Resource Consultant Training Program
Research Report No. 6

RCTP

Use of Two Metrics in Direct Observation for Evaluating Effectiveness of Special Education Programs

Richard Parker
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

Parker, Richard
Tindal, Gerald
Use of Two Metrics in Direct Observation for Evaluating Effectiveness of Special Education Programs
Research Report No. 6

Staff

Gerald Tindal, Program Director
Jerry Marr, Editor
Clarice Skeen
Denise Styer
Mark Baldwin

Acknowledgements

Preparation of this document was supported in part by the U.S. Department of Education, grant number G008715106. 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

Use of Two Metrics in Direct Observation for Evaluating Effectiveness of Special Education Programs

Richard Parker
Gerald Tindal

Abstract

In the past decade, a number of methodologies have been proposed for observation in the classroom. Generally, research has focused on the use of one instrument and has rarely reported results from validation investigations. The current study, however, employed two direct observation instruments concurrently within two reading programs—Breaking the Code and an Eclectic Program—in middle school resource rooms. A momentary time sample of task engagement and an event record of discrete student responses were used in six classrooms representing the two programs. A multi-method validation process was implemented that focused on treatment and criterion-related validities. Many findings were program-specific, with differences lost or diluted when data were combined. An argument is presented for structuring observation in a manner sensitive to classroom activity structures.

INTRODUCTION

In this study, two instruments were used to measure student behaviors within middle school (Grades 6 through 8) learning disabled classrooms: a 15-second momentary time sample (MTS) of task engagement and a 2-minute event record (ER) of discrete academic behaviors. The instruments employed two different time sampling techniques and two different metrics, namely "engagement rate" (percentage of class time) and "response rate" (occurrences per minute). The results from these instruments were used to distinguish between two remedial language arts programs which differed widely in curriculum and teaching procedures. In one program, Breaking the Code (Lebo, Hughs, Thomas, & Gurren, 1975), teachers employed (a) scripts, (b) a very controlled sequence of skills and activities, and (c) an integrated blend of reading, spelling, and writing word and letter combinations. In the other program, the Eclectic Program, an amalgamation of materials was employed, with no explicit sequence of skills and activities, and with an emphasis on oral and silent reading of text, rather than writing and spelling. Research questions, which were directed toward the validity of the two observational measures, addressed (a) their sensitivity to reflecting instructional differences beween the two programs, (b) their relationship with each other, and (c) their relationship with non-observation criteria of student reading achievement and workbook scores.

As Hoge (1985) noted in a review of validity of direct observation instruments, several methods for validating new instruments are available to researchers, although they are not attempted often. He defined three types of validity: treatment, criterion-related, and construct. Treatment validity is concerned with the sensitivity of the measure to reflecting different instructional interventions. Criterion-related validity relates to the degree of correlation with other, accepted measures. Construct validity is defined as the ability of the measure to confirm relationships among variables as-

sumed by accepted theory. This study addresses the first two of these three types of validity.

Time as a Process Variable

Of the classroom measures identified within the past 20 years, time has gained the widest acceptance. Influenced by Carroll's (1963) original Model of School Learning, the majority of later models presenting comprehensive conceptions of classroom learning have been largely or entirely time-driven (Haertel, Walberg, & Weinstein, 1983; Leinhardt, 1980). From these models of classroom learning has grown a sizable body of research showing that various indices of time do bear stable relationships to student learning outcomes, although the strength of relationship apparently varies, depending upon the learning context (Frederick & Walberg, 1980; Karweit, 1983; Karweit & Slavin, 1981). The relationship between instructional time and student outcomes also seems to increase as measures of time become more specific and proximal to the actual learning process (Greenwood, Delquadri, & Hall, 1984).

Among the various measures of time appearing in the literature over the past decade, the following have been employed in several studies:

1. Time-on-Task or Engagement Rate (Anderson, 1980; Bloom, 1980; Stuck, 1980), or teacher-allocated time. This is time in which the student is either passively attending to instruction or actively engaged in task performance.

2. Active Learning Time (ALT_1) (Harnishfeger & Wiley, 1985). This is time spent actively performing or practicing learning tasks that lead to mastery. The definition entails two criteria: Student behavior or task performance must be *observed*, and it must *accurately reflect* mastery of a terminal skill. This measure focuses on the student's active involvement in the learning process, which is viewed as a mediator of other instructional variables that are more directly under teacher control.

