by the total number of words written to calculate the "percent of legible words."

Total number of words written. A word was defined as a group of letters separated on a line by spaces. This count included (a) legible and illegible words and (b) words spelled correctly and incorrectly. Calculating the total number of words written was an intermediate step to arriving at other scores.

Number of correctly spelled words. Correct spelling plays an important role in the ability of writing to communicate with a reader. To measure this aspect of writing, all words were circled that were spelled correctly given the context of the sentence. This score was used to calculate students' percent of correctly spelled words.

Percent of correctly spelled words. The number of correctly spelled words was divided by the total number of words to obtain a score for percent of correctly spelled words.

Number of correct and incorrect word sequences. After legibility and correct spelling were assessed, teachers measured the grammatical accuracy of the writing sample by counting the correct and incorrect sequences of words. A correct word sequence was defined as a sequence of two adjacent, correctly spelled words that were syntactically and semantically correct (Videen, Deno, & Marston, 1982); sequences terminate at the end of sentences. A caret mark (^) was used between the two words of each correct word sequence. An inverted caret was placed between each *incorrect* sequence of words spelled either correctly or incorrectly. Each pair of words was then either a correct or incorrect sequence, marked by a regular or inverted carat.

Percent of correct and incorrect word sequences. The number of correct and incorrect word sequences was added, then divided into (a) the number of correct word sequences (for percent of correct word sequences) and (b) the number of incorrect word sequences (for percent of incorrect word sequences).

Mean length of correct word sequences. A widely accepted goal for written expression is that students use mature syntax by expanding their sentences (Isaacson, 1985). To score this aspect of students' writing, brackets were placed around all continuous strings of non-inverted carat marks. The number of carat marks per string was counted. These separate counts were then averaged to form "mean length of correct word sequences."

SUMMARY

The purpose of this article was to describe some of the research in reading and written expression assessment currently being conducted at the University of Oregon. The major focus of these investigations has been to develop procedures that teachers can use in their classrooms to make a variety of decisions. An important assumption of this work is that assessment must be linked closely to instruction and student activity in the classroom. Two important areas of research involve teacher decision-making (i.e., placement of students into reading groups) and monitoring student learning (i.e., establishing a sensitive metric for analyzing performance).

Note: The scoring procedures listed here were condensed because of space restrictions. A complete set of scoring procedures are included in: Hasbrouck, J. & Parker, R. (1988). *Objective and holistic scoring of writing: Procedures and practice materials (Module No. 2)*. Eugene, OR: University of Oregon, Resource Consultant Training Program. For information contact: Dr. Gerald Tindal, Director. Resource Consultant Training Program, College of Education, University of Oregon, Eugene, Oregon 97403).

REFERENCES

Burry, J., Catterall, J., Choppin, B., & Dorr-Bremme, D. (1982). *Testing in the nation's schools and districts: How Much? What kinds? To what ends? At what costs?* (CSE Report No. 194). Center for the Study of Evaluation, Graduate School of Education, UCLA.

Haller, E. J., & Waterman, M. (1985). The criteria of reading group assignments. *The Reading Teacher, 38*, 772-781.

Harris, L. A. ,& Lalik, R. M. (1987). Teachers' use of informal reading inventories: An example of school constraints. *The Reading Teacher, 40*(7), 624-630.

Isaacson, S. (1985). Assessing written language skills. In C. S. Simon (Ed.), *Communication skills and classroom*

success: Assessment methodologies for language-learning disabled students (pp. 403-424). San Diego: College-Hill Press.

Powell, W. R. (1968). Reappraising the criteria for interpreting Informal Inventories. In D. L. DeBoer (Ed.), *Reading diagnosis and evaluation: Proceedings of the 13th annual Convention of the International Reading Association*. Newark, DE: International Reading Association.

Rafoth, B. A., & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication, 1*(4), 446-457.

Salmon-Cox, L. (1981). Teachers and standardized achievement: What's really happening? *Phi Delta Kappan, 62*, 631-634.

Schwartz, S. (1984). *Measuring reading competence: A theoretical-prescriptive approach*. New York: Plenum Press.

Spandel, V. (1981). *Classroom applications of writing assessment: A teacher's handbook*. Portland, OR: Northwest Regional Educational Laboratory.

Stewart, M. F., & Grobe, C. H. (1979). Syntactic maturity, mechanics of writing, and teachers quality ratings. *Research in the Teaching of English, 13*, 207-215.

Strike, K.A. (1983). Fairness and ability grouping. *Educational Theory, 33*, 125-134.

Tindal, G. , & Parker, R. (in press). Assessment of written expression for students in compensatory and special education programs. *Journal of Special Education*.

Videen, J., Deno, S., and Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Research Report No. 84). University of Minnesota: Institute for Research on Learning Disabilities.

Wallace, G., Cohen, S. B., & Polloway, E. A. (1987). *Language arts: Teaching exceptional children*. Austin, TX: Pro-Ed.

Woods, M. L. & Moe, A. J. (1985). *Analytical Reading Inventory (3rd Edition)*. Columbus, OH: Charles E. Merrill.

Using Curriculum-Based Measures to Group Students in Reading

Tracey Hall
Gerald Tindal
University of Oregon

Throughout the school year, and particularly at the beginning, teachers place students into "homogeneous" reading groups, taking into account both the students' skills and the curriculum scope and sequence. Haller and Waterman (1985) report that teachers consider many attributes and circumstances to make placement decisions. "Reading ability, as perceived by the teacher, is the most important consideration, but not the only factor considered" (p. 779). Other criteria that enter into grouping decisions include general academic competence, behavior/personality attributes, work habits, and home background. Standardized published tests, administered either at the beginning or end of the year, often are administered to find grade level within which students are functioning. Frequently, Informal Reading Inventories (IRIs), either commercial or teacher-made, are given. Finally, teacher reports are an important basis for placement of students into instructional groups (Coleman & Harmer, 1982).

While published tests, IRIs, and teacher reports are purported to have validity for group placement decisions, they are far from perfect when used with specific curriculum materials. They require extensive teacher time to prepare and administer. More importantly, they are not amenable to *reliable* implementation on a very *frequent* basis; as a consequence, students are virtually locked into instructional groups for an entire school year unless conditions become extreme and they are perceived as "not fitting" into their class or instructional groups.

An alternative practice that may provide the benefits of these two procedures without their limitations is the use of oral reading fluency as defined under the rubric of curriculum-based measurement (CBM). As Deno (1985) has noted, this practice is becoming increasingly functional with a strong empirical base. This measure has the technical adequacy characteristics of published achievement tests (Deno, Mirkin, & Chiang, 1982; Deno, Fuchs, & Fuchs, 1982), focuses on relevant production behaviors like the Informal Reading Inventories (Johns, Garten, & Schoenfelder, 1977; Pikulski & Shanahan, 1982), and is consistent with teacher judgments (Marston, 1982).

While most of the previous research on CBM has focused on special education populations (Shinn, 1988), the study described here focuses on its application to general education. The purpose of this article is to report on (a) normative results in general education curriculum-based reading measures and (b) their relationship to the placement of students in reading groups. Two major sources of data, therefore, were student performance on oral reading and fluency and teacher assignment of students to reading groups.

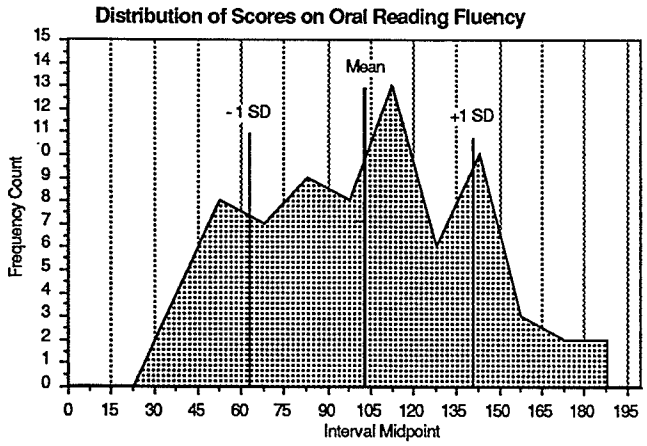
For the study, students in grades 2 through 5 at a school were given three 1-minute oral reading measures from randomly selected passages in grade-level readers. All students in the building were tested, including those receiving special education, Chapter I, and gifted and talented services. Readability levels of the passages within the texts varied greatly, sometimes as much as five grade levels. Therefore, grade-level passages with readability levels within a 1.5 range were selected. For example, passages with a readability of 3.5 to 6.5 were sampled from fifth-grade books. On each of the three reading assessments, misidentifications, omissions, reversals, substitutions, and 3-second hesitations were coded as errors; repetitions, insertions, and self-corrections were not considered errors. However, these latter responses impeded performance rate. Thus, the student's proficiency was negatively influenced. The normative results were based on the number of words read correctly per minute with performance from the two passages averaged for each student.

After students had been ability-grouped within and across grade levels for reading, procedures for making placement decisions were ascertained. Like the teachers described by Harris and Lalik (1987) and Haller and Waterman (1985), these fourth- and fifth-grade teachers administered yearly standardized measures, gave IRIs from the publisher, and used teacher reports and workbook materials to place students in reading groups. Students had been assigned to one of four across-grade groups (high, high-middle, low-middle, and low). Generally teachers from two grades shared students (i.e., first-second or fourth-fifth). Within each reading group, teachers ranked their students from low to high.

CBM scores for 67 fourth-grade and 126 fifth-grade students were sorted from low to high and by grade for subsequent analyses, ranging from the least proficient reader (lowest rate of words read per minute) to the most proficient reader (highest reading rate). This display of the grade level reflects a sensitive measure and a frequency distribution that closely approximates a normal curve. The mean reading rate for the fifth-grade students, represented in Figure 1, was 102 words read correctly per minute, with a standard deviation of 36.2.

The results were similar to previous school-based or local norming applications using CBM, in which such data were used extensively to make special education eligibility decisions (Deno et al.,1982; Deno, Marston, Shinn, & Tindal, 1983; Fuchs, Fuchs, & Maxwell, 1988). However, as noted

Figure 1: Mean Reading Rate



above, such local normative data has not been used extensively for instructional placement decisions within general education. Can curriculum-based normative data be used to help regular classroom teachers make instructional placement decisions in reading? To answer this question, student performance rankings from the curriculum-based measures were compared to student placement in reading groups using two procedures: (a) The average CBM reading performance for the reading groups was compared to each other and (b) placement into reading groups using CBM measures was compared to teachers' placement.

TEACHER PLACEMENT AND CBM PERFORMANCE

The four teacher-determined groups were regrouped for this analysis into three groups: one high group, one middle group (comprised of both the high- and low-middle), and one low group. In comparing the average performance for these three groups, significant differences were found. The low group (n=13) had an average rate of 50.6 correct words read per minute (SD=26.6), the medium group (n=45) had a mean score of 104 (SD= 23.3), and the high group (n=15) read an average of 154 words correctly per minute (29.1). These results clearly indicate that teacher judgment is consistent with reading proficiency: Teachers place students who are least proficient into the low group, those with more fluency into the medium group, and those with the greatest proficiency into high reading groups. It is important to remember that these placements had been made prior to independent measures of oral fluency.

The second analysis focused on the relationship between teachers' placement of students into reading groups (low, middle, and high) and their "placement" into groups using a trimarcation of the CBM normative distribution. Using the norming results, the distribution of scores was split into three groups—low, medium, and high. The low groups consisted of students with scores lower than 1 standard deviation below the mean; the medium group represented students whose scores were between plus or minus 1 standard deviation from the mean; and the high groups consisted of all students whose scores fell higher than 1 standard deviation above the mean. Of course, teachers also had placed students into three groups, with which these data were compared.

The relationship between CBM and teacher placement into groups was very strong (x^2=46.3, 4 df, p=.0001). The highest degree of consistency between teacher and CBM placement was for the two extreme groups, low or high (90% accuracy and above). Placement differences between CBM and teachers were greatest for students in the middle group, which had 14.8% of students not matched on the two measures.

ISSUES IN READING GROUP PLACEMENT

In both analyses, Curriculum-Based Measurements closely reflected teacher placement of students into reading groups. A high relationship was found between teacher placement of students in reading groups and their actual reading performance on curriculum-based reading measures. However, some discrepancies also occurred.

While the CBM data were limited to an average of two 1-minute reading samples, teacher judgment had been based on an amalgam of standardized published tests, IRIs, and teacher reports. Yet these two databases yielded similar results. The implications of such findings are widespread: Use of CBM to make reading placement decisions may be highly functional. With such a strong relationship, CBM normative data can be used by regular classroom teachers to establish groups at the beginning of the school year with relative ease and speed. These same measures are flexible and minimally intrusive; thus, they can help teachers make instructional placements for students coming into the school throughout the year. Most importantly, CBM in reading is easy to use and cost-effective: Teachers, instructional assistants, or volunteers can administer and score the reading passages in 3 to 4 minutes. Individual students' results can be easily plotted within a classwide distribution to help make decisions about placement into the appropriate curriculum and/or instructional group. For example, it took two 1-hour periods over 2 days to develop the entire normative sample for this study, with the data analyzed and plotted within 1 week.

Finally, although this study focused only on the use of CBM for making reading grouping placement decisions, such measures also can be used to make a host of other decisions, including re-evaluations of placement and measurement of student growth over time. In this way, teachers can have an empirical base to confirm their daily analysis of student performance. An interesting outcome was that students who had been identified as "difficult" in current reading groups were re-placed: Teachers searched the data to confirm their suspi-

cions about potentially misplaced children. In most cases, their intuition about student performance was very accurate. However, they also found that students for whom placement "seemed" inappropriate were indeed reading at a very different rate from their peers in that group; most of these students were re-placed into *different groups*.

In summary, oral reading fluency, although investigated in the context of special education decision-making, may well be applied in a general education context for grouping students into skill levels that are most appropriate.

REFERENCES

Coleman, M., & Harmer, W., (1982). A comparison of standardized reading tests and informal placement procedures. *Journal of Learning Disabilities*, 9, 396-398.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52,(3), 219-231.

Deno, S. L., Fuchs, L., & Fuchs, D. (1982). Reliability and validity of curriculum-based informal reading inventories. *Reading Research Quarterly*, 18, 16-26

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

Deno, S., Marston, D., Shinn, M., & Tindal,G. (1983). Oral reading fluency: A simple datum for scaling disability. *Topics in Learning and Learning Disabilities*, 2(4), 53-59.

Deno, S., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49(1), 36-45.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of curriculum-based informal reading inventories. *Reading Research Quarterly*, 18(1), 6-26.

Haller, E. J., & Waterman, M. (1985). The criteria of reading group assignments. *The Reading Teacher*, 38, 772-781.

Harris, L. A., & Lalik, R. M., (1987). Teacher's use of informal reading inventories: An example of school constraints. *The Reading Teacher*, 40(7), 624-630.

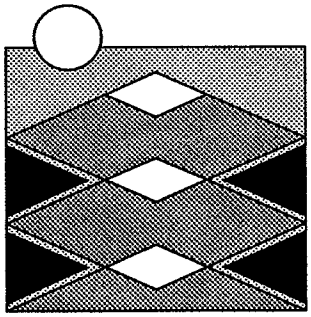
Johns, J., Garton, S., Schoenfelder, P., & Skriba, P. (1977). *Assessing reading behavior: Informal reading inventories—an annotated bibliography*. Newark, DE: International Reading Association.

Marston, D. (1982). *The technical adequacy of direct, repeated measurement of academic skills in low-achieving elementary students*. Unpublished doctoral dissertation, University of Minnesota.

Pikulski, J., & Shanahan, T. (Eds.). (1982). *Approaches to the informal evaluation of reading*. Newark, DE: International Reading Association.

Shinn, M. (Ed.). (1988). *Curriculum-based measurement*. New York: Guilford.

The Oregon Conference



1990

RReACS: A Model of Responsibly Reintegrating Academically Competent Special Education Students Into General Education

Mark R. Shinn
Kathleen Rodden-Nord
School Psychology Program
University of Oregon

Introduction

In a recent presentation (August, 1989) before a national special education services task force, Dr. Judith Schrag, Assistant Secretary of the United States Office of Special Education and Rehabilitative Services, discussed issues and priorities of the Office of Special Education Programs. Among her priorities, she cited the following:

1. The need for assistance for school systems to succeed with the Education for Handicapped Children Act (EHA, PL 94-142, 1975) mandates such as least restrictive environment (LRE). She cited evidence from the 10th Annual Report to Congress on the fact that many children, even with mild disabilities, are not integrated after years of 94-142.
2. The opportunity provided to special education to impact reform of general education.
3. The question of which children can be best integrated (mainstreamed), and when.

"Facts" about Special Education as Structured Currently

Special Education Continues to Grow

According to the Tenth Annual Report to Congress on the implementation of the Education of the Handicapped Act, over 4 million children now receive some form of special education. This number represents almost 7% of the total public school population. The greatest increase has been with respect to the number of students identified as learning disabled (LD) (Reynolds, 1984; Ysseldyke, Reynolds, & Weinberg, 1984). LD students now are the largest categorical group receiving special education in this country, accounting for 43.6% of all students served in special education in 1986-87 (U.S. Department of Education, 1988).

Low Achievers Are Being Certified

The increase in mildly handicapped (MH) students has led some to claim that too many students are being placed in special education (Gerber & Semmel, 1984). General educators may be referring more and more nonhandicapped, low-performing students for special education (Research for Better Schools, 1986; 1988). Furthermore, they also may be attempting to reduce achievement variability in their classrooms by referring the lowest-achieving students for special education (Gerber & Semmel, 1984). As it has been estimated that between 75-90% of referred children are found eligible, referrals to special education nearly guarantees placement (Algozzine, Christenson, & Ysseldyke, 1982).

Once Placed, Few Are Exited

Another explanation for the continued growth of special education is that handicapped students rarely are exited from special education back into general education on a full-time basis (Anderson-Inman, 1987). Many students enter the special education system, but few leave it. Although little data are available, it is estimated that as few as 2%

Note: The authors thank Susan Green, Ph.D., Victor Nolet, Todd Robicheaux, and W. David Tilly for their contributions to portions of the RReACS model and to Nancy Knutson for substantive contribution to the model's initial conception.

(Council for Great City Schools, 1986) to 8% (Shinn, 1986) of special education students are exited yearly. Some educators believe that the LRE mandate of the EHA is not being actualized for most MH students and therefore, they could be served best by receiving all of their instruction in general education settings (Hallahan, Kauffman, Lloyd, & McKinney, 1988). Finally, many states and districts are operating in a "policy vacuum" and do not have standards nor procedures established to document and justify exit from special education and reintegration into general education (Greenberg, 1989).

Dissatisfaction with Special Education Services as Typically Provided

In addition to the ever-growing numbers of students being served in special education, there has been growing dissatisfaction with the traditional "pull-out" model of special education service delivery, where handicapped students are served in segregated self-contained settings, or are "pulled out" to receive instruction in a resource room. Two major concerns are that few qualitative differences distinguish general and special education instruction and the efficacy of special education has yet to be established.

What Is Delivered May Not Be Different

Few differences have been found in the quantity or quality of reading or math instruction provided to handicapped and nonhandicapped students in general and special education setting (Haynes & Jenkins,1986; Wesson & Deno, 1989; Ysseldyke, O'Sullivan, Thurlow, & Christenson, 1989). If instruction in special education is virtually indistinguishable from that provided in general education, then students are being expected to learn by the same techniques that failed for them before. Consequently, handicapped students might as well receive those instructional techniques in less restrictive, and potentially less stigmatizing, environments, such as general education classrooms (Will, 1986).

What Is Delivered May Not Be Effective

Research examining the efficacy of the "pull-out" model of special education service delivery has been inconclusive and contradictory (cf. Carlberg & Kavale, 1980; Leinhardt & Palley, 1982; Semmel, Gottlieb & Robinson, 1979; Tindal, 1985). If special education programs are no more effective than general education programs in meeting the needs of handicapped children, then these children should be educated in general education settings. Madeline Will, former Assistant Secretary of the United States Office of Special Education and Rehabilitative Services, has stated that "the so-called 'pull-out' approach to the educational difficulties of students with learning problems has failed to meet the educational needs of these students and has created barriers, however unwittingly, to their successful education" (Will, 1986, p. 412).

Implications for Special Education: Current and Historical Practices

As a result of the increased dissatisfaction with the continued growth of special education and the lack of clear, demonstrable outcomes, some reformists have recommended that general educators assume greater responsibility for the education of all children, including handicapped students. Among the proposals put forth to accomplish this objective is the Regular Education Initiative (REI) (Reynolds, Wang, & Walberg, 1987; Stainback & Stainback, 1984; Will, 1986). Among the REI recommendations are: (a) reduction of emphasis on traditional "pull-out" programs for serving handicapped students, (b) an increased emphasis on adapting general education classrooms to meet the needs of all students, and (c) modification of special education regulations and funding laws to encourage and support innovative classroom-based programs. It has been suggested that the general education system needs to become more inclusive and willing to meet the needs of all learners and that the objectives of the REI could be met best by merging special and general education into a single system. REI advocates argue that such a partnership would encourage sharing of expertise and knowledge that ultimately would enable general education to educate virtually all handicapped children. Not only would the knowledge and expertise of special educators be shared with general education, it is likely that many special education resources would be shared as well (Fuchs & Fuchs, 1988).

The Problem: Who Should be Integrated into General Education? How and with What Effects?

Embedded within the REI is the implication that students currently labeled as handicapped (or those who would be labeled so in the future) would be (re)integrated into general education classrooms for instructional purposes. This implication, correct or otherwise, has generated a tremendous backlash (Kaufmann, 1989). Jenkins,

Pious, and Jewell (in press) caution against over reaction to what is, to date, only a broad policy statement. At a national level, specific procedures identifying which special education students would/should be (re)integrated, on what basis, and with what achievement or social/emotional outcomes, have not been identified.

Some special educators are urging a cautious or responsible approach (Fuchs & Fuchs, 1988). These authors, for example, argue that widespread reintegration of handicapped students should not become reality until there is an adequate data base, as well as clearly specified procedures, to support such efforts. They urge teachers, researchers, parents, and policy makers to demand further empirical studies of the feasibility of full-time mainstreaming programs, and more substantial documentation that such an approach would result in improvements in the education of handicapped students. Wang and Walberg (1988), vocal REI supporters, also maintain that reform should be based upon careful descriptive studies so that a research base on the effects of innovative programs can be established. Clearly, decisions about the structure of general and special education should be based on reliable data and careful logical analyses (Kauffman, 1989).

The Responsible Reintegration of Academically Competent Students (RReACS) Model

A model has been developed (see Figure 1) to represent the factors that must be considered and evaluated if potentially academically competent handicapped students are to be reintegrated responsibly into general education classrooms. The Responsible Reintegration of Academically Competent Students (RReACS) model incorporates known environmental factors in the conception of mildl handicaps such as: (a) academic expectations for success in specific instructional ecologies (Deno, 1989), and (b) teacher attitudes about handicapped students (Hersh & Walker, 1983; Walker & Rankin, 1983), as well as (c) technological advances in measuring student level and trend of improvement in general education curriculum (Curriculum-Based Measurement). New technology exists that may facilitate reintegration of academically competent handicapped students using a data base for decision making about who, when, and how to to reintegrate.

Environmental Variables

For responsible reintegration to occur, a number of elements within the general education (GE) environment must be identified clearly. The RReACS model suggests that three types of environmental data are crucial to responsible reintegration: (a) precise determination of minimal academic performance requirements in GE classrooms in terms of level and rate of progress, (b) teacher willingness to accept MH students in their classrooms (and knowledge of factors that can enhance their willingness), and (c) resources available to facilitate reintegration.

Defining "Level" of Minimal Academic Competency

A substantial number of handicapped students have academic skill levels comparable to their non-handicapped peers; few data are available, however, to document this assertion. This untested assumption would predict that handicapped students with sufficient skills would perform at a level equal to or greater than other GE students in the GE curriculum. The RReACs model first attempts to document the minimal academic skill levels required for success in the GE curriculum. To accomplish this, we must: (a) determine whether enough special education (SE) students with sufficient academic skills exist to warrant widespread reform, and (b) establish data-based, decision-making procedures so that educators can identify which of their SH students might be reintegrated, It is assumed that minimum academic skills would be determined on a local classroom-by-classroom basis by examining the range of students' academic skills in the lowest level of the general education curriculum. Therefore, an emphasis is placed on local norms, given that academic skill levels vary widely within and across localities (Shinn, 1986). Expectations for performance of SE students would be tailored to the setting in which they will be mainstreamed. Such efforts would determine the specific requirements for academic competence in the GE classroom into which specific SE students are reintegrated. At a minimum, these students should possess skills and academic behaviors that are within the range of skills possessed by the GE students in that classroom. It is, then, crucial to identify the level of academic skills in the curriculum that must be attained to be considered successful.

Defining Rate of Minimal Academic Competency

In addition to identifying the level of academic skill required for success in the general education curriculum, it is crucial to measure minimal rates of student growth or progress required for success in the GE setting. It is important that SE students maintain rates of progress equal to, or greater than,

other students in GE classrooms as they become reintegrated.

Teacher Willingness to Participate in Reintegration

Many authors argue that for successful reintegration of handicapped students to occur, general education teachers must be willing to instruct such students in their classrooms (Hannah & Pliner, 1983; Walker & Rankin, 1983). Identifying the factors that influence teachers' expectations for and willingness to work with handicapped students is, therefore, an important step in the reintegration process. General education teacher resistance must be considered and tempered, if necessary, to assure success of (re)integrated students.

To date, some limited studies have docu-

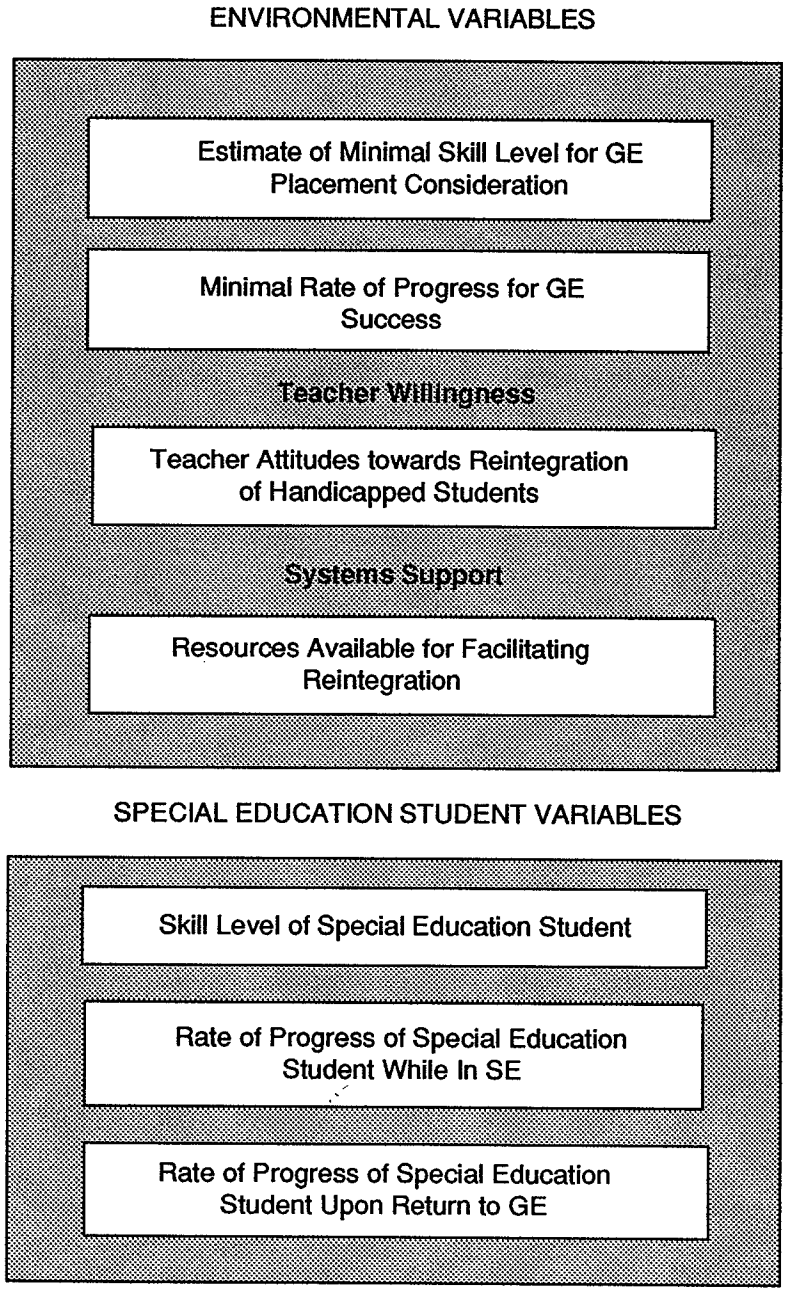


Figure 1. The Conceptual Model of Responsible Reintegration of Academically Competent Students (RReACS)

mented that general educators say they are generally unwillingly to teach children who do not conform to their academic or behavioral expectations. For example, nearly 80% of the general education teacher respondents in a survey conducted by Knoff (1985) indicated an unwillingness to accept handicapped students into their classrooms. Further, when general education teachers are willing to integrate special education students into their classrooms, they express concern that reintegration may reduce their instructional effectiveness and negatively impact the education of other students in their classes (Hudson, Graham, & Warner, 1979).

The conclusions from these negative, albeit limited, studies must be tempered by the contexts (or lack thereof) in which teachers were surveyed. Within the RReACS model, it is hypothesized that a more accurate picture of teachers' attitudes about reintegration will emerge as contextual variables are taken into account (Kauffman, Lloyd, & McGee, 1989; Ritter, 1989). Specifically, it is hypothesized that teacher attitudes are a function not only of their prior beliefs, but also a function of information about the skill competencies of specific students and additional resources that could be allocated. Thus:

Teacher Willingness = Prior Biases + Information about the Curriculum Performance of General and Special Education Students + Available Resources

Unfortunately, previous studies have examined teacher attitudes by presenting hypothetical situations, rather than actual information, on the achievement of students with whom the teachers may be familiar. Although teachers' unwillingness to mainstream might be attributable to their inherently negative view of handicapped students, it may be more likely that teachers simply view these students as unprepared for mainstreaming because they believe these students lack critical competencies and pertinent skills (Fuchs, Fuchs, Fernstrom, & Hohn, 1989). In the RReACS model, decision making emphasizes documented performance data of students in both general education and in special education, with respect to level and rate of improvement.

Systems Support

System support is a critical element to the success of any educational innovation (Sarason, 1982). In the RReACS model, systems support is a crucial component as well. In examining teachers' willingness to educate handicapped students in GE classrooms, Kauffman et al. (1989) found a significantly high number of teachers would consider taking responsibility for students who lack critical skills or exhibit unacceptable behaviors if they were also provided technical assistance. Thus, it is important to identify the resources a teacher requires to accept and instruct a handicapped student. Furthermore, it is possible that general educators may be unwilling to accept responsibility for handicapped students because the educators themselves believe that they do not possess the specialized instruction skills and adequate resources, to meet the needs of these students. Therefore, additional insight into teacher participation might be obtained by examining the modifications that may influence willingness to mainstream.

The RReACS model emphasizes the need to identify the requirements each mainstream teacher may have for reintegrating handicapped student. Given the unique interactions of student needs, teacher skills, and available resources, reintegration plans must be made on a student-by-student basis. Fuchs et al. (in press) agree that tailoring systems resources on a case-by-case basis rather than using a general approach to all reintegration cases is most effective. A consultation model (Curtis & Meyers, 1988), then, is the most effective way to deliver a variety of support services during the responsible reintegration process.

Special Education Student Variables

Within the RReACS model, it is crucial that students considered as candidates for reintegration into general education classrooms perform academically at minimal level that is comparable to other students in general education classrooms. Thus, it is important to not only assess the achievement expectations in terms of level and progress in the GE environment, but also to document the curricular skill levels of special education students within the general education curriculum.

The curricular skills would be assessed in the following, progressively more selective sequence. First, special education students considered for reintegration would have to perform at a level of success equal to, or greater than, other students in the lowest available level of the curriculum in the general education classroom. Second, the special education students would have to maintain rates of progress in the general education curriculum equal to, or greater than, other students in the lowest available level of the curriculum in the general education classroom while the special education student still receives special education instruction. This requirement would ensure that the students(s) learning rate is sufficient for adequate progress,

even though those exact conditions may not be replicable in the general education classroom. Third, the special education student(s) would have to maintain rates of progress in the general education curriculum equal to, or greater than, other students in the lowest available level of the curriculum in the general education classroom while receiving instruction in the general education classroom. Such instruction may or may not be supported by additional resources, based on an analysis of perceived teacher needs and the student's actual rates of progress in general education.

Operationalizing the RReACS Model

The key to creating researchable procedures for reintegration is an appropriate operationalization of each variable in the model. In Figure 2, the operationalization of the model is depicted. The variables represented in Figure 1 are displayed as the boxes on the left of Figure 2. Arrows are drawn to represent sequences of activities and relations how the variables would be measured in the research program.

Environmental Variables

Defining minimal level of academic competency and rate of progress. The RReACs model in this example is operationalized with special education students with severe reading problems. Identifying the minimal expectation for level of academic skills is operationalized by testing general education students in the lowest level of the general education reading curriculum in each grade. The scores of these students represent the teachers' expectations for minimal level of success in general education. These general education students would be assessed routinely and frequently (twice per week) to establish estimates of their rates of improvement. This standard serves as the yardstick against which reintegrated special education students must be monitored to ensure their growth is adequate.

Recent advances in educational assessment technology provide a method for operationalizing the expectations for success in the general education curriculum in terms of level and rate of improvement. Curriculum-based measurement (CBM), a technique for assessing student performance over time within a specific curriculum, is optimal for operationalizing these variables. CBM focuses on direct measurement of students' performance in their curriculum. This process involves observation of student basic skills activities (e.g., reading aloud from text) or products (e.g., worksheets or writing samples). When CBM is used, educational decisions are made with respect to what educators know best, their own students and their own curriculum. Thus, representativeness of testing content and norms, a problem inherent in most traditional achievement testing, is maximized.

Another unique premise of CBM is that specific assessment decisions are individually referenced (comparing a student to him/herself over time). When making decisions about student progress, for example, slope of academic improvement in the curriculum is used. This feature is particularly useful in the RReACS model because of its emphasis on reintegrating students on a case-by-case basis. In addition, CBM uses peer-referenced performance for decision-making. That is, the academic performance of any particular student is indexed against local normative performance in the curriculum. Such procedures are tailored to the information needed in the RReACS model for determining minimal academic competency in a given classroom and for selecting which SE students potentially qualify for reintegration.

Because minimum skill levels vary across localities, skill-level data must be collected in specific settings to determine which students have these competencies and may be eligible for general education placement. CBM measures are more sensitive than standardized achievement tests for determining student skill level (Marston, 1989). Additionally, it is crucial to have a technology for measuring student growth and providing feedback throughout the school year rather than just at the beginning or end. In addition to determining entry level skills when assessing academic achievement, it is also important to monitor change as it occurs over time (Deno, 1986). CBM measures provide a sensitive, reliable index of student gain over relatively short periods of time, unlike traditional standardized achievement tests (Marston, Fuchs, & Deno, 1986). Finally, CBM procedures are designed for frequent administration, so examiners can get a more reliable picture of pupil progress across time periods of varying length. Such accuracy is critical for developing appropriate estimates of the minimal rates of progress needed for success in general education.

Teacher Willingness

Teacher willingness is measured using the Teacher Attitudes Toward Reintegration questionnaire (TATR) (Shinn, Good, & Green, in preparation). A series of scales are being designed to identify attitudes about educating special education students who have academic skills at the level

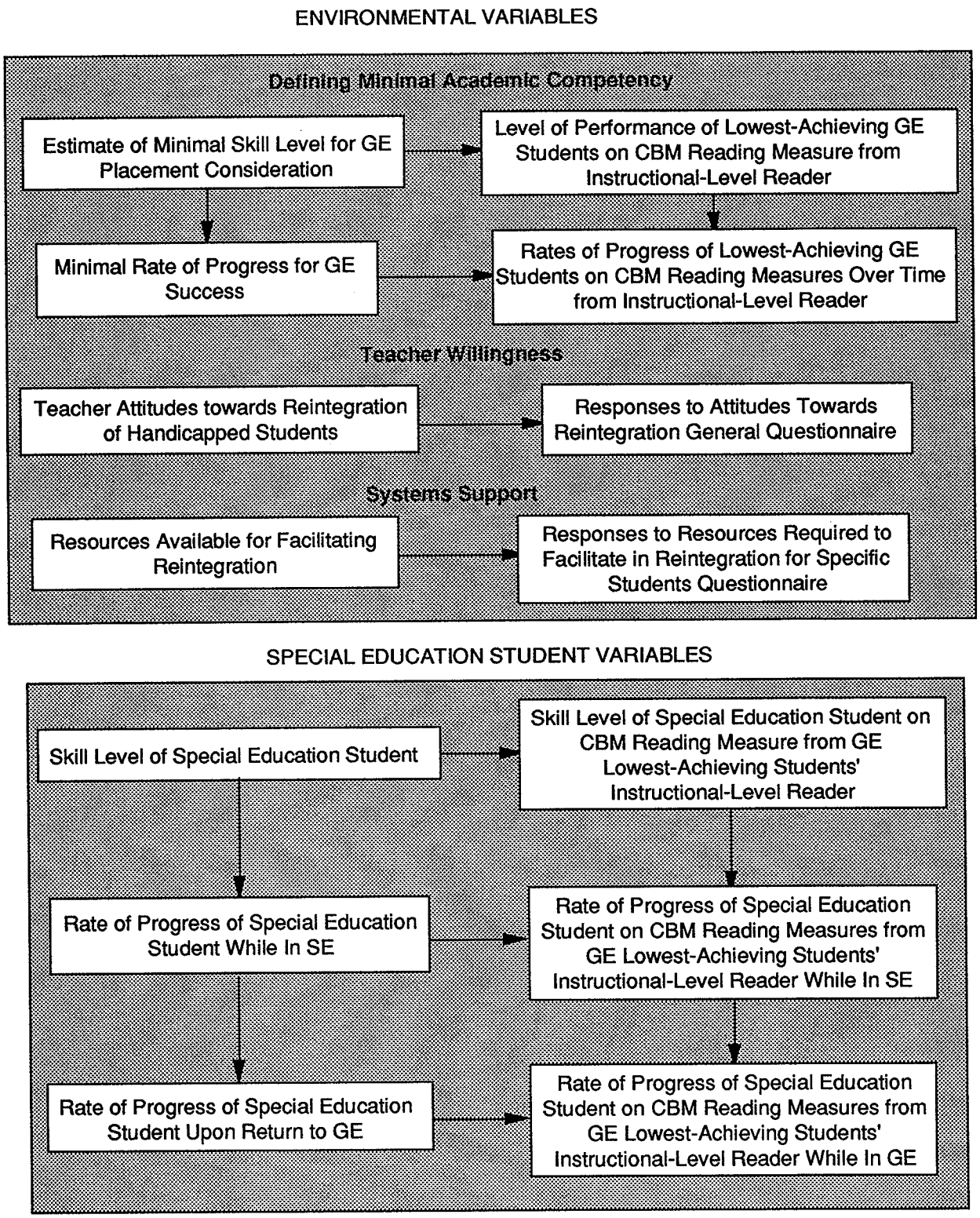


Figure 2. The Operational Model of Responsibly Integrating Mildly Handicapped Students into General Education Environments

of students in the lowest level of the general education curriculum. The scales are designed to assess general willingness towards the concept of reintegration and to identify variables (e.g., student performance information, available support resources) that may reduce unwillingness. Because the issues may be different for general and special education teachers, separate forms are being developed for each target group.

The TATR's initial development has been based on the work of Walker and the research he has conducted on teacher willingness to assume responsibility of students with severe problem behaviors (Hersh & Walker, 1983; Walker & Rankin, 1983). Walker developed and tested the Social Behavior Survival Scale (SBS) to assesses teacher attitudes about their willingness to assume responsibility for educating handicapped students and match student behaviors to teacher tolerances for those behaviors. The SBS identifies a list of competencies that a student must possess to facilitate integration success, and requires teachers to identify necessary resources that must be provided for successful reintegration.

Systems Support

Knowledge of resources perceived necessary for success in general education are operationalized in the RReACS model by having all participating parties (general education teachers, special education teachers, parents) complete an identical section of TATR. This section of the TATR is based on the knowledge base of effective teaching strategies and teachers' perceived needs with students with instructional problems.

Special Education Student Variables

CBM measures representing the lowest level of the reading curriculum in general education classrooms are administered in three stages to special education students. First, special education students who have an IEP objective in reading are tested in the lowest level of the reading curriculum in general education. Second, those special education students who perform in the range of reading skills of general education students are tested twice per week in the lowest level of the general education curriculum. Such measurement can then be used to determine if their rates of progress in special education are equal to, or greater than, general education students. Third, these students are considered for reintegration by multi-disciplinary teams. Reintegrated students' progress is then evaluated frequently using their basal reader in the general education classroom. CBM monitoring of student progress on a regular basis allows ineffective instructional programs to be modified and effective programs to be documented and maintained. Students who failed to maintain their rates of progress in general education would have additional resources provided, based on an analysis of the TATR.

Summary

Despite a lot of rhetoric advocating general education instruction for many handicapped students, few if any data are available on the magnitude of special education students who could be instructed successfully in general education classrooms. Not only are there no data on the numbers of students, there is no underlying model nor set of decision-making procedures to accomplish this activity. We have argued that identifying students who "may" be considered academically competent for general education instruction should be based on: (a) their level of performance, (b) their rate of progress on general education curriculum while receiving special education instruction, (c) their rate of progress upon reintegration, and (d) resources brought to bear in general education that may increase the success of reintegration. Further, information should be collected about teacher and parent attitudes about reintegration and what factors may mediate those attitudes. For successful reintegration to occur, positive proactive procedures should be identified and field tested on a small scale. The RReACS model allows for a data-based, testable set of procedures to accomplish this activity.

References

Algozzine, B., Christenson, S., & Ysseldyke, J. (1982). Probabilities associated with the referral to placement process. *Teacher Education and Special Education, 5*, 19-23

Anderson-Inman, L. (1987). Consistency of performance across classrooms: Instructional materials versus setting as influencing variables. *Journal of Special Education, 21*, 9-29.

Carlberg, C., & Kavale, K. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. The *Journal of Special Education., 14*(3), 295-308.

Curtis, M., & Meyers, J. (1988). Consultation: A foundation for alternative services in the schools. In J. Graden, j. Zins, & M. Curtis (Eds.), *Alternative educational delivery systems: Enhancing instructional options for all students* (35-48). Washington, DC: National Association of School Psychologists.

Deno, S.L. (1986). Formative evaluation of individual programs: A new role for school psychologists. School Psychology Review, 15, 358-374.

Deno, S. L. (1989). Curriculum-based measurement and alternative special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1-17). NY: Guilford.

Fuchs, D., & Fuchs, L. S. (1988). Evaluation of the Adaptive Learning Environments Model. *Exceptional Children, 55*(2), 115-127.

Fuchs, D., Fuchs, L.S., Fernstrom, P., & Hohn, M. (1989). *Achieving responsible reintegration of behavirorally disordered students.* Paper presented at the 1988 annual meeting of The Council for Exceptional Children, San Francisco.

Gerber, M., & Semmel, M. (1984). Teachers as imperfect test: Reconceptualizing the referral process. *Educational Psychologist, 19,* 137-148.

Graham, S., Hudson, T., Burdy, N.B., & Carpenter, C. (1980). Educational personnel's perceptions of mainstreaming and resource room effectiveness. *Psychology in the Schools, 17*(1), 128-134.

Greenberg, D. (1989). The Tenth Annual Report to Congress: One more ride on the merry-go-round. *Exceptional Children, 56,* (1), 10-13.

Hallahan, D. P., Keller, C. E., McKinney, J. D., Lloyd, J. W., & Bryan, T. (1988). Examining the research base of the regular education initiative. *Journal of Learning Disabilities, 21,* 29-35.

Hannah, M., & Pliner, S. (1983). Teacher attitudes toward handicapped children: A review and synthesis. *School Psychology Review, 12,* 12-25.

Hersh, R., & Walker, H. (1983). Great expectations: Making schools effective for all students. *Policy Studies Review, 2*(1),147-188.

Jenkins, J.R., Pious, C.G., & Jewell, M. (in press). Special education and the regular education initiative: Basic assumptions. *Exceptional Children.*

Kauffman, J. M. (1989). The Regular Education Initiative as Reagan-Bush educational policy: A trickle-down theory of education of the hard-to-teach. *Journal of Special Education, 23,* 256-277.

Kauffman, J. M., Lloyd, J. W., & McGee, K. A. (1989). Adaptive and maladaptive behavior: Teachers' attitudes and their technical assistance needs. *The Journal of Special Education, 23*(2), 185-200.

Knoff, H. M. (1985). Attitudes toward mainstreaming: A status report and comparison of regular and special educators in New York and Massachusetts. *Psychology in the Schools, 22,* 411-418.

Leinhardt, G., & Palley, A. (1982). Restrictive educational settings: Exile or haven? *Review of Educational Research, 52*(4), 557-578.

Marston, D., Fuchs, L. S., & Deno, S. L. (1986). A comparison of standardized achievement tests and direct measurement techniques in measuring student progress. *Diagnostique, 11,* 77-90.

Reynolds, M. C., (1984). Classification of students with handicaps. In E. W. Gordon (Ed.), *Review of Research in Education: Volume 11* (pp. 63-92). Washington, DC: American Education Research Association.

Reynolds, M.C., Wang, M.C. & Walberg, H.J. (1987) The necessary restructuring of special and regular education. *Exceptional Children 53* .

Ritter, D. (1989). Teachers' perceptions of problem behavior in general and special education. *Exceptional Children, 55,* 559-564.

Sarason, S. (1982). *The culture of the school and the problem of change* (2nd edition). Boston: Allyn and Bacon.

Shinn, M.R. (1986). Does anyone care what happens after the referral-test-placement process: The systematic evaluation of special education effectiveness. *School Psychology Review, 15,* 49-58.

Shinn, M.R.(1989) (Ed.), *Curriculum-based measurement: Assessing special children.* NY: Guilford Press.

Shinn, M.R., Good, R., & Green, S. (1989). *Teachers' attitudes towards reintegration scale.* Eugene, Oregon: University of Oregon

Stainback W., & Stainback, S. (1984). A rationale for the merger of special and regular education. *Exceptional Children, 51*(2) 102-111.

Tindal, G. (1985). Investigating the effectiveness of special education: An analysis of methodology. *Journal of Learning Disabilities, 18,* 101-112.

U. S. Department of Education (USDE), Office of Special Education and Rehabilitative Services. (1984). *Sixth annual report to Congress on the implementation of the Education for All Handicapped Children.* Washington, DC: Author.

Walker, H.M. & Rankin, R. (1983). Assessing the behavioral expectations and demands of less restrictive settings. *School Psychology Review 12* (3), pp 274-284.

Wang, M. C., & Walberg, H. J. (1988). Four fallacies of segregationism. *Exceptional Children, 55*(2), 128-132.