3. Academic Learning Time (ALT_2), as used in the Beginning Teacher Evaluation Study (Fisher, *et al.*, 1978; Rosenshine & Berliner, 1978) or teacher-allocated time containing engagement in learning tasks and high levels of student success (90% or better). Like ALT_1, this variable is defined by the two main criteria of observed student behavior and accurate reflection of terminal skill mastery.

4. Opportunity to Respond (Stanley & Greenwood, 1981). This measure is operationalized through a specific instrument, the Code for Instructional Structure and Student Academic Responding (CISSAR), wherein active and passive academic responding are coded. Active responding is organized into several narrow-band or molecular behavior classes. This system is the basis for all classroom observation research conducted at the University of Minnesota's Institute for Research on Learning Disabilities (c.f. Thurlow & Ysseldyke, 1983).

There are several major differences among these measures. The broadest measure, Time-on-Task, has no narrower behavior classes specified. ALT_2 (from BTES) operationalizes student accuracy criteria on tasks, while ALT_1 (from Harnishfeger & Wiley, 1985) requires that student products build toward mastery of a terminal skill. ALT_2 includes only time-on-task or engagement rate, while ALT_1 requires active task performance. ALT_2 uses the metric of "amount or proportion of class time" a student attends to instruction or is engaged in a particular class of activities, while the definitions by Harnishfeger and Wiley (1985) and Stanley and Greenwood (1981) allow the estimation of the rate of occurrence (per unit of time) of discrete behaviors within narrow behavior classes. Actual calculation of rate is not possible, but the summation of intervals allows a rate estimator.

While studies have addressed the relationship between different time sampling techniques (Green & Alverson, 1978; Powell, Martindale & Kulp, 1975; Powell & Rockinson, 1978; Repp, Roberts, Slack, Repp, & Berkler, 1976; Test & Heward, 1984) little research has been completed on the relationship between time sampling and event record data. For example, Greenwood, *et al.* (1984) found that engagement in academic behavior (writing and silent reading) significantly correlated with achievement, while attention to task alone did not. However, the broad class of "engaged time" was measured rather than narrow response classes. The measure of engaged time did not take into account the types of responses students were making.

To investigate the relationship between student engaged time and academic response rates, two observation systems were constructed. A momentary time sample (MTS) coded the percentage of occurence of several categories of academic engaged time. This was similar to the approach taken by Stanley and Greenwood (1981). An event record (ER), as used at the University of Washington (Jenkins & Stein, undated) coded the frequency of discrete behaviors occuring within sampled intervals to provide response rate (per minute) data. Both observation instruments are substan-

tially new measures, and are described more fully in the Method section.

Use of the MTS and ER instruments within a single observational session offers several advantages:

1. Molecular and molar categories of student behavior are considered within the broader context of Time-on-Task or student-engaged time.
2. Discrete behaviors are counted within these behavior categories.
3. The accuracy or correctness of responses can be expressed both in terms of correct responding rate (per minute) and error, or correct ratios.
4. Rate data on discrete student and teacher behaviors may facilitate the identification of potential contingent relationships between specific student behaviors and antecedent or subsequent classroom events.

This article describes validity data from the concurrent use of the engaged time metric (through the MTS) and response rate metric (through the ER) in coding classroom behavior. Two of the three types of validity referenced by Hoge (1985)—treatment and criterion-related—are presented. For treatment validity, data obtained on the MTS and ER in two widely different reading programs are compared. For criterion-related validity, observational data are correlated between the two instruments, as well as between each instrument and student performance on tests and workbooks.

METHOD

Instrumentation

Two direct observation instruments were designed for concurrent and complementary use in a classroom, a 15-second momentary time sample (MTS) and a 2-minute event record (ER).

Momentary Time Sample

The first instrument requires coding the presence or absence of one out of nine mutually exclusive behaviors for a single student at the end of a 15-second interval. At the end of each interval a different student's behavior is coded according to classroom seating. A classroom cycle is completed when every student in the class has been observed and coded once. Within a typical middle school class period, 6 to 12 cycles can be coded depending upon class size. Behavior categories for the 15-second MTS include the following:

1. Off-task: Class time allocated by the teacher for instruction or student performance, during which the student is not engaged (neither actively performing a task nor passively attending to instruction).
2. Non-task: Class time not teacher-allocated for instruction or student academic performance, and therefore the student has no opportunity to be on-task. Task organization, management, and transition times are included here.
3. On-task: (a) Passive Responding: Student is passively attending to an instructional presentation or a learning task; no student activity is observed; (b) Active Responding: Student is actively responding in a relevant manner to instructional presentation or learning task. Within Active Responding, the following specific tasks were differentiated: Oral Reading, Silent Reading, Spelling (units of less than one sentence), Writing (units of one sentence or more), Copying (from a complete model), and Other.