Wesson, C. L., & Deno, S. L. (1989). An analysis of long-term instructional plans in reading for elementary resource room students. *Remedial and Special Education, 10*(1), 21-34.

Will, M. (1986). *Educating students with learning problems: A shared responsibility.* U.S. Department of Education, Washington, D.C.

Ysseldyke, J. E., O'Sullivan, P. J., Thurlow, M. L., & Christenson, S. L. (1989). Qualitative differences in reading and math instruction received by handicapped students. *Remedial and Special Education, 10*(1), 14-20.

Ysseldyke, J. E., Reynolds, M. C., & Weinberg, R. A. (1984). *School psychology: A blueprint for training and practice.* Minneapolis, MN: National School Psychology Inservice Training Network, University of Minnesota.

Construct Validity in Published Achievement Tests

Victor Nolet
University of Oregon
Gerald Tindal
University of Oregon

Introduction

Scores from published achievement tests provide the basis for a variety of educational decisions, including student screening or eligibility determination and evaluation of Chapter 1 or classroom instructional programs. Recently, the technical adequacy of these tests for making educational decisions has been criticized (Ebel, 1978; Salvia & Ysseldyke, 1988). Six studies provide findings that exemplify this line of research:.

1. Jenkins and Pany (1978) analyzed five reading curricula along with reading subtests from five published achievement tests and found that student performance in a curriculum is dependent upon which specific achievement test was used. This study recently was updated by Shapiro & Derr (1987) and the major findings were replicated.

2. Leinhardt (1981) presented data from two projects that showed: (a) the amount of overlap between tests and either curriculum or instruction often is minimal, and (b) overlap is a powerful predictor of end-of-year test performance.

3. Good and Salvia (1988) conducted an investigation of content validity using actual student performance on different reading achievement tests after having been taught a specific curriculum. They found significant differences in test performance which could be predicted by differences in content validity (overlap with the curriculum).

4. Freeman, Kuhs, Porter, Floden, Schmidt, and Schwille, (1983) found the amount of overlap between five tests and four mathematics curricula was negligible. Their findings indicate that curricula present most of the test topics, but most tests do not assess the curriculum topics presented.

5. Knifong (1980) compared the computational requirements of word story problems and found great differences across five different published achievement tests.

6. Petrosko (1978) conducted an analysis of topics covered and strengths and weaknesses of standardized mathematics tests used in high schools. Four different topics were analyzed on 39 criteria of test quality. Many of the tests provided insufficient or unsupportive descriptions of item selection procedures, reliability, validity, and interpretive techniques.

Research pertaining to the technical adequacy of published norm-referenced achievement tests has focused almost exclusively on content validity, the manner and extent to which items on a test sample a domain of interest. As the studies listed above illustrate, published achievement tests lack content validity. Such a finding is not surprising because these tests do not sample a domain operationally defined by a particular curriculum. At best, published achievement tests can only hope to sample from the portion of a curriculum represented by a textbook; as the research summarized has shown, they perform this task poorly.

Recent research in this area has taken on a curiously lopsided focus. While we've been attending closely to evidence for content validity of published achievement tests, the more fundamental issue of construct validity of these instruments largely has been ignored. According to Cronbach & Meehl (1955), "Construct validity is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined" (p 282). Construct validation refers to development of a body of evidence that supports specific inferences about the meaning of scores obtained from a particular test.

Two studies have considered this aspect of validity, but their analyses were limited to inspection of the extent to which test publishers report data to support the construct validity of their tests. Petrosko (1978) found only 2% of the 322 general mathematics achievement tests reported any divergent validity data while none of the 26

applied mathematics, 122 algebra, or 52 geometry tests reported this information. Hall (1985) found only 4% of 37 tests examined reported construct validity information. To date, no one has examined the degree to which data reported by test publishers actually support the construct validity of their tests. In summary, the focus of recent research has missed the issue of construct validity in two ways: (a) by attending to content validity only, and (b) by examining how many test publishers report relevant data rather than analyzing the meaning of those data.

To appreciate the meaning of data reported by test publishers, a complete understanding of the nature of construct validity is required. A construct is a complex entity that is synthesized (or "constructed") from a number of subordinate elements. "Constructs" equate roughly with "concepts" which refer to classes of objects, events, or features. A key but subtle difference between concepts and constructs is that concepts tend to refer to classes whose defining attributes are context free, while constructs tend to derive meaning from the specific context in which they are used. For example, the concept "bicycle" refers to class of pedal powered vehicles that have two wheels. A relatively indisputable test for the presence of "bicycleness" (i.e., two wheeled, pedal powered vehicle) can be applied in a wide range of contexts such as touring bike, mountain bike, or tandem bicycle. On the other hand, the construct "ability" must be redefined anew with each use. Any test for the presence of "ability" requires explicit description of the context involved. Clearly "ability" in playing music is not the same as "ability" in playing baseball or "ability" in repeating the months of the year.

The contextual dependence of constructs has major implications in development of construct validity. Specific inferences about the meaning of scores obtained from a test are more or less valid, depending on the meaning ascribed to the construct the test is intended to measure. In this respect, validity refers to the inferences that can be made on the basis of a test, not the test itself. Construct validity, then is the shared responsibility of the test designer and the test user. The test designer is responsible for defining the construct and providing evidence that the test adequately measures it, and the test user must decide how to interpret scores obtained from the test on the basis of evidence provided by the designer.

Construct validity requires an array of multiple pieces and forms of information that, when combined, allow specific interpretations of scores obtained from a particular test. As Messick (1989) points out, validity is a matter of degree. No single piece of evidence can prove or disprove the validity of inferences made on the basis of a test. Rather, multiple pieces of evidence are evaluated in terms of the extent to which a conclusion is supported or weakened. By convention, evidence to support construct validity falls into two categories: convergent validity and discriminant validity. Convergent validity involves multiple forms of evidence that show a test adequately measures the construct of interest. Evidence to support the discriminant validity of a test shows that variables related to constructs other than the one the test is intended to measure do not influence scores on the test. In this respect, construct validation is a process of showing what a test does measure, as well as what it does not measure.

For example, assume scores obtained on a particular achievement test are the basis for inferences about the amount of mathematics learned by a group of students. High scores are interpreted to mean a test taker has learned a lot of math while low scores support the inference that not much math learning has occurred. If the test is an adequate measure of the construct "mathematics achievement," i.e., has convergent validity, students who obtain high scores on test would be expected have high scores on other measures of the construct "mathematics achievement," such as quiz grades, homework grades, or scores obtained on other math tests. Additionally, high scoring students also might be expected to perform well on such non-academic math related measures as computer programming or playing chess. On the other hand, because mathematics constitutes a different instructional domain than penmanship, students who have high scores on penmanship measures (such as teacher judgment or scores on penmanship tests) would not necessarily be expected to have high scores on this math test. If the math test measures a construct that is distinctly different from the construct "penmanship," no relationship between math and penmanship scores should be observed.

Obviously, some individuals could have good penmanship and high math scores, so the intercorrelation of math and penmanship scores from numerous students must be examined. The convergent validity of the math test would be suspect if scores obtained on this test fail to correlate positively with other measures of the construct "mathematics achievement." Similarly, the dis-

criminant validity of the test should be questioned, if scores obtained from the math test are correlated positively with penmanship measures. In the latter case, the test might be sampling a more inclusive construct, such as "tool movement," that could account for both math and penmanship achievement.

Published achievement test batteries are constructed on the premise that separate subtests measure achievement in domains that represent distinctly different constructs. For example, tests of reading achievement are considered tests of the construct "reading " while mathematics subtests are thought to sample the distinctly different construct "mathematics achievement." A further premise of published achievement tests is that few constructs of interest in school learning are unitary. For example, the construct "reading," often is characterized as involving at least three facets: reading comprehension, reading fluency, and vocabulary (Farr & Carey, 1986). Similarly, written language might involve variables such as syntax, semantics, and spelling.

In achievement test batteries, multiple tests often are used to measure achievement in various facets of multi-dimensional constructs. Achievement test batteries typically include tests of reading comprehension and word attack to test the construct "reading"; math computation and math problem solving to measure the construct "mathematics achievement," and language usage and language mechanics to measure the construct "written expression."

The construct validity of a published achievement test battery would be further supported by evidence that subtests included in the battery have both convergent and discriminant validity. The convergent validity of a single subtest in a battery would be supported by evidence that scores from the subtest correlate with other subtests measuring the same construct. For example, if a reading comprehension and word attack subtest each sample the construct "reading," we would expect a moderate correlation between scores obtained on the tests. At the same time, discriminant validity is supported by the fact that results from a particular subtest in a battery are not correlated with subtests intended to measure a different construct. A word attack subtest should not result in scores that are correlated with scores obtained from a math computation subtest.

In summary, the construct validity of an achievement test battery is supported by a pattern in which different subtests that are related, (i.e., intended to measure facets of the same construct) are more highly correlated with one another than they are with subtests to which they are unrelated, (i.e., intended to measure distinctly different constructs). The current study was conducted to investigate the existence of such a pattern of

Table 1. Tests Reviewed

Test Name	Year	Publisher
Group Administered Tests		
California Achievement Tests, Form E (CAT)	1985	CTB/McGraw-Hill
Comprehensive Test of Basic Skills, Form U (CTBS)	1982	CTB/McGraw-Hill
Circus/STEPIII 1979		Addison-Wesley
Iowa Test of Basic Skills (ITBS)	1986	Riverside Publ. Co.
Test of Achievement & Proficiency (TAP)	1986	Riverside Publ. Co.
Metropolitan Achievement Tests , Form 6 (MAT6)	1988	The Psychological Corporation.
SRA Survey of Basic Skills , Form H (SRA)	1985	Scientific Research Associates.
Stanford Achievement Test Series	1985	The Psychological Corporation.
Individually Administered Tests		
Diagnostic Achievement Battery (DAB)	1984	Pro-Ed
Woodcock-Johnson Psycho-Ed Battery (WJPEB)	1977	Teaching Resources
Peabody Individual Acheivement Test, Revised (PIAT)	1988	American Guidance Service

intercorrelations among subtests of published, norm-referenced achievement batteries currently used in schools.

METHOD

Technical manuals for eleven published norm-referenced achievement tests currently employed in educational settings were analyzed. Eight group and three individually administered achievement tests were reviewed (see Table 1). With only one exception, tests with national normative data collected within the last decade were considered. The latest version of the Woodcock-Johnson Psycho-educational Test Battery was not available at the time the study was conducted and the version examined is based on norms from 12 years ago.

This study was concerned with achievement test batteries intended to test multiple constructs, so tests aimed at a single construct such as oral reading, written language, or mathematics were not reviewed. Only tests and subtests focusing on three basic skill areas of reading, language arts, and mathematics were collected and analyzed. Content area subtests (such as science or social studies) were not evaluated as part of this study both because of our focus on basic skills and because of the inconsistent inclusion of such materials within many published tests. All levels of each test were examined. When multiple forms of a test were available, the current version that included levels across the most age or grade ranges was selected.

Data Collection and Analysis

For each battery, subtests were sorted into one of three categories: reading related, language arts related, and mathematics related. Subtests that were specified as components of a particular construct by the test publisher were considered related. For example, on California Achievement Test (CTB/McGraw Hill, 1985), scores from the Mathematics Computation and Mathematics Concepts and Application subtests contributed to a Total Math score, and these subtests were considered related. All subtests that were not identified by the publisher as a measure of a particular construct were considered unrelated. For example, on the CAT, any subtest that was neither Mathematics Computation nor Mathematics Concepts and Application was considered unrelated to either of the math subtests. With the exception of subtests that measured study skills, all subtests were clearly identified by publishers as belonging to one of the three constructs. When study skills subtests were not grouped by a publisher in a particular construct, they were evaluated as a component of reading because they are primarily related to the skill of reading (often including analyses of reading and reference materials), rather than a particular curriculum content area (Farr & Carey, 1986).

Correlation matrices showing the intercorrelation of all subtests within each battery were examined. All correlation coefficients reported for each subtest across all levels of the battery were analyzed, and for each subtest two distributions of correlation coefficients were developed—those associated with related subtests and those associated with unrelated subtests. Intercorrelations between total test scores and specific subtests were not used in this study because total test scores are comprised of subtest scores. Intercorrelation among these scores is unnecessarily redundant and results in artificially high coefficients.

Related and unrelated coefficient distributions for subtests were combined for each construct. This resulted in a final total of 6 distributions for each battery, (i.e., related and unrelated for reading, language arts, and math). The range and median were computed for each of these distributions. The median related coefficient was compared with the median unrelated coefficient and the amount of overlap between the two distributions was assessed. Visual comparison of medians and ranges was accomplished through use of Tukey's (1977) notched box plots. This method allowed estimation of the significance of difference between medians as well as comparison of overlap of related and unrelated distributions.

RESULTS

The results of the analyses are shown in Table 2. Tests are shown in the left column of the table with ranges and medians for related and unrelated subtests in the body. The percent of overlap of related and unrelated is shown in column on the far right. The TAP and PIAT included only one subtest each in language and mathematics so no intercorrelations among related subtests were available.

On all batteries, there was considerable overlap of the range of intercorrelations of related subtests with that of unrelated subtests. Correlations among related subtests ranged from .38 (DAB: Language) to .98 (W-JPEB Achievement: Language) while correlations among unrelated subtests ranged from .14 (Circus: Reading) to .85 (ITBS: Mathematics). For example, on the ITBS, mathematics related subtest intercorrelations ranged from .57 to .80 while unrelated subtest intercorrelations ranged

Table 2. Summary Data for 11 Test Batteries

	Related Range	Related Median	Unrelated Range	Unrelated Median	Difference Significant?	Percent Overlap
CAT						
Reading	.62 - .84	0.72	.45 - .82	0.68.	yes	98
Language	.64 - .80	0.69	.46 - .82	0.68	no	100
Mathematics	.62 - .83	0.72	.45 - .82	0.62	yes	91
Circus/STEP						
Reading	.74 - .85	0.79	.40 - .83	0.72	yes	94
Language			.14 - .83	0.71	na	na
Mathematics	.63 - .82	0.72	.40 - .78	0.68	yes	98
CTBS						
Reading	.64 - .82	0.77	.44 - .80	0.67	yes	99
Language	.54 - .78	0.67	.45 - .76	0.64	yes	99
Mathematics	.59 - .77	0.70	.43 - .77	0.61	yes	100
ITBS						
Reading	.47 - .83	0.71	.42 - .76	0.63	yes	88
Language	.45 - .76	0.68	.37 - .80	0.60	yes	100
Mathematics	.57 - .80	0.65	.36 - .85	0.60	yes	100
MAT 6						
Reading	.70 - .84	0.77	.42 - .79	0.63	yes	70
Language	.59 - .68	0.63	.44 - .80	0.65	no	100
Mathematics	.54 - .79	0.66	.42 - .79	0.58	yes	100
SRA						
Reading	.44 - .84	0.73	.34 - .84	0.65	yes	100
Language	.61 - .80	0.70	.36 - .83	0.68	yes	100
Mathematics	.54 - .83	0.74	.33 - .79	0.63	yes	68
Stanford						
Reading	.57 - .85	0.71	.42 - .83	0.64	yes	91
Language	.72. -.76	0.74	.43 - .83	0.66	yes	100
Mathematics	.66 - .81	0.72	.42 - .84	0.63	yes	100
TAP						
Reading	.79 - .82	0.80	.68 - .82	0.75	yes	100
Language			.70 - .78	0.73	na	na
Mathematics			.68 - .80	0.73	na	na
DAB						
Reading	.81 - .85	0.81	.36 - .83	0.69	yes	66
Language	.38 - .70	0.61	.33 - .80	0.60	no	100
Mathematics	.55 - .55	0.55	.27 - .76	0.66	na	100
PIAT						
Reading	.42 - .99	0.73	.43 - .90	0.70	no	93
Language			.30 - .90	0.66	na	na
Mathematics			.35 - .71	0.68	na	na
W-J PEB: Achmt.						
Reading	.52 - .80	0.67	.37 - .82	0.64	no	100
Language	.59 - .98	0.77	.41 - .80	0.63	yes	78
Mathematics	.53 - .74	0.66	.37 - .67	0.54	yes	80

from .36 to .85.

An hypothesis implied by the analysis conducted in this study is that intercorrelations among unrelated subtests tend to be lower than those among related subtests. As shown in Table 2, the hypothesis was supported. Intercorrelations among unrelated subtests were generally smaller than intercorrelations among related subtests. Unrelated medians ranged from 18% smaller (W-JPEB Achievement: Language) to 20% larger (DAB: Math) than related medians. Most differences were smaller than 15% (n=22), over one third of the differences were smaller than 10% (n=12) and on 9 batteries, a difference of 5% or less was observed. In only one battery (DAB: mathematics) were median unrelated coefficients more highly correlated than related coefficients. Ironically, only one related coefficient was reported for DAB: mathematics. This figure was lower than the median of unrelated subtests, resulting in a positive difference value. The statistical significance of differences was estimated by examining the 95% confidence interval represented by notches on the Tukey box plots. Overlapping notches on adjacent distributions implied no significant difference. The results of this procedure also are shown on Table 2. Figure 1 shows a set of box plots for the MAT 6. Correlation coefficients are shown on the vertical axis. The boxes show interquartile ranges with extreme scores plotted as circles. As can be observed in the figure, the related and unrelated medians for language arts are not significantly different. Also, on both language arts and mathematics, there is 100% overlap in scores.

Discussion

Published, norm referenced achievement batteries operate at three levels. At the subtest level, narrowly focused tests are intended to measure specific facets of a construct. At the test level, the combined results of subtests are intended to provide a measure of achievement for entire constructs and finally, at the battery level, performance across constructs is intended to provide an indication of general achievement.

Validity at the Subtest Level

The results obtained in this study indicate that at the level of specific subtests, publishers have not provided sufficient evidence to support the premise that distinctly different constructs are represented by individual subtests. We found that ranges of intercorrelations of related subtests overlapped with those of unrelated subtests and therefore we concluded that subtests purporting to measure distinctly different constructs were more

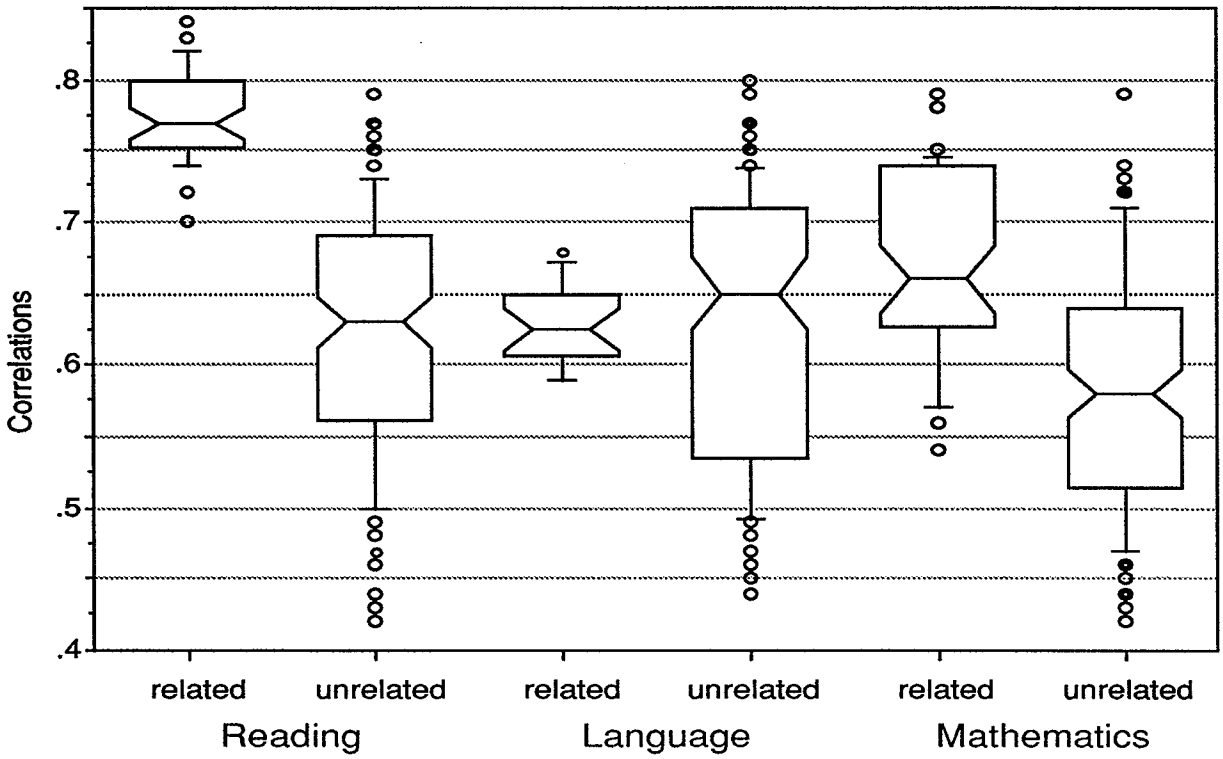


Figure 1. Sample Box Plots Showing Data for MAT 6

highly intercorrelated than subtests intended to measure facets of a single construct. This observation was consistent across all batteries in all three constructs, and often the overlap was substantial or complete. For example on the SRA, the range of intercorrelations among related language arts subtests was .61 to .80 while the range of intercorrelations of language arts subtests with subtests unrelated to language arts was .36 to .84. On the SRA, as well as numerous other batteries, subtests intended to measure facets of the construct "language arts" could as likely be measuring facets of the construct "mathematics."

Although no standard exists with which to evaluate the magnitude of differences between related and unrelated subtest intercorrelations, the meager differences obtained here do not offer compelling support for the premise that distinctly different constructs are represented by subtests. Differences of under 5% observed in almost all batteries provide further indication that many of the intercorrelations for related and unrelated subtests are of approximately the same magnitude. In sum, little evidence was found to support convergent or discriminant validity at the subtest level.

Validity at the Test Level

Issues of convergent and discriminant validity are extremely salient at the test level. For example, a "Total Language" score is based on two equally important premises. The first is that specific language arts subtests adequately represent facets of the construct "language arts," and the second is that behaviors other than those related to "language arts" do not influence the scores obtained on language arts subtests. These premises also underlie testing of the constructs of reading and mathematics and therefore bear some discussion.

The former premise relates to the issue of construct under-representation (Cook & Campbell, 1979). This phenomenon occurs when a test is "too narrow and fails to include important dimensions or facets of the construct" (Messick, 1989, p 34). The major implication of construct representativeness is readily apparent. To make valid inferences about a student's performance in a construct of interest on the basis of a test score, the test must fairly represent the construct. If the test is under-representative, inferences can only pertain to the explicit facets of the construct that were actually tested.

The issue of test format relates to the second premise underlying "Total Test" scores. Scores are not influenced by behaviors other than those related to the construct of interest. The use of a single response format across subtests forces all behaviors within a battery to look the same. On most batteries, to respond to word attack items, math computation items, or language mechanics items, the test taker performs the same task of choosing from among several choices the answer that best completes the item. This "mono-operation bias" (Cook & Campbell, 1975; Messick, 1989) permits irrelevancies such as test-taking behavior and motivation to influence scores. If a single dimension such as a response set or previous practice with the test format can influence scores across a variety of subtests or constructs, the validity of inferences based on scores obtained from the test becomes suspect.

The format employed in most published achievement tests forces construct under-representation All but two of the batteries examined relied exclusively on multiple choice selection type responses and the two batteries that did include production responses (DAB and WJPEB) did so rarely. Each of the constructs, reading, language arts, and mathematics, includes dimensions that require active production of behaviors. For example, one of the most important outcomes expected of language arts instruction is facility in written expression. Any test of language arts that fails to include a writing sample cannot possibly claim to include all important dimensions of the construct. Similar arguments can be made for the importance of oral reading fluency in the construct reading and problem solving with algorithms in the construct math. Clearly, the tests analyzed in this study fail to represent the wide range of classroom relevant behaviors that are components of each construct.

Validity at the Battery Level

Valid inferences can be made on the basis of battery level scores obtained from published, norm-referenced achievement tests. The tests seem to function as broad band, low fidelity measures of general achievement. They result in normal distributions, suggesting that they do capture the full range of achievement present in classrooms. Their wide use in schools indicates that they have at least face validity if not functional value for providing an overall indication of the level of achievement of groups of students.

Behaviors can't be separated within batteries with a sufficient degree of precision to allow scores below the battery level to support valid inferences about student achievement in any of the constructs the tests purport to measure. Therefore, published

achievement tests have minimal implications for use in classrooms for decisions about individual students. Similarly, the value of these tests for evaluating specific programs within classrooms or at grade levels is dubious. The most appropriate application of most published, norm-referenced achievement batteries is in assessing general achievement at a screening level. Educators need to evaluate whether the investment of time and expense involved in using these tests results in a satisfactory return.

References

Cook, T.D., & Campbell, D. T. (1979). *Quasi-experimental design and analaysis issues for field settings.* Boston, MA: Houghton Mifflin

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 581-302.

Ebel, R. L. (1978). The case for minimum competency testing. *Phi Delta Kappan, 59,* 546-549.

Farr, R., & Carey, R. F. (1986). *Reading: What can be measured?* (2nd Edition). Newark, DE: International Reading Association, Inc.

Freeman, D. J., Kuhs, T. M., Porter, A. C., Floden, R. E., Schmidt, W. H., & Schwille, J. R. (1983). Do textbooks and tests define a national curriculum in elementary school mathematics? *Elementary School Journal, 83,* 501-513.

Good, R. H., & Salvia, J. (1988). Curriculum bias in published, norm-referenced reading tests: Demonstrable effects. *School Psychology Review, 17,* 51-60.

Hall, B.W. (1985). Survey of the technical characteristics of published educational achievement tests. *Educational Measurement: Issues and Practices,* 4 6-14

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

Knifong, J. D. (1980). Computational requirements of standardized word problem tests. *Journal for Research in Mathematics Education, 11*(1), 3-9.

Leinhardt, G., & Seewald, A. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18,* 85-96.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and MacMillan.

Petrosko, J. M. (1978). The quality of standardized high school mathematics tests. *Journal for Research in Mathematics Education, 9*(2), 137-148.

Salvia, J., & Ysseldyke, J. (1988). *Assessment in special and remedial education.* Boston: Houghton-Mifflin.

Shapiro, E. S., & Derr, T. F. (1988). An examination of overlap between reading curricula and standardized achievement tests. *The Journal of Special Education, 21*(2), 59-67.6.

Tukey, J.W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Using Story Retell as a Measure of Comprehension

Mary Gleason
University of Oregon
Jane Krauss
Eugene School District 4J
Gerald Tindal
University of Oregon

Introduction

Teachers of mildly handicapped, Chapter I, and other at-risk students are concerned with two related aspects of teaching reading: effective intervention and assessment of student progress. In the area of reading comprehension, teachers typically have used two kinds of testing to determine whether the intervention has been successful and students have made progress. They either administer standardized achievement tests at the end of the year (Jenkins & Pany, 1978), or they collect data from informal observations throughout the year as they ask students comprehension questions about the material they've read (Salmon-Cox, 1981; Clark, 1982). Both methods are of limited value for measuring achievement for several reasons (Deno, 1985). No determination is made of comprehension of information that was not tested (Farr, 1969), questions are selective according to what the tester judged to be worthwhile (Hansen, 1979), questions frequently are not passage dependent (Hansen, 1979; Johnston, 1982), and questions may either lead students to the right answer (Goodman & Burke, 1972) or inadvertently lead students away from the right answer (Howell & Kaplan, 1980).

Teachers of special populations have been frustrated by their attempts to monitor student progress via standardized achievement tests, particularly in the area of reading comprehension. Many special education teachers have worked tirelessly throughout the year only to face standardized achievement test scores at the end of the year that appear to demonstrate no improvement. It has been well documented that there is little overlap between test items and curriculum content and objectives with standardized reading tests (Good & Salvia, 1987; Jenkins & Pany, 1978; Shapiro & Derr, 1987). In addition, reading comprehension in particular is not adequately measured by standardized tests because of the manner in which it is tested (Howell & Kaplan, 1980). Comprehension items on standardized tests typically are limited to questions with multiple choice answers. Because of the lack of congruence between the instructional intervention and the test used to measure progress, the global scores obtained from these tests, and the infrequency of administration of these tests, standardized achievement tests are not sufficiently sensitive to growth over time and consequently fail to reflect instructional interventions.

However, the alternative used by teachers is not adequate either. When evaluating reading comprehension in particular, teachers have relied on the use of informal questioning procedures (Clark, 1982; Salmon-Cox, 1981). In one study of the reliability and validity of this type of informal observational assessment, Fuchs, Fuchs, and Warren (1982) found that special education teachers consistently overestimated the progress of their students when they relied on informal observation. Questions used in basal readers and those designed by classroom teachers have the same drawbacks as questions presented in standardized tests; these drawbacks may be even more pronounced because informally designed questions tend to be technically inadequate (Johnston, 1982).

In summary, conventional strategies for assessing reading comprehension may be inadequate and sensitive classroom-relevant measures of reading comprehension urgently needed.

A potential alternative could be Curriculum-Based Measurement (CBM), begun by Stanley Deno and his associates at the University of Minnesota Institute for Research on Learning Disabilities (Deno, 1985). When Deno and company began their research approximately fifteen

years ago, the focus was on developing sensitive achievement measures that could be used to formatively evaluate instructional programs. A number of correlational and experimental studies were completed in the basic skill areas of reading, spelling, and written expression, using a variety of measures having potential relevance for use in the classroom.

In identifying relevant classroom behaviors that would be useful in frequently monitoring student's learning in reading, Deno, Mirkin, Chiang, and Lowry (1980) reported on five measures from curriculum material: (a) oral reading from passages (Oral Reading), (b) oral reading from word lists (Words in Isolation), (c) identifying missing words (Cloze), (d) defining key words from passages (Word Meaning), and (e) reading key words within passages (Words in Context). All five measures were designed for frequent use, necessitating brief measures; as a consequence, one-minute timed tasks were devised. In the initial three validation studies, high correlations were found between these measures, particularly oral reading fluency, and reading comprehension as measured by published achievement tests (Deno, 1985).

Curriculum-Based Measurement for reading (oral reading fluency) has received strong support in the research field but questionable acceptance in the schools. The biggest problem is that teachers question its face validity regarding comprehension. A typical response by teachers in workshops is that oral reading fluency fails to identify students who can read but don't understand what they read. These students are often labeled a "word caller". The arguments about the high correlations with published achievement tests generally do not satisfy teachers, in great part because of the lack of content validity of these criterion measures and their lack of relevance to classroom practices (Jenkins & Pany, 1978; Shapiro & Derr, 1987).

We wanted to investigate a measure of reading comprehension that would be both technically adequate and have great potential as a formative measure, useful for evaluating instructional programs for low achieving and mildly handicapped students. We wanted the measure to be sensitive to growth over time as a result of instruction and to have no floor or ceiling effect.

Oral retell is a measure that has considerable potential but almost no empirical support. Although retelling has been employed within instructional programs designed to increase regular education student comprehension, the technical adequacy of retelling has not been investigated (Gambrell, Pfeiffer, & Wilson, 1985); retell procedures have rarely been used with mildly handicapped students.

To date, two studies have looked at retell as a curriculum-based measure. Fuchs, Fuchs, and Maxwell (1988) found a high relationship between the number of words written and special education students' performance on other measures (e.g. Stanford Achievement Test, cloze, and oral retell). Parker and Tindal (1989), using regular education students from three grade levels (6, 8, 11) employed written retell along with two other criterion measures: a maze comprehension task and a creative writing sample. The retell measure was scored for the number of words written and idea units presented; students also completed a creative story as a measure of written expression. They found that the distributions for the retell measure were very narrow for the number of idea units produced, accounting for roughly 10% of the information in the text; furthermore, the average retell measures changed little over the grades. Finally, no relationship was found between retell and creative writing. In summary, the technical adequacy data supporting retell as a measure of comprehension are limited.

The purpose of our study was to assess the validity of oral retell as an informal reading comprehension measure. Students' retell responses were evaluated in three major ways: (a) holistically, according to quality (similar to procedures used to evaluate writing); (b) counting the number of ideas expressed in each retell response; and (c) counting the number of words in each retell. If these different indices of oral retell are related, perhaps they can assist in defining comprehension, and evaluating instruction as a unitary construct. If oral retell is a measure of comprehension, how highly correlated is it with oral reading fluency? In addition, if indices of retell did indeed measure comprehension, perhaps teachers would respond favorably to the use of oral retell as a curriculum-based measure.

Method

Subjects

Subjects were drawn from the various schools in which practicum students were teaching and included 35 third- and fourth-grade students from a school district located in a middle SES community on the west coast. University practicum students in special education served as the data collectors. In each school, teachers submitted class rosters totalling 115 students. Parents of 39 stu-

dents returned informed letters of consent.

The potential subjects completed a timed screening test to determine their eligibility. A ceiling level of 150 words per minute and a basal level of 21 errors was established. Students who performed between the basal and ceiling levels of the screening test became eligible for participation in the study. The final 35 subjects comprised three subgroups: 12 subjects were third-grade and 12 were fourth-grade students receiving regular classroom reading instruction; 11 were fourth-grade students receiving special instruction in reading (Chapter I or resource room services).

Materials

The study used short, expository passages of 105 to 175 words excerpted from *New Practice Readers*, Levels A through C (McGraw Hill Publishers, Webster Division, 1978). Each passage represented a single topic or theme from the natural and social sciences, similar to those taught in content area classes in the lower and intermediate grades. The passages were evaluated for readability levels with a computerized readability program (MECC Teacher Utilities, 1977). Spache and Fry Readability Test scores were used to derive a composite score and ten passages ranging from a grade equivalency of 2.1 to that of 4.1 were chosen. Passages were randomly ordered for presentation during a 10-week test period.

Procedures

Before the first test was administered, selected subjects participated in a practice session in which they were taught to orally retell information from a reading passage. As each student recalled information from the passage, the tester recorded the number of ideas the student expressed. If the student recalled two or fewer ideas from the passage, the tester demonstrated how to read and retell that same passage. The student then reread the passage and retold it again. If two or more ideas were recalled from the repeated passage, the student then read another passage and recalled it. This procedure continued until all subjects could read a new passage and orally retell more than two ideas.

Data collectors were trained to use scripted wording and prescribed administration procedures. Students read aloud from short expository passages once a week for 10 weeks. Two tests were administered each week: an oral reading fluency test, followed by an oral retell test. The data collector timed the oral reading of a passage and recorded the number of errors made. After reading, the student recalled all that he or she could remember from that same passage without prompts from the data collector. Each testing session was audiotaped for reliability checks and for scoring.

Scoring

Oral Reading Scoring Methods

When the test was administered, the data collector recorded the time in which the passage was read and the number of errors made. This information was then translated into the rate of words read correctly and incorrectly per minute.

Comprehension Scoring Methods

The students' oral retells were audiotaped during every session, then transcribed for scoring and analyzed for quantitative and qualitative scores.

Quantitative analyses were based on transcripts of retell responses using *Word Tools* (Clapp, 1986) to count the total number of words, as well as the number of unique words, adjectives, articles, and conjunctions occurring in the retell responses.

Two qualitative analyses were based on the number of ideas in the retell that reflected the content of the original passage and independent judgments using a holistic scoring procedure. An idea unit was defined as a simple T-unit or kernel sentence represented in each independent clause (Hunt, 1964). Information from subordinate clauses was disregarded. A loss of information would be an issue with more sophisticated prose, but the test passages were simple, and the sentences generally were made up of one or two clearly defined independent clauses. Student responses had to clearly express ideas from the original passage in order for those ideas to be credited. For instance, if the original idea unit was, "There is a tube that connects the middle ear to the inner ear" (an idea showing relationship), and the student stated, "There is a middle ear" (a single fact), this statement was not credited as representing the original idea unit.

Holistic judgments were made by three individuals who evaluated the retell responses using a 1 to 5 scale of quality. A score of 5 was assigned to the best retells, and a score of 0 was assigned to students who recalled nothing about the passage. The 350 retells were scored with an interscorer agreement of approximately .75.

Results

Relationships Among Oral Reading Scores and Indices of Retell

The following different relationships between oral reading and the different retell measures were

analyzed. First, we correlated the rate at which students orally read passages to the number of words in the retell response (range -.16 to .22), to the number of idea units retold (range .05 to .38), and to the holistic judgment scores (range .06 to .39). Then, we correlated the rate at which students made errors during oral reading to the number of words in the retell response (range -.21 to .35), to the number of idea units retold (range -.23 to .04), and to the holistic judgment scores (range -.35 to .09). See Tables 1, 2, and 3 for specific scores for each of the ten weeks.

Relationships Among Indices of Oral Retell

Two relationships between different indices of oral retell were analyzed. For each of ten weeks, we correlated the number of words in the retell response with the number of idea units retold (range .48 to .83) and with the holistic judgment scores (range .71 to .89). We also correlated the number of idea units with the holistic judgment scores (range .59 to .87). See Tables 1, 2, and 3 for specific scores.

Discussion

The purpose of the present study was to assess the validity of oral retell as an informal reading comprehension measure and to ask several questions. Were different indices of oral retell related to each other? If so, could they assist in defining comprehension as a unitary construct? If oral retell was a measure of comprehension, would it correlate with oral reading fluency? Is oral retell a practical approach to measuring comprehension?

Relationships Among Indices of Reading Retell

The results of qualitative and quantitative analyses indicate a strong relationship among the indices of oral retell. Because the holistic scoring and the idea-unit scoring were significantly correlated with the number of words in each retell, it appears that the quality of retelling is directly related to the quantity of it. This finding is consistent with that of Fuchs et al. (1988). In other words, students who retell more are also judged as having a higher level of comprehension.

The fact that such a simple measure as the number of words retold correlates with such well-accepted procedures as holistic judgments and idea-unit scoring is fortuitous. Holistic scoring and idea-unit scoring are evaluation methods that represent values teachers bring to the task of measuring student achievement. Holistic scoring is a tool that is accepted by teachers, as evidenced by its wide use for judging writing. Idea-unit scoring, which was significantly correlated with the holistic method in this study, is similar to what teachers often do when they assess reading comprehension. Just as teachers ask questions to determine how thoroughly students retain information, idea-unit scoring provides a way to quantify how much is retained. Holistic scoring and idea-unit scoring are not only valued by teachers, but they also offer objective control to the relatively subjective task of measuring comprehension. However, both scoring approaches are too unwieldy and time-consuming to be used for frequent progress monitoring or to be used practi-

Table 1. Correlations Among Total Number of Words and Other Indices of Reading Retell

Test Week	1	2	3	4	5	6	7	8	9	10
Correct words per minute	-.21	-.18	.19	-.02	.05	.18	.12	.35	.21	-.06
Erred words per minute	.02	.03	.11	-.02	.05	-.16	.11	-.15	.22	.06
Idea units per retell	.61	.63	.63	.82	.74	.48	.83	.67	.72	.72
Holistic judgment scores	.81	.80	.75	.84	.82	.78	.89	.78	.71	.71

Table 2. Correlations Among Idea Unit Scores and Other Indices of Reading Retell

Test Week	1	2	3	4	5	6	7	8	9	10
Correct words per minute	.62	.63	.66	.83	.74	.49	.83	.67	.70	.73
Erred words per minute	-.19	-.16	-.16	-.06	-.11	.03	.04	.01	-.21	-.23
Characters per retell	.62	.63	.66	.83	.74	.49	.83	.67	.70	.73
Words per retell	.61	.63	.63	.82	.74	.48	.83	.67	.72	.72
Holistic judgment scores	.77	.61	.59	.84	.86	.78	.87	.66	.87	.72

cally by classroom teachers. Given that the number of words in a retell response is significantly correlated with these otherwise acceptable indices, a simple word count may serve as a valid method for estimating reading comprehension. Ongoing measurement of the total number of words in a retell may serve as a useful means of monitoring growth in comprehension over time.

Retell and Comprehension

This study also sought to understand the various dimensions of recalling expository prose as measures of reading comprehension. In the past, retell has been positively correlated with comprehension subtests on standardized achievement tests and with other informal measures of comprehension such as cloze, maze, and question-based tests. Retell has also been positively correlated with an indirect index of comprehension, oral reading fluency. In this study, idea-unit scores were highly correlated with holistic scores. Whereas idea-unit scoring credits only information that matches the original passage (verbatim recall), holistic judgment can be more sensitive to how well the student integrates new information with

Table 3. Correlations Among Holistic Judgment Scores and Other Indices of Reading Retell

Test Week	1	2	3	4	5	6	7	8	9	10
Correct words per minute	.06	.28	.24	.06	.16	.09	.24	.11	.39	.32
Erred words per minute	.06	-.35	-.17	-.01	-.20	.01	.09	.08	-.18	-.15
Characters per retell	.83	.82	.76	.84	.81	.79	.91	.72	.69	.74
Words per retell	.81	.80	.75	.84	.82	.78	.89	.78	.71	.71
Idea units per retell	.77	.61	.59	.84	.86	.78	.87	.66	.87	.72

what s/he already knows. Memory and comprehension appear to be inexorably linked; it is difficult to remember what is not understood. As a curriculum-based measure, oral retell appears to have greater face validity than oral reading fluency.

Practicality of Oral Retell

Retell tests can be constructed inexpensively and with alternate forms from classroom reading materials. Retell administration procedures are simple; they take little time to administer, and can be used reliably with minimal training. However, as practical measures, oral retells must be easy to score. In this study, oral retells were audiotaped and transcribed for later scoring, which appears to make it unfeasible for frequent measurement. Written retell may be more practical for this use. Further research must look at a comparison of results between oral and written retells to examine whether information gained through written retell (ease of scoring) is economic in terms of results sought (measurement of comprehension). The most obvious problem is that teachers are in danger of losing information because of student's difficulties with writing and spelling. More able writers might achieve better retell scores even if their recall of information is similar. The results of this study indicate that at the very least, teachers using oral retell would simply record and count the number of words in each retell. It would not be necessary to use the more time-consuming procedures of counting idea units and conducting holistic scoring.

Relationships Among Oral Reading Fluency and Retell Scores

We did not find a significant relationship between oral reading rate and retell scores in this study. This finding is inconsistent with results from other studies and is probably related to a number of factors involving differences in subjects, reading material, procedures, etc. used in this study compared to other studies (Krauss, 1989).

Summary

That recall is an important component of comprehension is not disputed (Hansen, 1979), but appropriate scoring procedures for recall tests have yet to be fully explored and validated. Written retell tests may be more useful for measuring comprehension than oral retell tests (Fuchs et al., 1988), but the degree to which the expression of recall is influenced by writing ability needs to be explored. The number of words in a written retell response is positively correlated with oral reading fluency. In other studies, oral reading fluency correlates highly with formal and informal measures of comprehension, indicating that reading rate may be as good an index of reading, including comprehension, as any metric currently used. Methods for scoring retell tests that appear to be most appropriate are those that evaluate recall in both quantitative and qualitative ways (Kalmbach, 1986). The use of a combined scoring procedure may be appropriate for approaching the validation of recall as a measure of comprehension.

References

Clapp, D. (1986). *Word Tools*. Computer program, Aegis Development, Inc.

Clark, C. H. (1982). Assessing free recall. *The Reading Teacher, 35*, 434-39.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. (1980). *Relationships among simple measures of reading and performance on standardized achievement tests*. (Research Report No. 20). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Farr, R. (1969). *Reading: What can be measured*. Newark, DE: International Reading Association Research Fund.

Fuchs, L., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28.

Fuchs, L. S., Fuchs, D., & Warren, L. (1982). *Special education practice in evaluating student progress toward goals* (Research Report No. 21). Minneapolis: University of Minnesota: Institute for Research on Learning Disabilities.

Gambrell, L. B., Pfeiffer, W. R., & Wilson, R. M. (1985). The effects of retelling upon reading comprehension and recall of text information. *Journal of Educational Research, 78*, 216-220.

Good, R. H., & Salvia, J. (1987). Curriculum bias in published, norm-referenced reading tests: Demonstrable effects. *School Psychology Review, 17*, 51-60.

Goodman, Y., & Burke, C. (1972). *Reading miscue inventory: Procedure for diagnosis and evaluation*. New York: Macmillan.

Hansen, C. L. (1979). Chicken soup and other forms of comprehension. In J. E. Button, T. C. Lovitt, & T. D. Rowland (Eds.), *Communications research in learning disabilities and mental retardation* (pp. 67-91). Baltimore, MD: University Park Press.

Howell, K., & Kaplan, J. (1980). *Diagnosing basic skills: A handbook for deciding what to teach.* Columbus, OH: Charles E. Merrill.

Hunt, K. (1964). *Differences in grammatical structures written at three grade levels, the structures to be analyzed by transformational methods.* Cooperative Research Project #1998, Office of Education, U. S. Department of Health, Education, and Welfare.

Jenkins, J. R., & Pany, D. (1978). Curriculum biases in reading achievement tests. *Journal of Reading Behavior, 10*(4), 345-357.

Johnston, P. B. (1982). *Implications of basic research for the assessment of reading comprehension* (Tech. Rep. No. 206). Urbana-Champaign: University of Illinois, Center for the Study of Reading (ERIC Document Reproduction Service No. ED01 987).

Kalmbach, J. R. (1986). Evaluating informal methods for the assessment of retellings. *Journal of Reading, 30,* 119-126.

Krauss, J. (1989). *Measurement of reading comprehension using oral retell.* Unpublished master's thesis, University of Oregon, Eugene.

Parker, R., & Tindal, G. (1989). *Initial validation of two classroom materials-based measures of reading comprehension.* Unpublished manuscript, University ofOregon.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? *Phi Delta Kappan, 62,* 631-634.

Shapiro, E. S., & Derr, T. F. (1987). An examination of overlap between reading curricula and standardized achievement tests. *The Journal of Special Education, 21*(2), 59-67.

Teacher Use of Various Data Sources for Grouping, Placement, and Identification Decisions

Tracey E. Hall
Judith A. Gelbrich
Gerald Tindal
University of Oregon

Introduction

Ability grouping of students for instruction has been under study since the 1920s (Slavin, 1988; Anderson, Mason, & Shirey, 1984). Teachers use various forms of grouping to respond to student differences in knowledge, skills, and learning rate. To optimally present a lesson, teachers form instructional groups so that students may profit from the lesson, thus avoiding a presentation of skills that may be redundant for some learners and too difficult for others (Slavin, 1988; Carnine, Silbert, & Kameenui, 1990). Students must also receive instruction in materials that are at an appropriate level of difficulty or an instructional level.

Whether within classroom, within grade level, or across grade level, ability grouping is a very complex teacher decision-making phenomenon studied by many researchers to (a) evaluate grouping procedures (Wiesendanger & Birlem, 1981; Haller & Waterman, 1985; Wesson, Viertrhaler, & Haubrich 1989), (b) determine the function of grouping (Strike,1983), and (c) evaluate student outcome as a function of group instruction (Anderson, Mason & Shirey, 1984; Strike, 1983). Beyond the normal range of students a teacher must teach within the classroom, there are often instances of students performing at the extremes. Teachers must decide whether these students would best be served within the curriculum and instruction of the regular classroom or outside of the classroom. Our investigation evaluates how teachers use various forms of data for grouping, placement, and potential identification.