The Momentary Time Sample scoring sheet is presented in Appendix A.

Event Record

The second instrument requires coding for a single student on the frequency of each of 12 student behaviors and one related teacher behavior during a 2-minute interval. Start and stop times also are recorded for each category of behavior occurring within this interval. These start-stop times allow the calculation of response rate (per minute) for each behavior category. For each 2-minute interval, a new student is observed according to classroom seating. One cycle is completed when every student in the classroom has been observed and recorded over a 2-minute interval. Within one class period, two to four cycles typically are completed.

The five major ER behavior categories and subcategories are:

1. Oral Reading: (a) Correct Responses (number of words); (b) Error Responses (number of words); (c) Error Feedback by teacher or other source (per Error Response); and (d) Self-Correction of Errors (per Error Response).
2. Silent Reading: Correct Responses (number of sentences). The observer, standing directly behind the student, uses the student's eye focus on page, eye movements, finger cues, and page turning to determine the number of sentences read silently.
3. Spelling: (a) Correct Responses (number of words); (b) Error Responses (number of words); (c) Error Feedback by teacher or other source (per Error Response); and (d) Self-Correction of Errors (per Error Response).

4. Writing: (a) Correct Responses (number of sentences); (b) Error Responses (number of sentences); (c) Error Feedback by teacher or other source (per Error Response); and (d) Self-Correction of Errors (per Error Response).

5. Copying: (a) Correct Responses (number of words); (b) Error Responses (number of words); (c) Error Feedback by teacher or other source (per Error Response); and (d) Self-Correction of Errors (per Error Response).

The Event Record scoring sheet is presented in Appendix B.

In order to establish interrater reliability, two doctoral students in a special education program at the University of Oregon concurrently observed and coded individual student behaviors with the MTS and the ER. Reliability was established over a 15-minute period within one of the six target classrooms. For the MTS, interrater agreement was determined by pairing individual tallies of the two raters, and rating each pair as "identical" or "not identical." Tallies in an identical pair were required to fall in the same response category and apply to the same student. On this basis, 98% agreement on the MTS was obtained within the 15-minute period. For the ER, interrater agreement was calculated separately for (a) skill area of response (oral reading, silent reading, spelling, writing, copying), 100%; (b) skill area and total number of student responses, 81%; and (c) skill area, total responses, and coding of individual responses (correct, error, self-correct, and error feedback), 70%.

Subjects

Two middle schools (Grades 6 through 8) with populations of 475 and 650 were selected for this study from a West Coast suburban lower-middle SES school district of 9,000 students. These schools were chosen for accessibility. Also, their instructional programs varied little among classrooms within each school, yet varied greatly between schools. Subjects were students enrolled in language arts special classes for learning disabled students, taught by two fully trained, state-certificated teachers. Four classes existed in the school employing a Breaking the Code program (50 students), and two in the school employing an eclectic program (24 students). All students were on active IEPs for one or more language arts skill areas (Reading, Spelling, Written Expression). In the school employing a Breaking the Code program, the mean rate of words read correctly per minute (wcpm) in Grade-3 level Harcourt-Brace-Jovanovich Bookmark basal texts (Early, Cooper, & Sontensonio, 1979) was 92.7 (*SD* 39), compared with 88 (*SD* 30.7) for the school employing the Eclectic Program. Scores were not significantly different ($t_{(74)}$=.43, p>.66). The mean rate of words read correctly in a random sampling from the Harris-Jacobson (1972) corpus was 42.5 (*SD* 27.7) for students in Breaking the Code, and 40.5 (*SD* 22.2) for students in the Eclectic Program; no significant differences found between scores ($t_{(74)}$=.27, p>.6). Furthermore, no significant differences were found between the programs in reading and language subtests on the fall administration of the California Achievement Test (CTB/McGraw Hill, 1985).

Instructional Programs

Both schools offered resource room language arts programs that differed widely both in curriculum, which was composed of materials formally organized for teaching specific skills and knowledge, and teaching procedures, which included the selection and organization of curriculum content and its delivery to students.