Teachers have a variety of data sources available to help them in the grouping decision-making process. In a survey conducted by Haller and Waterman (1985), intermediate level teachers considered student ability, general academic competence, work habits, behavior, personality, and home background to place children in groups for instruction. By grouping students, teachers attempt to maximize instruction.

Although many factors appear influential, student ability and academic competence dominate the decision-making process. The main tool for documenting ability and achievement is standardized norm referenced tests. These tests are also the most prevalent tools available in American schools (Valancia & Pearson, 1987). While standardized achievement tests are rarely used as the sole criterion for admitting children to specialized programs, such as Talented and Gifted, they frequently serve as an important screening device. Generally, school districts set a standard for consideration, such as a score above the 95th or 98th percentile in reading and math (Eby & Smutny, 1990).

Recently, however, the usefulness of standardized tests for classroom teachers has been seriously contested (Salmon-Cox, 1981; Wesson et al., 1989). Test procedures have limited utility for placing pupils in specific groups because they lack content validity: test items rarely reflect the curriculum of instruction (Coleman & Harmer, 1982; Jenkins & Pany, 1978). Teachers value test scores , but they also search for instruments that are more diagnostic in nature, match the curriculum of instruction, and are timely (Salmon & Cox, 1981). Achievement tests were neither designed for, nor are they capable of carrying out such a function.

Teachers seem to be at an impasse, since predominant tests have serious technical problems. What choices, then, do teachers have available? Two familiar options are well known in reading. The first is published, teacher-made, and curriculum-based Informal Reading Inventories. The second alternative is to use diagnostic measures, such as the Grey Oral Reading Test (1986) and the

Gates MacGinite Reading Test (1978). Additionally, many diagnostic measures have been developed by teachers. However, these procedures also suffer from problems with validity and reliability (Fuchs, Fuchs, & Deno, 1982; Salvia & Ysseldyke 1985).

Another procedure that has been recently developed is Curriculum-Based Measurement (CBM) (Deno, 1989). CBM is a method of analyzing academic performance using systematic procedures with brief measures in specific academic areas. It is accurate in grouping students for instruction when scores are rank ordered and students are grouped based on similar results (Hall & Tindal, 1989; Wesson, et al., 1989). This procedure can assist teachers with student grouping and placement decision-making quite efficiently. Curriculum Based Assessment has recently been investigated as a screening measure in the identification of giftedness with kindergarten and first-grade children (Joyce & Wolking, 1988). Their findings suggest that Curriculum Based Assessment is at least as effective as the Metropolitan Achievement Test in identifying students for gifted programs within that age group.

In summary, CBM may be a viable alternative to the need for technically adequate measures that are relevant for classroom use. They overcome the problems of many extant tests (published, norm-referenced and diagnostic) and have some initial validity data on making placement decisions. Furthermore, they reflect the call by school personnel and researchers for procedures by which standards can be developed to determine student instructional level. Such measures should also be able to discriminate performance levels among students and be minimally time-consuming (Coleman & Harmer, 1982).

This study investigates how teachers value and actually use testing data to make placement decisions so that they can provide appropriate instruction to homogeneously grouped students. Teachers must also make reasonable referrals to special services at each end of the performance spectrum. Our investigation analyzes CBM to determine if it can serve as a tool for making qualitative decisions about instructional level, as well as the categorical decisions focusing on placement of students with extreme skills (either very low or very high). Since a considerable amount of research has already been done on the use of CBM for making placement decisions in special education (Shinn, Tindal, & Stein, 1988), we also looked at the use of CBM for making placement decisions with Talented and Gifted students. In this study, we were interested in both the process of decision making as well as the decision that was eventually made. In most previous research on CBM, only the outcome has been documented. To fully understand the context for decision making, however, more data need to be collected on how teachers make decisions and what types of information they value.

Method

Subjects

Eighteen teachers from a semi-urban school district in the Pacific Northwest participated in this study. These teachers represent one elementary building faculty serving approximately 380 students, in a district with four other elementary schools and a total student population of 1,787. This particular staff has been using Curriculum-Based Measurement procedures for two years on a school-wide basis, the teachers are familiar with the procedures and how to interpret normative data results. The emphasis of the decision making, however, has been confined to screening referrals for special education and writing IEP goals.

Teachers in this school have traditionally used ability grouping for reading instruction. The first and second grades maintain a self-contained structure and ability group within the homeroom setting. Each teacher had a wide range of student skills in the classroom. Third grade classes ability group across their grade level. Each teacher had a somewhat homogeneous group for that grade level. The fourth and fifth grades ability group across the two grades for reading. Each teacher had a relatively homogeneous group of students for both the fourth and fifth grades.

All teachers are state certified,with the majority (N=17) in elementary education. Several hold specialty endorsements in special education (N=4). One teacher is certified in secondary education, with a special education endorsement and another teacher has an administrator's certificate in addition to an elementary certificate. Several teachers hold advanced degrees (N=8), ten teachers have a BS or BA. The ages of teachers in this study range between 26 and 57, with the majority between 32 and 45 years of age (one teacher did not report in this field). Most teachers (N= 11) had between 8 and 20 years of teaching experience. Only three teachers had less than five years, and three teachers had greater than 25 years of experience. There were more female (N= 12) than male (N=6) teachers at this elementary school. This two-to-one ratio of females to males is typical at the elementary level in this school district.

Measures

The following procedures were investigated in this analysis: (a) direct observation of six teachers as they collaboratively grouped students for instruction in reading and math, (b) pencil-and-paper survey in which all teachers were asked to rank and assign a value to data sources available for decision making, (c) pencil-and-paper questionnaire in which all teachers described procedures and current practices for current grouping and instruction, (d) rank and sort of reading class roster, and (e) listing of referred and identified students for gifted and talented programs.

Direct Observation

A direct, non-interactive procedure was used to observe how teachers group students for reading and math. The trained observer sat in with one special education teacher and five fourth and fifth grade teachers during a one hour and twenty minute period as student instruction and placement decisions were made. Teachers were provided materials listing all students in the two grades as well as information from various data sources: (a) Stanford Achievement Test percentile scores by subject area, (b) Curriculum-Based Measurement percentile scores,(c) teacher recommendations, and (d) student's reading text placement from the previous spring.

The observer recorded the proceedings verbatim noting time, teacher speaking, and item of discussion. Thus, a complete transcript of the proceedings was produced. At a later time, a participating teacher was asked to read and confirm the transcript for accuracy. To quantify the transcript, each statement was coded according to a classification system developed following the observation. Inter-rater agreement for coding the transcript was 75.4% accuracy level, following two sessions of definition and clarification

Rank and Value of Data Sources

We administered a paper and pencil measure to evaluate how teachers assign a rank and value to specific data sources when making grouping decisions. Eight sources of student performance were listed alphabetically to avoid any ordering affect: (a) basal related tests, (2) behavior of a student, (c) fluency measures, (d) independent work, student performance, (e) informal reading inventorie, (f) published achievement tests, (g) recommendations from previous teachers, and (h) teacher observations. There was also space to write any additional source used. Teachers were asked to rank the data sources they felt were most helpful (1-8) and then assign a value of usefulness to each (1-4).

Questionnaire

A ten item questionnaire was given to each regular education classroom teacher in the school. The fundamental issues were related to decisions teachers made regarding placement of students into instructional groups, and the logistics, rationale, and procedures they used for the varying teaching structures in their individual classrooms.

Rank and Sort of Class Roster

Teachers were given their class roster for reading instruction with students listed alphabetically. We requested two tasks: first, teachers were asked to rank, by number, the students in their reading class from highest to lowest based on reading achievement; second, they were asked to sort the students according to reading ability into one of three categories, High, Medium, or Low. This procedure was completed prior to receiving CBM norm results for their classes.

Gifted Identification

To obtain information for this portion of the study, we asked the elementary teacher of the gifted program to identify the children at this school currently receiving services. In addition, the classroom teachers from grades two through five were asked to list students that they would refer to the gifted program. These two categories were compared to the academic performance of students on the Curriculum-Based Measures.

Procedures

Teachers completed the rating and value judging of information and written questionnaires between October 3 and 31, 1989. The sorting and ranking of students occurred before Fall CBM scores were made available to teachers. Throughout the collection of all information listed above, we attempted to be as unobtrusive as possible. Each teacher spent 20-30 minutes responding to the materials for this study. Data analysis included all of these sources of information as well as the Fall CBM norming data.

Results

Direct Observation

Teacher statements from this observation were categorized into six major headings: type of data, students, meeting organization, administrative concerns, placement decisions, and general comments (see Table 1). The student category comprised the largest number of comments, 40%. Within this category, the largest classification of statements was in reference to a specific child (69%). The second largest category was statements

in direct reference to the data files available (19%). Thirty-one percent of these statements were related to achievement tests, and 29% were comments specific to CBM information. Only 16% of the statements referred to teacher recommendations; and reference to level of reading material comprised only 5% of the comments in this category. The percentages in Table 1 sum to more than 100% due to rounding error.

Table 1. Summary: Coding and Classification of Comments in Grouping Meeting

Classification of Comments by Category	Number of Statements	%
Type of Data	77	19
Children	165	40
Meeting Org.	21	5
Admininistrative Concerns	60	14
Student Placement	74	18
General Comments	19	4
Total	416	100

Rank and Value of Data Sources

Table 2 presents a summary of the teachers' rank and value of the eight types of tools frequently used in placement and grouping decision-making, from highest ranked and valued to lowest. The teachers were consistent when assigning a rank order as they valued the measure as high, moderate, low, or no value.

Paired t-tests were used to determine if there were significant differences in how teachers value these sources of information. There were significant differences in how teachers valued direct observations (t (16)= -3.781, p=.0016), children's independent work (t (16) = -4.146, p=.0008)), and teacher recommendations (t (15)= -4.0012), p=.0012, in relation to CBM. Again, significant differences were apparent when observations (t (15) = 3.0, p=.0127), children's independent work (t (15) = 4.0, p=.0009), and teacher recommendations t ((15)= 4.0, p=.0006) were compared to the standardized achievement tests. However, when Curriculum-Based measures were compared to standardized achievement tests, there were no significant differences in how teachers valued these two measures. Nor were there any significant differences in the value of teacher recommendations compared to teacher observations and children's independent work.

Table 2. Teachers' Rank and Value of Data Sources

	Rank		Value	
	Mean	SD	Mean	SD
Student Indep. Work	2.7	1.53	3.78	.43
Teacher Observe	2.9	2.22	7.78	.73
Teacher Recom.	5.4	2.21	3.53	.72
Inform. Rdg Inv.	4.8	2.01	3.06	.77
Curric.-Base Msr	4.9	2.05	2.65	.93
Pub. Ach. Tests	4.9	2.32	2.75	1.07
Basal Tests	5.1	2.10	2.33	.76
Student Behavior	6.1	2.12	2.39	.85

Questionnaire

All regular education teachers from the school responded to the questionnaire (N=17), though not every teacher answered each question. These teachers ability-group students primarily because of teaching skills (N=6), followed by student characteristics (N=5), and finally, by administrative arrangements (N=4) such as class size,or traditional practice. When asked how assignments are made to group levels, most teachers listed that they worked together to make decisions, using such criteria as individual teacher skills, trading low and high groups by year or by subject, student characteristics, and volunteering to teach a particular group.

In response to questions regarding the procedures for grouping, five teachers could not specify a starting point to the sorting process. Four teachers reported starting the assignment of students to groups with low ability students. Only one teacher reported this process beginning with the middle group. No one listed the high group as their starting point.

Teachers who ability-group for reading reported spending more time on the grouping process (mode 2-3 hours) than teachers who used a self-contained model with a mode of less than one hour. Generally, these teachers seemed to be satisfied with grouping students for instruction (N=14). No one expressed any dissatisfaction, although one teacher expressed willingness to change from across-grade grouping to self-contained if the opportunity arose.

Teachers' instructional delivery systems varied once students were placed in classrooms for instruction. Some (N= 5) taught large groups for

reading since the class was ability-grouped. Those teachers (N= 10) who grouped students within the classroom selected ability level of the students as their primary concern for grouping, with size ranging from less than 5 to 15. Teachers with one instructional group listed alternatives for flexible instruction such as cooperative learning groups, learning centers, or large group instruction.

Rank and Sort of Class Roster

Data collected from teacher ranking and sorting of students into high, medium, or low groups were analyzed in relation to reading proficiency as measured by Curriculum-Based Measures. Norming results from CBM were sorted by grade from the most proficient reader (highest reading rate) to least proficient reader (lowest reading rate). This resulted in a frequency distribution, a common practice in analyzing CBM norming scores (Deno, 1989). The distribution of scores was split into three groups by calculating the mean and standard deviation for each grade-level distribution. Scores below minus one standard deviation were assigned to group one (low), scores between one standard deviation below and one standard deviation above the mean were assigned to group two (medium), scores above plus one standard deviation were assigned to group three (high). A three-by-three chi square test was used to compare the teacher sorting of students into low, medium, and high groups and the CBM assignment to one of three groups. Teacher placement of students into groups was highly correlated with CBM rankings at second and third grade levels, and somewhat less with the fourth and fifth combination (2nd grade, $x^2(4, N=64) = 40.429$, $p = .001$; 3rd grade $x^2(4, N=65) = 34.182$, $p = .001$; 4 & 5th grade $x^2(4, N=124) = 73.749$, $p=.001$).

Within-class and within-grade teacher ranking of students correlated highly when teacher rank and CBM ranks were compared using a Pearson correlation coefficient, $r = -.823$, , $r = -.60$, respectively. However this finding was not highly correlated for across grade grouping.

Table 3. Three by Three Comparison of Teacher and CBM Grouping

	Second grade (N=64)			Third grade (N=65)			4th&5th grade (N=124)		
	L	M	H	L	M	H	L	M	H
Low	9	6	0	8	9	0	17	17	0
Mid	4	28	4	2	21	2	2	51	4
Hi	0	4	9	0	11	12	1	15	17

Gifted Identification

Table 4 presents a matrix developed by using the percentile ranks of 80, 90, 95 and 98 for both Curriculum Based Measurement and the standardized achievement test. The students were separated into two groups, those receiving services for the gifted, and those who had been recommended to receive services, but who had not yet been identified as gifted. For each group, the number of students meeting the criteria at each percentile was calculated.

The majority of students identified as gifted received a score at or above the 90th percentile on CBM; this was also true of standardized test results. An equal number of students scored at or above the 98th percentile on both the CBM and standardized achievement tests. In this study CBM was comparable in accuracy for screening gifted students in relation to the less frequently administered standard achievement tests.

For the group that had been recommended and not yet identified as needing services for the gifted program, the majority of students scored at or above the 80th percentile on both the CBM and the achievement test. However, at the 90th and higher percentiles, less than half of the students scored above this level on both the CBM and the achievement test, once again, demonstrating comparable results for both measures.

Discussion

The interest in this study developed from three basic questions surrounding assessment procedures in elementary schools: How do teachers value and use available measures? Is CBM a functional tool for grouping and placement decisions within classrooms? Could CBM norm results assist in the identification or screening of gifted students?

Teachers clearly indicated that they value and make use of available data about their students. Of particular interest are tools the teacher develops or skills that can be directly observed. Our observation verified that teacher knowledge of the student is highly valued when available. This finding is very similar to Haller and Waterman's findings in 1985. However, direct observation also indicated that CBM and achievement tests were most useful for making grouping decisions. Teachers, when forced to rely on data sources because of little exposure to students at the beginning of a school year, used CBM first, and then published, norm-reference achievement test scores.

Table 4. Using CBM and Achievement Tests to Identify Students for Gifted Programs

Percentiles		80 PR				90 PR				95 PR				98 PR			
Test	N	CBM		Ach		CBM		Ach		CBM		Ach		CBM		Ach	
		+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
TAG Ident.	7	5	2	7	0	5	2	5	2	5	2	4	3	4	3	2	5
TAG Ref.	20	13	7	18	2	5	15	11	9	4	16	3	17	1	19	3	17

Key: + = TAG Identified by this criterion - = TAG Rejected by this criterion

We found some interesting results when comparing information from the questionnaire to direct observation. Only three teachers from the 4th and 5th grades reported that they began the sorting process for ability grouping at the low level. However, when observing this process, all six teachers began at the low level, using teacher information and test scores to get those students into a group and keep that group as small as possible. In other instances, direct observation verified the questionnaire findings when decisions about teacher assignment to group and administrative issues were addressed.

These teachers find CBM a functional tool for use in the grouping and decision making process as demonstrated through observation and survey instruments. The accuracy of CBM for grouping students is further demonstrated by the high correlation of teacher judgement of students and the CBM rank results. Furthermore, we found that CBM was functional as a screener for identifying gifted students with this population. However, we recognize that, in this study, the number of gifted students was too small to be conclusive, and we would encourage further research in CBM and gifted screening to examine other curriculum areas. Teachers have shown that CBM scores for reading are as useful as the published achievement tests for grouping and placing students. These teachers like CBM because it confirms their judgments about student performance. It is an efficient tool for grouping, and CBM tests represent student performance in the curriculum of instruction.

References

Anderson, R.C., Mason, J. & Shirey, L. (1984). The reading group: An experimental investigation of a labyrinth. *Reading Research Quarterly, 20,* 6-38.

Carnine, D., Silbert, J, & Kameenui, E. J. (1990). *Direct instruction reading* (2nd ed.) Ohio: Merrill.

Coleman, M., & Harmer, W.R. (1982). A comparison of standardized reading tests and informal placement procedures. *Journal of Learning Disabilities, 9,* 396-398.

Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum based measurement: Assessing special children* (pp. 1-17). New York: Gilford.

Fuchs, L. S., Fuchs, D. & Deno, S. L. (1982) Reliability and validity of curriculum-based informal reading inventories. *Reading Research Quarterly, 18,* 6-26.

Eby, J. W., & Smutny, J.F. (1990). *A thoughtful overview of gifted education.* White Plains, NY: Longman.

Hall, T., & Tindal, G. (1989) Using curriculum-based measures to group students in reading. In G. Tindal, K. Essick, N. George, M. George (Eds.), *The Oregon Conference '89 Monograph.* Eugene, OR: College of Education, University of Oregon.

Haller, E., & Waterman, M. (1985). The criteria of reading group assignments. *The Reading Teacher, 38,* 772-781.

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

Joyce, B.G., & Wolking, W.D. (1988). Curriculum-based assessment: An alternative approach for screening young gifted children in rural areas. *Rural Special Education Quarterly, 8,* 9-14.

MacGuilt, W. H., Kamons, J., Kowalski, R., MacGinite, R.T., & McKay, T. (1978). *Gates MacGinite Reading Test* (2nd ed.). Chicago: Riverside.

Salmon-Cox, L. (1981). Teachers and standardized achievement: What's really happening? *Phi Delta Kappan, 62,* 631-643.

Salvia, J., & Ysseldyke, J. (1988). *Assessment in special and remedial education.* Boston: Houghton Mifflin.

Shinn, M. R., Tindal, G., & Stein, S. (1988). Curriculum-based measurement and the identification of mildly handicapped students: A research review. *Professional School Psychology, 3*(1), 69-85.

Slavin, R. E. (1988). Synthesis or research of grouping in elementary and secondary schools. *Educational Leadership, 46*, 67-77.

Strike, K. (1983). Fairness and ability grouping. *Educational Theory, 33*, 125-134.

Valencia, S., Pearson, P. D. (1987). Reading assessment: Time for change. *The Reading Teacher, 40*, 726-732.

Wesson, C. L., Vierthaler, J. M., Haubrich, P.A. (1989). An efficient technique for establishing reading groups. *The Reading Teacher, 42*, 466-469.

Wiederhilt, J.L., & Bryant, B.R. (1986). *Gray Oral Reading Test*. Austin Texas: Pro-ed.

Wiesendager, K. D., Birlem, E. (1981). A critical look at the reading approaches and grouping patterns currently used in primary grades. *Reading Horizons, 22*, 54-58.

Identifying At-Risk Kindergarten and First-Grade Students: Recent Developments

Roland Good
Ruth Kaminski
Ilsa Schwarz
Catherine Doyle
University of Oregon

Although only a small number of children (2% to 5%) are identified as handicapped at entry into public school, by the age of 8 years, the percentage of children identified as handicapped and receiving special education services peaks at 11% or over 350,000 children (Education, 1989). Not all children experiencing reading difficulty are identified as requiring special education, however. Five million children currently receive services through Chapter 1 Programs (Guttman, & Henderson, 1987). In addition, many children spend an extra year in kindergarten or first grade through retention, participation in a transitional or pre-first grade, or by staying home an extra year until "ready" for school. Indeed, Haberman (1989), claims that "reading readiness has become the standard for kindergarten retention" (p. 285). Estimates of the incidence of retention vary dramatically from district to district, ranging from 10% to as many as 60% of children (Shepard, & Smith, 1988). And yet no children enter kindergarten with identified reading problems. Clearly, the early school years are a high-risk time for the development of reading problems in children.

Need for the Primary Prevention of Early Academic Problems

The focus on remediation of existing reading difficulties is problematic for a number of reasons: (a) reading difficulties tend to persist over time; (b) reading difficulties tend to become more severe over time; and (c) students must acquire reading skills at a faster rate than their peers to "catch-up" or reduce the discrepancy between their performance and that of their peers.

Persistence of Reading Difficulties

In general, even poor readers make progress as they mature. However, there is little evidence to suggest that they "outgrow" their reading disabilities completely or that they approximate the performance of proficient readers (Rourke, 1978). The persistence of reading problems was documented by Juel (1988), in a study investigating the reading and writing development of 54 children as they progressed from first through fourth grade. The probability of remaining a poor reader at the end of fourth grade given a child was a poor reader at the end of first grade was .88; while the probability of becoming a poor reader in fourth grade, given at least average reading skills in first grade, was .12. These results suggest that "the poor first grade reader almost invariably remains a poor reader by the end of fourth grade" (p. 12). Similar findings regarding the persistence of learning problems were presented by Fletcher, Satz, and Morris (1984b) who found little improvement in problem readers between second and fifth grade.

Increasing Severity of Reading Difficulties

Reading problems not only persist, they may actually increase in severity over time. Stanovich (1986), describes an escalating chain of side effects wherein children who have difficulty learning to read tend to fall further and further behind their peers.

Poor readers, for example, quickly begin to be exposed to less text and fewer opportunities to learn. Juel (1988) reports that by the end of first grade, good readers in her study had seen an average of 18,681 words in running text in basal readers. In contrast, poor readers had been exposed to 9,975, or about half as many, words. This difference in exposure to print between good and poor readers grew larger with each grade. Further exacerbating the discrepancy are wide differences in the amount of reading done out of school. For example, in third and fourth grade, reading after school became frequent for good readers but not for poor readers.

Catchup

By the time a poor reader is identified as needing support services, a severe discrepancy already exists between the child's reading abilities and the reading abilities of his/her peers. The child then must acquire reading skills at a faster rate than peers if the discrepancy between the child's performance and that of his/her peers is to be reduced. The child, however, already has displayed a slower rate of progress in order for the discrepancy to develop in the first place and is not experiencing success in the curriculum at a rate that will allow optimal progress. Further compounding the problem, as the discrepancy between the child's and peers' performance increases over time, the child's rate of progress must also increase. For example, to "catch up" in a year at second grade, a hypothetical poor reader's rate of progress, or slope, must be 200% of peers while at fourth grade, to "catch up" in a year his or her slope must be 300% of peers!

Primary Prevention of Early Academic Problems

Educational outcomes have been found to be more favorable for children when learning problems are identified early (Satz, & Fletcher, 1988). For example, Strag (1972) found that nearly 82% of students could be brought up to normal classroom work when the diagnosis of "dyslexia" was made in the first two grades of school. In contrast, 46% of the "dyslexic" problems identified in the third grade and only 10-15% of those observed in fifth to seventh grades were remediated successfully.

What is needed to stem the tide of increasing incidence of reading difficulties among children is the primary prevention of early academic problems (PPEAP). Students at risk for later reading problems need to be identified and provided with effective services and support before reading difficulties reach the "problem" level. If students at-risk for later learning problems can be provided with effective, non-intrusive services and support in regular education settings before reading difficulties reach the problem level, a much smaller change in the slope of pupil progress is necessary. With early intervention shown (middle of kindergarten), the discrepancy between a student with reading problems and the average peer's performance is not severe and they need only progress at a much lower rate to maintain adequate progress.

Assessment

For PPEAP to be feasible, there must be reliable and valid assessment procedures to identify those students who are "at-risk" for later reading problems. Currently, standardized, norm-referenced tests and teacher ratings are the procedures most often used for the early identification of reading problems (Lindsay, & Wedell, 1982; Mercer, Algozzine, & Trifiletti, 1989). These procedures are used as the basis for decisions regarding classification, retention, and promotion of students. According to Meisels (1989), the use of standardized, norm-referenced tests and teacher rating scales for the classification of students qualifies them as "high-stakes" tests. Unfortunately, research has not supported the efficacy of these high-stakes identification procedures (Adelman, 1982; Fletcher, & Satz, 1984a; Keogh, & Daley, 1983; Lindsay, et al., 1982; Satz, et al., 1988).

Standardized, Norm-Referenced Tests

Norm-referenced tests of readiness, achievement and intellectual functioning are similar in content and focus on language skills, visual and auditory perception, motor skill, perceptual-motor functioning, and letter recognition (Lindsay & Wedell, 1982). The best tests have modest predictive validity, with correlation coefficients between tests used as predictive measures and outcome measures ranging from .30 to .60. Although statistically significant, these results do not permit the classification of individual children according to outcomes. In fact, use of these tests for classification has been found repeatedly to result in many false positive and false negative errors (Adelman, 1982; Satz & Fletcher, 1988).

Teacher Rating Scales

The rationale for teacher rating scales is that a teacher who works closely with a child over a long period of time should be a reliable and accurate evaluator of risk. While teacher rating scales typically assess a wide range of academic and behavioral indicators of risk similar to those assessed by standardized tests, they are less time consuming than standardized, norm-referenced tests (Glazzard, 1977). Teacher ratings have been cited as being "as good as the best available psychometric procedures" (Adelman, 1982, p. 258) and "the best predictors of subsequent success or failure in school" (Algozzine, & Ysseldyke, 1986, p. 395). Unfortunately, little data exists to support claims of the efficacy of teacher predictions (Fletcher & Satz, 1984; Satz & Fletcher, 1988). In a review of four studies comparing teacher-based and test-based predictions of risk, Satz and Fletcher (1988) found that although the overall hit rate was almost identical between test and teacher predictions (approximately 75%), teachers made fewer risk predictions and generally missed more of the

true positive cases. These findings are similar to those reported by Fletcher and Satz (1984) who found that teacher predictions resulted in missing 87% of severely disabled readers.

Slope of Pupil Progress as a Measure of Risk

A fundamental flaw in the use of both standardized, norm-referenced tests and teacher ratings for the identification of children at-risk for academic difficulties is their exclusive focus on level of student performance, rather than on the rate or slope or student progress. Traditional tests and rating scales measure how an individual is performing at a given point in time rather than how an individual is changing over time. Howell (1986) provides a good description of the necessity of assessing change rather than simply level of performance:

> Achievement is the product of learning, it is not learning itself. Students' achievement levels are determined by the rate at which they learn material and the amount of available learning opportunity or instructional time. When a student is not achieving as expected, it is due to inadequate learning rate or opportunity to learn. Therefore, most effective corrective interventions are based on decisions about increasing the learning rate and/or providing more opportunities to learn. These decisions should not be based on static measures of what a student knows, but on dynamic measures of how s/he responds during instruction. (p. 326)

In addition, standardized tests and teacher rating scales are generally administered only once or twice a year, providing an infrequent measure of student performance. Such infrequent measures of performance are insensitive to environmental variables and emphasize only summative (i.e., pre- and post-intervention) strategies that, at best, identify ineffective programs only when they have been completed and when it is too late for modification (Deno, et al., 1982; Fuchs, & Fuchs, 1986; Howell, 1986). According to Adelman (1982) "It is clear that no currently available procedures intended for large-scale use can claim to identify a large number of problems without making many false positive errors" (p. 257). Alternatives to "high-stakes" testing clearly are needed.

One low-stakes alternative that has been suggested by Meisels and others (e.g., Adelman, 1982; Lindsay & Wedell, 1982; Mercer et al., 1988) is the monitoring of students' progress over time. Rather than making "high stakes" decisions based upon a student's performance at one point in time, the monitoring of pupil progress allows for sequential decision-making. Assessments of the rate of change over time can be used to differentiate children who are making adequate progress in school from children who are not making adequate progress. Individual instructional programs then can be designed to enhance the efficacy of instruction for those children who are not making adequate progress. Interventions can be non-intrusive and non-stigmatizing and the decision regarding whether to intervene can be reexamined on a frequent basis. If a student is making adequate progress, intervention can be discontinued. With continued monitoring of progress over time, the decision to intervene can also be made at any point a student demonstrates low slope of progress. As such, "the concept of sequential decisions is fundamental, permitting fallible data and resulting decisions to be evaluated over time, and modified as necessary, in an iterative fashion" (Macmann, Barnett, Lombard, Belton-Kocher, & Sharpe, 1989).

The need for frequent and ongoing monitoring of pupil progress is especially relevant when one considers the range of skill levels and abilities with which children enter kindergarten and first grade. Most have experienced extremely different learning environments. As described by Anderson, Hiebert, Scott, & Wilkinson (1984):

> Children enter a typical kindergarten class with very different levels of knowledge about printed language, and instruction needs to be adapted for these differences. One or two children, and sometimes more, may already be able to read simple stories. A handful may be totally unfamiliar with such basic concepts as a word, a sentence, and a letter, and may not even know that to read you hold a book right side up and turn the pages from front to back. (p 31)

Based upon measures of level such as standardized, norm-referenced tests of readiness, a prediction could be made that one student is at risk for reading problems while another is not. It is more difficult to make a prediction about some students whose performance is close to the mean classroom performance; a prediction could be made that they are not at risk.

Children can be expected to progress in the school curriculum at differing rates, however. Even when they test at the same starting level, some children learn at a faster rate than others (Babad, & Budoff, 1974; Campione, & Brown, 1987). For young children in particular, level of

skills and rate of progress may be unrelated. Thus, children who are at-risk for academic failure may initially be indistinguishable from children who are not at-risk.

Assume, for example, that Gabby is a child with learning difficulty who has attended a high quality preschool program. Although Gabby might display the expected level of skill development upon entering kindergarten, he would display a very different rate of academic progress in kindergarten and first grade from Betty, a nonhandicapped peer. Conversely, assume that Alfred is a child from a non-stimulating environment who enters school with low skills initially. Alfred is a child who may make adequate academic progress when provided with the opportunity to learn. In each case, it is the rate or slope of pupil progress that differentiates the child at risk for learning problems from his/her peers, not the level of skill development. "The point to be emphasized is that previous experience or opportunity is an important contributor to children's performance. To infer stable child deficit from initial test score does a disservice to a substantial number of children" (Keogh & Daley, 1983, p. 12).

The ongoing measurement of the slope of children's progress removes the need to make predictions about a child's future performance on the basis of an assessment at a particular point in time. As stated by Mercer et al. (1988), "If progress is monitored, false positives will progress rapidly, meet exit criteria, and be replaced in an appropriate program. False negatives, on the other hand, will progress slowly, meet entrance criteria, and be placed in the special intervention program" (p. 186). Information from the ongoing monitoring of progress can also be used to assist the teacher in identifying effective and ineffective interventions and in designing responsive educational modifications for individual students.

Criteria for Assessment of Slope

To measure the slope of pupil progress, assessments are needed that are relevant to and sensitive to instruction over time, and developmentally appropriate (i.e., kindergarten and first grade populations). Each of these criteria will be addressed below.

Relevance to Instruction

To accurately monitor student progress, assessment procedures are needed that are relevant to instruction and intervention. The assessment procedures should sample adequately the content of what the teacher is teaching and should be linked to later reading skill. If procedures are not relevant to instruction/intervention and to later reading, measures of change may be invalid and conclusions drawn about progress may be erroneous.

Sensitivity to Instruction

Assessment procedures that are sensitive to instruction allow for the validation of instructional program changes on an individual basis. Even effective interventions do not work equally well for all children. Deno (1986) presents data demonstrating differential effects of the same intervention program for two different children. Given the same intervention introduced into the children's programs at the same time, one child's performance increased dramatically while the second child's performance improved very little. In fact, program changes positively affecting one student might negatively affect another.

Appropriateness for Kindergarten and First Grade

It is important that assessment measures used for monitoring progress have adequate range and variability for kindergarten and first grade children. Limited range and variability could result in assessment materials being too difficult for younger and lower functioning children and/or too easy for older and higher functioning children. Such floor and ceiling effects would render the assessment insensitive to differences in performance among kindergarten and first grade children.

Pilot Study

Unfortunately, assessment procedures that meet the above criteria currently do not exist. Although curriculum-based measurement reading (CBM-R) procedures have been shown to be technically adequate (i.e., reliable and valid) and effective in identifying students who are experiencing academic difficulties and monitoring progress in grades 2 through middle school (Deno, et al., 1982; Deno, Marston, Shinn, & Tindal, 1983; Shinn, & Marston, 1985; Shinn, Tindal, & Stein, 1988), existing CBM-R procedures have limited utility in the kindergarten and first grade. Student performance on CBM-R tasks is measured by having the students read a passage from their basal reader aloud for 1 minute and recording the total number of words read correctly. Most kindergarten children do not read; the median reading performance of beginning first graders is typically low. For example, Shinn (1988) describes local norms in which the median performance of first grade readers in a school district is 2 words read correctly per minute at the beginning of the year. Thus, at the kindergarten and first grade levels, current CBM-R tasks have floor effects and limited vari-

ability. As a result, current CBM-R measures are insensitive to differences in performance among kindergarten and first grade children and, in fact, are too difficult for young students to be monitored. Thus, existing CBM approaches are insufficient to meet the needs of PPEAP.

This pilot study examined the utility of slope estimates based on three curriculum-based measures of pre-reading skills (CBM-P) that permit frequent, repeated assessment. Subjects were 13 children between 4 and 5 years of age who were enrolled in a University-based early intervention program for language delayed kindergarten children.

Measures

The first two CBM-P measures, letter naming fluency and number naming fluency, are adaptations of CBM readiness tasks described by Marston and Magnusson (1988). These measures were selected because of the finding that letter and number identification are highly related to later school achievement (Simner, 1983). The third measure, picture naming fluency, measures expressive language skills, which play a critical role in achieving school success (Simner, 1983). Language deficits are a common characteristic of children who are later identified by schools as handicapped (Schiefelbusch, & Bricker, 1981). Therefore, development of adequate language skills is frequently a focus of early intervention programs for handicapped preschoolers (Scruggs, Mastropieri, Forness, & Kavale, 1988). Because measures for progress monitoring need to be administered frequently and repeatedly, measures were developed to meet the following criteria (Deno, 1985):

1. Are easy to administer by teachers, parents, and students.
2. Have many parallel forms that are frequently administrable to the same student.
3. Are time efficient.
4. Are inexpensive and easy to produce.
5. Are unobtrusive with respect to routine instruction.
6. Are simple to teach to teachers, parents, and children.
7. Are linked to target/goal competencies in the student's local curriculum.

Letter Naming Fluency

Two sets of the alphabet (one upper and one lower case) were randomized and printed in block form on an 8.5" by 11" sheet of white paper. Following standardized procedures, the examiner showed the sheet of letters to the student and directed the student to name as many letters as he/she could within one minute. The number of

Table 1. Correlation Coefficients Between Level and Slope of CBM-P Measures

CBM-P Measure	Number Level	Number Slope	Letter Level	Letter Slope	Picture Level
Number Slope	-0.00 (11)				
Letter Level	0.80** (10)	-0.13 (10)			
Letter Slope	-0.22 (10)	0.52 (10)	-0.58 (10)		
Picture Level	0.93** (11)	-0.34 (11)	0.74* (10)	-0.34 (10)	
Picture Slope	-0.20 (11)	0.64* (11)	-0.51 (10)	0.95** (10)	-0.14 (13)

*$p < .05$ **$p < .01$

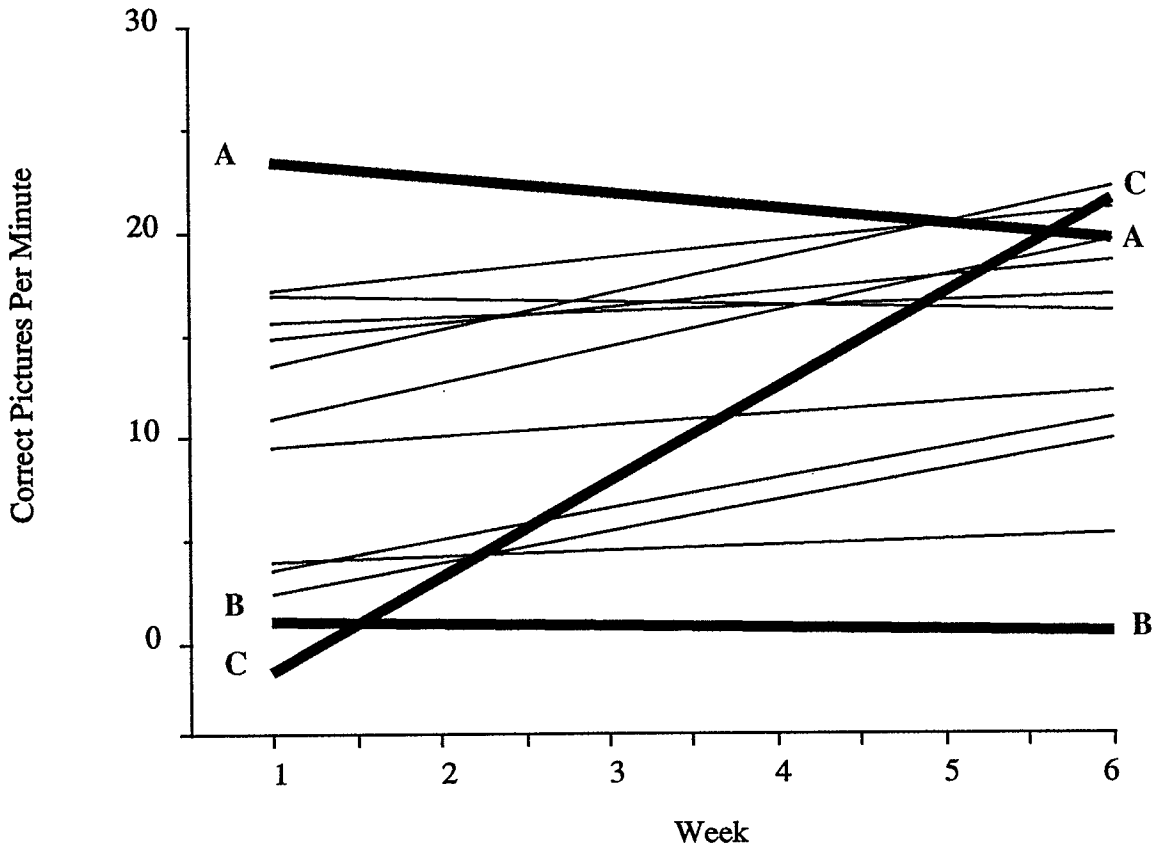


Figure 1. Skill Level and Slope of Pupil Progress for Picture Naming Fluency

correct letter names per minute was calculated for each letter naming task

Number Naming Fluency

A set of numerals from 1 to 20 were randomized and printed in block form on an 8.5" by 11" sheet of white paper. Following standardized procedures, the examiner showed the sheet of numbers to the student and directed the student to name as many numbers as he/she could within one minute. The number of correct numbers per minute was calculated for each number naming task.

Picture Naming Fluency

Pictures of words selected from the Harris-Jacobsen word list were randomized and arranged in block form on an 8.5" by 11" sheet of white paper. Following standardized procedures, the examiner showed the sheet of pictures to the student and directed the student to name as many pictures as he/she could within one minute. The number of correctly named pictures per minute was calculated for each picture naming task.

Procedures

All measures were administered to each subject two times a week over a six-week period. The rate of pupil progress was estimated by the slope of the least-squares regression line fit to the repeated measurements. Criterion measures included the Metropolitan Readiness Test (MRT), the Peabody Picture Vocabulary Test-Revised (PPVT-R), and the Test of Nonverbal Intelligence (TONI).

Results

The correlations among the CBM-P estimates of slope and level are reported in Table 1. With few exceptions, the estimates of level were significantly correlated, as were the estimates of slope. However, the slope estimates were not significantly correlated with the level estimates.

The independence of the slope and level estimates is illustrated in Figure 1 for the Picture Naming Fluency task: Level is portrayed by the height of the line, and the rate of pupil progress is portrayed by the slope of the line. Children with a low level of skills displayed both positive and

Table 2. Correlation Coefficients Between CBM-P Measures and Criterion Measures

	Criterion Measure		
CBM-P Measure	PPVT-R	TONI	MRT
Number Level	-0.02 (9)	0.11 (7)	0.78* (8)
Number Slope	-0.76* (9)	-0.43 (7)	-0.46 (8)
Letter Level	0.23 (8)	0.46 (6)	0.66 (7)
Letter Slope	-0.46 (8)	-0.32 (6)	-0.07 (7)
Picture Level	0.55 (11)	0.27 (7)	0.88** (8)
Picture Slope	-0.24 (11)	-0.41 (7)	-0.23 (8)

Note: PPVT-R = Peabody Picture Vocabulary Test - Revised. TONI = Test of Nonverbal Intelligence. MRT = Metropolitan Readiness Test.
*$p < .05$. **$p < .01$

negative slopes, as did children with a high level of skills.

Correlations with criterion measures are reported in Table 2. The MRT was significantly correlated with the Number Naming Fluency and Picture Naming Fluency estimates of level, but not with any of the estimates of slope.

Conclusions

Measures of the slope of pupil progress provided information that was not provided by the measures of level, including existing measures of risk (i.e., readiness and intellectual functioning). Thus, obtaining an estimate of the slope of pupil progress may contribute substantially to the evaluation of risk. In Figure 1, for example, students B and C are initially performing at the same level. However, student C is making adequate progress and is not at risk for academic problems. In contrast, student B is not making adequate progress and is at risk for academic difficulty. In fact, student A would appear not to be at risk based only on initial skill level; however, the negative slope provides reason for concern. These findings support the notion that level of skills and slope of progress may be unrelated to each other for young children and that risk may be defined best by the slope of student progress rather than the level of student skills.

In addition, the CBM-P measures were sensitive to the learning and growth that occurred during this six week study. Repeated assessments of skill were able to provide an estimate of the slope of pupil progress over the short period of time involved in this study. Procedures that are sensitive to learning are critical to the assessment of pupil progress.

Further research is indicated to determine if the CBM-P measures provide an accurate identification of children at risk for academic problems. In particular, it is important to investigate whether the CBM-P measures are related to meaningful outcome measures, whether they are sensitive to the effects of interventions, and whether improvement on the CBM-P measures is associated with a reduction of risk. Successful outcomes of this research will make a significant contribution toward the primary prevention of early academic problems. This will be accomplished by the

provision of valid "low stakes" assessment procedures that are characterized by monitoring the slope of student progress and sequential decision-making.

References

Adelman, H. S. (1982). Identifying learning problems at an early age: A critical appraisal. *Journal of Clinical Child Psychology, 11*(3), 255-261.

Algozzine, B., & Ysseldyke, J. E. (1986). The future of the LD field: Screening and diagnosis. *Journal of Learning Disabilities, 19*, 394-398.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1984). *Becoming a nation of readers: The report of the commission on reading*. Washington, D.C.: U.S. Department of Education.

Babad, E. Y., & Budoff, M. (1974). Sensitivity and validity of learning potential measurements in three levels of ability. *Journal of Educational Psychology, 66*, 439-447.

Campione, J. C., & Brown, A. (1987). *Linking dynamic assessment with school achievement*. New York: Guilford Press,

Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The use of standard tasks to measure achievement in reading, spelling, and written expression* (87). University of Minnesota, Institute for Research on Learning Disabilities.

Deno, S. L., Marston, D., Shinn, M. R., & Tindal, G. (1983). Oral reading fluency: A simple datum for scaling reading disability. *Topics in Learning and Learning Disability, 2*, 53-59.

Education, U. S. D. (1989). *Eleventh annual report to congress on the implementation of the Education of the Handicapped Act*. Washington, D.C.: Author.

Fletcher, J. M., & Satz, P. (1984a). Test-based versus teacher-based predictions of academic achievement: A three-year longitudinal follow-up. *Journal of Pediatric Psychology, 9*, 193-203.

Fletcher, J. M., Satz, P., & Morris, R. (1984b). *The Florida longitudinal project: A review*. New York: Praeger, 280-304.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199-208.

Gaddes, W. (1976). *Learning disabilities: Prevalence estimates and the need for definition*. Baltimore: University Park Press, 3-24.

Gickling, E., & Thompson, V. (1985). A personal view of curriculum-based assessment. *Exceptional Children, 52*, 153-165.

Glazzard. (1977). The effectiveness of three kindergarten predictors for first-grade achievement. *Journal of Learning Disabilities, 10*(2), 36-40.

Guttman, B., & Henderson, A. (1987). *A summary of state Chapter 1 participation and achievement information for 1984-85*. Washington, DC: Decision Resources.

Haberman, M. (1989). Thirty-one reasons to stop the school reading machine. *Phi Delta Kappan*, December, 284-288.

Howell, K. (1986). Direct assessment of academic performance. *School Psychology Review, 15*, 324-335.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437-447.

Keogh, B. K., & Daley, S. E. (1983). Early identification: One component of comprehensive services for at-risk children. *Topics in Early Childhood Special Education, 3*(3), 7-13.

Lindquist, G. T. (1982). Preschool screening as a means of predicting later reading achievement. *Journal of Learning Disabilities, 16*, 331-332.

Lindsay, G. A., & Wedell, K. (1982). The early identification of educationally "at-risk" children revisited. *Journal of Learning Disabilities, 15*, 212-217.

Macmann, G. M., Barnett, D. W., Lombard, T. J., Belton-Kocher, E., & Sharpe, M. N. (1989). On the acturial classification of children: Fundamental studies of classification agreement. The *Journal of Special Education, 23*, 127-149.

Mann, V. A., & Brady, S. (1988). Longitudinal prediction and prevention of early reading difficulty. *Annals of Dyslexia, 34*, 117-136.

Marston, D., & Magnusson, D. (1988). *Curriculum-based measurement: District level implementation*. Washington, DC: National Association of School Psychology, 137-172.

Meisels, S. J. (1989). High-stakes testing in kindergarten. *Educational Leadership*, April, 16-22.

Mercer, C. D., Algozzine, B., & Trifiletti, J. (1989). Early identification - an analysis of the research. *Learning Disability Quarterly, 11*, 176-188.

Rourke, B. (1978). *Reading, spelling, and arithmetic disabilities: A neuropsychological perspective*. New York: Grune and Stratton, 97-120.

Satz, P., & Fletcher, J. M. (1988). Early identification of learning disabled children: An old problem revisited. *Journal of Consulting and Clinical Psychology, 56*, 824-829.

Schenk, B. J., Fitzsimmons, J., Bullard, P. C., Taylor, H. G., & Satz, P. (1980). *A prevention model for*

children at risk for reading failure. Baltimore: University Park Press,

Schiefelbusch, R. L., & Bricker, D. (1981). *Early language: Acquisition and intervention.* Austin, TX: PRO-ED.

Scruggs, T. E., Mastropieri, M. A., Forness, S. R., & Kavale, K. K. (1988). Early language intervention: A quantitative synthesis of single-subject research. *Journal of Special Education,* 22(3), 259-283.

Shepard, L., & Smith, M. (1988). Escalating academic demand in kindergarten: counterproductive policies. *The Elementary School Journal, 89,* 135-145.

Shinn, M. R. (1988). Development of curriculum-based local norms for use in special education decision making. *School Psychology Review, 17,* 61-80.

Shinn, M. R., & Marston, D. (1985). Differentiating mildly handicapped, low-achieving and regular education students: A curriculum-based approach. *Remedial and Special Education, 6,* 31-45.

Shinn, M. R., Tindal, G., & Stein, S. (1988). Curriculum-based assessment and the identification of mildly handicapped students: A research review. *Professional School Psychology, 3,* 69-85.

Simner, M. L. (1983). The warning signs of school failure: An updated profile of the at-risk kindergarten child. *Topics in Early Childhood Special Education.*

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360-406.

Strag, G. A. (1972). Comparative behavioral ratings of parents with severe mentally retarded, special learning disability, and normal children. *Journal of Learning Disabilities, 10,* 52-56.

Wang, M. C., & Reynolds, M. C. (1985). Avoiding the "Catch 22" in special education reform. *Exceptional Children, 51,* 497-502.

Wang, M. C., Reynolds, M. C., & Walberg, H. J. (1987). State of the art and future directions of research, practice, and policy related to the education of students with special needs. *Educational Researcher, 16.*

Ysseldyke, J. E., & Marston, D. (1982). A critical analysis of standardized reading tests. *School Psychology Review,* (11), 257-266.

Analysis of a Grade 3-5 Elementary Math Curriculum for Production of Valid Math Tests: Work in Progress

Richard Parker
Ed Kameenui
Asha Jitendra
University of Oregon

Introduction

A recent review of research on effective teaching with low achieving students concluded that there is no "effective instruction" unique to this population (Christenson, Ysseldyke, & Thurlow, 1989). Drawing from earlier reviews (e.g., Walberg, 1984; Good and Brophy, 1986) and their own research, the authors identify ten critical factors for achievement of handicapped and non-handicapped learners in both regular and special education environments. Directly relevant to the present project are four factors which create a supportive framework for effective instruction, which is present when there is alignment or congruence of (a) the curriculum, (b) learning goals/objectives, (c) assessment, and (d) instruction.

Christenson et al. (1989) state that learning goals/objectives should be short-term, clearly articulated, and closely related to the curriculum. Assessment of student learning should provide frequent feedback, guide instruction, and be closely linked to both mastery goals/objectives, and the curriculum. Instruction should be oriented toward curriculum goals/objectives and should occur at those points in the curricular sequence where students can show regular improvement. Figure 1 depicts the supportive framework of Goals/Objectives, Curriculum, Assessment, and Instruction. The figure is an expansion of Nitko's (1989) discussion of "tripartite congruence" between instruction, objectives, and test items. In the figure, major influence among the four elements are represented as dark arrows; the light arrows describe minor or less frequent influence. The limited focus of this model necessitates excluding other influences on instruction such as teacher variables, class composition, and the program structure and resources (Smylie, 1988).

Unfortunately, Goals/Objectives, Curriculum, Assessment, and Instruction often are poorly aligned in elementary education (Nitko, 1989). Mismatches are common between Goals/Objectives and Instruction, between Assessment and Goals/objectives , and between Assessment and Curriculum. The aim of this project was to develop procedures for improving the alignment of the Goals/objectives, Curriculum, and Assessment in elementary mathematics within regular classrooms. Better alignment of these factors could improve performance of students at-risk and those with handicaps within regular classrooms.

Project Rationale

The arrows in Figure 1 depict three main influences on instruction: the curriculum, assessment results, and goals and objectives from within or outside the curriculum. Of these influences, assessment occupies a pivotal position, as it influences and is influenced by each of the other three variables. Assessment content and format should be strongly influenced by the curriculum used, the goals and objectives pursued, and the subject taught. The light arrows show lesser influence in the opposite direction; assessment results may influence selection/adaptation of curriculum materials, or modification of instructional goals. This project focused on only three of the critical variables—curriculum, assessment, and goals/objectives. The study did not attempt to account for the fourth variable—instruction. It was expected that improving alignment of the other three variables would provide at least a supportive framework for effective instruction. The approach taken for aligning curriculum, assessment, and goals/objectives was to develop and apply a single taxonomy for classifying both curriculum objectives and activities, and test items. That approach is not new; Instructional Quality Inventory (IQI) procedures developed for U.S. military training (Merrill, Reigeluth & Faust, 1979) have been successfully applied to public school curricula and

assessment (Roid & Haladyna, 1982). In addition, three mature methodologies currently exist to help improve congruence among elements in the above model: CurriculumAnalysis, Item Construction, and Test Construction.

After a brief review of relevant literature, this paper will describe procedures currently being piloted for integrated curriculum analysis, item creation, and test construction in elementary mathematics. As the title states, the development project is ongoing. This report offers little hard data; instead, the focus is describing procedures and their utility.

Mismatch Problems

Assessment Mismatch

No single national math curriculum can be defined by tests or basal programs. Content analyses of five nationally standardized math tests and four major basal math programs demonstrated wide differences at the level of specific objectives (Freeman, Kuhs, Porter, Floden, Schmidt, & Schwille, 1983). These differences prevent standardized test scores from being interpreted in a straightforward manner, as they represent "opportunity to learn" as well as actual student learning (Romberg & Carpenter, 1986).

Another type of misalignment occurs when achievement is tested through inappropriate item presentation and response formats. Most standardized math tests rely on the multiple choice response format, although it seldom occurs in instruction and curriculum-based assessment (Murnane & Raizen, 1988; Alexander & James, 1987). Problem-solving applications especially require free response formats; multiple choice selections reflect different abilities. Other disadvantages of the multiple-choice test format are its influences on how teachers present math content, and on how students study for tests (Frederiksen, 1984; Kirkland, 1971).

The mismatch between curriculum and assessment has two undesirable results with standardized tests being (a) relatively insensitive to achievement within a particular curriculum, causing consistent underestimation of student improvement (Porter, Schmidt, Floden, & Freeman, 1978), and (b) standardized tests are differentially sensitive to achievement in different curricula. As a consequence, successes obtained by an instructional program may be overlooked—simply not measured. In evaluating competing instructional programs, an unfair edge will be obtained by the program with content which overlaps most with that of the test (Airasian & Madaus, 1983). Because standardized achievement tests are "content-biased" (Schmidt, 1983), they permit and encourage unfair evaluations of instructional programs.

Curriculum Mismatch

Although teachers often are allowed to depart from the content and sequence of basal texts, they seldom do so (Stake & Easley, 1978; Stephens,

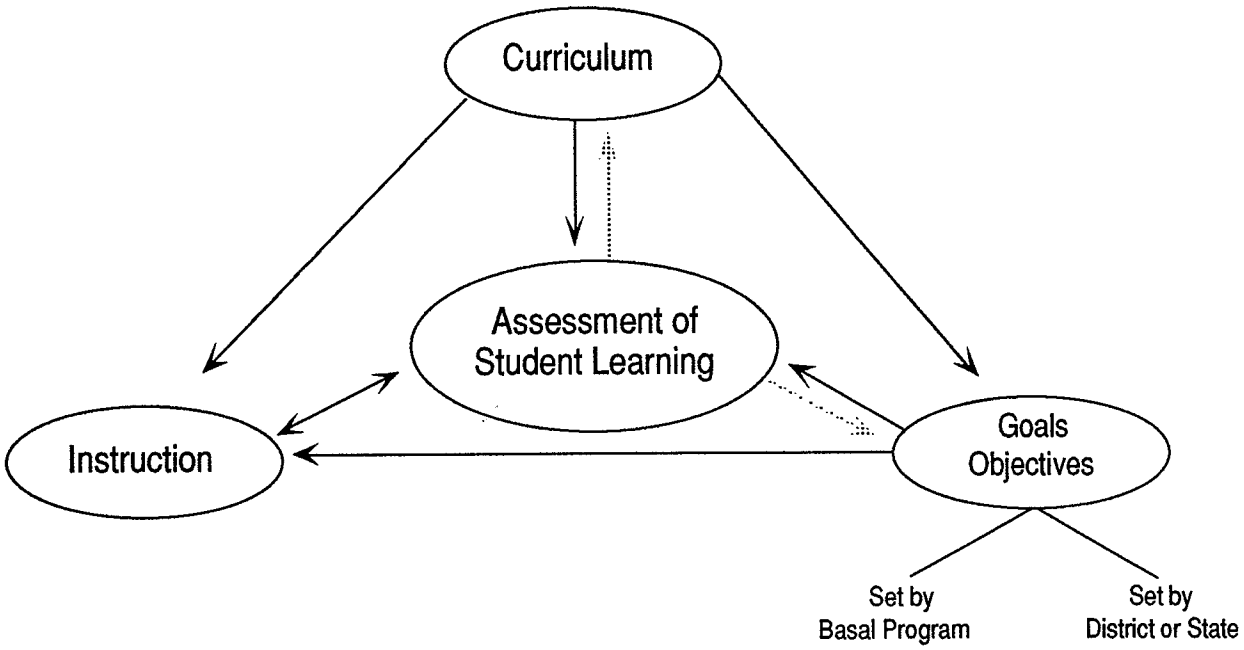


Figure 1. Influences on Instruction

1982). The math textbook is perceived by teachers as "the authority on knowledge and the guide to learning" (Romberg & Carpenter, 1986). However, basal texts often suffer from problems of internal misalignment. Their listed "scope and sequence" objectives may be too broad, too ambiguous, or simply too inaccurate to reflect prescribed lesson activities (Popham, 1984; Roid & Haladyna, 1982). It is also well-known that "different curricula are associated with different patterns of achievement" (Walker & Schaffarzick, 1974). A second major problem is that learning activities often provide only exposure, not measurable skill growth toward mastery: "a very large percentage of the topics taught receive only brief, perhaps cursory, coverage" (Porter, 1989, p. 12). When teaching for "exposure" and "review" replace teaching for skill development and mastery, the alignment of goals/objectives, curriculum, and instruction becomes tenuous. Furthermore, assessment lacks a satisfactory foundation—should tests be based on goals/objectives or on actual activities? Because of the difficulties in assessment, accountability for student learning is also problematic (Nitko, 1989; Porter, 1987).

Another type of curriculum mismatch is that between basal learning activities and learning objectives mandated by outside authorities on social/political bases (Nitko, 1989; Jaeger, 1989). The extent to which state and district level core curriculum goals or competencies appear to be achieved will vary according to the particular basal program in use (Freeman, et al., 1983). An unbiased or "curriculum-fair" test can be developed, but it must be based upon detailed curriculum analyses to identify common content areas. Besides the need for logical justification through curriculum analysis, an unbiased test must also be justified empirically. After administering the test to students instructed through different basal programs, individual item-types may be examined for curriculum bias in the same way that racial and sexual bias is assessed (Cole & Moss, 1989).

Curriculum Analysis

Although in this project curriculum analysis is used to assist in test development, it may serve a range of other purposes, including basal comparisons, identifying "friendly" texts, adapting texts for slow learners, and rewriting deficient instructional sequences. The increasing popularity of curriculum analysis can be ascribed to three recent trends. First, the basal text has been identified as a strong influence on classroom instruction in the curriculum areas of reading, mathematics, social studies, and science (Durkin, 1978-79; Komoski, 1985). Second, technical advances in criterion referenced testing (CRT) have provided empirical support for analyses of specific curriculum areas (Hsu & Yu, 1989; Nitko, 1989). Finally, theory and research have recently provided new insights into the cognitive demands of these curriculum areas (Kameenui & Griffin, 1989; Snow & Lohman, 1989).

Curriculum analysis helps define a curriculum-referenced domain of behaviors for CRT test construction and a strategy for item sampling (Popham, 1984; Nitko, 1980). At least three curriculum features must be tabulated and described for valid test construction: (a) the subject content presented, (b) the activities and required student performance (cognitive and behavioral), and (c) the points where they occur in the curriculum. The first two features assist in item construction, and the third in systematic item sampling for test construction (Roid & Haladyna, 1982). The second feature specifies important characteristics of learning activities: presentation formats, learner response modes, and required cognitive operations or "reasoning" (Hively, Patterson, & Page, 1968; Osburn, 1968).

At least three different curriculum taxonomies have been developed recently. In mathematics education, there is considerable agreement on 12 major content areas (Denmark & Kepner, 1980). By crossing content areas with levels of cognitive process (Bloom, Hastings, & Madaus, 1971) a taxonomy or item classification matrix can be created, one of which has served as the basis for the National Assessment of Educational Progress (NAEP) math tests (1983). An alternative math content description has been offered by Glennon and Wilson (1972), consisting of seven hierarchically organized domains. These domains have been crossed with Williams and Haladyna's (1982, pp 161-173) LOGIQ matrix (including level of abstraction, intellectual operation, and response mode) to create a complete test item typology (Tindal, 1989). A third alternative taxonomy for math content analysis has been developed at the Institute for Research on Teaching at Michigan State University (Kuhs, Schmidt, Porter, Floden, Freeman, & Schwille, 1979). The three dimensional taxonomy for test items includes "the general intent of the item (e.g. conceptual understanding or application), the nature of content presented to students (e.g. fractions or decimals), and the operation the student must perform (e.g., estimate or multiply)" (Freeman, Kuhs, Porter, Floden, Schmidt, & Schwille, 1983, p. 502). The taxonomy

proved reliable in application to both basal texts and standardized tests.

Item Creation and Indexing

Curriculum analysis provides information on curriculum content, characteristics of learning activities, and their location and focus within the curricular sequence. This information is required for the next two procedures—item construction and test construction. With few modifications, the taxonomy used in curriculum analysis should also serve as a typology for item creation, which requires specification of the subject content, the presentation and response formats, and the reasoning or cognitive operations involved in task completion. These specifications are essential to "amplified" behavioral objectives (Popham, 1978), items construction rules (Bormuth, 1970; Millman, 1980), or item templates to guide item creation (Nitko, 1980).

Once items are created, they can be classified according to these same characteristics and strategically selected to create tests for different purposes. For example, items could be selected by content: "word problems involving subtraction of 2 to 4 digit numbers, with regrouping." Items could also be selected by a combination of presentation and response formats: "math problems presented verbally, requiring a written response within a time limit, and without scratch pad or calculator." It would also be possible to select items by cognitive operations: "items requiring only rote memory." Finally, items could be selected according to any sensible combinations of these characteristics.

A curriculum analysis alone does not guarantee good items, as item writing requires both familiarity with the particular basal program and expertise in the content area. Specific item-writing skills have been identified by Roid and Haladyna (1982), Hambleton and Eignor (1978) and Haladyna and Downing (1989). Roid (1989) recommends either hiring expert item writers or conducting a summer workshop for a group effort, and requiring item critique and interchange between content experts and psychometric experts.

Test Construction

Curriculum-based tests are used with students at various points in the curriculum, who have demonstrated varying degrees of content mastery, and for whom the published curriculum goals/objectives, lesson activities, or instructional materials may have been adapted. Teachers therefore need to flexibly select items by content, lesson, unit, grade level, instructional material (basal text, supplementary practice book, etc.), or a combination of these. Although skill sequences are important, we know that the lesson and unit sequence of a basal program is the primary consideration in lesson planning for most teachers.

Teachers also plan lessons on the basis of skill content, but usually from within the basal curriculum. Therefore, they need to know the instructional "focus," where in the curriculum sequence particular skills are introduced, practiced, and reviewed. This information allows a smaller number of items to efficiently cover a broad segment of the curriculum. For test production, information on curriculum location and focus may be combined with the informationon subject content, task format, response mode, and cognitive level (Hambleton & Swaminathan, 1985; Stone, 1989).

Item selection for test production is increasingly accomplished by electronic database. Items are "drawn" on computer and individually saved as small graphic files. The items are indexed according to curriculum content, activity characteristics, curriculum location and focus, etc. These characteristics are used as search codes for strategic item selection from the database to create tests for various purposes. Recent reviews of item banking technology are provided by Roid (1989) and Stone (1989). Practical limitations to computer-assisted local test development are described by Hiscox (1985) and Millman and Arter (1984).

Method

This project involves integrated curriculum analysis, item creation and indexing, and test production in elementary mathematics for the purpose of producing sensitive and valid curriculum-based tests. The *Open Court Math* basal program for grades 3, 4, and 5 was the object of the curriculum analysis. This program is being used by approximately 220 students at each of the three grade levels, across thirty-one classrooms in six elementary schools in a Northwest district.

Curriculum Analysis

All Grade 3, 4, and 5 *Open Court Math* program lessons (140 per grade level) were analyzed, utilizing mainly the teacher's manuals, which included reprints of student workbook pages. Individual activities were coded using a taxonomy for (a) *Curriculum Content*, (b) *Curriculum Location and Focus*, and (c) *Activity Characteristics*. The coding categories of the taxonomy are presented in Table 1.

In addition to the above taxonomy, lesson activities were classified according to the State of

Oregon's 73 core math competencies established for Grades 3-5 (Oregon Department of Education, 1987). Activity coding was completed by four graduate students in education and required approximately four minutes per lesson, or nine hours per grade level. Only those activities were coded which had clear performance expectations for students. Lectures or discussions which did not contain or result in explicit student responses were excluded. On the average, each lesson yielded about five scorable activities.

Initial reliability estimates were calculated across the raters for a random sample of 20 activities selected across the three grade levels. Pearson's Phi was used as an index of strength of association for categorical data. Cramer's V scales Phi from 0 (no association) to +1 (perfect association) (Hays, 1981). Cramer's V reliability coefficients were: *Grade level, Lesson #, and Basal Source, 1.00; Task Intent, .59; Cognitive Level, .55; Task Format, .73; Response Mode, .69; Response Type, .70; Cues/Hints, .73.* Consistency in coding Curriculum Content was calculated separately for three raters on 140 items, using a "Proportion Agreement Index" which considers both occurrences and non-occurrences (Suen & Ary, 1989). The resulting score for Curriculum Content was 71% agreement. After initial reliability estimates were calculated, individual categories were further clarified. Work continues on clarifying the taxonomy and improving guidelines for curriculum coders.

All data from the curriculum analysis were computer-analyzed on an electronic spreadsheet. Cross-tabulating the data provided frequency and proportion summaries at Unit, Grade, and Cross-grade Levels, for *Curriculum content,* for five *Curriculum Location and Focus* variables, and for five *Activity Characteristics* variables. Both simple independent summaries and more complex conditional summaries were conducted. (Space limitations prevent display of results here, but they are available on request from the second author.)

Item Creation & Indexing

Once the curriculum had been analyzed, the development team created items for pilot-testing. The first test to be constructed was a broad, survey instrument for measuring student growth over the entire year's curriculum through fall, winter and spring assessments. It was originally predicted that the *Activity Characteristics* and *Curriculum*

Table 1. Coding Categories

Curriculum Content	Curriculum Location and Focus	Activity Characteristics
a. Numeration	• Grade Level	Cognitive Level
b. Addition		a. Memorey/Rote Learning
c. Subtraction	• Lesson Number	b. Skill/Procedural
d. Multiplication		c. Conceptual understanding
e. Division	• Basal Source	d. General understanding
f. Multiple/Basic Facts	a. Student Pages	e. Problem-solving applications
g. Multiple/Multi-digit	b. Mental math	Task Format
h. Fractions	c. Thinking Story	a. Mental Math
i. Decimals	d. Demonstration	b. Paper/Pencil
j. Word Problems		c. Manipulative
k. Measurement	• Task Intent	d. Discussion
l. Geometry	a. Introduced	Response Mode
m. Percent	b. Practiced	a. Oral
n. Applications	c. Reviewed	b. Written
o. Algebra		c. Show/Demonstrate
p. Relationships	• Assessment	Response Type
q. Reasoning	a. Unit Test	a. Selection
r. Statistics	b. Review Test	b. Production
s. Calculations		Cues/Hints
		a. Examples from problem
(115 Content Codes within the above 19 main categories)		b. Pictures

Content information would be sufficient for the creation of individual test items. However, some of the curriculum coding data proved more valuable than others in item creation.

Cognitive Level

Cognitive Level was valuable as a screener for identifying certain general understanding and memory/rote learning activities which could not be assessed in a group paper and pencil test. Skills, problem-solving applications, and conceptual understanding proved to be very useful descriptive categories; the latter two were later integrated into the *Curriculum Content* codes.

Task Format

Task Format proved to be of little use in item construction. The *discussion* and *mental math* categories provided information already available in other categories, and *manipulative* activities had to be omitted or converted to*paper/pencil* exercises for the group survey test.

Response Mode & Type

Response Mode was also of little use because it provided redundant information, and the oral and show/demonstrate descriptors applied to activities which could not easily be group-tested. Both *Task Format* and *Response Mode* could be useful for a more thorough, exploratory curriculum analysis (also in progress), but not for item construction. *Response Type* coding produced little variation; nearly all activities in the basal required *Production* responses. Only *Production* responses were included in the group test.

Cues/Hints

Cues/Hints were of little use in preparing a Grade-level survey test with the scope of a full year. These data would be more useful in constructing items for narrower diagnostic skills tests or Unit-level tests. Most lesson activities which included cues or hints for students were followed by activities where these cues/hints were removed. Therefore, no cues/hints were included with items prepared for the survey test; instead, Items represented more terminal performance.

In summary, item construction relied mainly on *Curriculum Content* and *Cognitive Level* data. As well as relying on these two sources, it was necessary to skim the actual lessons. Items were produced in two stages. They first were sketched in 2x3 inch boxes, and then were created as individual graphic files on a personal computer. The inclusion of diagrams, special symbols, and pictures in the items resulted in items which were faithful to the curriculum (Stone, 1989).

Test Production and Scoring

The frequency and proportion data from the curriculum analysis summarized on the electronic spreadsheet helped ensure that the survey test closely reflected the curriculum in content and emphasis. Items were selected for each Grade-level survey test according to the relative frequency of occurrence of *Curriculum Content* and *Cognitive Level* categories. Other more detailed information on *Task Intent (introduced, practiced, reviewed) and Lesson #* was not required, as each test spanned a full year of work.

The survey test produced for each grade level was designed for untimed administration during three sessions of approximately 20-30 minutes. The number of items per test were: Grade 3, 155; Grade 4, 147; and Grade 5, 147. Tests were administered to approximately 250 students at each grade level. Analyses of test protocols provided two types of results: a summary test score for each student, and item-specific information. Summary scores could be provided efficiently (two weeks), while item analyses were more time consuming.

For the summary scores, deciles were chosen instead of percentiles to acknowledge the lack of test precision prior to item screening. The item analysis (classical and one-parameter Rasch analysis) was performed using BILOG® PC software (Mislevy & Bock, 1984). Summary scores on survey tests provide little guidance for teachers' instructional decisions. However, individual item analysis provides three useful functions: (a) identifying and screening out ambiguous and non-discriminating items, (b) providing scores for several item clusters which are tied to separate topics or objectives, and (c) providing both *norm-referenced* (percentile rank among grade peers) and *criterion-referenced* or mastery (percent of problems correct) information for each student on these item clusters (Reckase, 1989). These results do offer clear instructional guidance when the tests are congruent with the curriculum activities and learning objectives. Item analysis also permits the production of multiple skill probes which have sufficient reliability that they can be administered frequently to monitor skill growth.

Item analysis methods also have their disadvantages. They are time-consuming, requiring that item scores for each student be computer-input. In addition, the more powerful IRT analyses require large sample sizes (200 to 500 students minimum), a normal ability distribution, and tests or subtests which deal with only a single ability. The benefit of item analysis is that it may need to be conducted

only once per item. Use of the item in subsequent tests will benefit from the same initial item information, especially if IRT analyses were conducted (Hambleton, 1989). In this project, mastery information on item clusters is currently being prepared for return to teachers.

Conclusions

An obstacle to long-term improvement of instruction is the lack of congruence of learning goals/objectives, lesson activities, and classroom tests. This paper has described the implementation of integrated curriculum analysis, item creation, and test construction for elementary mathematics. Instead of relying on the publisher's list of objectives or scope-and-sequence, each lesson was examined and coded according to a taxonomy that included *Curriculum Content, Curriculum Location and Focus* variables, and *Activity Characteristics.* The same taxonomy was used to create, index, and select test items for a grade-level survey test.

Curriculum coding with taxonomy categories was completed efficiently and with moderate intercoder reliability. In applying the taxonomy it was apparent that some categories should be moved from *Activity Characteristics* to *Curriculum Content* for more efficient coding. Several of the *Activity Characteristics* which were coded proved to be of little use in creating test items, although they could be valuable in a more complete curriculum analysis conducted for a different purpose.

The curriculum analysis provided data for three subsequent activities: (a) item creation, (b) item indexing, and (c) item selection to produce a test for a particular purpose. The taxonomy was not sufficient for item-creation; it also was necessary to scan the actual lessons in a teacher's manual to obtain a sufficiently clear idea of the activity to produce items. The taxonomy did appear adequate for item indexing, given decision-rules for multiple classifications. The taxonomy also appeared adequate for item selection, although such a judgment needs to be validated over a range of testing purposes. The feasibility of item selection for different testing purposes also depends upon obtaining item-specific information from classical or IRT analyses, which currently are being conducted. Item analyses that are skill-specific and are referenced both to the student's grade level norm group and to (percentile rank) and to skill mastery (percent correct) are being used to produce results for teachers. The classroom utility of these results has not yet been assayed.

References

Airasian, P. W. & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20,* 103-118.

Alexander, L., & James, H.T. (1987). *The nation's report card.* Cambridge, MA: National Academy of Education.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook of formative and summative evaluation of student learning.* NY: McGraw-Hill.

Bormuth, J. R. (1970). *On the theory of achievement test items.* Chicago: University of Chicago Press.

Christenson, S.L., Ysseldyke, J.E., & Thurlow, M.L. (1989). Critical instructional factors for students with mild handicaps: An integrative review. *Remedial and Special Education, 10*(5), 21-31.

Cole, N. S., & Moss, P. A. (1989). Bias in test use, In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-220), NY: Macmillan, Inc.

Denmark, T., Kepner, H. S., & Jr. (1980). *Basic skills in mathematics: A survey. Journal for Research in Mathematics Education,* (March), 104-123.

Durkin, D. (1978-79). What classroom observation reveals about reading instruction. *Reading Research Quarterly, 14,* 481-533.

Durkin, D. (1981). Reading comprehension instruction in five basal reader series. *Reading Research Quarterly, 16,* 515-544.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 193-202.

Freeman, D.J., Kuhs, T. M., Porter, A.C., Floden, R.F., Schmidt, W.H., & Schwille, J.R. (1983). Do textbooks and tests define a national curriculum in elementary school mathematics? *The Elementary School Journal, 83*(5), 501-513.

Glennon, V. J., & Wilson, J. W. (1972). Diagnostic-prescriptive teaching. In *National Council of Teachers of Mathematics, The slow learner in mathematics: The 35th yearbook.* Reston, VA: NCTM.

Good, T., & Brophy, J. (1986). School Effects. In Merlin C. Wittrock (Ed.), *Handbook of research on teaching* (3rd Edition). (pp. 570-602). New York: Macmillan Publishing Company.

Haladyna, T.M., & Downing, S.M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement In Education, 2,* 37-50.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L.

Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200, NY: Macmillan.

Hambleton, R. K., & Eignor, D. R. (1978). *Competency test development, validation, and standard-setting*. Paper presented at the Minimum Competency Testing Conference of the American Education Research Association.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hays, W. L. (1981). *Statistics*. New York: Holt, Rinehart and Winston.

Hiscox, M. D. (1985). Computer-based testing systems: Much ado about nothing? *Educational Measurement: Issues and Practice, 4*(1) 27-28.

Hively, W., Patterson, H. L., & Page, S. A. (1968). A "universe defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

Hsu, T., & Yu, L. (1989). Using computers to analyze item response data. *Educational Measurement: Issues and Practice, 8*(3), 21-28.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 147-200) New York: Macmillan.

Kameenui, E. J., & Griffin, C. C. (1989). The national crisis in verbal problem solving in mathematics: A proposal for examining the role of basal mathematics programs. *The Elementary School Journal, 89*(5), 575-593.

Kirkland, M. C. (1971). The effects of tests on students and schools. *Review of Educational Research, 41*, 303-350.

Komoski, P. K. (1985). Instructional materials will not improve until we change the system. *Educational Leadership, 12*, 31-37.

Kuhs, T., Schmidt, W., Porter, A., Floden, R., Freeman, D., & Schwille, J. (1979). *A taxonomy for classifying elementary school mathematics content* (Research Series No. 4). East Lansing: Institute for Research on Teaching, Michigan State University.

Merrill, M. D., Reigeluth, C. M., & Faust, G. W. (1979). The instructional quality profile: A curriculum evaluation and design tool. In H. F. O'Neil, Jr. (Ed.), *Procedures for instructional systems development*. NY: Academic Press.

Millman, J., & Arter, J. A. (1980). Issues in item banking. *Journal of Educational Measurement, 21*, 315-330.

Millman, J. (1984). Individualized test construction and administration by computer. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 78-96). Baltimore, MD: Johns Hopkins University Press.

Mislevy, R. J., & Bock, R. D. (1984). *BILOG: Maximum likelihood item analysis and test scoring with logistic models*. Mooresville, IN: Scientific Software.

Murnane, R. J., & Raizen, S. A. (1988). *Improving indicators of the quality of science and mathematics education in grades K-12*. Washington, DC: National Academy Press.

Nitko, A. J. (1980). Distinguishing the many varieties of criterion-referenced tests. *Review of Educational Research, 50*(3), 461-485.

Nitko, A. J. (1989). Designing tests that are integrated with instruction, In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 447-474), NY: Macmillan.

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement, 28*, 95-104.

Popham, J. (1978). As always, provocative. *Journal of Educational Measurement, 15*, 297-300.

Popham, W. J. (1984). Specifying the domain of content or behaviors. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 29-48). Baltimore: Johns Hopkins University Press.

Porter, A. C. (1989). A curriculum out of balance: The case of elementary school mathematics. *Educational Researcher, 18*(5), 9-15.

Porter, A. C. Schmidt, W. H. Floden, R. E., & Freeman, D. J. (1978). Practical significance in program evaluation. *American Educational Research Journal, 15*, 529-539.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practices, 8*(3), 11-16.

Roid, G., & Haladyna, T. (1982). *A technology for test item writing*. NY: Academic Press.

Roid, G. H. (1989). Item writing and item banking by microcomputer: An update. *Educational Measurement: Issues and practices*, 8(3), 17-20.

Romberg, T. A., & Carpenter, T. P. (1986). Research on teaching and learning mathematics: Two disciplines of scientific inquiry (pp. 850-873). In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.), NY: Macmillan.

Schmidt, W. H. (1983). Content bias in achievement tests. *Journal of Educational Measurement, 20*, 165-178.

Smylie, M. (1988). The enhancement function of staff development: Organizational and psychological antecedents to individual teacher change. *American Educational Research Journal, 25*(1), 1-30.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 447-474), NY: Macmillan.

Stake, R., & Easley, J. (Eds.) (1978). *Case studies in science education.* Urbana: University of Illinois.

Stephens, W. (1982). *Mathematics knowledge and school work: A case study of the teaching of developing mathematical processes* (Project Report). Madison: Wisconsin Center for Education Research.

Stone, C. A. (1989). Testing software review: MicroCAT Version 3.0. *Educational Measurement: Issues and Practices, 8*(3), 33-38.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tindal, G. (1989). Evaluating the effectiveness of educational programs at the systems level using curriculum-based measurement. In M. Shinn (Ed.), *Curriculum-based measurement.* New York: The Guilford Press-School Practitioner Series.

Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership, 41*, 19-30.

Walker, D., & Schaffarzick, J. (1974). Comparing curricula. *Review of Educational Research, 44*, 83-111.

Williams, R. G., & Haladyna, T. M. (1982). Logical operations for generating intended questions (LOGIQ): A typology for higher level test items (pp. 161-185). In G. Roid & T. Haladyna (Eds.), *A technology for test item writing,* New York: Academic Press.

Prediction of Reading Growth in Low-Performing Students

Stephanie Stein
Central Washington University

Overview

Frequent monitoring of academic progress has been accepted widely as an important component of instructional programs for low performing students (Mirkin & Deno, 1979; Howell, 1986; White & Liberty, 1976). The importance of progress monitoring also has been recognized formally through federal legislation. Public Law 94-142 mandates that every handicapped student must have an Individualized Education Program (IEP) that specifies short-term objectives and long-term goals for the student. The law further requires that some method for ascertaining student progress towards these objectives and goals be employed.

Many of the current methods used to monitor the progress of low-performing students, such as standardized achievement tests and teacher judgment, have problems that limit their effectiveness as accurate progress monitoring tools. Some of the problems with using standardized achievement tests to monitor student progress are the tests' insensitivity to student growth (Carver, 1974), their lack of content validity, and the unreliable difference in scores obtained through pre-posttest comparisons of student performance (Salvia & Ysseldyke, 1985). Likewise, the practice of using teacher judgment to monitor student progress often results in inaccurate assessments, especially in making specific judgments about students' reading problems (Stern & Shavelson, 1983) and the achievement of low-performing students (Coladarci, 1986).

In an effort to overcome many of the weaknesses of the current ineffective progress monitoring procedures, educators and researchers have proposed numerous models that rely on the use of direct and frequent measurement of student progress over time (Blankenship, 1985; Deno & Mirkin, 1977; Glickling & Thompson, 1985; Lindsley, 1964; 1971; White & Haring, 1980). Gathering of time-series data through direct and frequent measurement has proven to be a reliable and valid procedure for repeatedly indexing a student's level of performance in curriculum areas such as reading (Deno, Mirkin, & Chiang, 1982), spelling (Deno, Mirkin, Lowry, & Kuehnle, 1980), and written expression (Deno, Marston, & Mirkin, 1982).

In the current professional literature, direct and repeated measurement commonly is referred to as Curriculum-Based Measurement or CBM (Germann & Tindal, 1985; Tucker, 1985), a term that grew out of the research on alternative measurement and evaluation procedures at the University of Minnesota Institute for Research on Learning Disabilities. Curriculum-Based Measurement procedures have many desirable characteristics of a progress monitoring system, including relevance, sensitivity to growth, flexibility, ease of administration and availability of alternate forms (Jenkins, Deno, & Mirkin, 1979). Research on the use of CBM procedures by special educators has shown that teachers can gather and graph accurately the results of time-series data on academic performance (Skiba, Wesson, & Deno, 1982). In addition, CBM procedures take up very little instructional time once educators are trained in the efficient administration of these measures (King, Wesson, & Deno, 1982; Marston & Magnusson, 1985).

However, research also has shown that teachers rarely use the student performance data gathered on a regular basis to make instructional changes in students' programs (King et al., 1982; Skiba et al., 1982; Skiba, Marston, Wesson, Sevcik, & Deno, 1983). Skiba et al. (1982) hypothesized that teachers don't make better use of CBM data because they are unsure how to analyze and evaluate the student progress information. Similarly, a survey by King et al. (1982) of teachers who are familiar with CBM procedures, but are reluctant to use them, has shown the main inhibiting factors to be a lack of knowledge on how to implement the procedures and a concern that they may be too time-consuming to utilize on a daily basis.

The purpose of this study was to investigate characteristics of time-series progress monitoring data gathered through direct and frequent measurement of reading performance by elementary students. In particular, I wanted to determine whether (a) the frequency of measurement and (b) the number of passages sampled per administra-

tion during monitoring had differential effects on the accuracy of predicting short- and long-term student reading achievement. In addition, I examined two different methods of analyzing progress monitoring data to determine which was more accurate in predicting student reading performance over different lengths of time.

Method

Subjects

The subjects in this study consisted of 46 fourth and fifth grade elementary students who were receiving Chapter 1 or special education services in reading. Two schools in an urban school district and two schools in a semi-rural school district participated in the study. Approximately two-thirds of the subjects were male. The students were split almost equally across the two grades, (57% fourth graders and 43% fifth graders) and the two districts (urban=22, rural=24). About half of the subjects were identified as Chapter 1 students (54%) and half as Learning Disabled (46%).

Data Collection Procedures

To determine the appropriate level of long-range goal (LRG) material for monitoring progress, students were asked to orally read passages taken from several different levels of the basal reading text used in their schools. The students read for 1 minute per passage from two randomly selected passages from each reading level administered. The number of words read correctly and number of errors were averaged across the two passages at each level. The LRG level for each student was defined as the highest level in the curriculum in which the student read from an average of 40 to 80 correct words per minute with 8 or fewer errors. I selected these criteria to ensure that the material was easy enough to allow progress to occur, but difficult enough that the student still had room to improve. An error limit was used to ensure a minimum level of reading accuracy.

The subjects were assigned randomly to one of two data collection conditions. In Condition #1, the students were monitored once a week with three randomly selected passages from LRG material. The students read out loud from each passage for 1 minute and the number of words read correctly for each passage was recorded. The median number of words read correctly from the three passages then was recorded on an equal-interval graph for the individual student. This procedure was repeated once a week for 12 weeks.

The students in Condition #2 were monitored 3 times a week with one passage each day selected randomly from LRG material. The correct number of words read out loud in 1 minute was recorded on the individual student's graph. This procedure was repeated for 12 weeks.

Four weeks after the end of the regular data collection, each student orally read three randomly selected passages from LRG material. The median number of words read correctly was recorded on the individual student's graph.

Procedure for Determining Quarter-Intersect Lines

I completed a graph for each student in the study, recording the number of words each student read correctly in 1 minute on the ordinate and the days and weeks of data collection on the abscissa. I photocopied each graph after 8 weeks of student reading performance had been recorded and drew a slope of progress on each photocopied graph using the quarter-intersect procedures described by White (1972, 1974). The quarter-intersect procedure involves drawing an intersection through each half of the student's reading data at the middle rate and day for that half. The trend line was drawn through intersecting points in each half of the data. I extended the quarter-intersect trend line past the 16th week and then drew three vertical indicators along the bottom of the graph to represent the middle days of weeks 10, 12, and 16. The values of the quarter-intersect line-vertical line intersection represented the predicted score for the student that week.

Procedure for Determining Linear-Regression Lines

To determine the predicted values for weeks 10, 12, and 16, I used a one-factor, linear-regression computer program (Coffman & Steiber, 1987) that was designed for use on Statistical Package for the Social Sciences (SPSSx). The computer program, which was modified to use either 8 data points (1x week condition) or 24 data points (3x week condition) in making predictions for the three time periods, generated a line of best fit (slope) through each set of individual student data using a simple linear-regression formula.

Results

Actual Student Performance

In the 1x week measurement condition, the increase in average student reading performance ranged from 1.1 words per minute between weeks 11 and 12 to 7.1 words per minute between weeks 10 and 11. The overall average student growth between weeks 1 and 12, when monitoring occurred regularly, was 24.5 words per minute or about 2 words per minute/per week. Between

weeks 12 and 16, there was an increase of 2.2 words per minute, which is equivalent to approximately .6 words per minute/per week.

In the 3x week measurement condition, increases in student reading performance ranged from 1.3 words per minute between weeks 3 and 4 to 10.6 words per minute between weeks 10 and 11. Between weeks 1 and 12, the average student growth was 26.7 words per minute, which is equal to an increase of about 2.2 words per minute/per week. No student growth was evident between weeks 12 and 16, with the decrement in student performance equal to an average of about .5 words per minute/per week.

In both measurement conditions, student performance decreased the most between weeks 9 and 10 and increased the most between weeks 10 and 11. The variability in student performance increased over time in both measurement conditions.

Procedures for Determining Accuracy in Prediction

Quarter-intersect predictions. The first 8 weeks of student reading data were used to predict student reading performance for weeks 10, 12, and 16 using the quarter-intersect procedure. In both measurement conditions, the quarter-intersect method predicted that the students would increase their reading performance from weeks 10 to 12 to 16. In the 1x week measurement condition, average student performance was predicted to increase about 2.9 words per minute/per week. Average student performance in the 3x week measurement condition was predicted to increase about 2.4 words per minute/per week. Variability in predictions of student performance was greater for the 1x week condition and increased more over time than in the 3x week condition.

Linear-regression predictions. The first 8 weeks of student reading data also was used to predict student reading performance for weeks 10, 12, and 16 using the least squares linear-regression procedure. This method predicted that student reading performance would increase from weeks 10 to 12 to 16, regardless of measurement condition. Average student performance in the 1x week measurement condition was predicted to increase about 2.8 words per minute/per week. In the 3x week measurement condition, average student performance was predicted to increase about 2.3 words per minute/per week. The variability in linear-regression predictions of student performance followed the same pattern as in the quarter intersect predictions, with both greater and more rapidly increasing variability noted in the 1x week condition than in the 3x week condition.

Accuracy in predictions. I determined the accuracy in prediction separately for each method of analysis procedure by subtracting individual student's median predicted number of words read correctly for each week from the student's median actual number of words read correctly for the week. The smaller the obtained difference score, the more accurate the prediction. A positive score indicated an underprediction and a negative score indicated an overprediction. All of the mean difference scores were negative which indicates that the majority of my predictions over-predicted actual student performance. Further, the difference scores appeared to be greater in the once-a-week measurement condition than the three-times-a-week measurement condition. In addition, the predictions for week 12 appeared to be more accurate (i.e., smaller difference score) than either the 10-week or the 16-week prediction. For example, the mean difference score for week 12 was -4.2, whereas the mean difference scores for weeks 10 and 16 were -9.5 and -14.1, respectively.

Statistical Analysis

I used a repeated-measure multiple analysis of variance (MANOVA) to address the major research questions in this study. Measurement condition, method of analysis, and length of prediction time served as the independent variables. The residual between the predicted and actual scores served as the dependent variable. The prediction residual for each time factor was nested within method of analysis, (quarter-intersect and linear-regression), which was a repeated factor. Results of the analysis are summarized for each independent variable.

Measurement condition accuracy. I analyzed the effects of the type of data collection strategy, either 1x week or 3x week. As I stated earlier, it appeared that there were differences in the mean residuals between the two measurement conditions (1x week = -12.7, 3x week = -5.7). However, the results yielded an $F(1,43)=3.016$, $p=.09$, indicating that there was not a significant main effect for measurement condition. Regardless of the time of prediction and method of analysis, there were no reliable differences between the accuracy of the 1x week and 3x week measurement conditions.

Method of analysis accuracy. To test the effects of method of analysis, I compared the mean difference scores obtained from the quarter-intersect (-9.8) and linear-regression methods (-8.7). The comparison between methods of analysis across measurement conditions yielded an

$F(1,43)=.84$, $p=.365$, indicating that there was not a significant main effect for method of analysis. In addition, analysis of the interaction between method of analysis and measurement group, $F(1,43)=.275$, $p=.603$), was not significant. That is, there were no reliable differences between the accuracy of prediction obtained through the quarter-intersect method and the linear-regression method, regardless of the data collection strategy.

Prediction time accuracy. Because the prediction time factor was nested within the method of analysis factor, I compared the three prediction times (weeks 10, 12, and 16) within each separate method of analysis. For example, within the quarter-intersect method, the means for week 10 (-9.9), week 12 (-4.8), and week 16 (-14.7) for the total sample were compared. The results yielded an $F(1,43)=6$, $p=.018$, which indicated a reliable difference at the .05 level of significance. The comparison of the prediction time means within the linear-regression method (week 10 = -9.0, week 12 = -3.6 and week 16 = -13.5) for the entire sample yielded an $F(1,43)=19.710$, $p<.001$, which indicated, once again, there was a significant main effect for prediction time. I explored the possibility that the prediction times were differentially accurate depending on the measurement condition and found no significant interaction effect between prediction time and measurement condition within either method of analysis (quarter-intersect: $F(1,43)=.994$, $p=.324$; linear-regression: $F(1,43)=1.21$, $p=.277$).

I further explored the main effect for prediction time by conducting several post-hoc t-tests to determine whether the significant difference in accuracy of prediction was between weeks 10 and 12, weeks 10 and 16, or weeks 12 and 16. The results of the post-hoc t-tests indicated that there was a significant difference in accuracy between each of the prediction times for both the linear-regression and quarter-intersect methods of analysis. The prediction for week 12 was more accurate than both the 10-week and 16-week predictions and the 10-week prediction was more accurate than the prediction for week 16.

Discussion

Average student reading performance improved overall during the course of the study, regardless of the measurement condition. During the first 12 weeks, when student reading progress was regularly monitored, the students in the 1x week condition improved an average of 2 words a week and the students in the 3x week condition improved an average of 2.2 words per week. The overall mean gain in words correct per minute between week 1 and week 16 was 26.7 words in the 1x week condition and 24.8 words in the 3x week condition. The average variability in student performance also increased over time in both measurement conditions.

My visual examination of the mean accuracy of the 1x week and 3x week measurement conditions for predicting future student performance suggested that the 3x week condition was more accurate in making predictions. That is, the difference score between actual and predicted student performance was consistently smaller in the 3x week condition than in the 1x week condition. However, the results of the statistical comparison between the two measurement conditions indicated that there were no reliable differences found in terms of the accuracy of predicting either short-term or end-of-year student performance as a function of frequency of measurement. This finding should be interpreted with caution, however, because the statistical difference between the two measurement conditions "approached significance," indicating that my lack of significant findings may be more a function of insufficient power in the study (i.e., small effect size and/or small sample size) rather than a real lack of difference between the accuracy of the 1x week and 3x week conditions.

Similarly, my comparison of the quarter-intersect and linear-regression methods of analyzing the student performance data did not reveal significant differences in the accuracy of predictions about student performance. Therefore, the results of this study suggest that, on average, the two methods of analyzing time-series data do not differ in the accuracy of the predictions. However, a similar study, (Shinn, Good, & Stein, 1989) found that, although there weren't reliable differences in the average residuals for these two analysis methods, the linear regression method did give more precise predictions. In other words, in comparison to the hand-drawn trend line, the linear regression procedure resulted in less prediction error for individual students. Therefore, further research in this area may be needed to determine whether one of these analysis methods is actually superior in making predictions about student reading performance.

In contrast to the effects of the data collection strategy and method of analysis, I found significant differences in accuracy of prediction depending on what point in the future predictions were made. Based on 8 weeks of student progress data in

reading, the prediction of student performance in the 12th week of progress monitoring (4-week prediction) was more accurate than either the prediction of the 10th week (2-week prediction) or the 16th week (8-week prediction). Further, the prediction of the 10th week was more accurate than the 16th week prediction. This pattern was found in both the quarter-intersect predictions and the linear-regression predictions.

I expected that a prediction 2 or 4 weeks into the future would be more accurate than a prediction 8 weeks into the future because the longer the time period, the more likely that uncontrolled variables, such as change in instruction, could affect student performance in unpredictable ways. However, the superiority of a 4-week prediction over a 2-week prediction was unexpected. If student improvement in reading is linear and the prediction lines of progress are linear, the most that I could expect is that predictions over two weeks and four weeks would be equally accurate. It is more likely, however, that the predictions would be less accurate over time rather than more accurate over time.

In examining student growth between the initial monitoring week and the prediction times, it is evident that, in both measurement conditions, average student growth was greatest between weeks 10 and 12, less between weeks 1 and 10 and the least between weeks 12 and 16. As previous research has suggested (Fuchs, 1986), the regular monitoring of student reading may have affected growth, in which case it is not surprising that students improved the least during the period when they were not being systematically monitored (between weeks 12 and 16). However, this explanation does not account for why students appeared to have improved more per week between weeks 10 and 12 than they did between weeks 1 and 10.

One of several possible explanations for the unexpected difference in accuracy between the predictions for weeks 10 and 12 is the presence of unanticipated major changes in instruction during the study. I purposely withheld the student progress data from teachers until the end of the study to ensure that teachers did not use the data to alter instruction and thereby change the within-instructional phase nature of the study into several distinct instructional phases. However, unspecified instructional changes may still have occurred. Teachers may have changed the level of the instructional material used or altered the amount of daily instruction provided over the course of the study.

A second possible explanation for the unexpected differences in prediction accuracy is that the implementation of the measurement strategies may have changed over time. A total of six data collectors participated in the study and, although the data collectors were initially trained to a high level of agreement, some drift in measurement procedures may have occurred over the course of the study. I didn't conduct any follow-up reliability checks during the study, leaving the administration and scoring of measurement procedures a possible source of error of unknown magnitude.

A third possible explanation for the inaccurate predictions is that student reading progress may *not* be linear over time. The lack of linearity in student growth has been suggested by Precision Teaching proponents (Lindsley, 1964; White & Haring, 1980), who recommend the use of semi-logarithmic charts to accurately portray normal student progress. However, a study by Marston (1988) comparing the accuracy of predictions based on semi-log charts and equal-interval graphs found that the predictions of student achievement based on trend lines from equal-interval graphs were more accurate than predictions based on semi-log chart trend lines. Nevertheless, the differences found in week-to-week word growth, ranging from -3.0 words to +7.1 words in the 1x week condition and -6.1 to +10.6 words in the 3x week condition, support the finding that, in this study, student growth was not consistent and linear across time.

However, after reviewing my findings, I came up with the fourth and most likely explanation, called the "spring break" factor. It turns out that spring break occurred between instructional weeks 9 and 10 for all of the students in the study. Student performance may have been unusually low during the 10th week due to the fact they had not received reading instruction for at least a week prior to the monitoring. In addition, students (and teachers) still may have been preoccupied with vacation activities. Whatever the reason, it appears that the 1-week break in instruction functioned as a major change in instruction for the students in this study. Therefore, week 10 was not a valid indicator of typical student reading growth over time.

One consistent feature across methods of analysis and measurement conditions was that student performance was overpredicted in my study. On the average, students tended not to perform as well at the three prediction points as their previous 8 weeks of performance suggested they should. This overprediction contrasts with findings from Shinn et al. (1989) where the residuals between predicted and actual scores for both

methods of analysis did not differ from zero. Variables such as differences in teacher training, curriculum materials, student populations, and the number of data points in the two studies may account for the dissimilar findings.

References

Blankenship, C. (1985). Using curriculum-based assessment data to make instructional decisions. *Exceptional Children, 52*, 233-238.

Carver, R. P. (1974). Two dimensions of tests: Psychometric and edumetric. *American Psychologist, 29*, 512-518.

Coffman, D., & Steiber, S. (1987). *Simple linear regression model program for SPSSx [computer program]*. Eugene, OR: University of Oregon, College of Education.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*, 141-146.

Deno, S. L., Marston, D., & Mirkin, P. K. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children, 48* (2), 368-371.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Minneapolis, MN: University of Minnesota, Leadership Training Institute/Special Education.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49* (1), 36-45.

Deno, S. L., Mirkin, P. K., Lowry, L., & Kuehnle, K. (1980). *Relationships among simple measures of spelling and performance on standardized achievement tests* (Research Report No. 21). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Fuchs, L. S. (1986). Monitoring progress among mildly handicapped pupils: Review of current practice and research. *Remedial and Special Education, 7*, 5-12.

Germann, G., & Tindal, G. (1985). An application of curriculum-based assessment: The use of direct and repeated measurement. *Exceptional Children, 52* (3), 244-265.

Glickling, E. S., & Thompson, V. P. (1985). A personal view of curriculum-based assessment. *Exceptional Children, 52*, 205-218.

Howell, K. W. (1986). Direct assessment of academic performance. *School Psychology Review, 15* (3), 324-335.

Jenkins, J. R., Deno, S. L., & Mirkin, P. K. (1979). Measuring pupil progress toward the least restrictive environment. *Learning Disability Quarterly, 2*, 81-91.

King, R., Wesson, C., & Deno, S. (1982). *Direct and frequent measurement of student performance: Does it take too much time?* (Research Report No. 67). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Lindsley, O. (1964). Direct measurement and prosthesis of retarded behavior. *Journal of Education, 147*, 62-81.

Marston, D. (1988). Measuring progress on IEPs: A comparison of graphing approaches. *Exceptional Children, 55* (1), 38-44.

Marston, D., & Magnusson, D. (1985). Implementing curriculum-based measurement in special and regular education settings. *Exceptional Children, 52* (3), 266-276.

Mirkin, P. K., & Deno, S. L. (1979). *Formative evaluation in the classroom: An approach to improving instruction* (Research Report No. 10). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

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

Shinn, M., Good, R. & Stein, S. (1989). Summarizing trend in student achievement: A comparison of methods. *School Psychology Review, 18* (3), 356-370.

Skiba, R., Marston, D., Wesson. C., Sevcik, B., & Deno, S. (1983). *Characteristics of the time-series data collected through curriculum-based reading measurement* (Research Report No. 125). Minneapolis, MN: University of Minnesota, Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. ED 236-197).

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, MN: University of Minnesota Institute for Research on Learning Disabilities.

Stern, P., & Shavelson, R. J. (1983). Reading teachers' judgements, plans, and decision making. *The Reading Teacher, 37*, 280-286.

Tucker, J. A. (1985). Curriculum-based assessment: An introduction. *Exceptional Children, 52* (3), 199-204.

White, O. R. (1972). *A manual for the calculation and use of the median slope-technique of progress estimation and prediction in the single case.* Eugene, OR: University of Oregon, Regional Resource Center for Handicapped Children.

White, O. R. (1974). The "split-middle": A "quickie" method of trend estimation. Univer-

sity of Washington, Seattle, WA.: Experimental Education Unit, Child Development and Mental Retardation Center.

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

White, O. R., & Liberty, K. A. (1976). Behavioral assessment and precise educational measurement. In N. G. Haring & R. L. Schiefelbusch (Eds.), *Teaching special children* (pp. 31-111). New York: McGraw-Hill.

Educational Validity in Social Studies Instruction

Victor Nolet
University of Oregon

With recent interest in special education on issues such as training high school students in functional life skills and transition to post-high school environments, social studies classes seem to hold tantalizing possibilities as a mainstream environment for low-achieving students. The domain of social studies involves the kind of knowledge and skills Reynolds & Birch (1982) describe as "cultural imperatives," i.e., the fundamental tools of our culture. These include "acceptable behavior in citizenship" and "preparation for an economical useful life" (p 101). A minimum expectation of schools is that all students will be afforded an opportunity to learn the cultural imperatives necessary to function as adults in our society.

The decision to place a low achieving student into general education high school classes such as social studies is indefensible unless there is reasonable expectation that an incremental benefit over placement in any other setting will be realized. Such an expectation must be based on a clear understanding of specific outcomes that can be anticipated from social studies instruction. Obviously, such an understanding is only possible when the characteristics of the social studies domain are known.

Although the database is incomplete, some salient features of social studies have been described in three particular areas: (a) the goals of social education (Armento, 1986; Cogan, 1989), (b) the characteristics of social studies textbooks (Armbruster & Gudbrandsen, 1986; Beck, McKeown, & Gromell, 1989) and, (c) the nature of instruction in social studies (Shaver, Davis, & Helburn, 1979; Superka, Hawke, & Morrisett, 1980). This growing body of research has made it clear that thoughtful consideration of each area (goals, materials, and instruction) is necessary whenever placement of low achieving students into middle or high school general education social studies classes is contemplated.

A valid placement of special education students into general education social studies classes requires that three conditions be met. First, the goals of instruction must be defined explicitly. Second, the content of the curriculum and instruction must be clearly identified. Third, the desired outcomes of instruction must be delineated and measured. Without this systematic instructional decision making, placement of handicapped students into general education social studies cannot be supported.

This paper has two purposes. First, I will present a conceptual framework for evaluating the efficacy of instructional practices in social studies, particularly as they pertain to low achieving students. The basis for this framework is the concept of *educational validity*. Second, I will present a decision-making model aimed at ensuring the educational validity of social studies instruction.

Educational Validity in Social Studies Instruction

The concept of validity generally is associated with test or experimental design. In the context of test design, validity pertains to the extent to which an assessment procedure measures a domain of interest. Generally, validity pertains to the question, "Does the test really measure what it is intended to measure?" More precisely, validity refers to the appropriateness of inferences that can be made on the basis of the results of an assessment procedure (Messick, 1989). Validity, then, refers not to the quality or characteristics of a particular procedure but to the *decisions* that can be made on the basis of information obtained. Validity is therefore bound to the context in which a procedure is used. Scores obtained from a particular assessment procedure may be useful for one type of decision but not applicable to others.

With experimental design, issues of validity traditionally have been separated into two categories: external and internal validity. External validity pertains to generalizability and asks, "To what populations, or situations can this effect be generalized?" (Campbell & Stanley, 1963, p 5). In other words, do the results obtained in one particular situation apply in other situations? Internal

validity is the fundamental basis for interpreting the results of an intervention and asks, "Did the [procedure] in fact make a difference in this specific . . . instance?" (Campbell & Stanley, 1963, p 5). For internal validity to be supported, an effect must be demonstrated to be the direct result of a particular procedure. Internal validity is supported when the occurrence of a particular event co-varies with a specific procedure.

Issues of validity, however, need not be limited to tests and research methodology but can be raised in reference to instructional interventions. For example, social validity refers to the social value and acceptability of an educational intervention (Kazdin, 1977; Wolf, 1978). Socially valid educational programs focus on skill or knowledge that is important to the learner in current or future environments.

Educational validity is a broader term that encompasses all these different concepts of validity (Voeltz & Evans, 1983). For an intervention to have educational validity, three criteria must be met. First, a behavior change must occur as a function of the intervention. Second, the intervention must be implemented as planned, and third, the intervention must be beneficial to the student, i.e. socially valid. Each of these criteria is necessary, and no one is sufficient alone. With educational validity we ask, "Did the instructional intervention teach the skills or knowledge it was intended to teach, and is that content important to the student in current or future environments?"

Validity involves examination of multiple pieces of evidence to decide whether a particular inference is supported (Messick, 1989). Educational validity pertains to the specific inference that, as a direct result of instruction, a student has learned something useful.

Although educators assume that students learn important skills and knowledge as a function of instruction, the educational validity of our instructional interventions receives little systematic scrutiny, particularly in content areas such as social studies. Consequently, teachers are held accountable to a far less rigorous standard than test developers or scientists. Yet, daily classroom instructional interventions have a far greater potential impact on students than any single test or experiment. The assessment technology typically used in schools provides little information about the actual effects of instruction. Instruction is only loosely aligned with goals and these in turn are rarely analyzed for their functional value to the learner. The effects of this misalignment of goals, instruction, and outcomes can be subtle but very real, as the following example illustrates.

Recently, the National Geographic Society (NGS), in collaboration with the Gallup Organization conducted a survey of the geographical knowledge of adults in 20 countries (Grosvenor, 1989). A world map was displayed on which 57 geographical features were identified by number. When presented with a list of names of 16 topographical features, respondents were directed to associate a name with a number corresponding to its map location. The list contained such prominent world features as France, the Pacific Ocean, the Soviet Union, Central America, and Canada. The task combined map reading skills with basic knowledge of world geography and generally was assumed to be in the range of abilities expected of most high school graduates.

Americans obtained the lowest score among respondents aged 18 to 24. Only 25% of Americans in all age groups correctly located Sweden or the Persian Gulf and only 29% identified West Germany. What conclusions can be drawn from these results? Are the results of this survey yet another indictment of American schooling and further evidence that students aren't taught even the most basic skills and knowledge? Or are other explanations possible for the poor performance of the Americans? Educational validity provides a framework with which to evaluate these results. Consider a hypothetical student attending school in the state of Oregon. Would the task involved in the NGS/Gallup survey be considered important for the student? Although the task probably would not be considered vital for social or economic functioning in future adult environments, the ability to locate geographic features on a world map is crucial for interpreting important events around the world and likely is an index of a deeper understanding of global citizenship. Therefore, we could classify this goal as a moderately important outcome expected of social studies education.

Next we need to evaluate whether the student actually will be taught the content required to perform the specific task required on the NGS/Gallup survey. According to guidelines recommended by the National Council for the Social Studies (NCSS, 1989), most training in locating places on a map and globe should occur in grades 4 and 7-9, with some instruction continuing throughout high school. To assess the extent to which instruction on this task is presented, we can either examine the textbooks in which the student might be instructed or observe in classrooms. Often, it is

assumed that the role of the teacher is to interpret or supplement information presented in textbooks while the text serves as the foundation for instruction. If this characterization is accurate, examination of textbooks will facilitate an estimate of the instruction likely to occur in a given domain.

A 4th grade text that appears on the state adoption list (*World Regions*, Bacon & Fairweather, 1983, Follett Social Studies series) contains a world map showing all the features in the NGS/Gallup survey. However, the map is displayed in a different projection (Modified Van Der Grinten projection) than the NGS/Gallup map (Robinson projection). The difference between these two particular maps is not drastic, but the task demands are nevertheless different. Also, the map is displayed across two pages of text so that much of Europe and Africa are obscured. The most important point, however, is that the lessons in the text do not focus on locating specific features. For example, no specific instruction pertaining to either Central America or Vietnam is included. If the 4th grade student uses this social studies text, instruction in locating the specific geographic features in the NGS/Gallup task will need to be provided by some other means.

If the student is in middle school, he or she might be instructed in *Nations of the World* (Jarolimek, Lefferts, & Soifer 1985) in the MacMillan Social Studies series. This text includes one world map shown in Eckert projection which is similar to the Robinson projection shown in the NGS/Gallup survey. However, the map in this text also is displayed on two pages and most of western Europe is obscured. Also, Central America and the Persian Gulf are not labeled. Again, the textbook fails to provide complete information necessary to perform the NGS/Gallup task.

If our hypothetical student is in high school, she or he might be enrolled in a world geography class and use *World Geography Today* (Helgren & Sager, 1985), a state adopted text for advanced secondary students. The only world map in this book that includes labels of major geographical features is shown in Goode's Interrupted Projection, which is markedly different than the projection used in the NGS/Gallup survey. This map also is displayed on two pages but no major features are obscured. However, the labels for many features on the map are displayed in list form with arrows connecting feature names to locations on the map, with considerable loss in precision. Additionally, some geographical features included on the NGS/Gallup survey such as the United Kingdom, Central America, and the Persian Gulf are not labeled. Even this advanced high school world geography text provides insufficient information for our hypothetical student to perform adequately on the NGS/Gallup survey task.

The NGS/Gallup survey was not aligned with a particular curriculum nor was such an alignment ever intended. However, the central role of the social studies teacher in delivering key information is obvious. Unfortunately, the situation presented in the example is not unlike the misalignment that occurs in current practice in social studies instruction and assessment. Expected outcomes are poorly defined, instruction is not matched to these outcomes and assessment procedures are not sensitive to goals or instruction.

The conclusion that can be drawn from this example is that ensuring educational validity in social studies requires teachers to complete an integrated planning process. Although broad goals can be established in state or district guidelines, specific goals for instruction can only be based on teacher knowledge of the content area and the needs of the learner. There must follow from these goals an explicit definition of the student performance outcomes expected as a result of instruction. Therefore, instruction must be outcome oriented, not textbook driven. Clearly, this process requires systematic decision-making regarding the most fundamental variables in teaching: goals, instruction and assessment.

Decision-Making for Educational Validity

Development of educationally valid instruction requires a systematic planning process that includes these steps: select goals, define behaviors, design and deliver instruction, and assess student performance outcomes. The components of a decision making model to ensure educational validity are illustrated in the diagram in Figure 1.

Selection of goals precedes implementation of the other three major components; however, once goals are selected, the other components function in consort. Design of instruction and assessment procedures depends on explicit definition of behaviors, while student performance data informs instruction and facilitates measurement of behavior change.

Educational validity demands that three questions, each associated with an instructional component, be addressed. Behaviors selected for intervention must be important to the learner in

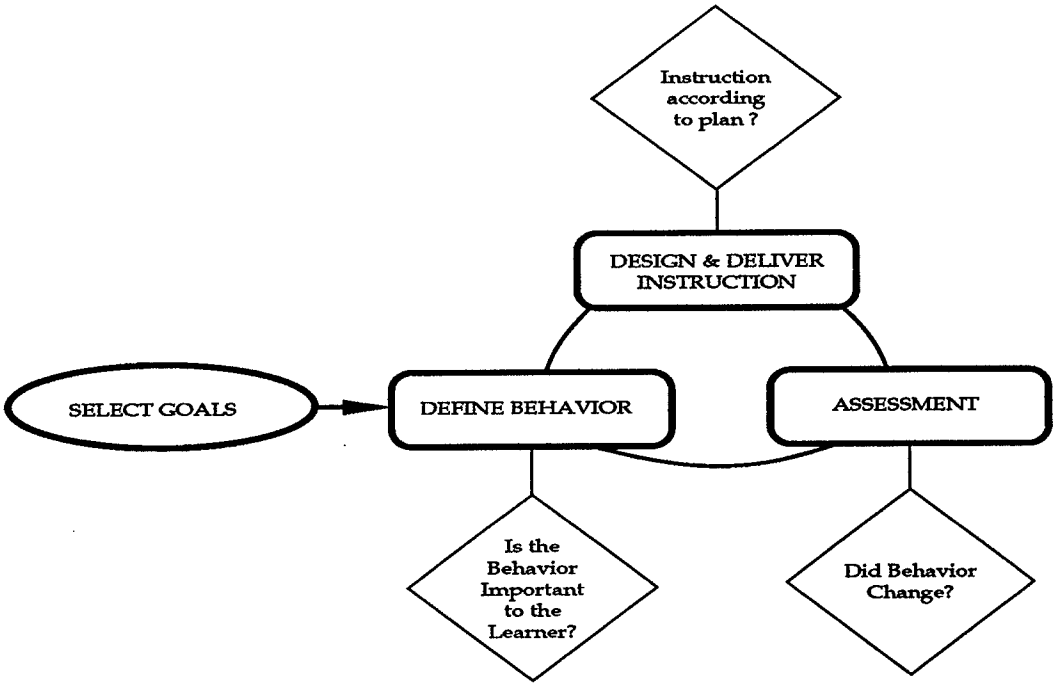


Figure 1. Decision-Making Model to Ensure Educational Validity

current or future settings. Instruction must be delivered in a manner consistent with instructional goals, and assessment must be designed to show that a behavior change occurred.

Identifying Goals in Social Studies

The importance of clearly identified goals in instruction has been well documented (Rosenshine, & Stevens, 1986). In the context of educational validity, goals need to specify behaviors that (a) are useful to the learner, (b) facilitate instruction, and (c) can be measured. Unfortunately, social studies does not readily lend itself to development of useful goal statements. Goals found in curriculum guides often take the form of broad conceptual statements that pertain to a complex array of skills and knowledge. Such statements provide little information for designing instruction or assessing student achievement. For example, the phrase "understand individual responsibility for the democratic system" (California History-Social Science Curriculum Framework, 1988 p 23) can be operationalized in a variety of instructional or assessment formats and the difficulties one may encounter in teaching or measuring "understanding" should be obvious.

Much of social studies instruction may be structured by textbooks (EPIE, 1977; Shannon, 1982). However, social studies as it is currently operationalized in textbooks is a broadly defined domain without a clearly identifiable body of content that is linked to measurable goals. As a source of goals for instruction, textbook series provide minimal support for teachers. Examination of goals specified in one textbook illustrates this point.

The scope and sequence chart in the teacher's manual of the Follett Social Studies Series grade four book, *World Regions* (Buggey, 1983) includes seven skill areas; map and globe use, reading, communication, time, math, citizenship, and thinking. Within each of these skill areas is listed from two to ten specific skills to be taught as part of the social studies program during the year. To address this array of skills, nine major objectives are provided. None of the nine objectives specify application of any of the seven skill areas. Although they generally pertain to the broad topic areas specified in the scope and sequence, the objectives use vague, non-observable terms such as "understand," "appreciate," or "recognize." Unit objectives employ the same format. Those for Unit 2 are typical. They take the form of the stem "to understand," followed by such phrases as "the major landforms," "climate," and "climatographs."

To some extent, broad, ill-defined goal statements may be unavoidable. There is a long-standing tradition in social studies of emphasis on such vague notions as "understand democracy" or "appreciate diversity" (Cornbleth, 1985; Cogan, 1989). Restricting instruction to those behaviors that can be easily summarized in an operationally defined behavioral objective may be unrealistic.

Teachers must assume an active, central role in goal selection. Therefore, the characteristics of the social studies domain require teachers to become experts at analyzing, selecting, and modifying goals to facilitate educationally valid instruction. The most important criteria for selecting goals is importance to the learner.

Evaluating goal importance requires consideration of a range of variables about which the teacher is the most authoritative source of information. These issues include time allocated for social studies instruction, the learner's current performance, teacher knowledge of the content area, the scope and sequence of the local social studies curriculum, and the teacher's knowledge of local needs and priorities. A process in which a goal is selected because it is the next one in the textbook cannot possibly be sensitive to these variables.

Once goals have been selected they must be broken into teachable components. For example, the statement, "Students will learn an appreciation of the diversity of American cultures" is a reasonable goal that might be judged important and subsequently selected. However, teaching "appreciation" will be difficult because the term has a wide range of possible interpretations and reliable measurement of "appreciation" may be impossible. Therefore terminal behaviors associated with learning the goal must be defined.

Defining Behaviors in Social Studies

A key discrimination in defining behaviors for social studies instruction is distinction between skill and knowledge (Tindal & Marston, in press). Generally, skills involve a motoric response such as running, writing letters or reading orally. Knowledge is information specific to a particular content domain and is manifested in verbal behavior. Any behavior exists on a continuum that ranges from all skill to all knowledge. Most of the domain of social studies falls on the knowledge end of this continuum.

Knowledge can be categorized into four forms: facts, concepts, principles and strategies (Tindal & Marston, in press; Kameenui, in press). Facts, which include only one example, are associations of a specific response with a symbol or object. An example of a fact that might be taught in social studies is the statement, "Egypt is on the continent of Africa." Concepts are classes of objects, events, actions, or situations that can be grouped together on the basis of an attribute they share in common. The word "grasslands" is an example of a concept. All grasslands are distinguished on the basis of common attributes such as vegetation, rainfall, temperature, etc. Principles express relationships among at least two facts, concepts or other principles. Principles are expressed as rules or if-then statements, such as, "If the climate in a region is very dry, certain kinds of plants cannot survive." Often principles in social studies express complex relationships such as, "The climate and geology of a region determine the economic prosperity of people who live there." Strategies are multiple step procedures that involve combinations of any of the other knowledge forms. Locating places on a map is a strategy commonly taught in social studies. This task minimally requires manipulation of facts about the way maps are constructed, and concepts such as continents or land masses.

Learning involves fluency building in performance of a skill or manipulation of a knowledge form. Performance of a skill can be observed directly and specification of the behavioral component is a straightforward process. Manipulation of a knowledge form involves a range of intellectual operations including reiteration, summarization, illustration, prediction, evaluation, and predication (Williams & Haladyna, 1982). Overt demonstration of these intellectual operations requires the learner to either speak or write.

Defining behaviors in social studies requires identification of the specific skill or knowledge form involved and specification of the motor or intellectual operation the student will be expected to demonstrate following instruction. For example, suppose a teacher has established the instructional goal, "Students will demonstrate a knowledge of economic conditions associated with the Great Depression." Behaviors associated with this goal might include reiteration of facts concerning the dates of the depression, summarization of concepts associated with the stock market, and application of principles pertaining to the relationship of industrial production with economic variables. Clearly, specification of behaviors in this fashion permits a robust treatment of instructional goals in a manner that can be sensitive to a wide range of learner abilities.

Once a goal has been broken into definable behaviors, the value of each behavior for a particular learner or group of students can be assessed. For example, suppose a teacher whose class includes students with a wide range of abilities has selected the goal, "Students will demonstrate an understanding of Japan's role in the world economy." The teacher might decide that, for some students in the group, reiteration of facts about Japan's economy, such as its major trading partners

or type of currency, are appropriate behaviors, while for other students, evaluation of principles related to world trade or factors affecting industrialization following World War II are important behaviors. Both sets of behaviors relate to the instructional goal, yet there are clear differences in the intellectual operation required.

Design and Deliver Instruction

Definition of terminal behaviors leads directly to design of instruction. Decisions related to instructional validity involve selection of materials and examples to teach each knowledge form and clarification of assumptions about the learner (Kameenui, in press). Each knowledge form makes specific demands on the learner and implies specific instructional strategies . Detailed discussion of design of these instructional strategies has been provided elsewhere and is beyond the scope of this article (see for example Kameenui, in press). However, two important issues will be addressed here: assumptions about the learner and use of textbooks.

Assumptions About the Learner

Design of instruction involves making assumptions about the learner, and these assumptions force certain decisions in the planning process. Returning to the example above, we are given the goal, "Students will demonstrate an understanding of Japan's role in the world economy." The teacher has defined a terminal behavior related to this goal as, "Tell how supply and demand can explain Japan's trading relationship with the U.S." This behavior requires application of the principle, "If supply goes up, then demand goes down." The teacher needs to decide (a) whether the learners have sufficient background knowledge about the principle to apply it to the situation at hand, (b) if the learners can be assumed to construct the principle given a range of examples, and (c) whether component concepts and facts need to be taught. Clarification of assumptions about the learner takes into account the nature of the knowledge form, previous instruction and instructional design. Importantly, it removes the focus of attention from student ability to teacher planning.

Textbook Use

Although textbooks can be a critical resource for teachers, they must be used skillfully and with full knowledge of their nature. Social studies textbooks tend to contain a high density of facts and concepts, and often the exact relationship expressed in principles is implied rather than stated explicitly (Beck, McKeown & Grommell, 1989). Teachers must evaluate the extent to which textbooks teach knowledge forms and make explicit decisions about the manner in which textbooks will be used. In order to plan instruction, a teacher must analyze the range of examples and explicitness of relationships presented in the textbook and then decide how to use the text to teach the explicit intellectual behaviors planned. A textbook is not the same as a curriculum. Curriculum involves the full range of content organized and delivered by the teacher, and textbooks should represent only a portion of the total instructional package. Use of a textbook to implement curriculum should involve systematic planning.

Decisions involving to the use of textbooks relate to the criteria for educational validity that instruction be delivered as planned. Clearly, if the terminal behavior calls for application of a principle and a textbook provides only a listing of facts, the teacher has either to assume the learner can construct the principle, given the facts, or explicitly teach the principle. Therefore, to evaluate the fidelity of instruction, the teacher must compare instruction delivered with assumptions about the learner and the terminal behavior expected.

Educational validity is not supported unless there is a sufficiently high probability the behavior can be demonstrated by the learner, given instruction and assumptions.

Assessment

Educational validity is only supported if the learner demonstrates a change in behavior as a function of instruction. The teacher must show that the student didn't have the ability to perform the task before instruction and knows how to perform the task after instruction. This implies an effective assessment technology that is sensitive to growth.

Unfortunately, much of the assessment technology currently in use in schools is of little use in time series design. Tests embedded in textbooks reflect only the content they cover and, as has been emphasized throughout this article, educationally valid instruction demands that the teacher supplement or modify textbook content, particularly for low achieving students. Also, textbook embedded tests generally lack technical adequacy necessary to show behavior change. Published norm-referenced achievement tests are not at all linked to instruction and therefore lack content validity.

A number of curriculum-based assessments have been offered as alternatives to prevailing assessment procedures. But, only Curriculum-Based Measurement (CBM) has been demonstrated empirically to be technically adequate. To date,

CBM has not been used in content areas such as social studies for a number of reasons. First, the measures employed in CBM are designed for the most part to test fluency in basic skill areas such as reading, math computation, written expression, and spelling. They are not designed to directly test intellectual operations involving knowledge forms. A second issue in the use of CBM relates directly to educational validity. Some of the CBM measures, specifically reading and spelling, involve sampling from basal textbooks. Their validity as curriculum-based measures is only supported if the textbook comprises the curriculum. Given that research shows 90-97% of reading instruction (EPIE, 1976) is structured by the textbook, this may be a safe assumption. However, such textbook-based procedures likely would not be useful in social studies, particularly in the model described here, where much of the instruction is structured by the teacher.

Teacher-made tests offer the a viable alternative for assessment in social studies. Procedures have been described (Tindal & Marston, in press) for construction of technically adequate teacher developed tests. Only the most salient issues will be discussed here.

The primary decision relevant to designing test items to measure terminal behaviors in social studies is specification of the terminal behavior expected to result from instruction. In assessing knowledge, a test is defined by the intellectual operation performed on each knowledge form (for example, reiteration of a fact, illustration of a principle, application of a strategy, etc.). The remaining element in designing test items is specification of the response format, i.e. selection or production.

To assess growth, two types of procedures can be used: pre-post and progress monitoring. In a pre-post design, a test of the terminal behaviors is administered to the learner prior to instruction, and an alternate form of the test is administered following instruction. The extent to which behavior change occurs as a result of instruction is indicated by differences in the scores on the two tests. Pre-post testing relies heavily on the equivalence of the alternate forms and only provides information about the student's growth in the narrow domain of content taught during the time between the two test administrations.

In a progress monitoring design, tests that sample from a large domain representing the range of performance expected to result from instruction are administered on a repeated basis throughout the school year or term. Growth is evaluated by analysis of the trend shown in data over time. Although this procedure provides less information about student achievement in a particular lesson or unit, it is more closely linked to assessment of growth toward terminal behaviors specified by goals. In this respect, progress monitoring can be a valuable tool in evaluating educational validity. Major concerns in using a progress monitoring approach center on the practicality of designating a broad domain of content from which to sample items and then developing a sufficient bank of test items to generate enough alternate forms of tests that can be administered on a repeated basis. Such long range planning probably is beyond the realm of practicality in a content area such as social studies. Currently, there is a need for development of new assessment procedures that are sensitive to growth in acquisition of knowledge forms. Research under way at the University of Oregon has been exploring the use of written retell as a curriculum based measure in content area classes and this procedure holds promise for progress monitoring (Tindal & Parker, 1989). However, technically adequate teacher-made tests can be constructed using existing technology and it is incumbent upon content area teachers to become skilled at using these.

Conclusion

We should not expect the nature of social studies education to change in the foreseeable future. This is a content area that is driven by strong forces over which teachers have little control. National goals for social studies will continue to be influenced by such vagaries as reports issued by "blue ribbon" commissions, "innovative" guidelines promulgated by influential textbook adoption states, surveys that show American students are deficit in yet another "critical" knowledge area, and popular catch phrases such as "cultural literacy" or "global perspective." Furthermore, textbooks and achievement tests will continue to be designed and marketed by for-profit companies whose primary concern is the bottom line of their balance sheets, not student learning.

The nature of social studies has clear implications for instruction as well as establishment of a research and training agenda—teachers must take control of the instructional process. Decisions about what and how to teach cannot be delegated to textbook publishers or absentee planners. With increased control must come increased responsibility. Teachers need increased skills at selecting

goals, and defining expected outcomes in terms that can result in instruction and assessment. Instruction must be designed around the demands of knowledge forms and assumptions about learners. Teachers must be provided with information about how to design and use technically adequate assessment procedures to plan and evaluate instruction in content areas. Finally, research must continue in areas such as the nature of social studies textbooks, effective ways to measure student learning in content areas, and design of content area instructional strategies for students with diverse skills and learning histories.

References

Armbruster, B. B., & Gudbrandsen, B. (1986). Reading comprehension instruction in social studies programs. *Reading Research Quarterly, 21*(1).

Armento, B. J. (1986). Research on teaching social studies. In A. Whitrock (Ed.) *Third handbook on research in teaching* (pp. 942-951). Boston: MacMillan Publishers.

Beck, I. L., McKeown, M. G., & Gromell, E. W. (1989). Learning from social studies texts. *Cognition and Instruction*, 6(2), 99-158.

Buggey, J. (1983). *World regions* (Teacher's Annotated Edition), Follett Social Studies, Boston: Allyn and Bacon.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally

Cogan, J. J. (1989). Citizenship for the 21st century: Observations and reflections. *Social Education, 53*(April/May), 243-245.

EPIE Institute (1977). *Report on a national study of the nature and quality of instructional materials most used by teachers and learners* (Report No. 76.). Author.

Grosvenor, G., M. (1989). Superpowers not so super in geography. *National Geographic, 176*(6), 816-821.

Helgren, D. M., & Sager, S. R. J. (1985). *World geography today.* New York: Holt, Rinehart and Winston

History-Social Science Curriculum Framework Committee (1988). *History-Social Science Framework for California Public Schools.* Sacramento: California State Department of Education.

Jarolimek, J., Lefferts, W., & Soifer, I. (1985). *MacMillan Social studies: Nations of the world.* New York: MacMillan.

Kazdin, A. (1977). Assessing the clinical or applied importance of behavior change through social validity. *Behavior Modification, 1*(4), 427-451.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and MacMillan.

National Council for Social Studies (1989). Report of the ad hoc committee on scope and sequence. *Social Education, 53*(6), 375-387.

Reynolds, M. C., & Birch, J. W. (1986). *Teaching exceptional children in all America's schools.* Reston, VA: The Council for Exceptional Children; 1982.

Rosenshine, B., & Stevens, R. (1986). Teaching functions. In A. Whitrock (Ed.), *Third handbook on research in teaching* (pp. 376-391). Boston: MacMillan Publishers.

Shannon, P. (1982). Some subjective reasons for teacher's reliance on commercial reading materials. *The Reading Teacher* (May), 884-889.

Shaver, J. P., Davis, O. L., & Helburn, S. (1979). The status of social studies education: Impressions from three NSF studies. *Social Education* (February), 150-153.

Superka, D. P., Hawke, S., & Morrisett, I. (1980). The current and future status of the social studies. *Social Education* (May), 362-369.

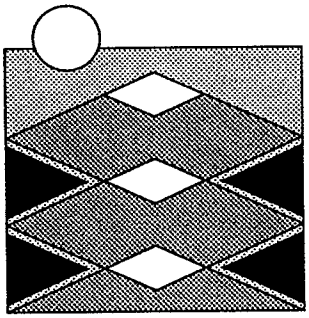
Tindal, G., & Parker, R. (1989). Development of written retell as a curriculum-based measure in secondary programs. *School Psychology Review, 18*(3), 326-341.

Voeltz, L. M., & Evans, I. M. (1983). Educational validity: Procedures to evaluate outcomes in programs for severely handicapped learners. *Journal of the Association for the Severely Handicapped, 8*(Spring), 3-15.

Williams, R. G., & Haladyna, T. M. (1982). Logical operations for generating questions (LOGIQ): A typology for higher level test items. In G. H. Roid & T. M. Haladyna (Eds.), *A technology for test-item writing* (pp. 161-186). New York: Academic Press.

Wolf, M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*(2), 203-214.

The Oregon Conference



1991

Toward a Dynamic Assessment of Complex Mathematical Operations: An Examination of Current Models and Directions for Educational Practice

Asha Jitendra
University of Oregon

There is a growing need to examine assessment techniques as viable tools that provide information about a learner's knowledge of mathematics skills (Bransford, Delclos, Vye, Burns, & Hasselbring, 1987; Campione, Brown, & Connell, 1990; Ferrara, 1987). Traditionally, assessment of mathematics has relied heavily on norm-referenced tests. Performance data from standardized measures or norm-referenced mathematics tests are used to screen students at risk for mathematics failure. Norm-referenced approaches to assessment typically focus on differences between individuals or the causes of learners' problems (e.g., general intellectual inability, a perceptual deficit, or specific disability) (Meyers, 1987). In this approach, the learner is the primary focus of assessment and little regard is given to the learner's ability to modify his/her behavior. The importance of shifting the focus of causality from sources within the individual (e.g., learning disability, mental retardation) to forces outside the individual (e.g., the instructional environment) is a recent phenomenon (Heller, Holtzman, & Messick, 1982; Kameenui & Simmons, 1990).

Traditional assessment measures do not address specific intervention strategies for remediating learning deficits (Kameenui & Simmons, 1990; Meyers, 1987). Traditional assessment procedures consider learning to be best assessed as a summative product. In this context, the opportunity to influence learning is overlooked (Meyers, 1987). In sum, traditional assessments fail to recognize the learner's potential to succeed with adequate environmental support. The failure of traditional assessment has prompted researchers to search for better assessment tools.

Traditional and Dynamic Assessment Techniques

Bransford, Delclos, Vye, Burns, and Hasselbring (1987) provide three arguments that move away from traditional assessment tools and toward dynamic or interactive assessment procedures.

First, traditional assessment is concerned with the products rather than processes of learning. In conventional testing situations, an individual's failure to produce a correct solution is considered a final product and suggests the absence of a skill. Such testing fails to consider the process that generates the inadequate answer. In contrast, dynamic assessment permits the examinee to clarify his/her response. For example, an analysis of the three phases (i.e., input, elaboration, and output) of the mental act in one dynamic assessment model allows the examiner to note whether the failure to respond correctly is a result of inappropriate data gathering or data integration (Feurestein, 1979).

Second, traditional assessment fails to address the responsiveness of an individual to instruction. It is based on the premise that prior learning adequately predicts future performance, but this notion of assessment disregards the differential quality of individuals' prior learning experiences. In an assessment process that is dynamic, examining how a child responds to instruction is deemed as important as predicting a child's ability to learn based on his/her prior learning.

Third, traditional assessment does not provide prescriptive information for designing potentially effective instruction. Instead, broad generalizations of individual learning problems that do not consider the quality of instruction are rendered. On the other hand, dynamic assessment techniques yield systematic information about the kinds of tasks and teaching strategies that are beneficial for individual students (Feurestein, 1979).

Ysseldyke and Regan (1980) contend that assessment should incorporate educational program planning. They consider educational program planning inherent to the assessment process and view assessment as a dynamic rather than a static process. As noted by Bethge, Carlson, and Wiedl (1982), "Dynamic assessment permits analysis of how individual differences are affected through modifications in testing conditions" (p. 89). To this end, dynamic assessment is a valuable

alternative to traditional psychometric approaches because, in the analysis of cognitive performance, the importance of human factors is considered. Although numerous models of dynamic assessment are available in the research literature (Embretson, 1987; Ferrara, 1987), the salient feature that characterizes the various approaches is that of guided learning in determining a learner's potential for change (Campione, 1989; Meyers, 1987).

Models of Dynamic Assessment

Five distinct models of dynamic assessment have been gleaned from the research literature: (a) Test-train-test assessment, (b) Learning Potential Assessment Device (LPAD): Mediational assessment, (c) Testing the limits assessment, (d) Graduated prompting assessment, and (e) a continuum of assessment model - Mediated and graduated prompting. The types of tasks utilized in dynamic assessment range from general to specific academic content areas such as reading, math, or science. Whatever the selected task, the major goal of dynamic assessment is "to determine and modify the reasons responsible for failure" (Jensen & Feuerstein, 1987, p. 391).

Test-Train-Test Assessment

A test-train-test procedure of assessing the learning potential of educable retarded children is based on a psychometric model of assessment developed by Budoff (1974). In this procedure, a child's present level of functioning is determined by a pretest measure that correlates positively with IQ and other socioeconomic indices. According to Budoff, a child's optimal level of performance is reflected by his/her posttest performance. A measure of the child's responsiveness to training is then obtained from the residualized gain score derived by adjusting the posttest score for the pretest level of performance. Budoff utilized these gain scores to identify three types of learners: (a) *high scorers* whose performance on the pretest was comparatively good, (b) *gainers* whose performance on the pretest was not adequate, but who after training demonstrated substantial gains on the posttest, and (c) *non-gainers* whose performance both before and after training was inadequate. Within this classification, *gainers* and *high-scorers* in special education classes were seen to profit from optimal curricula and classroom instruction.

Learning Potential Assessment Device (LPAD): Mediational Assessment

Feuerstein's approach to assessment is based on the theory of cognitive functioning, in which a lack of mediated learning experiences results in cognitive deficiencies. The Learning Potential Assessment Device (LPAD) developed by Feuerstein assesses the nature and extent of an individual's deficiencies as well as the amount and type of training (i.e., mediated learning) necessary for an individual to profit from direct learning. A cognitive map that consists of seven parameters is used to identify and clarify a learner's deficiencies for the purpose of modifying the deficiencies. The parameters include content, modality, phase, operations, level of complexity, level of abstraction, and level of efficiency. The LPAD device, like Budoff's approach, utilizes a test-train-test procedure. Based on clinical observations, the training phase provides mediation by systematically varying the parameters of the cognitive map.

Testing the limits assessment

Testing the limits approaches to assessment incorporate the integration of specific interventions within the testing process (Carlson & Wiedl, 1976, 1978, 1979). In this approach, various procedures that lead to higher levels of performance are used to test the limits of the child's abilities. They include (a) simple feedback, (b) prompting the child to verbalize how they solved the problem, (c) prompting the child to verbalize while solving the problem, (d) providing elaborated feedback that explains the principles involved in completing the task, and (e) a combination of prompting the child to verbalize during and after solution and providing an explanation of the principles needed to complete the task.

Graduated Prompting Assessment

Central to Campione, Brown and their colleagues' (Brown & French, 1979; Campione, Brown, & Ferrara, 1982) test-train-transfer-test method of assessment is the use of progressive prompts to enable successful solution and, therefore, assess a learner's potential. Pretesting is conducted to determine the starting performance on the assessment task and general intellectual ability. Dynamic components of this assessment method involve training and transfer phases. During the training and transfer tasks, a series of prompts based on task analysis are given. The prompts are sequenced from general and abstract to more explicit, specific, and concrete. The number of prompts needed for problem solution is an inverse function of learning/transfer ability (i.e., the higher the number of prompts required to solve a problem, the less likely transfer will occur). Transfer problems typically involve near, far, or very far items. Improvement in independent performance is assessed by a posttest similar to the pretest.

A Continuum of Assessment Model—Mediated and Graduated Prompting

This model (Bransford, Delclos, Vye, Burns, & Hasselbring, 1987) incorporates the assessment procedures of Feuerstein' mediation assessment and graduated prompting assessment of Campione, Brown, and associates. The technique involves the initial administration of static measures followed by the graduated prompting procedure described earlier. Children who perform below criterion in this phase are then

provided *mediation* dynamic assessment. The components of the mediation assessment include (a) familiarization of materials and basic cognitive functions, (b) teaching of specific rules and procedures necessary for task completion, and (c) feedback.

In sum, most models of dynamic assessment attempt to link assessment and instruction. Proponents of different models of dynamic assessment are in agreement that dynamic assessment provides a fair estimate of an individual's abilities, predicts how well an individual will perform independently in the future, and prescribes effective instructional strategies for remediating cognitive deficiencies (OSEP Research Project Directors' Conference, Palincsar, 1990).

Dynamic assessment has been a subject of much interest and research for several years. Despite the noted merits of dynamic assessment, such assessments have not been used extensively in educational practice. In fact, translating dynamic assessment models into assessment procedures is still in the early stages of research. Savell, Twohig, and Rachford (1986) note the difficulty in evaluating the alleged claims of dynamic assessment techniques because, in most cases, these techniques have not been empirically tested by researchers other than the ones who originally developed the techniques. Although the techniques developed by Feuerstein and his colleagues (Feuerstein, Rand, Hoffman, & Miller, 1980) are an exception to such empirical testing, nevertheless the data reported for their techniques raise a number of questions (See Savell, Twohig, & Rachford, 1986). The adequacy of dynamic assessment procedures is further limited by relatively sparse research in specific academic skill areas. The present discussion examines the research on dynamic assessment in the context of solving math problems.

Strategies for Facilitating Mathematical Problem-Solving Skills

A great deal of attention in recent years has focussed on mathematics instruction in schools. This emphasis is based in part on the National Assessment of Educational Progress (NAEP) report that children's problem-solving skills are deficient (Carpenter, Lindquist, Brown, Kouba, Silver, & Swafford, 1988; Kouba, Brown, Carpenter, Lindquist, Silver, & Swafford, 1988). Although this allegation seriously questions the problem solving ability of children, the use of traditional, static measures of assessment to evaluate data in NAEP may in fact understimate students' knowledge and the potential for modifying children's problem-solving skills with appropriate instructional procedures. Traditional assessment methods, as noted previously, fail to yield accurate estimates of students' learning potentials. They also ignore the differential quality of students' prior learning experiences in responding to adequate instructional environments. These problems with traditional assessment procedures call for reconsidering the NAEP results. Arguably, a new analysis using dynamic assessment tools that provide more sensitive estimates of children's problem-solving abilities could be both intriguing and informative.

Central to the process of problem solving is the ability to interrelate pieces of information within a knowledge base and to effectively connect understanding of concepts with procedures (or algorithms) needed to solve problems. A recent research synthesis by Prawat (1989) notes that the relationships interlinking various elements in a cognitive structure influence accessibility of knowledge. However, to access these relationships it is important to have an adequate knowledge base, which is the basis for connecting units of information and determining instructional strategies.

Prawat maintains that acquiring an adequate knowledge base should parallel efficient organization of knowledge. The significance of organizing knowledge for efficient recall of information is especially relevant when one considers individuals' limited capacity for processing information (Hasselbring, Goin, & Bransford, 1988). As such, it is pertinent that memory requirements placed on learners be minimized by adequately promoting understanding and connections between conceptual and procedural knowledge.

In characterizing the knowledge base, Prawat stresses the importance of linkage or connectedness that frames conceptual understanding. This is in direct contrast to understanding as an isolated piece of information. A learner's ability to gain quick and reliable access to his or her knowledge is seen to be based on the "elaborateness or richness between units of knowledge" that comprise that knowledge base (Chi & Koeske, 1983, cited in Prawat, 1989, p. 4). An important aspect of the knowledge base is the role of key ideas. As such, Prawat suggests that effective teaching should focus on a limited set of major ideas to enhance connectedness.

Prawat (1989) recommends a model of concepts that builds from the familiar to the unfamiliar. He describes instructional strategies that foster conceptual understanding of key ideas. One of these strategies emphasizes developing representational links by using concrete materials, analogies, or metaphors. Representations are construed as interpretations of ideas and can be verbal, pictorial or diagrammatic, or physical. They are conducive to solving problems and, with meaningful manipulation of concrete materials, representations can clarify an otherwise abstract concept (Lesh, Post, & Behr, 1987).

A second instructional strategy involves explicitly connecting elements of the knowledge base. This requires comparing and contrasting a new concept with

quires comparing and contrasting a new concept with previously learned concepts to promote connectedness and accessibility of the knowledge being acquired (Steinberg, Haymore & Marks, 1985). Also important in mathematics is connecting concepts and procedures that are pertinent to problem solving. Nesher (1986) notes that when procedures are conceptually understood they are more likely to be accessed and transferred to other similar problems. Prawat (1989) identifies two techniques that connect concepts, networking (i.e., breaking a whole into parts and then relating those parts) and concept maps (i.e., "diagrammatic or spatial representations of the relations between concepts," p. 11).

The final strategy entails utilizing students' informal knowledge and connecting it appropriately to formal, or school, knowledge. Teaching students to make connections between prior and new knowledge is seen as essential in facilitating meaningful learning. In addition, students must be taught to evaluate whether the prior knowledge is consistent or incompatible with the new knowledge and act accordingly. Novak and Gowin (1984) recommend utilizing concept maps to help students' identify unseen or erroneously linked concepts in their informal knowledge structure.

Given this framework, it is clear that developing conceptual connections within a knowledge base is important for knowledge acquisition and utilization. However, as Von Glaserfeld (cited in Prawat, 1989) notes, "Rarely, if ever, is there a hint, let alone an indication, of what one must do in order to build up the conceptual structures that are to be associated with the symbols" (p. 6). Although research discussions abound concerning connectedness, they are, for the most part, definitional. This research does not explicitly demonstrate how skills or concepts can be presented as semantic networks in which the various elements or components are interrelated. The lack of an explicit strategy to connect the various elements within a knowledge base makes it difficult for practitioners to teach how concepts should be interconnected. Students are, therefore, forced to make these connections on their own.

The research literature on dynamic assessment and the importance of linkages in facilitating conceptual understanding make it important to assess connectedness in students' mathematical comprehension of complex operations. It would be relevant if the assessment method that is utilized to examine complex mathematical operations (e.g., problem solving) incorporates the techniques suggested by Meyers (1987) to assess the process of learning within a model of dynamic assessment. The assessment tool must include a careful observation of the child's behavior during problem-solving situations and document the distinction between important and irrelevant behaviors. A facilitative questioning procedure must be utilized to examine how students solve problems, and detailed instructions about think-aloud procedures should be provided. In summary, the child's problem-solving strategies are of primary interest in evaluating mathematical skills. Furthermore, such evaluations must incorporate dynamic assessment procedures because dynamic assessments provide valuable insight about a child's ability to learn with effective instruction.

References

Bethge, H., Carlson, J. S., & Wiedl, K. H. (1982). The effects of dynamic assessment procedures on Raven Matrices performance, visual search behavior, test anxiety and test orientation. *Intelligence, 6*, 89-97.

Bransford, J. D., Delclos, V. R., Vye, N. J., Burns, M. S., & Hasselbring, T. S. (1987). State of the art and future directions. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 479-496). New York: Guilford Press.

Brown, A. L., & French, L. A. (1979). The zone of potential development: Implications for intelligence testing in the year 2000. *Intelligence, 3*, 255- 277.

Budoff, M. (1974). *Learning potential and educability among the educable mentally retarded.* (Final Report Project No. 312312). Washington, DC: National Institute of Education.

Campione, J. C. (1989). Assisted assessment: A taxonomy of approaches and an outline of strengths and weaknesses. *Journal of Learning Disabilities, 22*(3), 151-165.

Campione, J. C., Brown, A. L., & Connell, M. L. (1990). Metacognition: On the importance of understanding what you are doing. In R.I. Charles & E. A. Silver (Eds.), *Research agenda for mathematics education: Teaching and assessment of mathematical problem solving.* Hillsdale, New Jersey: Erlbaum.

Campione, J. C., Brown, A. L., & Ferrara, R. A. (1982). Mental retardation and intelligence. In R. Sternberg (Ed.), *Handbook of human intelligence.* Cambridge, England: Cambridge University Press.

Carlson, J. S., & Wiedl, K. H. (1976). Modes of presentation of the Raven Coloured Progressive Matrices Test: Toward a differential testing approach. *Trierer Psychologische Berichte, 3*, 7.

Carlson, J. S., & Wiedl, K. H. (1978). Use of testing-the-limits procedures in the assessment of intellectual capabilities in children with learning difficulties. *American Journal of Mental Deficiency, 82*, 559-564.

Carlson, J. S., & Wiedl, K. H. (1979). Toward a differential testing approach: Testing-the-limits employing the Raven Matrices. *Intelligence, 3*, 323-344.

Carpenter, T. P., Lindquist, M. M., Brown, C. A., Kouba, V. L., Silver, E. A., & Swafford, J. O. (1988). Results of the fourth NAEP assessment of mathematics: Trends and conclusions. *Arithmetic Teacher, 36*(4) 38-41.

Embretson, S. E. (1987). Toward development of a psychometric approach. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 141-170). New York: Guilford Press.

Ferrara, R. A. (1987). *Learning mathematics in the zone of proximal development: The importance of flexible use of knowledge.* Unpublished doctoral dissertation, University of Illinois: Champaign.

Feuerstein, R. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques.* Baltimore: University Park Press.

Feuerstein, R., Rand, Y., Hoffman, M. B., & Miller, R. (1980). *Instrumental enrichment: An intervention program for cognitive modifiability.* Baltimore, Maryland: University Park Press.

Hasselbring, T. S, Goin, L. I., & Bransford, J. D. (1988). Developing math automaticity in learning handicapped children: The role of computerized drill and practice. *Focus on Exceptional Children, 20*(6), 1-7.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.) (1982). *Placing children in special education: A strategy for equity.* Washington, D. C.: National Academy Press.

Jensen, M. R., & Feuerstein, R. (1987). The learning potential assessment device: From philosophy to practice. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 379-402). New York: Guilford Press.

Kameenui, E. J., & Simmons, D. C. (1990). *Designing instructional strategies: The prevention of academic learning problems.* Merrill Publishing Company.

Kouba, V. L., Brown, C. A., Carpenter, T. P., Lindquist, M. M., Silver, E. A., Swafford, J. O. (1988). Results of the fourth NAEP assessment of mathematics: Number, operations, and word problems. *Arithmetic Teacher, 35*(8), 14-19.

Lesh, R., Post, T., & Behr, M. (1987). Representations and translations among representations in mathematics learning and problem solving. In C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (pp. 207-214). Hillsdale, New Jersey: Erlbaum.

Nesher, P. (1986). Are mathematical understanding and algorithmic performance related? *For the Learning of Mathematics, 6*, 2-9.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn.* London: Cambridge University Press.

Meyers, J. (1987). The training of dynamic assessors. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 403-425). New York: Guilford Press.

Palincsar, A. M. (1990, July). *Dynamic assessment.* Paper presented at OSEP Research Project Directors' Conference, The Council for Exceptional Children, Washington, D. C.

Prawat, R. S. (1989). Promoting access to knowledge, strategy, and disposition in students: A research synthesis. *Review of Educational Research, 59*(1), 1-41.

Savell, J. M., Twohig, P. T., & Rachford, D. L. (1986). Empirical status of Feurestein's "Instrumental Enrichment" (FIE) technique as a method of teaching thinking skills. *Review of Educational Research, 56*(4), 381-410.

Steinberg, R., Haymore, J., & Marks, R. (1985, April). *Teachers' knowledge and structuring content in mathematics.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Ysseldyke, J. E., & Regan, R. (1980). Nondiscriminatory assessment: A formative model. *Exceptional Children, 46*, 465-466.

Content Analysis of Geography Textbooks Using a Taxonomy of Knowledge Forms

Victor Nolet
University of Oregon

Published textbooks are the primary instructional tool in social studies classes (Shaver, Davis & Helburn, 1979), with as much as 75% of instructional time structured by printed curriculum materials (EPIE Institute, 1978). Social studies textbooks have been the subject of considerable debate and criticism in the last decade, with attention focused on both their content and format. They have been characterized as banal compendiums of topics that treat few issues in depth (Tyson & Woodward, 1989) and described as "inconsiderate" of poor readers (Armbruster, 1984) because they often fail to integrate ideas across chapters, sections, or events within paragraphs.

Research that focuses on improving textbooks has shown that when textbook passages are rewritten to be more considerate, students show improved performance. For example, Konopak (1988) reported that average and high ability high school students showed improved vocabulary learning when they were instructed with passages from U.S. history textbooks that had been revised to be more coherent and explicit in presentation of information. Others have shown positive effects by manipulating text features (Herman, Anderson, Pearson, & Nagy, 1987; Meyer, Brandt, & Bluth, 1980) and reorganizing content within passages (Gold & Fleisher, 1986). However, problems may arise in implementing textbook modifications in classrooms.

First, conducting extensive rewrites of textbook passages probably is unrealistic for classroom teachers with limited time and financial resources. The practicality of teachers making passage modifications throughout an entire textbook has not been investigated systematically.

Second, students are not well served by curriculum materials that are to content-area textbooks what *Reader's Digest Condensed Books* are to literature. Textbook adaptations may provide incomplete or inaccurate information when conducted strictly for the purpose of improving readability (Armbruster, Osborn, & Davison, 1985). For example, Beech (1983) presents a number of guidelines for simplifying textbooks for poor readers, including sequencing information in chronological order, placing the main idea first in paragraphs, and eliminating extraneous or irrelevant information. Adaptations such as these might be difficult to implement in courses such as geography where information can't always be organized chronologically and discrimination of relevant from irrelevant information requires considerable domain-specific expertise and is largely context dependent.

To overcome potential difficulties associated with textbook modifications, Lovitt (in press) suggests a variety of instructional strategies such as advance organizers, study guides, graphic organizers, spatial displays, precision teaching vocabulary sheets, and teaching keywords. Strategies for using textbooks may be particularly powerful for low-achieving students characterized as "strategy deficient" (deBettencourt, 1987) because they involve skills that can be generalized across content areas and can be used with curriculum materials that vary widely in reading difficulty. Interestingly, while social studies teachers seem to endorse use of reading strategies, they seldom teach or encourage students to use them (Tixier y Vigil & Dick, 1987). Possibly, one reason for the reluctance of content-area teachers to incorporate strategies into their instruction is the amount of expertise and planning these approaches require.

Most strategies require the instructional planner to select key information from curriculum content to be formatted as study guides or graphic organizers, taught as background vocabulary, or incorporated into in-class discussions. Minimally, teachers must identify specific knowledge forms (e.g., facts, concepts and principles) contained in a passage that will be the focus of the strategy. Unfortunately, effective procedures for evaluating the knowledge forms contained in textbooks have not been validated.

The educational importance of describing and categorizing forms of knowledge in textbooks is considerable. For example Kameenui and Simmons (1990) describe instructional strategies designed for specific knowledge forms. Such procedures are a refinement of the works of Bruner (1966), Gagne (1985), and Englemann and Carnine (1982). The effectiveness of this approach to instructional design has been documented in a wide range of instructional applications (c.f., Gersten & Keating, 1987; Markle, 1975; Moore, 1986). Identification of knowledge forms also is important from the perspective of assessment. For example, classroom-based assessment in content domains requires explicit description of the knowledge forms to be tested to distinguish among items that require different intellectual operations (Roid & Haladyna, 1980; Tindal & Marston, 1990; Williams & Haladyna, 1983).

One of the most widely-used frameworks for categorizing information contained in instruction is the taxonomy of educational objectives proposed by Bloom, Englehart, Furst, Hill, and Krathowohl, 1956. Bloom presents a hierarchical arrangement of cognitive operations that are intended as outcomes for education in the cognitive domain. However, Bloom's taxonomy cannot be easily employed for analyzing textbook content. The idea that learners interact with different forms of information in different ways is implicit in Bloom's framework, but the taxonomy is not explicit in describing specific forms of knowledge (Becker, 1986; Kameenui & Simmons, 1990). Furthermore, educators tend to disagree about the specific behaviors associated with the various levels of Bloom's taxonomy (Fairbrother, 1975; Seddon,1978).

Recent attempts to develop operational definitions of the various forms of knowledge have resulted in descriptions that do not demand the high levels of inference for interpretation required by the Bloom, et al. framework. The taxonomies proposed by Kameenui and Simmons (1990) and Williams and Haladyna (1983) are particularly successful in this regard. The Kameenui and Simmons taxonomy arranges knowledge on a continuum of complexity that includes three levels of facts, concepts, rule relationships, and cognitive strategies. In the Williams and Haladyna framework, knowledge is formatted as facts, concepts, principles, or strategies. The value of these taxonomies is that they separate the intellectual operations that learners perform from the format of the knowledge that learners manipulate. Therefore, the information contained in curriculum materials can be analyzed systematically for the purposes of both instructional design and assessment.

To date, social studies textbooks have not been analyzed using a taxonomy of operationally defined knowledge forms. In the study described here, such an analysis was conducted in geography textbooks. The focus of the study was on developing a set of procedures that could be used by content area teachers to identify facts, concepts and principles contained in textbook passages to facilitate design of instruction and assessment.

Method

The work described here was undertaken in conjunction with a study in which instruction was observed for three weeks in a seventh grade world geography classroom. During this observation period, three topics were taught: "India and Its Neighbors," "North Africa," and "The Middle East." In the current investigation, chapters corresponding to these topics were examined in three world geography textbooks.

Materials

Analyses were completed on the world geography textbook used in the observed classroom and two additional textbooks randomly selected from a list published by the Geographic Education National Implementation Project (GENIP) (St. Peter, 1990). All three textbooks were the most recent annotated teacher's editions; all were published by major textbook publishers, and all are in wide distribution throughout the United States.

Procedures

Tables of contents and unit summaries were inspected to identify the chapters in each book that covered the topics of India, North Africa, and the Middle East. To control for differences among the textbooks in inclusion of graphic material (such as graphs, maps, and photos) and because of difficulty in analyzing the large quantity of information presented in graphic material (Mosenthal & Kirsch, 1990) only text was examined.

All three textbooks employed a two-column format in which columns of text were three inches wide by eight inches long, however, few pages contained only text. On most pages, sections of text were separated by graphic material placed within columns or across both columns. Therefore, a linear measure was obtained as an estimate of the total amount of text contained in each chapter. Column-wide segments of text were measured from the top of the first line of print to the bottom of the last line of print in the section. To facilitate interpretation, linear quantities of text were summarized as full page equivalents (FPE), that is, the amount of text on a single page that contains no graphic material. Full-page equivalents were estimated by dividing the total inches of text in each chapter by the number of inches of text in two full columns (16 inches). Type size was judged to be approximately comparable in all three textbooks and two books, all of which were formatted with 5 lines of type per inch. The third textbook was formatted with six lines per inch so estimates of full page equivalents obtained from this book were multiplied by 1.2 to correct for this difference.

All text contained within the main body of each chapter was inspected, including supplemental material, such as extended passages set in a italics or printed on a shaded background. To avoid recording redundant information, advance organizers and chapter summaries presented at the beginning of chapters were not included in the analysis. Similarly, vocabulary words set in margins or in boxes at the beginning of sections were excluded unless they were (a) accompanied by a definition and (b) not defined again within the text of the chapter. Material found at the end of chapters also was excluded, such as review questions, vocabulary quizzes, homework assignments, or enrichment exercises.

Three categories of knowledge forms were defined: facts, concepts, and principles. Textbook chapters were examined to determine how many of each type were present. Data collectors read through the chapters and each time they encountered a fact, concept, or principle they marked a recording protocol accordingly. Category definitions were adapted from taxonomies provided in Williams and Haladyna (1983) and Kameenui and Simmons (1990). The definitions developed for the three categories of knowledge forms, as well as examples and non examples of each are shown in Table 1.

Data were collected by three graduate students in the teacher education program at the University of Oregon. Prior to initiation of the study, all data collectors received approximately 3 hours of training pertaining to use of the knowledge forms definitions and procedures for recording data. Inter-judge agreement estimates obtained at the conclusion of training were 93% for facts, 85% for concepts and 95% for principles.

Table 1. Definitions, examples, and Non-examples of Knowledge Forms Counted in Textbook Passages

	FACTS
Definition	A simple association of a name, event, object, date, or place with a specific stimulus. Facts entail only one example or instance.
Examples	• Pierre is the capital of South Dakota. • The first Congress of the United States was composed of 26 senators and 59 representatives.
Non-examples	• state capitals • members of congress
	CONCEPTS
Definition	A class of events, names, dates, objects, places, etc., that share a common set of defining attributes or characteristics. Concepts involve three components: (a) a label, (b) the set of defining attributes, and (c) multiple examples or exemplars. At least the name and the defining attributes must be provided in a textbook passage for a concept to be recorded.
Examples	• A state is one of the fifty internally autonomous territorial and political units comprising the federation known as the United States • A capital is a town or city that is the official seat of government in a state, nation or other political entity. • An island is a body of land surrounded on all sides by water.
Non-examples	• Mississippi • Cairo • Vanuatu
	PRINCIPLES
Definition	A consistent relationship among events, objects, or behaviors. Principles generally can be stated as "if-then" or "when-then" statements. Principles generally involve multiple applications in which the fundamental relationship is constant across examples. To be coded in a textbook passage, the fundamental relationship must be stated explicitly, regardless of whether or not examples are provided.
Examples	• When demand goes up, then supply goes down. • If vital resources in a region become scarce, the inhabitants will move away. • If soil is not fertile, then it is poorly suited for farming.
Non-examples	• The most valuable resource in the Middle East is water. • Many plains Indians were nomadic. • In many countries, the land is not suitable for agriculture.

Following data collection, randomly selected passages from chapters analyzed by each observer were analyzed by a second observer and results were compared to determine inter-judge agreement. All such estimates were between 84% and 96%.

Results

The textbooks were generally similar in their treatment of the three topics ("India and Its Neighbors," "North Africa," and "The Middle East") but some notable differences were observed. In two of the textbooks, each of the topics was presented in separate chapters containing information about the physical and cultural geography of each region. In the third textbook, instruction about the region containing both North Africa and the Middle East was distributed across three short chapters while "India and Its Neighbors" was covered in a separate chapter. Also, each textbook included a slightly different combination of countries within each region. The manner in which each textbook taught the topics is shown in Table 2. The number of chapters devoted to each chapter is shown in parenthesis in the center column and the countries included in the region is shown in the column on the right.

The number of pages, total inches of text, and number of full page equivalents contained in each chapter of the textbooks are shown in Table 3.

The number of facts, concepts, and principles per chapter was computed by dividing the total number of each knowledge form by the number of chapters analyzed. Similarly, the number of facts, concepts, and principles per full page equivalent was computed by dividing the total number of each knowledge form by the number of FPE's in all chapters analyzed. These data are summarized in Table 4.

Discussion

The data indicate that in each chapter, geography textbooks present from 184 to 470 facts, but explicitly teach less than one principle. Concepts are taught explicitly on only one out of two full pages of text. The paucity of concepts and principles found in any of the textbooks must prompt examination of the accuracy of the analysis system used here. One possibility that must be considered is that the definitions of knowledge forms employed in this study (shown in Table 1) may have resulted in fewer concepts and principles being recorded than the textbooks actually presented. For example, more concepts might have been recorded if presentation of defining attributes was not a necessary condition in the definition. The following example illustrates this point.

All three textbooks introduced the term *cottage industry* in the context of "India and Its Neighbors," In textbooks A and B, the term was presented in boldface

Table 2. Treatment of Topics in Each Textbook

Textbook	*Topic*	*Countries Included Associated Chapters*
A	India and Its Neighbors (1 chapter)	India, Pakistan, Bangladesh, Afghanistan, Nepal, Bhutan, Sri-Lanka
	North Africa (1 chapter)	Morocco, Algeria, Tunisia, Libya, Egypt
	Middle East (1 chapter)	Bahrain, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, Yemen (Aden), Yemen (San'a)
B	India and Its Neighbors (1 chapter)	India, Pakistan, Bangladesh, Nepal, Bhutan, Sri-Lanka
	North Africa (1 chapter)	Morocco, Algeria, Tunisia, Libya, Egypt, Sudan
	Middle East (1 chapter)	Afghanistan, Bahrain, Cyprus, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, Yemen (Aden), Yemen (San'a)
C	India and Its Neighbors (1 chapter)	India, Pakistan, Bangladesh, Nepal, Bhutan, Sri-Lanka
	North Africa/Middle East (3 chapters)	Afghanistan, Algeria, Bahrain, Cyprus, Egypt, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco,Oman, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, Turkey, United Arab Emirates, Yemen (Aden), Yemen (San'a)

Table 3. Lengths of Chapters Analyzed

Textbook	Topic	Total Pages	Total Inches of Text	Full Page Equivalents
A	India and Its Neighbors	12	109.5	6.8
	North Africa	10	108.75	6.8
	Middle East	13	135	8.4
	Total Analyzed	35	353.25	22.0
B	India and Its Neighbors	14	120	9
	North Africa	12	114.8	8.5
	Middle East	18	181	13.5
	Total Analyzed	44	415.8	31.0
C	India and Its Neighbors	7	67	4.17
	North Africa/Middle East	20	163.5	10.2
	Total Analyzed	27	230.5	14.37

type, and followed by a set of defining attributes ("families use their own tools and work at home") plus examples (e.g., rug and fabric weaving, jewelry making, and pottery). In textbook C, the term *cottage industry* was included on a list of vocabulary words at the beginning of the chapter and underlined in red in the teacher's edition, however, no defining attributes were provided in the text. Furthermore, the reader was not instructed to refer to the glossary when "cottage industry" first appeared in the text, even though the term was defined there. Therefore, a concept was recorded for the presentation of "cottage industry" in textbooks A and B, but not in textbook C. If the definition of concept used in this study had required only inclusion of the concept label and positive examples, a concept would have been recorded for the presentation of "cottage industry" in all three textbooks.

The definitions of knowledge forms employed in this study are consistent with an extensive body of research that has investigated the characteristics of concepts (Bourne, Ekstrand & Domineski, 1971; Ehrenberg, 1981), strategies for teaching concepts (Jitendra & Kameenui, 1988; Klausmeier & Feldman,1975; Tennyson & Cocchiarella, 1986; Yoho, 1986), strategies for presenting concepts in texts (Kameenui, Simmons & Darch, 1987), and the properties of causal relationships contained in principles (Kelly, 1973). Findings consistently reported in this literature support the conclusion that the educational validity of using less rigorous definitions than those employed here may be dubious. Likely, presentations of concepts or principles that were invisible to the analysis employed in this study (e.g., *cottage industry* in textbook C) also would be invisible to many students reading the texts.

Table 4. Average number of Facts, Concepts, and Principles per Chapter in Each Textbook

Textbook	Knowledge Form	Average per Chapter	Average per FPE
A	Facts	339	46.3
	Concepts	3.6	.50
	Principles	1.3	--
B	Facts	470	45.5
	Concepts	4.6	.54
	Principles	<1	--
C	Facts	184	51.2
	Concepts	1.5	.41
	Principles	<1	--

Further inspection of textbook C revealed that the range of examples provided (e.g., weaving, spinning, and furniture making) did not illustrate the critical features that distinguish cottage industries from other types of industries. Despite the attention called to the term in the vocabulary list at the beginning of the chapter, and in the teacher's edition, the functional effect of the absence of defining attributes and effective examples was that, no instruction was provided to teach "cottage industry" in textbook C. In contrast, the presentations of "cottage industry" provided in textbooks A and B, albeit very lean (Markel, 1983), contained more of the features considered minimally necessary to teach concepts (Jitendra & Kameenui, 1988). To the extent that the analysis used in this study is sensitive to such educationally meaningful differences among textbooks, the validity of the system is supported.

Learners operating within a content domain need a foundation of domain-specific knowledge to effectively solve problems (Alexander & Judy, 1988). In geography, facts are employed in a range of problems such as summarizing trends of cultural diffusion, predicting effects of climate changes in a region, or evaluating patterns of urban growth (Jordan & Rowntree, 1990).

Given that middle or high school students generally take only one world geography class (O'Neil, 1989), the textbooks analyzed in this study are most often used by novice geographers who are likely to possess little domain-specific knowledge. Therefore, to some extent, encyclopedic presentation of hundreds of facts per chapter is at least understandable, if not defensible. However, the profile of geography textbooks that emerges in this study suggests that both students and teachers are faced with the problem of organizing a huge quantity of information. Extrapolated over the length of an entire year, the shear volume of information presented in these textbooks is staggering. Consider textbook B, which contains 28 chapters at 470 facts per chapter or approximately 13,160 facts, not counting graphs, maps or photos.

These findings present intriguing research and instructional challenges. Our understanding of the manner in which students construct meaning when bombarded with this much information is limited and the design characteristics for instruction aimed at teaching content areas such as geography, have not been well defined.

References

Alexander, P. A., & Judy, J. E.,(1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research, 58*(4) 375-404.

Armbruster, B. B. (1984). The problem of "inconsiderate text." In G.G. Duffy, L. R. Roehler, & J. Mason (Eds.) *Comprehension instruction: Perspectives and suggestions.* New York: Longman.

Armbruster, B. B., Osborn, J. H., & Davison, A. L. (1985). Readability formulas may be dangerous to your textbooks. *Educational Leadership,* (April) 18-25.

Becker, W. C. (1986). *Applied psychology for teachers.* Chicago: Scientific Research Associates.

Beech, M. C. (1983). Simplifying text for mainstreamed students. *Journal of Learning Disabilities, 16*(7) 400-402.

Bloom , B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives. *The classification of educational goals: Handbook I, cognitive domain.* New York: Longmans, Green.

Bourne, L. E., Ekstrand B. R., & Domineski, R. L. (1971). *The psychology of thinking.* New York: Prentice Hall.

Bruner J. S. (1966), *Toward a theory of instruction.* Cambridge, Massachusetts: Belknap Press of Harvard University.

deBettencourt, L. U. (1987). Strategy training: A need for clarification. *Exceptional Children, 54*(1), 24-30.

Ehrenberg, S. (1981). Concept learning: How to make it happen in the classroom. *Educational Leadership,* (October) 36-40.

Englemann, S. and Carnine , D. (1982). *Theory of instruction: Principles and applications.* New York: Irvington.

EPIE Institute (1978). *Report on a national study of the nature and quality of instructional materials most used by teachers and learners.* EPIE Report Number 76. Author.

Fairbrother, R. W. (1975). The reliability of teachers' judgements of the abilities being tested by multiple choice items. *Educational Research, 17,* 202-210

Gagne R. M. (1985). *The Conditions of learning and theory of instruction* (4th ed.). New York: Holt, Rinehart, & Winston.

Gersten, R. & Keating, T. (1987). Long-term benefits from direct instruction. *Educational Leadership* (March), 28-31.

Gold, J. & Fleisher, L. (1986). Comprehension breakdown with inductively organized text: Differences between average and disabled readers. *Remedial and Special Education, 7*(4), 26-32.

Herman, P., Anderson, R., Pearson, P., & Nagy, W. (1987). Incidental acquisition of word meaning from four expositions with varied text features. *Remedial Research Quarterly, 22*(3), 263-284.

Jitendra, A. & Kameenui, E. (1988). A design-of-instruction analysis of concepts teaching in five basal language programs: Violations from the bottom up. *The Journal of Special Education, 22* (2) 199-219.

Jordan , T. G. & Rowntree, L. (1990). *The human mosaic: A thematic approach introduction to cultural geography.* New York: Harper & Row.

Kameenui, E. J., & Simmons, D. C. (1990). *Designing instructional strategies: prevention of academic learning problems.* Columbus, Ohio: Merrill.

Kameenui, E. J., Simmons, D. C. & Darch, C. B. (1987). LD children's comprehension of selected textual features: Effects of proximity of information. *Learning Disability Quarterly 10,* 237-248.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28,* 107-128.

Klausmeier H. J., & Feldman, K. V. (1975). Effects of a definition and a varying number of examples and nonexamples on concept attainment. *Journal of Educational Psychology, 67*(2), 174-178.

Konopak, B. C. (1988). Effects of inconsiderate vs. considerate text on secondary students vocabulary learning. *Journal or Reading Behavior, 20*(1), 25-41.

Lovitt, T. C. (in press). Adapting materials for handicapped learners. G. Stoner, M. Shinn, & H. Walker (Eds.), *A school psychologist's interventions for regular education.* Washington D. C.: National Association of School Psychologists.

Markle, S. M. (1975). They teach concepts, don't they? *Educational Leadership, 4,* 3-9.

Markle, S. M. (1983). *Designs for Instructional Designers.* Champagne Illinois: Stipes.

Meyer, B. J. F., Brandt, D. M., and Bluth, G. J.,(1980). Use of top structures in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly, 16*(1), 72-103.

Moore, J. (1986). Direct instruction: A model of instructional design. *Educational Psychology 6*(3), 201-229.

Mosenthal P. B. & Kirsch, I. S. (1990) Understanding general reference maps. *Journal of Reading 34*(1), 60-63.

O'Neil, J. (1989, November). Social studies: Charting a course for a field adrift. *ASCD Curriculum Update,* 1-7.

Roid, G. H., & Haladyna, T. M. (1980). *A technology for test-item writing.* New York: Harcourt, Brace, Jovanovich.

Seddon, G. M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of Educational Research, 48,* 303-323.

Shaver, J. P., Davis, O. L., & Helburn, S. (1979). The status of social studies education: Impressions from three NSF studies. *Social Education, 43*(2), 150-153.

St. Peter, P. (1990). *Text Assessments in Geography: Interpretive Analyses of Standard Geography Textbooks, 7-12.* Indiana, Pennsylvania: Geographic Education National Implementation Project GENIP.

Tennyson, R. D., & Cochiarella, M. J. (1986). An empirically based instructional design theory for teaching concepts. *Review of Educational Research, 56*(1), 40-71

Tindal, G. R, & Marston, D. (1990). *Classroom-based Assessment: Evaluating Instructional Outcomes.* Columbus, Ohio: Charles Merrill.

Tixier y Vigil, Y. & Dick, J. (1987). Attitudes toward and perceived use of textbook reading strategies among junior and senior high school social studies teachers. *Theory and Research in Social Education, 15*(1), 51-59.

Tyson, H. & Woodward, A. (1989). Why students aren't learning very much from textbooks. *Educational Leadership 47*(3), 14-17.

Williams, R. G., & Haladyna, T. M. (1982). Logical operations for generating intended questions (LOGIQ): A typology for higher level test items. In G. H Roid & T. M. Haladyna, *A Technology for test-item Writing.* New York: Academic Press.

Yoho, R. F. (1986). Effectiveness of four concept teaching strategies on social studies concept acquisition and retention. *Theory and Research in Social Education, 14*(3), 211-223.

The Portfolio Concept: Applications in Curriculum-Based Measurement

Tracey Hall
Gerald Tindal
University of Oregon

Testing is an educational tradition, used to evaluate programs and classify people. However, the institutional benefits of assessment have become further and further removed from the learner (Martinez & Lipson, 1989). Educators have recently begun to argue that because current assessment methods have "not served learners directly," they must question current practices and call for alternatives to standardized assessment procedures (Wolf, 1989). We cannot eliminate testing altogether (Salmon-Cox,1981), yet the addition of new procedures designed to meet teachers' needs could flood the classroom with testing, taking away valuable instructional time. The field appears to be searching for alternatives to traditional assessment, specifically, standardized tests (Neill, & Medina, 1989).

A variety of options have been introduced to educators, one of which is the concept of an educational portfolio (see special issue in *Educational Leadership*, April, 1989). Using the idea of an artisan or craftsman, self-created products are assembled to reflect both the diversity and proficiency of skills. For example, architects develop portfolios that include the variety of structures they have drafted, demonstrating their proficiency in design. This concept is now being adopted in education.

The portfolio concept in education would translate into a folder of products reflecting academic proficiency. In an academic area such as writing, a folder might include samples of student poetry, journal entries, essays, or reports. Over the course of the year, writing samples should reflect the improvement a student has made as a function of instruction. We would hope to see samples near the end of the year that are better organized, develop a unique statement, and reflect better and more consistent language and mechanics. Such entries could be used both to display the diversity of students' skills as well as their improvement over time. Construction of an educational portfolio acknowledges the unique focus of the individual student without the contrived formats that accompany mass-production assessment.

Learning portfolios have the potential to be highly diverse, and they may incorporate any performance in any subject area. With a slight variation in administration, many products that reflect any instructional outcome generated in the classroom could be included within a portfolio. Using learning portfolios, student performance could be evaluated as work samples directly or be summarized in the form of progress charts. In the end, a teacher could have access to student products that encompass the entire range of classroom behaviors for which they have focused their instruction.

While it sounds as if learning portfolios may solve all of our assessment needs, caution is in order. We don't want an educational portfolio to become an incoherent scrapbook. Without careful regulation, the concept could lose its' potential for providing a representative sample of student performance. For example, without information in all academic areas, we would not want to draw inappropriate conclusions about overall student performance.

Comparing work samples would be difficult with a portfolio, if conditions under which the samples were collected were not controlled. Questions such as, "How much prompting or teacher student interaction was present?" must be addressed. If teachers differentially helped students (i.e., helped some students more often than others), then the products would become difficult to compare over time. To the degree that the products reflect differential learning, nothing can be said about improvement. Variations that are problematic include the following: (a) stimulus formats for student responses, (b) administration procedures, such as directions by the teacher, written directions on forms, timing, etc., (c) scoring and marking procedures, and (d) summarizing and reporting techniques. Such variations can occur within a single teacher's practices, within and across students over time, across products, or across all three—teachers, students and products. If there is an absence of systematic procedures, teachers must be cautious when rating students on any measure.

In the same manner that standardization is important in assessment delivery, standardizing the language of communication about students is important across educators. Portfolios also have the potential to provide a common basis for professionals to discuss student performance. However, specialists may have their own view or angle from which to interpret a student's work, as well as communicate using their

specialized technical language. This scenario is very much like the phone game: Everyone evaluating the student may have a different interpretation, even though the product that informed the inference was the same. Unless we structure the language to be used with the portfolio concept, we may lose the dialogue among professionals.

Another advantage of standardized procedures is the ability to capture the concept of individual or group growth over time. While change is occurring on a daily basis in classrooms, the anchors that are used to reflect those changes also move over the course of a school year. Teachers are teaching many different skills over time, and capturing growth implies some consistency. In order to solve this dilemma, we can establish a common task and watch for improvement over time. In this way, the behaviors measured show the change in student performance on a task or set of skills that has been carefully controlled. The products therefore represent student improvement on that specific skill, the change occurring as a result of instruction, and maturation.

In summary, learning portfolios appear to be a good idea, but they lack specifics regarding development. Until the problems noted above are resolved, what we may have is simply a permanent product version of anecdotal information. This information can be used in communicating with parents and helping teachers inform instruction. It is unlikely, however, that it could be useful in formatively and substantively evaluating instructional programs. The question, then, is whether we can capture the essential features of a portfolio assessment and infuse some kind of standardization in their creation.

While most of the literature on learning portfolios has come from the general education environment, a fair amount of research and development has occurred in special education. A format to address the above concerns with application to a portfolio is Curriculum-Based Measurement (CBM) (Deno, 1989). CBM is a method of analyzing performance using systematic procedures with brief measures in specific academic areas. Although CBM has traditionally been investigated in the context of special education, these measures have more recently been applied in a general education context (Wesson, Vierthaler, & Haubrich,1989; Hall & Tindal, 1989). As reflected in the name, CBM emphasizes on curriculum items using a production response format, in contrast to selection response, typical of most standardized achievement tests. In the basic skill areas, this includes production responses such as reading passages, spelling words, creating sentences with written words, and calculating math answers.

CBM procedures are standardized, meaning that conditions under which the measures are collected are always the same. This assures a data base for individual evaluation that is reliable and valid. CBM measures have been designed to spread student performance scores across a distribution for the purpose of peer comparison. The issue of technical adequacy is sufficiently addressed when comparisons are made between an individual's performance and the performance of that student's peers (Deno, 1989). Additionally, because there is a permanent product with production responses, CBM measures may be analyzed both quantitatively and qualitatively.

A final consideration for using CBM to collect examples of student performance in a portfolio is the timeliness of the measures in today's classroom. The procedures required to collect samples using CBM are very brief and unobtrusive. Each measure requires between one to four minutes to administer. Since these measures are so easy to collect, teachers can quickly administer a desired measure to a class or individual to contribute to their portfolio.

Method

We implemented Curriculum-Based Measurement with the students in the fall and spring of the 1989-90 school year to illustrate the student portfolio as a document representing progress or change in performance over time, within and across classes, and across abilities, The two testing periods were early October and late May. The methodology dealt with three issues: subjects, instrumentation, and data analysis. All students were tested in four basic skill subject areas, two of which are described below. Results are reported for two students.

Subjects

Students attended two elementary schools in a small, urban school district in the Pacific Northwest. Twelve students were selected for this project by their teachers, to represent a range of ages and skills. The ages and ability levels across subjects were as follows: three regular education students of high ability, (Grade 1 = 2, Grade 5 = 1); three regular education low ability students, (Grade 1=2, Grade 5=1); two students receiving services in Chapter One reading (Grade 3 = 2); four identified special education learning disabled students (Grade 2 = 2, Grade 4 = 2). This report will focus on two of those students.

Results are reported using fictitious names. In reading, results are reported for one "low performing" first grader, Jake. An illustration of spelling measures is provided by a fourth grade mildly handicapped student named Tom, who is receiving services in reading, math, and spelling.

Instrumentation

Students were tested in four basic academic areas: reading, written expression, spelling, and math. The measures require a minimal amount of time to administer and are representative of the instructional curriculum. Following is an explanation of the measures reported in this article.

Reading. All students were asked to orally read passages selected from the school-adopted reading series. Readability levels from this text varied greatly, sometimes as much as five grade levels within the text. Therefore, only passages with readability levels within + 1.5 grade-level range were identified, and a random selection from this pool was chosen for the portfolio assessment.

Students were tested individually, away from classroom distractions in an office or lobby area outside the classroom. Standard administration procedures were used for each test. Each student's oral reading was timed for one minute per passage. Identical procedures were used in the fall and spring testing periods.

Spelling. The spelling assessment was given to second through fifth graders. Students had to produce their responses by writing words dictated from a list. The spelling test was developed by obtaining a random selection of words from the grade level spelling curriculum. The examiner used a "rolling dictation" to present 16 to 18 words to students, depending on their grade level. The tester dictated each spelling word three times, following standardized procedures: The word was stated, then used in a pre-planned phrase, and finally stated a final time. The words were dictated with 8 to 10 second intervals, depending on the grade level. The tests were administered individually because of the small number of participating students.

Data Analysis

Initially, all tests were scored using quantitative procedures which yielded a numerical score. Students were evaluated individually, using measures of rate, percent correct, and percent improvement. A final method of quantitative evaluation is norm-referenced. In this investigation, we used box plots to describe the normative data. Box plots graphically represent the distribution of scores, specifically, the middle 50% of the population (see Figure 1). Within the box, the bottom line the 25th percentile, the box top the 75th percentile, and the line mid-box represents the 50th percentile. The bottom and top T's represent the 10th and 90th percentiles, respectively.

The second type of analysis employed was qualitative, with student performance evaluated on variables that are less countable and somewhat subjective. Qualitative procedures may easily be described for reading and written expression. There are few, if any, qualitative measures used in spelling and math scoring. One might be able to evaluate the student's neatness, in relation to handwriting or alignment of numerals or letters, however, these attributes are less informative with regard to growth in those subject areas. Therefore, no qualitative measures will be described for spelling.

Reading. In scoring individual reading, the following quantitative error types consistently counted as an error:

- Mis-identification—the student said the wrong word.
- Substitution—the student said a synonymous word instead of the printed word (e.g., house for home).
- Omission—the student completely skipped a word or words printed on the page.
- Reversals— the student transposed a word or words (e.g., 'was' for 'saw,' 'said she' for 'she said').
- Three-second hesitations— the student did not decode the word within three seconds of reading the previous word.

The following were not considered errors: repetitions—a word or words reread; insertions—the addition of words not on the page; and self-corrections—the student corrected a decoding error within 2-3 seconds. The number of errors was subtracted from the total number of words read to determine the number of words read correctly per minute.

Audio recordings of each student's reading were analyzed for voice quality, reading expressiveness, attention to end marks, and other punctuation.

Spelling. Spelling tests were analyzed using two scoring procedures. First we counted a word correct when the entire word was written using the appropriate letters in the proper order. Second, we counted Correct Letter Sequences (CLS), pairs of letters within the word that are in the appropriate order. For example, if the student spells the word myself, "miself," the correct letter sequence would be counted as 5, (one for starting the word correctly with *m*, one count for the *s* to *e*, one count for the *e* to *l*, one count for the *l* to *f*,and a final count for ending the word with an *f*. (The word "myself" would receive 6 CLS if spelled entirely correct).

Results

The results of this study will be reported to illustrate the flexibility in using a CBM portfolio. Student scores can be evaluated in relation to individual growth or improvement over time, as well as in comparison of the individual's performance to a peer population, (in this case a district norm group). Student reading performance has been both quantitatively and qualitatively judged. Select cases in reading and spelling have been discussed to exemplify the CBM portfolio.

Table 1. First Grade Chapter Regular Education Student, Jake, Reading.

Words Read				
Fall		Spring		% Improvement
Correct	Incorrect	Correct	Incorrect	
4	10	100	1	71%

Reading

In reading there are a number of variables that can be counted, heard, and observed as a student reads. Selections of student reading have been included to illustrate the use of CBM measures in reading while evaluating both quantitative and qualitative data. The student's reading was audio taped to provide a permanent product.

TheCase of Jake. Jake was not being considered for any special services, yet the teacher was aware that his performance was below average in this first-grade class at the beginning of the school year. In the fall, Jake

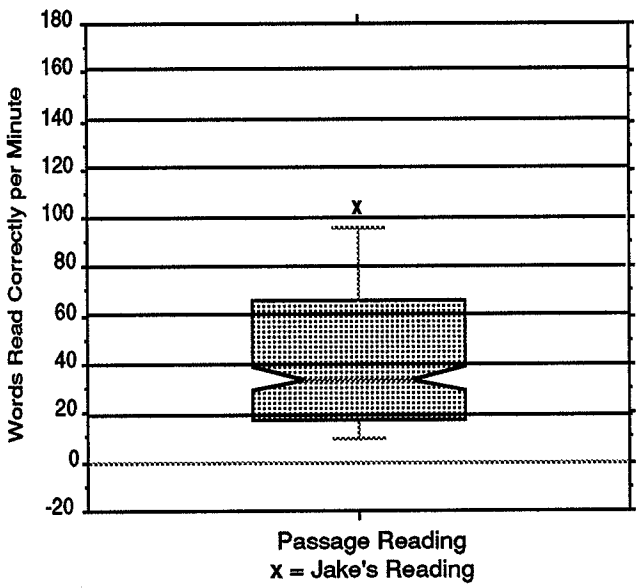


Figure 1. Box Plot for First Grade Spring Reading Scores

read at a rate of four words correct in the one minute; 10 words were read incorrectly, which is 28% correct words per minute of reading. In contrast, he read with 99% accuracy in the spring, the rate of correct words per minute having improved to 100 and decreasing errors to one (see Table 1). Jake made an individual improvement of 71% from October to May of his first-grade year. Percent improvement is determined by calculating the percent accuracy of correct words read in the fall and spring and finding the difference of the two scores.

Reading performance also can be quantitatively compared to a peer group. In this school, the entire first grade was assessed in reading in the winter and spring using the procedures described above. There was no fall assessment of the entire first grade. Jake's scores can be compared to that of his peers for the spring of this school year. His score of 100 words read correctly per minute is slightly above the 90th percentile for the first grade (See x on the box plot in Figure 1). By spring, Jake, who had been rated by his teacher as a lower-performing student, ranked above the 90th percentile when compared to his first grade peers in the district.

Jake's performance also was evaluated qualitatively. First graders read from the following from a passage in the fall:

> *It was winter. Ice was on the lake. Otter got out his skates. He got out his helmet and pads. He put on his cape and mittens. He held a red cane in his hand.*
>
> *Otter was a careful skater. He skated up and down the lake.*

A phonetic transcription of Jake's reading of this passage appears below. The periods between words indicate a time lapse as the student read each word inthe passage; more periods represent more time. Examiner-assisted, and self-correct errors are also noted in the transcription. A key is provided below to explain each symbol used in this transcription.

Key
.... = time between words, greater with each period
] = end of one minute time period
( ) = examiner said word
/ = quickly self-corrected

...It.....(was).....(winter).....(ice).....o..on..the.....
(lake).....(Otter).....(got).....(out).....p..putt...hims...
.....(skates).....h..he.

Jake's reading was very choppy, he separated each word, and actually decoded very few words in the passage. The tester, in following procedures, told him eight words on which Jake had hesitated for more than three seconds. It is questionable as to whether Jake could understand any of what he had read as a story. In listening to the audio tape voice quality, inflection, expression, and intonation do not exist.

In the spring there is a significant contrast in Jake's reading as described above in the quantitative analysis and as can be seen in the following transcription of his reading. First graders read from the following passage in the spring:

> *"Wake up, Shawn," Mom called as she tugged on the sheets. "It's time for Grandpa to come."*

Shawn was up in a flash! He put on shorts and a T-shirt.

Shawn sat on the front steps and waited for Grandpa. At last Shawn saw Grandpa's car drive up. He dashed down the steps and jumped into the car.

"are you all set to go to the shore?" asked Grandpa.

"Let's go," said Shawn, as he waved to his mom and dad.

Shawn loved the seashore. He liked to splash in the warm waves. He liked to find little fish in the water near the rocks.

Below is a transcription of Jake's reading in the spring. Note the smaller time lapse between words. Jake also used more voice intonation to express the questions and exclamations in the passage.

Overall, this example of spring reading is much easier for an audience to hear, and Jake appears to be understanding the words he read through expression and intonation.

"..Wake.up.Shawn"..Mom.called..as.she..tugged.on.
the..sheets.
"It's..time..for..Grandpa..to..come".Shawn..
wake.up.in..a.flash..He..put.on.on..his..shorts.and.a.T-
shirt..
Shawn.ssat.on.the.front.porch.steps..and..waited.for..
Grandpa..At.last.Shawn.saw.Grandpa's.car..drive.up.he..
dashed.down..the.steps..and.jumped..into.the..car.

"All..are.you.all.set.to.go.to.the.shore.yet?"asked..
Grandpa.
"Let's..go!!....said..Shawn..as.he.wwaved.to..his..
mom..and.dad.
Shawn.l.loloved
the..seassshore...He..licked/liked..to..
splashsh..in.the.waves/warm
waves..He..liked..to.find.. little..fish.in.the]

Spelling

The Case of Tom. Tom had been receiving special services in spelling and his other IEP areas for three years. For this study, he was assessed in fourth grade spelling material. His results were evaluated based on individual performance and improvement.

Using the quantitative measures of correct letter sequences and total words (see Table 2), Tom made some individual growth in this school year. In the fall, he was able to spell 2 of the possible 18 total words correctly. Using the more sensitive measure of Correct Letter Sequences, Tom scored 56 CLS which indicates 49% accuracy CLS in the fall assessment.

Table 2. Fourth Grade Special Education Student, Tom, Spelling

	Fall		Spring		% Improvement
	Correct	Incorrect	Correct	Incorrect	
Total Correct Words	2	16	6	10	22%
Correct Letter Sequences	57	59	89	36	22%

To assess growth in the school year we then compared Tom's fall spelling scores to those he obtained in the spring. Here we looked for numerical change and calculated percentage of improvement. In this case, Tom made a 22% improvement in both total words and CLS.

Discussion

As can be seen with the two examples presented here, the notion of a learning portfolio has tremendous transportability. Products within a portfolio can accommodate a wide range of academic behaviors that are useful for making a number of educational decisions. They can also be used by a wide range of professionals across general education and special education environments. Because the products are classroom-focused, within the school, year-long portfolios have a great potential for showing progress over time. Furthermore, they can be useful across all grade levels. Finally, a great diversity exists in the manner in which portfolios are generated and/or maintained.

As teachers become familiar with students in their classes, and come to understand the improvements children are making (or not making), they should be better able to make instructional decisions. For some students this might mean simply providing more structure, more practice, and more guidance. For others, more intensive interventions may be needed, such as working one-to-one with peers and counselors, or providing guided practice. There may be evidence that extreme interventions requiring assessments for special education placement are needed. This was clearly demonstrated in the first example of Jake's reading. The teacher had targeted Jake as a low reader at the beginning of the school year. With CBM portfolio data, she was able to make a conscious decision to devote special time to Jake and students with similar deficits to help their reading performance. In this case, Jake made tremendous improvement ranking above the 90th percentile when compared to his peers by the end of the school year.

In this scenario, CBM portfolios are very transportable across the general education and special education environments. They provide the focus from which professionals can interact in developing individualized instruction.

In a larger sense, CBM portfolios are quite timeless. Any particular work sample included within the portfolio simply represent a "snapshot" of what a student has done at one point in time. They need little explanation, other than the time and context in which they were constructed. To illustrate, parents take pictures of their children and explain that this is what they looked like when they were five. As such, portfolios provide a powerful reporting function. Teachers can generate learning products any time during the year and show them to parents to say, "This is what your child 'looked like' (how he performed) in November. Now that it's the end of the year, look at the great improvement that has occurred. Notice how her handwriting has improved, she's using higher level vocabulary. Your child has clearly mastered several language and writing skills."

The very same reporting function can be accomplished within a school among teachers. As students move across the grades, teachers can inform the next grade level teacher about specific student tendencies (i.e., skill specific strengths and weaknesses) as illustrated through a CBM portfolio.

A number of issues must be examined with respect to the CBM portfolio. For example, is there an ideal number of samples to collect within a school year? Should the collection of work samples under standardized conditions be restricted to reading, math, spelling, and written expression? How much is enough to capture change over time? Should every child have a portfolio in all areas? Before schools jump ship from current assessment practices and adopt a portfolio concept using CBM, questions like these and many more need to be researched and answered empirically to ensure good practices for educational decision-making.

References

Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1-17). New York: Guilford.

Hall, T., & Tindal, G. (1989). Using curriculum-based measures to group students in reading. In G. Tindal, K. Essick, N. George, & M. George (Eds.), *The Oregon Conference '89 Monograph*. Eugene, OR: College of Education, University of Oregon.

Martinez, M. E., & Lipson, J. I. (1989). Assessment for Learning. *Educational Leadership*, April, 73-75.

Neill, D. M., & Medina, J. J. (1989). Standardized testing: Harmful to educational health. *Phi Delta Kappan, May*, 688-697.

Salmon-Cox, L. (1981). Teachers and Standardized Achievement Tests: What's really Happening? *Phi Delta Kappan, 62*, 631-634.

Wesson, C. L., Viertaler, J. M., & Haubrich, P.A. (1989). An efficient technique for establishing reading groups. *The Reading Teacher, 42*, 466-469.

Wolf, D. P. (1989). Portfolio assessment: Sampling student work. *Educational Leadership. 46*, 35-39.

Floating Percentile Norms for Progress Monitoring Data

Richard Parker
University of Oregon
Stephanie Stein
Central Washington University
Gerald Tindal
University of Oregon

Frequent assessment of progress on long-term-goal tasks is commonly recommended for students with low basic skills. Long-term-goal (LTG) tasks permit students to demonstrate increasing speed and accuracy on terminal skills, such as oral reading from a basal text, or creative writing from a story starter. Monitoring of progress on a semi-weekly, weekly, or bi-weekly basis yields a data series which can be used to judge the success of the instructional program and help decide on improvements needed in the instructional program (Deno & Mirkin, 1977; White & Liberty, 1976; Salvia & Hughes, 1990).

Student progress can be judged from line graphs with time on the horizontal axis and performance scores on the vertical. This type of data comparison has been termed *individually referenced* (IR) because the student's current performance is compared with his/her own previous performance, rather than with the performance of peers, the latter defining "norm referenced" (NR) data comparisons (Tindal & Marston, 1990). The combination of progress monitoring and IR data comparisons has considerable empirical support as a method for improving students' basic skills (see review by Fuchs & Fuchs, 1986).

Despite the empirical support, progress monitoring and IR data comparisons are not widely used by special education or remedial teachers (Wesson, King, & Deno, 1984). The methods may be too time-consuming, too technical, or may involve test-tasks which are not similar to tasks being performed in class (Coffman, 1983; Yeh, Herman, & Rudner, 1981). Another reason for lack of use may be that IR data comparisons are foreign to common testing practice. Most teacher-created or curriculum-embedded tests yield only *number correct* or *percent correct* scores without normative comparisons. Published achievement tests (individual and group administered) yield NR scores such as group means, percentile ranks, and grade equivalents. These NR scores are familiar to teachers, but are not a part of IR time series data comparisons.

To increase the acceptability and utility of IR time series data comparisons, we suggest overlaying NR data standards on time series graphs. Although generally not a part of progress monitoring, including NR group-comparisons has at least four rationales: (a) serving group-based program delivery, (b) permitting stronger hypotheses for program change, (c) reporting to parents, and (d) serving accountability and group placement. First, although progress monitoring is nearly always conducted for individual students, instruction usually occurs in groups. Instructional changes which must be made for the entire group should be supported by NR group performance data.

The second reason for including NR standards with time series graphs is the additional guidance it would provide for deciding what type of instructional change to make. For example, if one student's scores show marked deterioration over three weeks, we might hypothesize that (a) events elsewhere in the student's life are interferring, (b) the student needs increased incentives, (c) the performance goal should be lowered, (d) a normal, temporary "plateau" has occurred in the learning curve, (e) the instructional program is deficient and should be modified. However, comparing this student's performance graph with those of peers in the same instructional group better focuses our hypotheses. If the three-week performance drop is shared by most other students, then we are less likely to consider (a) interferring life events, (b) increasing individual incentives, (c) lowering the performance goal, or (d) dismissing the drop as a temporary skill plateau (since peers differ in skill levels). We would be more likely to consider (b) increasing group incentives, and/or (e) modifying the instructional program.

Combining NR comparisons with IR graphs also should improve the reporting and documenting of student performance. As are teachers, parents are accustomed to performance summaries of either the criterion referenced ("Has she mastered it?" "How well can she do it?) or NR ("Is he at the bottom of the group?" "Is he at least average?") variety. Instead of choosing between time series data or unrelated NR group test scores, normative comparisons could be overlaid onto the IR graphs.

The final occasion when IR time-series data may not be suitable or sufficient is when progress must be documented. Documentation of student progress fulfills the requirement for teacher and school accountability, and often helps make promotion decisions and group or class placements in the future. These purposes require simple summaries, easily interpreted by others (including those outside the school) at a later date. Both of these purposes also require a normative comparison with peers. Again, IR progress monitoring data can meet these needs when NR comparisons are overlaid on time series graphs.

In this paper we will show how dynamic or *floating* percentile norms can be fitted to graphs of weekly basic skills assessments using LTG reading passages. This percentile line fitting technique is most efficient when only three (25th, 50th, 75th) to five (10th, 25th, 50th, 75th, 90th) different percentile benchmarks are desired, and with small-to-medium size instructional groups (e.g. 5 to 25). The technique will be described and applied to oral reading fluency data collected over 12 weeks from students in three small remedial reading groups. Finally, two aspects of reliability of the floating percentiles will be discussed—their descriptive accuracy, and their measurement error.

Time Series Data Summary

A group of test scores commonly is summarized by the mean, median, range, and standard deviation. Test scores in a time series more often are summarized by a trend line drawn along the time axis through the center of the data. The line's level and slope indicate generally whether and how much a student is improving over time. The line may be created by computer, through common least-squares regression. Simpler pencil-and-ruler methods also have been available to practitioners for nearly a century, and are particularly warranted when the data are markedly skewed or show unequal variance over time (Hoaglin, Mosteller, & Tukey, 1983).

Wald's (1940) seminal line-fitting article recognized earlier techniques by C. F. Roos (1928) and R. J. Adcock (1877). Several of these robust methods which have survived today are summarized by Hoaglin, Mosteller, & Tukey (1983). The various methods are similar. First, the data are sectioned in halves (Wald, 1940; Brown & Mood, 1951) or thirds (Bartlett, 1949; Gibson & Jowett, 1957; Tukey, 1970) by vertical line(s). Next, each section is again vertically divided in half, forming quarters or sixths. Finally, on these latter dividing lines are marked the means (Wald, 1940; Bartlett, 1949; Gibson & Jowett, 1957) or medians (Brown & Mood, 1951; Tukey, 1970) for each half or third of the data. These data points can be used to calculate an equation for a line which best fits the data. The data points also can simply be connected by ruler to form the best-fit line.

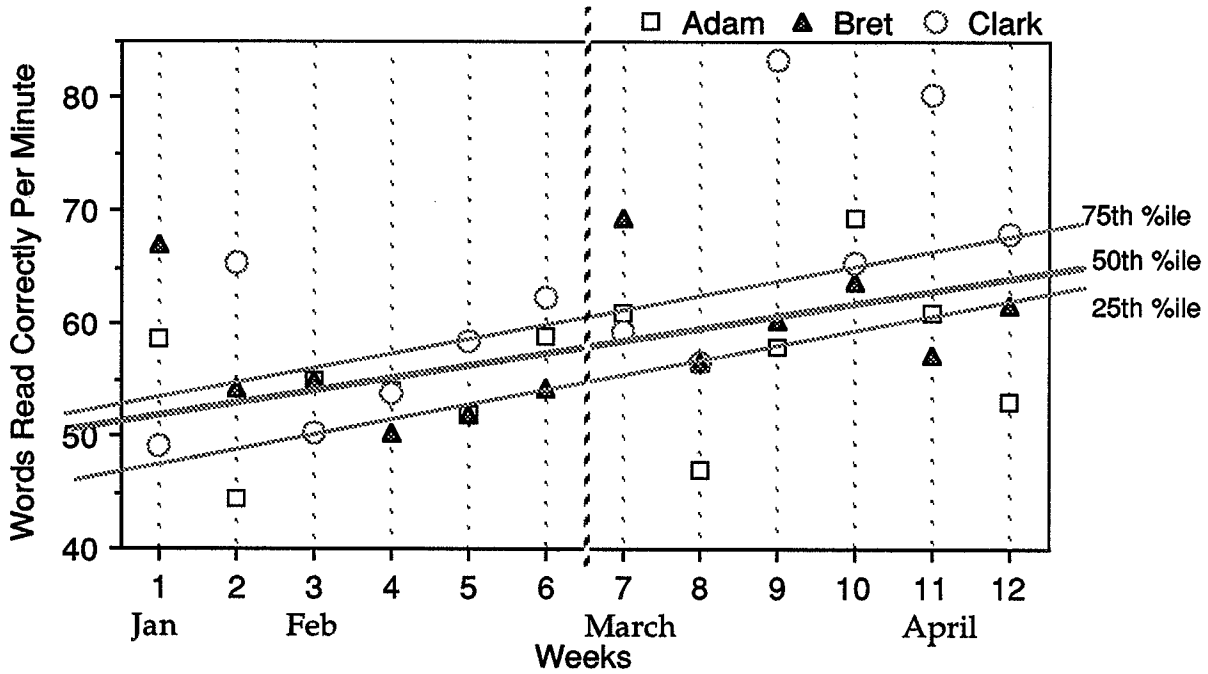


Figure 1. Three Percentile Lines Fit to Reading Progress Data Collected Over 12 Weeks from Three Students

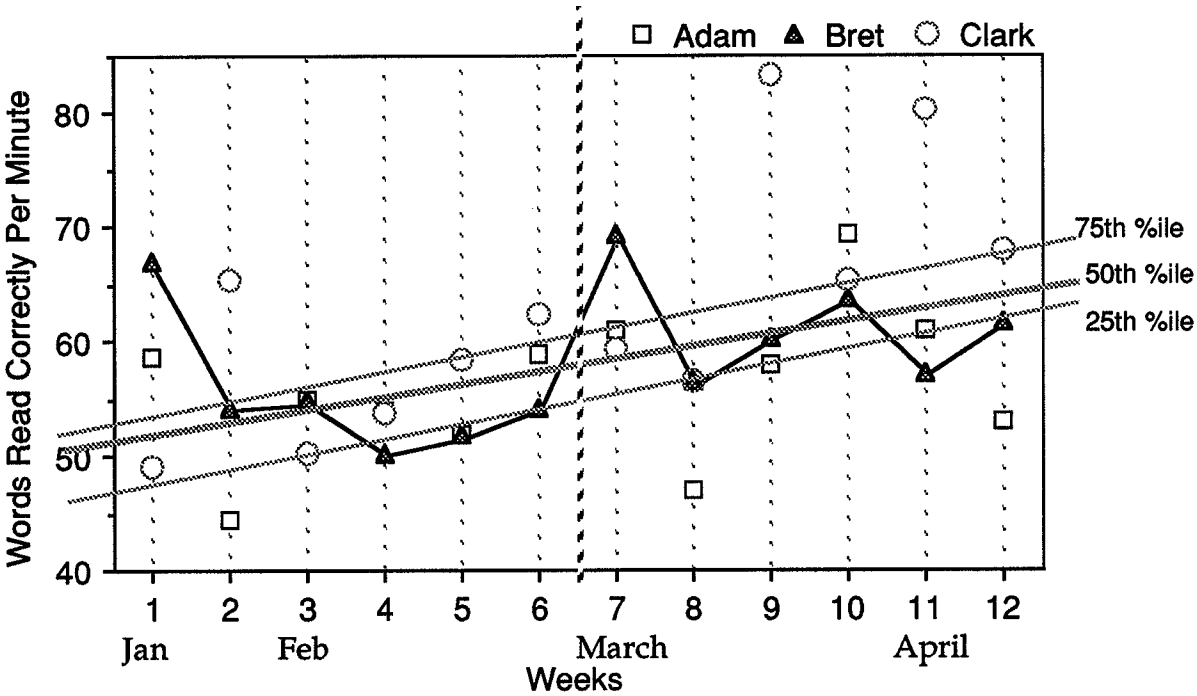


Figure 2. Comparison of Bret's Reading Scores with Three Group-based Percentile Standards

In educational literature, the fitting of median (50th percentile) lines has been popularized by White's *split-middle* application (White & Liberty, 1976) of Koenig's (1972) "quarter-intersect" method. The 50th percentile line is but one of a family of percentile lines. This paper demonstrates how other percentile lines as well can be fit to the data to provide within-group NR comparisons for the time series data.

Method

The median-based, two-section method for fitting 50th percentile lines (Brown & Mood, 1951; Mood, 1950) has been extended by Hogg (1975) to fit other percentile lines as well. Hogg's procedure is demonstrated here on data from a small instructional group of only three students, for clarity. Figure 1 is a plot of weekly oral reading fluency data (words read correctly

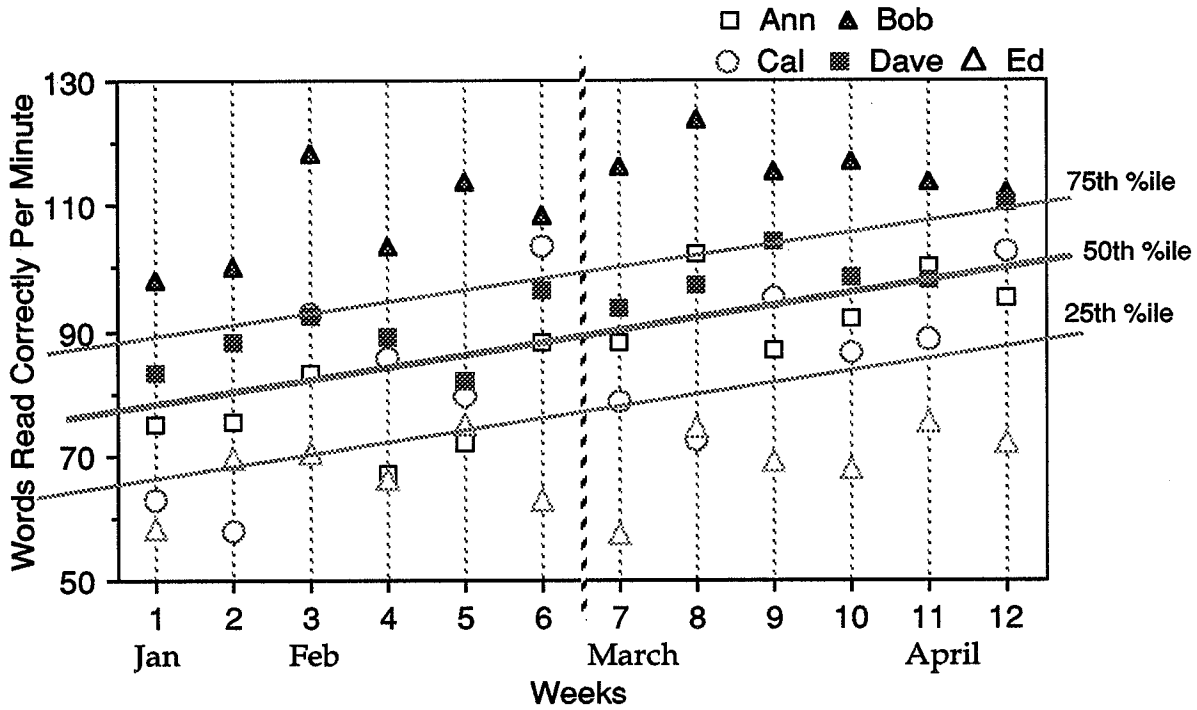


Figure 3. Three Percentile Lines Fit to Reading Progress Data of Five Students

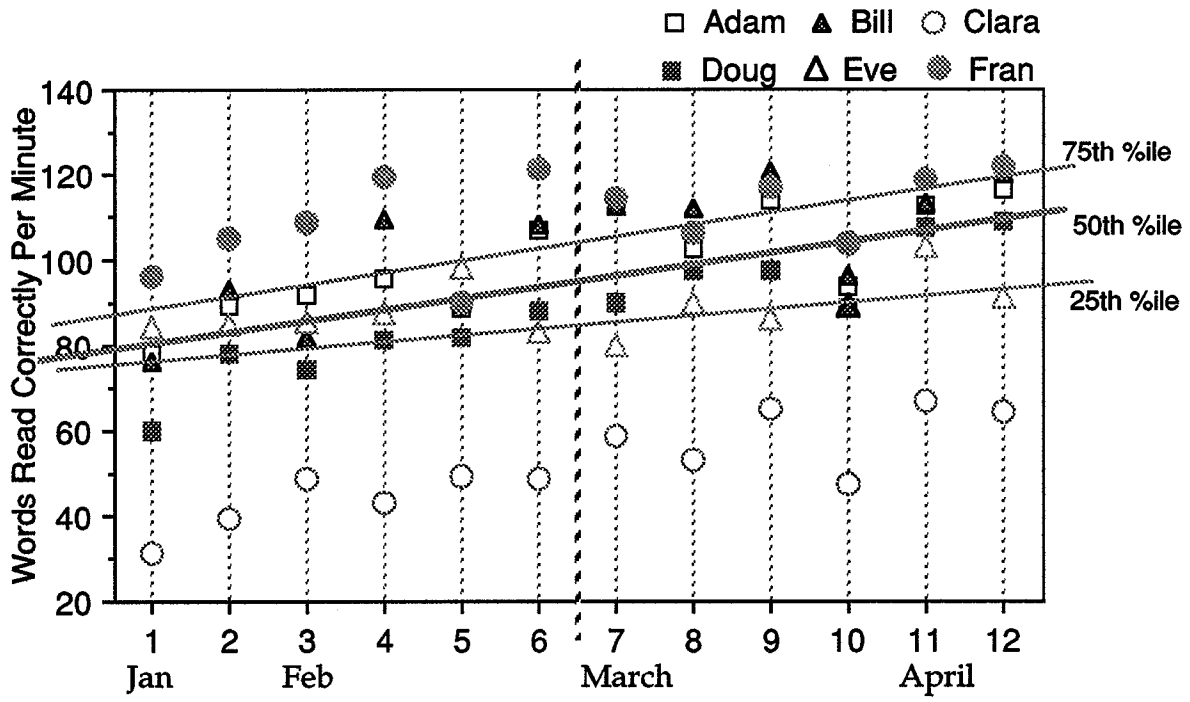


Figure 4. Three Percentile Lines Fit to Reading Progress Data of Six Students

per minute) collected over twelve weeks, from three Grade 4 students with reading disabilities. Each data point is the average of three one-minute timed readings from long-term goal material. Ten passages of the same readability were selected from a basal text. Each week three of these passages were randomly selected for testing. Following are the directions for applying Hogg's (1975) dynamic percentile norms to these reading data.

Directions

First the plot is divided vertically into two sections, with an equal number of data points in each section [see (a) in Figure 1]. This plot contains a total of 36 data points (12 for each student), so each section contains 18 points. Hogg recommends laying a transparent ruler horizontally through the data so that 50% of the data points *in both left and right sections* are above the ruler edge. On the plot in Figure 1, 9 points should be above the line in both right and left sections. The resulting line (see Figure 1) marks the 50th percentile. Begin with an approximate ruler position, and adjust by trial and error.

The 75th percentile line is formed in the same manner. This line will separate the top 25% of scores in both right and left sections (41/2 data points in each section) (see Figure 1). The 25th percentile line separates the bottom 25% of scores in both right and left sections (the bottom 4.5 data points in each section) (see Figure 1). The interval bounded by the 25th and 75th percentiles, termed the "interquartile range," contains the middle half of all scores.

Interpretation

The 25th, 50th, and 75th percentile lines serve as "floating" normative standards for judging a student's progress over any period of time. These floating standards are based on averages from twelve different test periods, so they are more stable over time than percentile ranks calculated from just one or a few assessment periods. We can use these percentile lines to help interpret Bret's test scores (joined by lines in Figure 2).

Over the four months, Bret performed below the group median (7 scores) about as often as above it (5 scores). The variability of his performance was about the same as for the group, with half of his scores (6) outside the interquartile range. Bret's ratio of high (above the group's 75th percentile) to low (below the 25th percentile) test scores is 2:4, worse than the group average of 1:1.

The percentile lines also help us summarize progress of the group as a whole. We can say that the more stable middle range of scores (the interquartile range) spanned about 6 points throughout the four months. This middle range improved from 47-53 in early January, to 55-60 in early March, to 60-66 in early April.

The smaller the group of students, the less meaningful are comparisons of a student's scores with percentile norms, as the individual composes a large part (1/3 in Figure 2) of the normative group. Hogg's (1975) techniques were applied to two slightly larger remedial reading groups at Grade 5 ($N = 5$) and Grade 4 ($N = 6$) (see Figures 3 & 4). The three groups had different

teachers and were all from different elementary schools. Students in the two latter groups also were assessed in oral reading fluency on randomly selected LRG material on a weekly basis. As for the first group, each data point in the two larger groups represents the average of three one-minute timings.

The parallel percentile lines in Figure 3 indicate a constant raw score spread from low (25th percentile) to high (75th percentile) benchmarks over time, whereas the diverging lines in Figure 4 indicate that the raw score spread is increasing over time. The distances between percentile lines cannot be compared between Figures 3 and 4, because the two graphs differ in vertical scales.

Reliability

The usefulness of the floating percentile method for teachers will probably depend on three factors: (a) whether it can be easily and quickly produced and interpreted (efficiency), (b) whether the information it yields is relevant to important decision needs (validity), and (c) whether the information it yields is accurate (reliability). We focus here on the third criterion. Two questions about accuracy can be asked. First, do the percentile lines accurately describe the plotted data? Second, how confident can we be in a percentile line's location, given the day-to-day variability of responding, and the limited number of students and assessments plotted? The first question is related to descriptive accuracy, and the second to measurement confidence or inferential adequacy.

To answer the first question, 25th, 50th, and 75th percentiles were computed directly from raw scores and averaged for left and right sections of each plot. These computed scores were compared to the values read from the middle of the left and right segments of the plotted percentile lines. KaleidaGraph® (Synergy

Table 1. Comparison of 25th, 50th, and 75th Percentiles Computed From Raw Scores with Those Read From Left and Right Percentile Line Segments

	Left Section			Right Section		
	25th Percentile	50th Percentile	75th Percentile	25th Percentile	50th Percentile	75th Percentile
Group of 3						
Computed	51.3	55.4	59	57.4	60.9	68.5
	(1.1)	(1.1)	(1.8)	(2.0)	(1.1)	(3.0)
Line-based	51.2	54.7	56.1	58	61	64.6
Difference	0.1	0.7	2.9	0.6	0.1	3.9
Group of 5						
Computed	71.7	83.9	94.	81.	95.4	105.2
	(3.2)	(3.6)	(2.7)	(2.5)	(2.0)	(1.7)
Line-based	70.1	83.	93.	82.5	95.	104.
Difference	1.6	0.9	1.	1.5	0.4	1.2
Group of 6						
Computed	76.3	87.6	96.1	89.8	103.3	111.1
	(3.5)	(2.8	(4.1)	(3.1)	(3.0)	(3.5)
Line-based	80.	86.5	95.	89.	103.	111.
Difference	3.7	1.1	1.1	0.8	0.3	0.1

Note: Bracketed scores are standard errors of the mean.

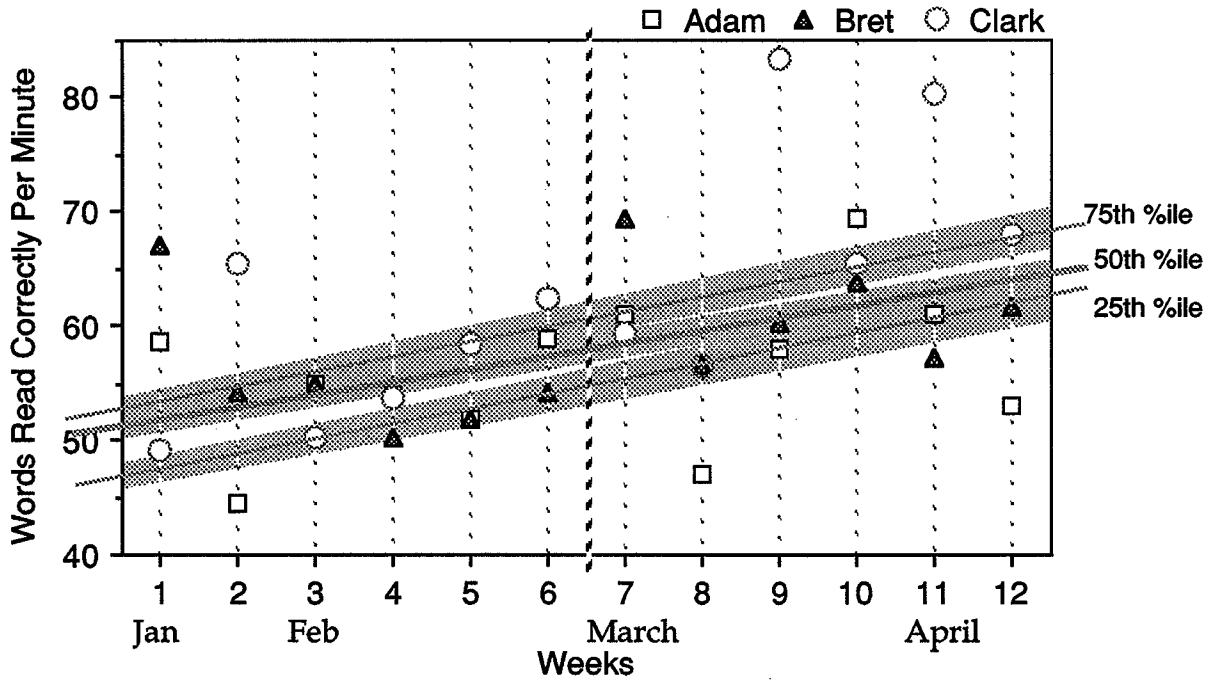


Figure 5. One-SEM-Confidence-Interval Envelopes Around Percentile Lines for Group of Three Students

Software, 1990) software was used to electronically read line segments values. Computed values and percentile line-based values are compared in Table 1. Most differences between the computed and the percentile line-based values were very small. Differences ranged from 0.1 to 3.9 points, with an average of 1.2 points. The accuracy of the percentile lines was not dependent on the number of students in a group.

The second accuracy question was, what confidence co we have in the location of a percentile line, given within-group variability and the limited number of students and data points? To answer this question, confidence interval *envelopes* were constructed for each percentile line. The envelopes represent ±1 standard error, enclosing the 68% confidence interval for that percentile. We can be 68% certain that the true percentile

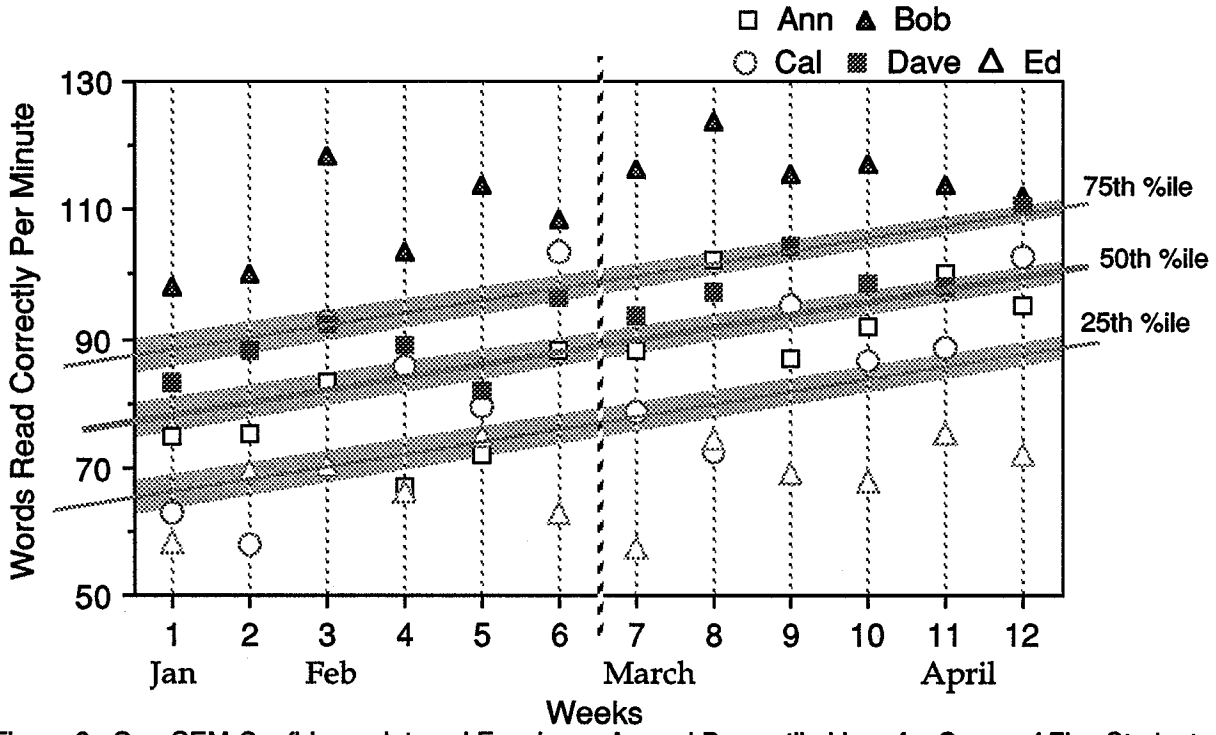


Figure 6. One-SEM-Confidence-Interval Envelopes Around Percentile Lines for Group of Five Students

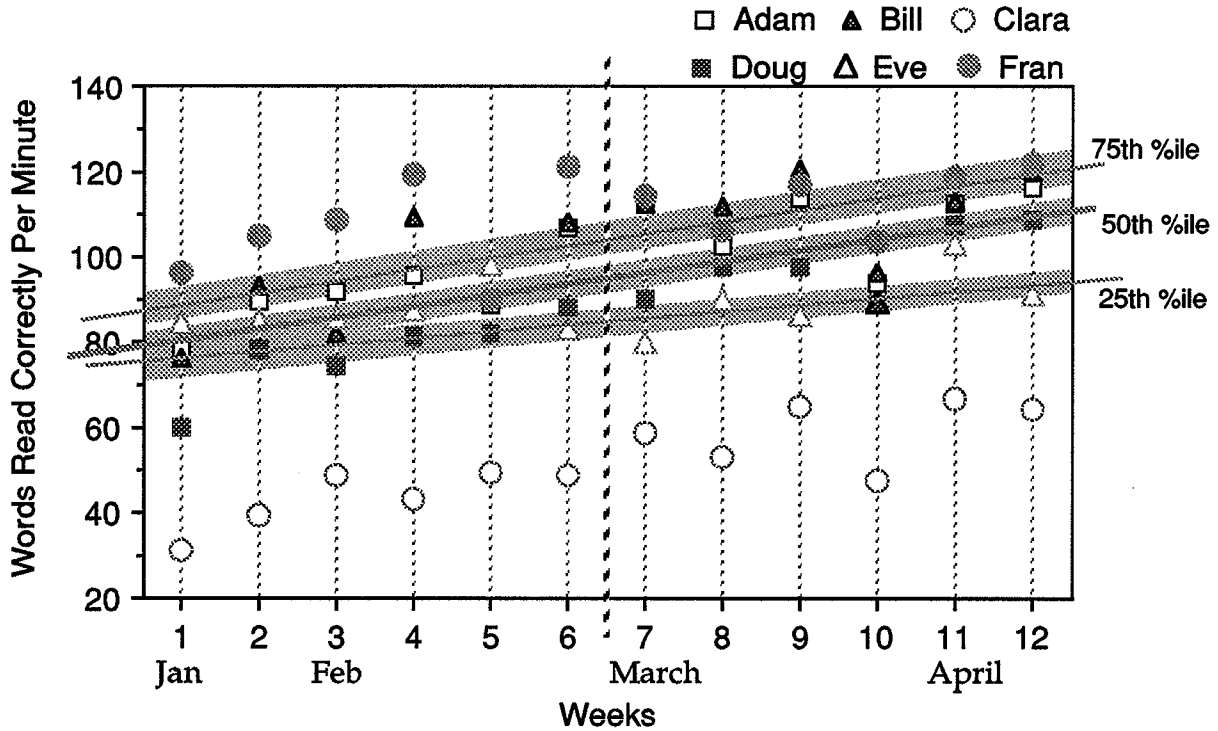


Figure 7. One-SEM-Confidence-Interval Envelopes Around Percentile Lines for Group of Six Students

line is somewhere within that envelope. Two standard errors—for right and left segments—were calculated for each line. For example, for a 75th percentile line the standard errors were computed as the standard error of the mean (SEM) for the six 75th percentile scores in each data section. These SEM are presented in brackets in Table 1. The confidence interval envelopes are shown as shaded areas around the percentile lines in Figures 5, 6, and 7. In Figure 5, for the smallest group (3 students) there is considerable overlap of the confidence interval envelopes for the 75th and 50th percentile lines, and for the 50th and 25th percentile lines. This means that we cannot distinguish these percentiles with reasonable certainty. Only the 25th-75th percentile distinction can be made reliably.

For the group of 5 students (Figure 6) there is no overlap of confidence intervals; the three percentile lines are clearly distinguished from each other. The graph for the group of 6 students (Figure 7) shows reliable separation of the percentile lines, except for the 50th and 25th percentiles in the first data segment (January and February). From examining these three figures, it is clear that there is better separation of percentile lines when there is (a) more data (more students and test periods), and (b) a greater spread of test scores within the group. We have been investigating differences between percentile lines 25 points apart—smaller differences of only five or ten points clearly are not warranted with datasets such as these.

The problem of measurement error is not peculiar to percentile line fitting procedure. Though not always acknowledged, the problem is a part of all educational assessment. In fact, unreliable measurement is so prevalent as to call into question many of the decisions we make for students based on assessment data (Salvia & Ysseldyke, 1978). The problem is greatest when we wish to infer from a student's test scores to his or her (a) relative position in a broader population of students, (b) achievement on other similar tests or items, or (c) predicted achievement at some time in the future. When the emphasis is on describing existing data, and not inferring beyond it, this type of measurement error is not as important.

Conclusions

Hogg's (1975) procedure for fitting floating percentile norms has been applied to progress monitoring data from three small remedial reading groups. The lines provide within-group, norm-referenced (NR) data comparisons, for which several uses have been described. Following Hogg's procedures three percentile lines per graph were drawn in only a few minutes, through counting plotted points, trial and error ruler adjustment, and recounting. Where 10th and 90th percentiles are desired, they can be added with little additional time. However, they appear not to be warranted for datasets as small as these.

Is the procedure appropriate for groups of any size? A group of only three students may be too small for reliable distinctions between percentiles, and for distinguishing the individual from the group. The procedure appears better suited to groups of four and five or more. These graphs contained 12 assessments

for each student; the effect of fewer assessments per student is not known. Application of the procedure to a classroom group may be inefficient because of the many ruler adjustments required. In that case, White and Liberty's (1976) procedure can help by initially placing the 50th percentile line. Other procedural improvements and limitations of Hogg's technique will become apparent from applying it to a variety of progress monitoring data.

References

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics, 5*, 207-212.

Brown, G. W. & Mood, A. M. (1951). On median tests for linear hypotheses. In J. Neyman (Ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probbability*. Berkeley and Los Angeles: University of California press.

Coffman, W. E. (1983). *Testing in the schools: A historical perspective*. Paper presented at the annual invitational conference of the Center for the Study of Evaluation, University of California, Los Angeles.

Deno, S. L., & Mirkin, P. K. (1977). *Data based program modification: A manual*. Reston, VA: The Council for Exceptional Children.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199-208.

Gibson, W. M., & Jowett, G. H. (1957). 'Three-group' regression analysis, part I. Simple regression analysis. *Applied Statistics, 6*, 114-122.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). Resistant lines for y versus x. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 129-163), New York: John Wiley & Sons.

Hogg, R. V. (1975). Estimates of percentile regression lines using salary data. *Journal of the American Statistical Association, 70*, 56-59.

Koenig, C. H. (1973). Charting the future course of behavior. *Dissertation Abstracts International, 33*, 4200A. (University Microfilsms No. 72-1850).

Mood, A. M. (1950). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Salvia, J. & Hughes, C. (1990). *Curriculum-based assessment: Testing what is taught*. New York: Macmillan.

Salvia, J. & Ysseldyke, J. E. (1978). *Assessment in special and remedial education*. Boston: Houghton Mifflin.

Tindal, G. R., & Marston, D. (1990). *Classroom-based assessment: Evaluating instructional outcomes*. Columbus, OH: Charles Merrill.

Tukey, J. W. (1970). Straightening out plots (using three points), in *Exploratory data analysis* (pp. 169-181), Menlo Park, CA: Addison-Wesley.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of mathematical Statistics, 11*, 284-300.

Wesson, C., King, R., & Deno, S. L. (1984). Direct and frequent measurement: If it's so good for us, why don't we use it? *Learning Disability Quarterly, 7*(1), 45-48.

White, O. R., & Liberty, K. (1976). Behavioral assessment and precise educational measurement. In R. L. Schiefelbusch and N. G. Haring (Eds.), *Teaching special children* (pp. 31-71). New York: McGraw-Hill.

Yeh, J., Herman, J., & Rudner, L. (1981). *Teachers and testing: A survey of test use*. Center for the Study of Evaluation Report No. 166, Los Angeles: University of California.

Curriculum-Based Measurement: An Additional Measure for Screening Academically Able Students for Gifted and Talented Programs

Tracey Hall
Judith Gelbrich
University of Oregon

In 1987, the state of Oregon passed the mandate for identification ofand specialized education for talented and gifted (TAG) school-aged children. According to the Oregon Talented and Gifted Education Act, gifted children must be identified, and their abilities assessed, by the beginning of the 1991-92 school year. Children who "score at or above the 97th percentile on a nationally standardized test of mental ability" are considered intellectually gifted [OAR 581-22-403 (1) (a)]. Those who "score at or above the 97th percentile on one or more tests of academic achievement on a nationally standardized test battery" are considered academically talented [OAR 581-22-403 (1)(d)].

Under the recommended procedures for identification as outlined in the Oregon Technical Assistance Paper #1 (Cohen, Sheperd, & Balzer,1990), school districts are urged to create a screening pool by nominating or referring students. These students will then be evaluated and considered for inclusion in the gifted program. This screening pool is developed through a process in which school personnel review students' skills through extant data sources including formal test scores, grades, and academic abilities demonstrated in the classrooms. By law, the initial screening process must include more than one measure of "intelligence, achievement, or aptitude" [OAR 581-21-030 (2)].

Realistically, not all children should be assessed formally for talented and gifted programs. Measurement procedures are quite involved in terms of the volume, time, and resources required. Therefore, a referral or nomination procedure is-through a pull-out program, children who are identified as gifted potentially miss instruction in the regular classroom when receiving TAG services. Classroom teachers are responsible for monitoring a TAG student's academic performance in and outside of their classrooms.

School districts in the state of Oregon are in the process of developing procedures to meet requirements for the identification, assessment of learning rate, and current performance level of TAG students. Our focus is to investigate the referral and screening process for identification of Talented and Gifted children. More specifically, we examine the utility of curriculum-based measurement (CBM) as a screening tool for TAG nomination.

Referral Process for TAG Identification

As with most public schools, gifted program screening and eligibility procedures have been formalized for the school district in this study. Nomination or referral originates at the classroom-teacher level. Although it occurs less frequently, parents may also initiate a referral. Prior to referral by the teacher, an informal evaluation of a student's academic ability must take place. This evaluation usually includes a review of the student's classroom academic performance and standardized test data.

In order to determine whether a student has the requisite academic ability to benefit from the gifted program, extant data are used to find out if a student's standardized achievement test scores are at or above the 90th percentile in reading and math. Students who meet the achievement data criteria become part of the screening pool. Teachers and parents continue the process by completing abehavioral checklist. Finally, the student is administered an intelligence test (see Figure 1). These three data sources are then used in coordination by the schooldistrict's team of TAG teachers to make a placement decision in the gifted program.

Standardized achievement tests are critical in the decision-making process for identifying academically able students. They have two functions. First, standardized achievement tests influence the classroom teacher's decision to nominate students into the screening pool for the referral process. Secondly, they are the instruments for establishing criteria that screen students for program consideration. Using a criterion measure for screening students in the process of identifying academically able students is a common practice in gifted education (Feldhusen, Asher, & Hoover, 1984). A criterion is necessary due to limited resources,

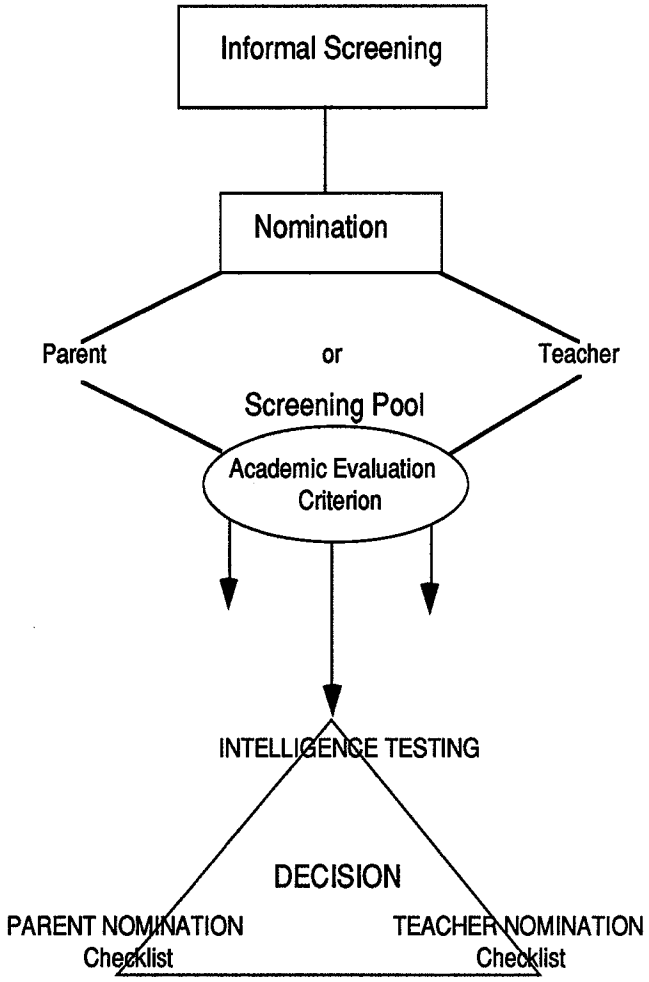


Figure 1. Referral Process for Placement of Children in Gifted Programs

and the fact that not all students will benefit from a specialized program (Borland, 1989). It has been suggested that the most common criterion of the 90th percentile (as used in this district) may be too stringent, and recommendations have been made to consider referral for students who score at or above the 80th percentile (Eby & Smutny, 1990).

Problems with Current Assessment for Identification

Standardized achievement tests are valid and reliable assessment tools when used to assess general academic ability. However, the usefulness of standardized tests has been challenged in recent years (Salmon-Cox, 1981). Test items provide little assessment Standardized achievement tests were designed to measure school achievement in basic skills. They were not created or validated for identification of children for gifted programs; they do not distinguish individual differences well at the extreme high or low ends of a distribution (Eby and Smutny 1990). Standardized tests were not designed to be efficient discriminators (Ebmeier, & Schmulbach, 1989).

Ideally, a TAG identification procedure should include information about how target students compare to their classmates, rather than to national norms (McFarland, 1980). The typical referral process described here has no consideration for student performance in relation to their peers or in the instructional material.

Finally, standardized achievement tests are typically given once during a school year (Harrington, 1982). As such, students have only a single opportunity to demonstrate their ability and performance on a high-stakes assessment tool that may determine their consideration for the gifted program.

Requirement for Additional Information

Concerns regarding the utility of standardized tests inability to: (a) represent the curriculum, (b) compare performance to peers, (c) effectively discriminate high performers, and (d) be frequently administered, should be addressed at the screening level. An assessment and measurement system developed over the last decade, known as curriculum-based measurement (CBM) (Deno, 1985), may fill this role.

CBM is a method of analyzing academic performance using systematic procedures with brief measures in specific academic areas. It is accurate in distributing students based on obtained scores. This measurement system has been found helpful to assist teachers with student grouping for classroom instruction (Hall & Tindal, 1989; Wesson, Vierthaler, & Haubrich, 1989). Curriculum-Based Measurement also has been investigated as a screening measure in the identification of giftedness with kindergarten and first grade children (Joyce & Wolking, 1988). CBM measures can be administered at any time of the school year, in a very brief and timely fashion.

Teachers need measurement systems that represent the behaviors they teach. As reflected in the name, curriculum-based measurement emphasizes curriculum items using a production response format, in contrast to selection response typical of most norm referenced tests, including standardized achievement tests. In the basic skill areas, this includes production responses like reading passages, spelling words, writing words in sentences to communicate new ideas, and calculating math answers to computation and story problems. Although CBM has traditionally been investigated in the context of special education decision-making, these measures have more recently been applied in a general education context (Hall & Tindal, 1989; Hall, Gelbrich, & Tindal, 1990).

CBM procedures for data collection are standardized, meaning that conditions under which the measures are collected are always the same. This assures a data base for individual evaluation that is reliable and valid. CBM measures have been designed to spread

student performance scores across a distribution for the purpose of peer comparison. The issue of technical adequacy is sufficiently addressed when comparisons are made between an individual's performance and the performance of that sutdent's peers (Deno, 1989). Therefore, two benefits have been identified for adding CBM to the TAG screening process: (a) comparison of students to their own peer groups versus the national norms, and (b) a comparison in the curricula of instruction, in which students must maintain growth over time.

A third consideration in the addition of CBM to the TAG screening process is the timeliness of the measures in today's classrooms. The procedures reguired to collect data using CBM are very brief and unobtrusive. Each measure takes from 1 to four minutes to complete. Since these tools are so easy to collect under standardized conditions, teachers can quickly administer CBM tests to a class or individual at virtually any time.

Our study investigated the effectiveness of adding CBM procedures to the nomination and screening process to assist educators in identifying students for TAG specialized services.

Methods

Subjects

The investigation took place in a somewhat small school district in the Pacific Northwest serving approximately 5000 students in five elementary level buildings with kindergarten through fifth grade classrooms. Four of the district's five elementary buildings were involved, the fifth building was not using CBM measures; therefore, no student participation was solicited. Parents of students identified for TAG and students referred for TAG in the 1989-90 school year were contacted by mail to request permission to access test data. As a result, 45 elementary students voluntarily participated in the study, representing 66% of the potential students in the two areas. The two categories were: (a) students who were referred (but not found eligible) to the district TAG program (n= 21), and (b) students currently identified as TAG and receiving specialized services (n= 24). Students represented Grades 2 through 5, and ranged from 7 to 11 years in age.

Measures

Two measures were used in the investigation: the Stanford Achievement Test (SAT) (Gardner, Rudman, Karlsen, & Merwin,1984) total reading and total math subtest scores; and CBM reading and math computation scores. This district gave the entire SAT to all grade levels in May of 1990.

The CBM measures used in this study were administered in the school district as a part of an on-going special services project to implement the assessment procedure. Specific descriptions of each measure, as well as administration and scoring procedures, are described in detail by Deno and Fuchs, (1987), and Shinn, (1989). CBM data were collected at the elementary level in fall, winter, and spring of the 1989-90 school year, including students of all grades and ability levels. We analyzed only the spring collection of CBM scores to compare results with SAT scores in two subject areas.

Procedures

Extant data were gathered on participating students at each building. This included the national percentile scores and standard scores for the total reading and total math items on the SAT. A conversion table was use to transpose percentile scores to z-scores. Raw scores from the CBM reading and math results were also converted to z-scores, thereby making a comparison possible by having a standard scale for the scores. A Pearson correlation coefficient was applied for reading and math, respectively.

Results

All subjects for whom SAT and CBM data were available in the spring of 1990 were included in the analysis (n= 45). Percentile scores were blocked in increments to describe this population on both SAT and CBM measures (see Table 1). The majority of students scored above the 50th percentile on math and reading tests of the SAT and CBM measures. On the curriculum-based measure, 65% of the students were above the 80th percentile based on the local norms created at the spring testing period. On the nationally normed Stanford Achievement Test, 76% of students scored above the 80th percentile. The population selected from TAG-referred and TAG identified students represents high performing students.

Table 1. Description of Population by Percentile Scores on Each Measure

Percentile	SAT Rdg	SAT Math	CBM Rdg	CBM Math
90	n = 26	n = 30	n = 18	n = 14
80	n = 9	n = 4	n = 13	n = 3
50-79	n = 9	n =12	n = 10	n = 19
< 50	n = 0	n = 0	n = 1	n = 5

Table 2 summarizes the correlation between student test scores on the SAT and CBM tests in reading and math. These results indicate a positive correlation in both subject areas, reading r= .42, and math r= .45.

Discussion

Unlike the Joyce & Wolking (1988) study, where the Metropolitan Achievement Test (MAT) was found

Table 2. Correlation of SAT and CBM for Reading and Math

	Reading Tests	Math Tests
$n = 43$	$r = .42$	
$n = 43$		$r = .45$

to be essentially equivalent to curriculum-based assessment, for kindergarten and first graders, we did not find that SAT and CBM were equivalent in predicting academic performance for second through fifth graders. We did, however, find that there is at least a moderate correlation between SAT and CBM test in reading and math at this age level. One of the factors that may be moderating this relationship is the inability of the SAT to adequately discriminate abilities at the high and low ends of the population. In terms of diversity of the percentile scores, our results indicate that the CBM has more of a tendency to spread students in this population across the distribution, thus providing a finer discrimination of student performance when compared to their peer groups. The moderate correlation could also be attributed to the small sample size of this study. We recommend an additional study of this sort with a larger sample size.

CBM offers useful assessment information to assist in the process of screening. The mandate of the Oregon Talented and Gifted Education Act requires the use of more than one measure of "intelligence, aptitude or achievement" [OAR 581-21-030 (2)] to create a screening pool for the identification of gifted children. Unlike standardized achievement tests, CBM can be administered at any time, quickly and efficiently. Scoring procedures are relatively easy, so teachers have almost instant access to the results. Thus, the data are immediately available to assist in making decisions in screening for gifted programs.

Talented and gifted identification procedures must include information beyond general academic ability, and standardized achievement tests by themselves are not capable of this task. Curriculum-based measures represent the curriculum of instruction allowing teachers to evaluate student performance based on what students have actually been taught. Additionally, CBM is closely associated to teacher judgement of student performance (Hall, Gelbrich, & Tindal 1990). Thus, the addition of CBM to the nomination and screening process can confirm and provide specific academic information for teachers regarding their students.

Identification procedures should include peer comparisons to reflect performance in a school's program (McFarland, 1980). CBM offers teachers a means to compare gifted students' performance to that of their peers by assessing students in the curriculum of instruction and scaling their performance on a local, rather than national, distribution. This factor is particularly critical when a school district has a large minority population or low socioeconomic status (Eby & Smutny, 1990).

CBM can also assist teachers in determining whether students have the necessary mastery of an instructional curriculum to be able to compensate if they miss instruction. This knowledge is important in making referrals for gifted students, since pull-out programs are one of the more common methods used to deliver instruction in gifted programs. CBM provides teachers with a timely and efficient method to evaluate their students and insure that they maintain academic proficiency so as not to suffer if instruction is missed.

In the future, the school district in which we conducted this investigation will not be administering the Stanford Achievement Test to all grade levels each year. This situation could be true of other schools and districts as assessment procedures are scrutinized. Therefore, schools may be evaluating students for TAG screening and nomination with standardized test data that is several years old, thus strengthening the argument for the addition of CBM as a tool in the screening process. As stated previously, CBM can be administered at any time convenient for the teacher, making the procedure highly functional for up-to-date information on academic performance.

In the final analysis, CBM appears to be a useful tool in assisting with the referral or nomination process in identifying students for gifted programs. Multiple measures are necessary to ensure that students are given proper consideration, and with the addition of CBM assessment procedures, teachers will have more complete and up-to-date information on which to base this critical decision.

References

Borland, J. H. (1989). *Planning and implementing programs for the gifted*. New York: Teachers College Press.

Coleman, M., & Harmer, W. R., (1982). A comparison of standardized reading tests and informal placement procedures. *Journal of Learning Disabilities, 9*, 396-398.

Deno, S. L., (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.

Deno, S. L., Fuchs, L. S. (1987). Developing curriculum based measurement systems fro data-based special education problem solving. *Focus on Exceptional Children, 19*, 1-16.

Deno,S. L., (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1-17). New York: Guilford.

Ebmeier, H., & Schmulbach, S. (1989). An examination of the selection practices used in the talent search program. *Gifted Child Quarterly, 33*, 134-141.

Eby, J. W., & Smutny, J. F. (1990). *A thoughtful overview of gifted education.* White Plains, New York: Longman.

Feldhusen, J. F., Asher, J.W., & Hoover, S.M. (1984). Problems in the identification of giftedness, talent or ability. *Gifted Child Quarterly, 28*,149-151.

Gardner, E. F., Rudman, H.C., Karlsen, B., & Merwin, J.C. (1984). *Stanford Achievement Test.* The Psychological Corporation; Harcourt Brace Jovanovich, Publishers: U.S.A.

Hall, T., Gelbrich, J., & Tindal, G. (1990). Teacher use of various data sources for grouping, placement and identification decisions. In , J. Marr, N. George, M. George, & G. Tindal (Eds.), *The Oregon Conference '90 Monograph.* Eugene, Oregon: College of Education, University of Oregon.

Hall, T., & Tindal, G. (1989). Using curriculum-based measures to group students in reading. In G. Tindal, K. Essick, N. George, & M. George (Eds.), *The Oregon Conference '89 Monograph.* Eugene, Oregon: College of Education, University of Oregon.

Harrington, R.G. (1982). Caution: Testing may be hazardous to the educational programs of intellectually gifted children. *Education, 103*, 112-117.

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

Joyce, B. G., & Wolking, W.D. (1988). Curriculum-based assessment: An alternative approach for screening young gifted children in rural areas. *Rural Special Education Quarterly, 8*, 9-14.

McFarland, S. L. (1980). The young gifted child: Guidelines for the identification of young gifted and talented children. *Roeper Review, 3*, 5-7.

Oregon Administrative Rules, §581-20-403, § OAR 581-21-030

Cohen, L. M., Sheperd, G., & Balzer, C. (1990). *The identification of talented and gifted students K-12*, Oregon Technical Assistance Paper #1.

Parker, R., Kameenui, E. & Jitendra, A. (1990). Analysis of a grade 3-5 elementary math curriculum for production of valid math tests: Work in progress. In J. Marr, N. George, M. George, & G. Tindal (Eds.), *The Oregon Conference '90 Monograph.* Eugene, Oregon: College of Education, University of Oregon.

Salmon-Cox, L. (1981). Teachers and Standardized Achievement Tests: What's really happening? *Phi Delta Kappan , 62*, 631-634.

Shinn, M. (Ed.), (1989). *Curriculum-based measurement: Assessing special children.* White Plains, New York: Guilford.

Stanley, J. C. (1984). Use of general and specific aptitude measures in identification: Some principles and certain cautions. *Gifted Child Quarterly, 28*, 177-180.

Van Tassel-Baska, J. (1987). The ineffectiveness of the pull-out program model in gifted education: A minority perspective. *Journal for the Education of the Gifted, 10*, 255-264.

Wesson, C.L., Viertaler, J. M., & Haubrich, P. A. (1989). An efficient technique for establishing reading groups. *The Reading Teacher, 42*, 466-469.

Construction and Use of Behavior Rating Scales

Changnam Lee
Gerald Tindal
University of Oregon

Some published rating scales or checklists are developed using systematic procedures, while others are constructed in an arbitrary fashion (Rie & Friedman, 1978), which may result in inadequate reliability or validity. In addition, most scales are designed to assess people of a specific age range, or specific characteristics, and are inappropriate for other groups of people. Therefore, the purpose of the current review study is to provide (a) systematic procedures to construct both reliable and valid rating instruments, and (b) guidelines for selecting one or more instruments among published ones.

Scale Construction

A *behavior rating scale* is a collection of descriptive items, the responses to which are scored with numerical values and are combined to yield a summative score (adapted from Dawis, 1987). Dawis (1987) divides the scale construction procedure into three stages: *scale design, scale development,* and *scale evaluation.*

Scale Design

Content

Content is related to the source of items and the method of collecting items. Items are most frequently collected from two sources: (a) existing literature, and (b) information provided by prospective users of the instrument. In the *critical incident technique* proposed by Flanagan (1954), only qualified observers who are well-acquainted with a specified job performance provide critical facts pertaining to behavior in defined situations, through interviews using carefully provided questions.

Format

While scale content is pivotal in establishing content validity, an appropriate scale format can minimize errors in rating and enhance reliability of rating. Generally, more scale points are better than fewer, in terms of psychometric properties and variability. An even number of scale points can also be effective to avoid excessive use of the middlemost scale point (Dawis, 1987).

Guilford (1954) classified commonly-used rating scale formats into five broad categories: (a) numerical, (b) graphic, (c) standard, (d) cumulated points, and (e) forced choice.

Numerical scales. The typical numerical scale illustrated by Item 1 in Figure 1 provides the observer or rater with a sequence of defined numbers paired with descriptive cues. The rater selects an appropriate number that is associated with an appropriate definition or description for each statement. Some numerical scales do not provide overt numbers for the rater (e.g., Item 2 in Figure 1). In this case, the rater responds to the items by choosing the most appropriate descriptive cues, and the researcher assigns numbers to them. Numerical scales are sometimes called "Likert-type" formats (Dixon, Bobo & Stevick, 1984; Latham & Wexley, 1977).

How did you feel about the math classes?

1. Very unpleasant
2. Moderately unpleasant
3. Indifferent
4. Moderately pleasant
5. Very pleasant

Does the baby reach for familiar persons?

Never Seldom Sometimes Generally Always

Figure 1. Two Types of Numerical Scale: With or Without Overt Numbers for Response Alternatives. (The numbered discriptors are adapted with permission from Guilford, 1954 [p. 263].)

Graphic scales. The graphic scale format has numerous variations with a common feature: A straight line is displayed with various cues combined with it to aid the rater (Guilford, 1954). The line can be segmented into a varied number of units, or it can be continuous (Figure 2), and it can be placed horizontally or vertically.

Standard scales. Like the scales for assessment of handwriting skills, standard scales provide some standard specimens or examples of performance ability that have previously been calibrated on a commonly used scale of excellence. The rater can equate a new sample of performance to one of the standards, or judge it as being between two standards.

In social conversation, how have you been?

talkative	an easy talker	talked when necessary	preferred listening	refrained from talking

Is he slow or quick thinking?

extremely slow	sluggish, plodding	thinks with ordinary speed	agile-minded	exceedingly rapid

Figure 2. Examples of Graphic Scales (Reprinted with permission from Guilford, 1954 [p.265]).

Cumulated points. In this type of format, the rater assesses an individual by the sum of scores derived from a number of ratings, whether they are weighted or un-weighted. This format is commonly-used in check-list scaling methods.

Forced choice format. In a forced choice scale format, two pairs of statements or descriptive adjectives (one pair with high preference value and the other with low preference value) are combined to form an item (e.g., Figure 3). Essentially, the rater must say whether an individual has more of one trait than another for each pair.

The child is

careful
clumsy
tidy
slovenly

Figure 3. Example of a Forced Choice Scale

Scale Development

This stage includes selection, arrangement, and calibration of items. Six types of method are described in this section: Thurstone method, Likert method, Guttman method, Behaviorally Anchored Rating Scales (BARS), Behavioral Observation Scales (BOS) and the Mixed Standard Scales (MSS).

Thurstone Method

In Thurstone and Chave's (1929) procedure, to develop a scale of attitudes, items (i.e., statements) are selected based on the following criteria: They should (a) be as brief as possible, (b) be either endorsed or rejected according to their agreement or disagreement with the respondents' attitude, (c) include indicators regarding the reader's attitude about the issue in question, (d) not be "double-barreled" (i.e., having two or more contradicting values in one statement) or ambiguous descriptions (p. 37). Also, a majority of the statements really belong together on the attitude variable to be measured.

Thurstone's equal-appearing-interval method is based on the assumption that every statement chosen for the scale is placed in one of the eleven intervals on the psychological continuum, and that the distance between successive intervals is equal. The scale format for this approach does not require judges to respond to every item, but simply to check the ones they endorse as their own.

For the paired-comparison method (Thurstone, 1927), the items are presented to the rater in all possible combinations and the rater is required to decide which pair has more quantity of the defined construct. This process is laborious and therefore not widely used.

Likert Method

Likert (1932) proposed a method of selecting and developing statements for attitude scales based on the following criteria: All statements should be (a) described as desired behaviors (i.e., the present attitude of the subject), but not as facts (i.e., some past attitude), (b) stated clearly, concisely, and straightforwardly to avoid any kind of ambiguity, and (c) worded in such a way that the modal reaction to it may be approximately in the middle of the possible responses and that about one-half of them have one end of the attitude continuum corresponding to the left or lower part of the reaction alternatives, and the other half have the same end of the attitude continuum corresponding to the right or upper part of the reaction alternatives. Finally, they should include those different alternatives pertaining to only a single attitude variable, not several, if multiple choice statements are used.

More statements should be selected than are actually needed; inappropriate items will be deleted in the process. Then, for each statement, a cluster of response choices is provided, and numerical values are assigned to the response choices in an arbitrary order. The order of the numerical values in a cluster of alternatives plays a more important role than the numbers themselves.

Item analyses ought to be performed by calculating the correlation coefficient of each statement with the entire battery for two purposes: (a) to check objectively if the numerical values are properly assigned, and (b) to see if each statement measures the same construct that the rest of the statements measure (i.e., differentiating). Only those items with high positive correlations and those with high negative correlations, but with the numerical values reversed, should be selected for the final score.

Likert (1932) initially used a laborious method called sigma units for the scoring system, but later tested a simpler rating system with numerical units, as outlined above. The scores obtained by these two methods "almost perfectly" correlated, justifying the use of the simpler method for obtaining a total score by summing scores across items—hence, the term, the *method of summated ratings* (Bird, 1940, p. 159).

Guttman Method

According to Guttman (1944), a genuine scale exists when homogeneity or unidimensionality is virtually complete. For a unidimensional scale, all the statements that constitute the scale must be related to the attributes that define the concept or behavior to be measured. The scale must ensure the unidimensionality in such a way that an individual with a higher rank or score than another individual on the same set of statements must also rank just as high or higher on every statement in the set. This condition is called *perfect reproducibility* (Edwards, 1957).

The Guttman method is a scale analysis, rather than a scale construction technique (Edwards, 1957; Guilford, 1954). It presents an excellent system to monitor the logical nature of responses and consistency of rating, but it has some weaknesses. Guilford (1954, p. 461) summarizes these as follows: (a) The scalability criterion is hard to achieve; (b) we cannot be sure if the scale is unidimensional or if it has more than one variable; and (c) response popularity, when responses in one category pile up, results in high reproducibility.

The Method of Behaviorally Anchored Rating Scales (BARS)

The original BARS procedure proposed by Smith and Kendall (1963) is highly systematic and has well-designed psychometric properties. However, it is laborious and time-consuming; therefore, some variations also have been suggested to reduce the time factor. Three procedures for BARS are presented below: (a) the traditional BARS, (b) the shortcut BARS, and (c) the optimal procedure.

The traditional BARS procedure. Essentially designed to encourage and standardize direct observation of behaviors, and to familiarize raters with ratings for future scale use (Bernadin & Smith, 1981), this procedure requires that two groups of people with sufficient knowledge and experience on the job be investigated. It includes three steps: (a) the critical incident procedure, (b) retranslation, and (c) scaling (Smith & Kendall, 1963).

The *critical incident technique* requires one group of participants to (a) list performance qualities of the target job, (b) describe specific examples representing definitions of high, low, and acceptable performance for each quality, and (c) cluster these examples into a smaller set of performance dimensions or qualities that they typically define.

In the *retranslation* procedure for eliminating inconsistent and inappropriate examples or qualities, the second group reallocates critical examples to the same set of dimensions as in the previous step. Typically, an example is retained if a certain percentage of the group assigns it to the same dimension as did the first group (Schwab, Heneman III & DeCotiis, 1975).

In *scaling* the examples, the second group rates the behavior described in each example in terms of its effectiveness in representing performance on a specific dimension. The average rating assigned to an example determines its degree of performance effectiveness. The final instrument is a series of vertically arranged graphic scales, one for each dimension, anchored by the examples that meet the retranslation and the standard deviation criteria (i.e., 1.50 or less).

The shortcut BARS procedure. Green, Sauser, Fagg and Champion (1981) attempted another method for developing BARS cheaply and efficiently, while retaining rater participation in the process. In this method, sufficient rater training, through models and examples, replaces the retranslation procedure.

Champion, Green and Sauser (1988) suggested that the shortcut BARS can be an appropriate substitute for traditional BARS with little loss in the measurement quality and with less cost.

The optimal procedure. Kinicki and Bannister (1988) designed their scale for rating teaching effectiveness, based on the recommendations of Bernadin, LaShells, Smith, and Alvares (1976). The procedure is described as follows:

1. The first group of participants generated and defined the dimensions of teaching effectiveness. Then they illustrated behavioral examples, based on their judgment of good and poor performance on each dimension.
2. Using the list of performance dimensions along with their definitions, and a randomized list of behavioral examples, the participants in the second group independently sorted the examples into the dimensions best represented by those examples. Items were eliminated if they did not meet the 60% retranslation criterion.
3. With the remaining items, raters in the third group independently rated the level of effectiveness of each behavioral example on its dimension on a 7-point scale. Items with standard deviations greater than 1.50 were eliminated.
4. To obtain a measure of the behavioral specificity of each item, the fourth group independently rated each item on two scales. To reduce the interrater variability usually found in this procedure, Kinicki and Bannister (1988) proposed that the raters be trained to observe and document specific examples of performance.

Method of Behavioral Observation Scales (BOS)

The procedure of Behavioral Observation Scales (BOS) developed by Latham and Wexley (1977) is described as follows:

1. The target behavior is clearly defined.
2. Items are collected by the critical incident technique.

3. Among the incidents thus collected, those which describe essentially the same behavior are grouped into one cluster, which is to be a behavioral item.

4. Similar clusters are grouped together to form a BOS.

The scales may take the form of a dichotomous rating, but, a Likert-type rating format is recommended for more accurate assessment of frequency. To select most discriminating items, item analysis is employed, as in the Likert method. A total score for each individual is calculated by summing the rater's responses to all the items.

The Method of Mixed Standard Scales (MSS)

The mixed standard scale (MSS) proposed by Blanz and Ghiselli (1972) minimizes such common errors in rating, such as halo and leniency errors (defined below), and provides a useful index of the accuracy of rating. The logic of the scoring system is similar to that of the Guttman method: Any illogical responses of the rater are regarded as errors, and the error counts help to identify erroneous raters and ambiguous dimensions, which is a useful indicator of reliability. If the errors are unique to a particular rater, they can be attributed to the rater's lack of training. Yet if a certain error is consistent across raters, the dimensions for the scales should be examined for any ambiguity. Errors can also occur when an individual's performance is not appropriately known to the raters, or when a great deal of inconsistency exists in an individual's behavior. The MSS also is characterized by its disguised items and dimensions, which reduce halo and leniency errors (Saal & Landy, 1977).

Scale Evaluation

Jacobs, Kafry and Zedeck (1980) proposed three broad categories of criteria as the requisite properties for a scale evaluation system: (a) quantitative criteria, (b) qualitative criteria, and (c) utilization criteria.

Quantitative Criteria

Reliability

Reliability refers to the consistency of ratings across conditions. Four types of reliability are usually specified: (a) interrater reliability, (b) reliability across formats, (c) reliability over time, and (d) internal consistency.

Interrater reliability. This term means the consistency of ratings across raters, or the degree to which ratings on an individual from two different raters tend to converge. This type of reliability is addressed by correlating the ratings by one rater with those by another rater.

Reliability across formats. Reliability across formats or indices reflects the degree to which the assessments of an individual from two types of measures are in agreement. It is determined by correlating the ratings on one format with ratings on another format.

Reliability over time. Reliability over time is the degree of stability in evaluating an individual at two different points in time. This type of reliability is estimated by the "test-retest" method.

Internal consistency. Internal consistency is related to whether or not the statements or items are measuring the same behavior dimension. This type of reliability is addressed by correlating the scores on each item with the scores on the whole battery.

Validity

Generally, the term validity means the extent to which a rating instrument measures what it is supposed to measure.

Content validity. The extent to which the behaviors specified in a set of scale items are related to, and represent, the domain of behaviors they are designed to measure is referred to a content validity. To examine a behavior measurement instrument for content validity, (a) the domain of behaviors must be clearly defined, (b) a panel of experts decides which items fit into which behavior categories, (c) the experts investigate the degree of match between items and behavior categories using a systematic process, making judgments about the extent to which the whole set of items represents all the aspects specified in the definition of the behavior domain, and (d) the researcher collects and summarizes the results of this process (Crocker and Algina, 1986).

Criterion-related validity. Two measures are compared to investigate this type of validity: a measure of interest under investigation (i.e. the rating scale), and a measure on some standard (the criterion), which may be pre-established. This type of validity usually has two variants: predictive validity, and concurrent validity.

Predictive validity indicates how accurately a specific test or rating on one behavioral dimension estimates the future occurrence of another behavioral dimension. *Concurrent validity* refers to the extent to which we can estimate the performance on one measure (the criterion) using the performance on another measure (the scale under investigation). We can establish the concurrent validity of a rating instrument by comparing the scores on the rating with direct observations of performance on the criterion.

Construct validity. *Construct validity* refers to the extent to which a rating instrument measures an individual's standing on a specific construct (Ghiselli, Campbell, & Zedeck, 1981). Two types of construct validity that are frequently addressed are convergent validity and discriminant validity.

Convergent validity is usually defined as the correlation between the scores on two or more different instruments on the same construct (Ghiselli, Campbell, & Zedeck, 1981; Green, Sauser, Jr., Fagg & Champion, 1981). Campbell and Fiske (1959) suggested that, for

convergent validation, this correlation should be significantly different from zero and sufficiently large. *Discriminant validity* is defined as correlations between tests that should differ from each other. Tests are invalidated if they are too highly correlated with other tests from which they should differ (Campbell & Fiske, 1959; Dickinson & Zellinger, 1980).

Accuracy

Halo error. A halo error refers to the tendency of a rater to appraise an individual similarly across traits or dimensions because of the rater's general impression of the individual.

Leniency error. Leniency error refers to the tendency of raters to assess individuals too high or too low, resulting in negatively or positively skewed distributions of ratings.

Central tendency error. Here, raters tend to hesitate in making positive or negative extreme judgments and try to be safe by rating individuals around the mean of the group. We can avoid this tendency by using an even number of scale points or, in graphic scales, by spacing the descriptive phrases around the middle farther apart (Guilford, 1954).

Logical error. An error also occurs when the rater thinks that certain pairs of dimensions or traits in a rating instrument are similar, and thus gives ratings in a similar way.

Contrast error. This type of error occurs when raters use themselves as referents when evaluating others, and consequently rate the others as opposite from themselves in trait or behavioral dimensions.

Similar-to-me error. This kind of error occurs when raters perceive individuals to be similar to themselves, and rate the individuals more favorably. Similar-to-me errors can be minimized by training raters to be properly acquainted with those errors.

Proximity error. A proximity error refers to the tendency to rate a person on adjacent traits or dimensions in a similar fashion. This tendency may be avoided by placing similar traits or dimensions farther apart and more obviously different ones close together, or by rating one trait at a time with greater time intervals between them (Guilford, 1954).

Qualitative Criteria

Qualitative criteria are those rules or guidelines by which we evaluate performance-rating scales on their adequacy, usefulness and benefits. These criteria include: (a) relevancy of behaviors specified in the rating scale for successful performance of a task, (b) data availability, (c) practicality in terms of time and cost for both developing and administering the scales, (d) existence of equivalent frames of reference for raters to use in evaluating individuals.

Utilization Criteria

Utilization refers to the purpose for which a rating is conducted. Criteria for this includes (a) identifying individuals who need more training or practice, (b) providing feedback on the present level, (c) identifying individuals for promotion to a higher level, or for award, (d) validating a program, and (e) providing objectives for training and supervision.

Using Published Rating Scales

This section provides teachers or researchers with guidelines for selecting published rating scales. For comprehensive review of published rating scales and checklists, readers may refer to Rie and Friedman (1978), Edelbrock (1988), Witt, Cavell, Heffer, Carey, and Martens (1988), and Wilson and Bullock (1989).

Based on reviews by Edelbrock (1988) and Witt, et al., (1988), we suggest the following considerations for selecting a published rating instrument:

1. Read the manual for the rating instrument to find what type of behavior and what domain of the behavior it is designed to measure. Some rating scales focus on global attributes, whereas others are designed to measure smaller, or more specified ranges of behavior.
2. Investigate the extant psychometric properties such as reliability and validity.
3. Ensure that the rating scale is appropriate for the subject's and/or informant's age, cognitive or educational level, and language capacity. To be certain of the contents, examine the individual items.
4. Examine the response format to determine the difficulty and/or burden of the measure. It should be simple to use, and clear to avoid any errors in rating. Also the number of items should not be too large, to save time and efforts for rating and scoring.
5. The financial costs of the assessment materials should be taken into consideration.

Conclusion

Behavior-rating scales provide useful information that can be incorporated in screening and grouping individuals and in designing and validating programs for those with inappropriate behaviors. In many cases, published rating instruments can be used or adapted, for which we have provided guidelines. However, new instruments must be developed if (a) the individuals to be rated significantly differ from the normative group in terms of their age, ecological or cultural situations, degree of retardation, etc., and (b) no published instrument is found that contains the target behavioral constructs or domains. The procedures and methods specified in this paper can be used or adapted to construct valid and reliable instruments that reduce rating errors.

References

Bernadin, H. J., LaShells, M. B., Smith, P. C., & Alvares, K. M. (1976). Behavioral expectation scales: Effects of developmental procedures and formats. *Journal of Applied Psychology, 61*(1), 75-79.

Bernadin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). *Journal of Applied Psychology, 66*(4), 458-463.

Bird, C. (1940). *Social psychology*. New York: Appleton-Century-Crofts.

Blanz, F., & Ghiselli, E. E. (1972). The mixed standard scale. *Personnel Psychology, 25*, 185-199.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Champion, C. H., Green, S. B. & Sauser, Jr., W. I. (1988). Development and evaluation of short-cut derived behaviorally anchored rating scales. *Educational and Psychological Measurement, 48*(1), 29-41.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*(4), 481-489.

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology, 65*(2), 147-154.

Dixon, P. N., Bobo, M. & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychologica Measurement, 44*(1), 61-66.

Edelbrock, C. (1988). Informant reports. In E. S. Shapiro, & T. R. Kratochwill (Eds.), *Behavioral assessment in school* (pp. 351-383). New York: The Guilford Press.

Edwards, A. I. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*(4), 327-358.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. H. Freeman and Company.

Green, S. B., Sauser, Jr., W. I., Fagg, J. N., & Champion, C. H. (1981). Shortcut methods for deriving behaviorally anchored rating scales. *Educational and Psychological Measurement, 41*, 761-775.

Guilford, J. P. (1954). *Psychometric Methods*. New York: McGraw-Hill.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology, 33*, 595-640.

Kinicki, A. J. & Bannister, B. D. (1988). A test of the measurement assumptions underlying behaviorally anchored rating scales. *Educational and Psychological Measurement, 48*(1), 17-27.

Latham, G. P., & Wexley, K. N. (1977). *Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 30*, 255-268.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, No. 140.

Rie, E. D., & Friedman, D. P. (1978). *A survey of behavior rating scales for children*. Ohio Department of Mental Health and Mental Retardation.

Saal, F. E., Landy, F. J. (1977). The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance, 18*, 19-35.

Schwab, D. P., Heneman III, H. G, & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*, 549-562.

Smith, P. C. & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*(2), 149-155.

Thurstone, L. L. (1927). A law of comparative judgment. *The Psychological Review, 34*, 273-286.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Wilson, M. J., & Bullock, L. M. (1989). Psychometric characteristics of behavior rating scales: Definitions, problems, and solutions. *Behavioral Disorders, 14*(3), 186-200.

Witt, J. C., Cavell, T. A., Heffer, R. W., Carey, M. P., & Martens, B. K. (1988). Child self-report: Interviewing techniques and rating scales. In E. S. Shapiro, & T. R. Kratochwill (Eds.), *Behavioral assessment in school* (pp. 384-454). New York: The Guilford Press.

Worksamples: A Focus on Pupil Learning

David V. Myton
Teacher Standards and Practices Commission
Nancy Nagel
Pacific University
Dean Osterman
Teacher Standards and Practices Commission

Oregon's Teacher Standards and Practices Commission (TSPC) regulated the wrong aspect of teacher education. That was the assessment of the TSPC's approval of teacher preparation programs in 1986 by a Legislative Interim Committee on Education. The Interim Committee, co-chaired by House Speaker Vera Katz and Senate President John Kitzhauber, believed that the curriculum of teacher preparation programs was too prescribed by Rules for Certification. Instead of regulating such program "inputs," the Interim Committee stated that TSPC should establish standards for program approval that evaluate the "outcomes" of preparation programs.

This paper reviews the changes that have resulted from redirecting the emphasis of program approval and describes the fifth-year teacher education programs that have been approved under new "outcome" standards. The new standards employ "worksamples" to focus the attention of university and school district teacher educators on whether elementary and secondary pupils are meeting district learning objectives when instructed by student teachers.

Background on Program Approval in Oregon

The Interim Committee's criticism of the approved program approach to certification called into question a process that had been considered state-of-the-art for the past quarter-century. Oregon's approach to certification was based on administrative rules that specified the curricular content of teacher education programs. Responsibility for developing and assessing teacher candidates' knowledge and skills in the specified areas was delegated to colleges and universities whose preparation programs were approved by TSPC. Teaching certificates were issued upon recommendation of the preparing institution without further evidence of the candidates' knowledge or teaching skills.

It became apparent to the Interim Committee that fundamental changes would be required in teacher preparation, teacher certification, and public school practice to solve what the Committee believed to be an imminent crisis in elementary and secondary schools. The Committee's recommendations included several reforms that had been advanced in other states:

- Abolish the undergraduate major in education.
- Establish a fifth-year professional preparation program leading to initial certification and a master's degree.
- Increase the amount and quality of field-based experiences and student teaching.
- Require passage of tests in subject matter and professional knowledge prior to granting an initial teaching certificate.
- Create a forgivable loan program for minorities and teachers in fields with insufficient applicants for teacher education.
- Provide mentors to assist first-year teachers' adjustment to the conditions and requirements of the workplace.
- Fund continuing professional development activities that are directed by school building committees composed primarily of classroom teachers.
- Limit the numbers of candidates admitted to teacher education programs, to better balance the supply of personnel with needs of public schools.
- Raise standards for approval of programs to insure that candidates receive better preparation for teaching and to grant institutions greater flexibility in designing that preparation.

New Standards for Approval of Programs

The last recommendation prompted TSPC to create new standards for fifth-year teacher education programs in 1987. TSPC also modified in 1988 the standards under which four-year programs will be approved after 1993. The new and modified standards incorporate many of the recommendations of the Interim Committee on Education, such as required tests of subject matter and professional knowledge, in addition to the California Basic Educational Skills Test (CBEST), which has been required since 1985. The most far-reaching change in the new and modified standards is the requirement that student teachers complete three

worksamples which report the extent to which their pupils achieved the objectives of their lesson plans.

Worksamples: What and Why

A worksample consists of a series of related lessons, of 2 to 5 weeks in duration, which is a part of the school's regular curriculum during the period the student teacher is assigned. Each worksample includes the following:

- Goals for the unit of study
- Instructional plans for each lesson in the series
- Information on pupils' knowledge and skills in the area prior to instruction
- Data on learning gains resulting from instruction
- Interpretation and explanation of the learning gains or lack thereof
- Description of the uses to be made of the findings on learning gains in planning further instruction and in reporting pupils' progress

Prospective student teachers are encouraged to begin planning lessons that will be a part of their first worksample during field experiences that precede student teaching. This type of lesson planning is typical in most teacher education programs and provides training in developing objectives, selecting instructional materials and media, estimating time needed for instruction and practice exercises, and planning for evaluation of learning. What is different about the worksample is the emphasis on pre- and post-teaching assessment data, interpretation of learning gains, and use of data on learning gains in planning further instruction and in reporting pupil progress.

Cooperating teachers take an active role in worksamples and assist student teachers in selecting appropriate topics of study, developing lesson plan ideas, selecting and utilizing evaluation procedures, and interpreting learning gains. Thus, the worksamples stimulate a common focus on teaching and pupil learning for the cooperating teacher and the student teacher.

Worksamples Contribute to Effectiveness

TSPC believes that worksamples will make a difference in the effectiveness of student teachers. The ability to plan instructional activities is an important part of effective teaching (Jackson, 1966; Eisner 1967). Clear content development is consistently positively correlated with pupil achievement (Brophy & Good, 1986). Proper sequencing of instruction assists pupils to order new learning or perform new tasks (Gagne and Briggs, 1979; Brophy, 1982). Worksamples insure that units of instruction are carefully designed and reviewed before the student teacher attempts to teach the lessons.

Worksamples focus unit planning on the curriculum objectives of the school and district. Beginning teachers are primarily concerned about maintaining order in the classroom (Fuller, 1969). These concerns may be reflected in their plans by efforts to reduce their own uncertainty and anxiety (Yinger, 1978). They may be tempted to create their own objectives and curricula or to pick and choose parts of textbooks, while ignoring other sections altogether (McNergney, 1985). One of the important functions of worksamples is the sense of security and direction that they provide for student teachers.

An accurate assessment of the levels of achievement of pupils prior to instruction is essential in worksamples. To provide effective learning experiences, teachers must match the difficulty of content to the pupils' current achievement levels and needs (Brophy & Evertson, 1976; Tyler, 1965). Bloom's (1976) work suggests that diagnostic skills are among the most important for teachers to master. From a study of teachers' planning decisions, Borko, Cone, Atwood, and Shavelson (1979) found that effective teachers estimate pupils' abilities and learning needs, when planning instruction, and that they revise their instructional plans when additional information about the pupils is provided. Worksamples require close working relationships between cooperating teachers and student teachers to match content to pupil achievement levels and needs.

Worksamples assist student teachers to estimate the time needed for instruction and seatwork assignments, a task that is difficult for the beginning teacher. Timing is important for effectiveness, because the amount of time spent on academic tasks affects achievement (Squires, Huitt, & Segars, 1984). Spending and/or allocating insufficient learning time has a direct negative effect on achievement (Gettinger, 1985). Effective teachers provide sufficient activities and enough time for each activity to be completed (Brophy, 1983; Good & Grouws, 1977; Hawley & Rosenholtz, 1984; Rosenshine & Stevens, 1986). Worksamples provide a means for cooperating teachers to assist student teachers with time estimates.

Student teachers are expected to go beyond demonstrating what subject matter they know or what teaching skills they possess. Through worksamples they demonstrate what pupil learning gains they can accomplish. The inherent weakness of the traditional approved program approach is reliance on prescribed courses, grade point averages, and test scores as demonstration of "mastery" of basic skills or content to be taught (Schalock, 1990). Some teaching skills are demonstrated only in conjunction with college courses through micro-teaching exercises or in simulated lessons taught to the prospective student teacher's peers. Worksamples provide a means to apply knowledge and skills in a real-life setting.

The use of worksamples changes the goal of student teaching from performing like an effective teacher to demonstrating that one is an effective teacher. The limitation of observing "effective" teaching characteristics in student teachers is that no research has shown that effective teachers have specific characteristics in all teaching situations. The relationship between specific teaching behaviors and pupil learning is difficult to defend empirically. Worksamples provide more concrete evidence of effectiveness than do observations of performance alone.

Finally, worksamples bring forcefully to attention the context in which teaching occurs. Teaching behaviors that foster learning in pupils vary not only by grade level and subject area, but by high- and low-ability pupils within grades and subjects as well. Student teachers analyze achievement of district learning goals by quartiles of pupils prior to instruction. Such analysis is intended to focus attention on the subtle and complex relationship between context of teaching and effectiveness of teaching (Schalock & Myton, 1988).

Worksamples Are Not the Sole Basis for Certification

The research base is not adequate to support pupil learning gains as the sole basis for determining student teacher competence. Evidence of positive learning gains is not required for successful completion of a worksample. In fact, a worksample containing an interpretative essay convincingly explaining the lack of learning gains might be judged as worthwhile as a worksample with mediocre learning gains, but without an insightful explanation of the results.

Although the standards require that student teachers be evaluated on the worksamples by both the university supervisor and the cooperating teacher, grades on worksamples are not reported to TSPC. Instead, the worksamples form a part of the documentation for two of the four objectives assessed in student teaching: planning for instruction and evaluating pupil achievement. The other two objectives are evaluated largely on observation of teaching performance: establishing a classroom climate conducive to learning and implementing plans for instruction. Both the university supervisor and the cooperating teacher must provide verification to the institution and to TSPC that the student teacher has demonstrated minimum competence on each of the four objectives in order for a teaching certificate to be issued.

Achieving the Potential of Worksamples

Although all teacher education programs will be required to incorporate worksamples by 1993, only the State's new fifth-year programs were required to have worksamples beginning in 1988-89. During that year a total of 19 student teachers completed fifth-year programs for secondary teachers offered at three independent liberal arts colleges and universities.

The authors reviewed each of the 1988-89 student teachers' worksamples and observed over half of them at their student teaching sites. Each of the programs was evaluated on site during April 1989 by a TSPC team. By all accounts, the implementation of worksamples was successful. As with any program with such small numbers, success was insured by the enthusiasm for the new approach and the intense collaborative efforts of institutions and school districts to make it work. The question Oregon now faces is: Will worksamples be effective on a larger scale?

In 1990-91, eight Oregon colleges and universities will have implemented fifth-year programs. Two of the independent institutions approved in 1988-89 for secondary teacher education added elementary programs in 1990-91. One state college is expected to revise its four-year program in advance of new standards that apply in 1993, thus bringing its programs under the requirement of worksamples. The total number of elementary and secondary student teachers completing worksamples in 1990-91 is expected to be approximately 1200.

When worksamples were first proposed by TSPC, college and university faculties were not enthusiastic about them as "outcome" measures of teacher education programs. They were concerned that pupil learning would become the sole criterion for certification. But those who have tried the worksamples with their student teachers are much more supportive, especially since the worksamples are now seen as contributing to evaluation of student teaching. Several university supervisors have stated that the worksample is a powerful tool to shift the focus from the student teacher as performer to student teacher as coach and the focus from pupil as audience to pupil as worker.

Cooperating teachers are also optimistic about the potential of worksamples. Several commented, however, that they lacked knowledge of evaluation strategies needed for the pre- and post-teaching assessments in worksamples. Most cooperating teachers felt that the worksample assignments placed very high expectations on student teachers, more like those placed on first-year teachers.

School district administrators are supportive of the higher requirements in student teaching, because the State Department of Education requires evidence that schools are meeting district and state curricular objectives. But most are concerned about the amount and depth of training that cooperating teachers will need to perform the evaluation assignments connected with worksamples and student teachers. Some larger districts have taken upon themselves responsibility for such inservice, but the financial ability of districts to provide

such inservice varies greatly. All of the approved colleges and universities are also providing some inservice opportunities to prepare cooperating teachers.

Student teachers are concerned about the time necessary to complete the worksamples, but nearly all stated that the assignment was beneficial. One 1988-89 student teacher wrote that her "knowledge and skills increased due to the larger picture of teaching presented to me because of the requirements of the worksample." Another student teacher stated, "The worksample prepared me to organize lessons that would clearly match the unit of study's goals and helped me become aware of the learning occurring in the classroom."

Worksamples Contribute to Program Improvements

One major weakness has been identified in the preparation of student teachers to complete the worksamples—skill in pre- and post-teaching assessment of pupil achievement. To correct this deficiency, the Oregon Association of Colleges for Teacher Education arranged for Richard Stiggins of the Northwest Regional Educational Laboratory to conduct a three-day teacher-training workshop on evaluating pupil achievement. Each teacher training institution had a team of faculty and cooperating teachers participate in the training workshops. Dissemination of the workshops within the institutions and the districts with which they cooperate in student teaching is being conducted by these trainers.

The new standards and the site evaluation procedures dispense with the traditional institutional report to the visiting team. Instead, the institution is required by the standards to produce and provide for its advisory consortium and TSPC a series of four reports. In the first year of program implementation, a site visit is conducted during the student teaching phase. A site visit also is made during Year 5 of the program and every five years thereafter. The majority of a visiting team's time is devoted to evaluation of the field-based component of the program, including substantive observations of student teachers and review of worksamples. Although programs under the new standards have not been in operation long enough to have gone through the entire cycle, it is expected that worksamples will figure prominently in the data analyzed throughout the cycle of reports and evaluations.

Summary

In response to recommendations of a 1986 Legislative Interim Committee on Education, the Teacher Standards and Practices Commission created new standards for teacher education programs. One distinctive change is the addition of worksamples to student teaching requirements. Worksamples are 2- to 5-week units of study during which the student teacher is to demonstrate ability to foster pupil learning.

Implementation of the worksample requirement in 1988-89 was deemed successful, however only 19 student teachers were placed under the new standards. Several problems have been identified that must be addressed in order to achieve the objectives of lengthened student teaching with worksamples when the full complement of 1200 student teachers a year is reached in 1990-91.

Procedures for evaluation of programs under the new standards will draw upon evidence from worksamples to guide future program improvements. In the end, elementary and secondary pupils are expected to benefit from the new approach to program approval through increased learning gains.

References

Bloom, B. S. (1976). Human characteristics and school learning. New York: McGraw-Hill.

Borko, H., Cone, R., Atwood, R., & Shavelson, R. J. (1979). Teachers' decision making. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp.136-160). Berkeley, CA: McCutchan.

Brophy, J. (1982). Effective teaching strategies for the inner city child. *Phi Delta Kappan, 64*, 527-530.

Brophy, J. (1983). Classroom organization and management. *The Elementary School Journal, 83*, 265-284.

Brophy, J. & Evertson, C. (1976). *Learning and teaching: A developmental perspective*. Boston: Allyn and Bacon.

Brophy, J. & Good, T. L. (1986). Teaching behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook on research on teaching* (3rd ed., pp. 328-375). New York: Macmillan.

Eisner, E. W. (1967). Educational objectives: Help or hindrance. *School Review, 75*, 250-266.

Fuller, F. F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6*, 207-226.

Gagne, R. M, & Briggs, L. J. (1979). *Principles of instructional design* (2nd Ed.). New York: Holt, Rinehart and Winston.

Gettinger, M. (1985). Time allocated and time spent relative to time needed for learning as determinants of achievement. *Journal of Educational Psychology, 77*, 3-11.

Good, T. & Grouws, D. A. (1977). Teaching effects: A process-product study in fourth grade mathematics classrooms. *Journal of Teacher Education, 28*, 49-54.

Hawley, H. D. & Rozenholtz, S. J. (1984). Good schools: What research says about improving student achievement. *Peabody Journal of Education, 61*, 15-52.