Breaking the Code Program

One school employed Breaking the Code (Lebo et al., 1975), a published language arts curriculum with a synthetic phonics approach to spelling and reading individual words and word parts. Oral spelling and writing of individual words accompanied word reading. The curriculum emphasized reading in response units equal to or smaller than a single sentence. Similarly, the curriculum prescribed writing in response units equal to or smaller than a single word. The authors describe the program as basic skills re-teaching to prepare students for a subsequent, broader language arts program, which would include larger reading and writing response units. Teaching procedures were strongly teacher-directed, involving scripted presentations that included cues for group choral responding. Instruction was always at the whole-class level and teaching procedures were constant from day to day.

Eclectic Program

The Eclectic school curriculum consisted of a blend of high interest readers, workbooks, and published and teacher-made handouts. Oral and silent reading were emphasized; no activities were observed for which spelling was the primary focus. Teaching procedures typically included a sequence of vocabulary preview, round-robin oral reading, class discussion, comprehension questions, and chalkboard presentations on word structures. Silent

reading often occurred within and at the end of the period, if time remained. The particular materials used and the order in which they were presented varied greatly from one day to the next.

Procedure

One of the two graduate students used the MTS instrument, while the other used ER. They concurrently observed and coded individual student behaviors within the four Breaking the Code classes and two Eclectic Program classes. Each of the six classes was observed for approximately 4 hours over 4 days spanning 3 weeks during the months of January and February. Total observation time was 48 hours in each class.

Research Questions

As stated above, research questions focused on two types of validity of the observation instruments, *treatment validity* and *concurrent validity*. Treatment validity was assessed by observation system sensitivity to differences in curriculum and teaching procedures between the two reading programs at individual classroom and aggregate program levels. In particular, the observational systems should reflect (a) differences between the two programs in variability of instruction across time and (b) differences in content focus—spelling and reading letter combinations in the Breaking the Code Program, and reading of text, orally and silently, in the Eclectic Program.

Assessment of concurrent validity included the following analyses. First, performance scores on similar behavioral categories across the two observational measures were intercorrelated. The second analysis correlated (a) active engagement scores (MTS) and proportion of correct responses and (b) response accuracy (ER) with beginning, middle, and end-of-year performance on combined passage and wordlist reading scores. Predictive validity is referenced in the correlation between fall test performance and winter observation of student responses, and between this measure and spring test performance. The third analysis compared student observational data on the MTS and ER measures with student performance on workbook quiz scores compiled over the school year.

RESULTS

Treatment Validity

Differences Between Programs in Variability of Instruction

Because of wide program differences, variability in student performance—both across and within classes—was expected more for Eclectic Program students than for students in the Breaking the Code Program. Means and standard deviations of five MTS and two ER variables are presented by program in Table 1. Scores first are aggregated at a school level, then ranges of class-level scores within each school are presented. At the aggregate school level, standard deviations are larger for the Eclectic program for all seven measures, despite the fact that Breaking the Code students showed greater variability in oral reading scores on tests administered early in the year, as reported earlier.

At the class level those classrooms receiving the Eclectic Program demonstrated greater variability on all five MTS and both ER scores than classes receiving Breaking the Code. The Eclectic Program's greater variability is reflected in (a) a greater range of mean scores among classes, (b) greater standard deviations for individual classes, and (c) more variability among standard deviations for different classes (for six of the seven measures). Because of the greater number of students and classes in Breaking the Code, and the larger *SD*'s for entry-level reading skills among those students, chance alone would dictate greater variability in scores for that program; the opposite was found to be true.

Table 1 presents the results on differences between the two programs in variability of instruction, and offers other interesting information as well. For example, at the aggregate school level, students in the Eclectic Program show much higher *active* engagement levels (a full *SD* unit difference), but little difference when *passive* engagement is included. When the measure focuses on non-engagement, the same result occurs, with students in the Eclectic Program not actively engaged less (by one *SD*) than the students in Breaking the Code. On the ER measures, students in the Eclectic Program completed five times as many oral reading responses and about half as many spelling responses as those in Breaking the Code.

Differences in Content Focus

Treatment validity also was addressed by determining whether the measurement system was sensitive to the differences in content focus — spelling/reading letter combinations and isolated words in Breaking the Code and oral/silent reading of text in the Eclectic Program. Because the major program differences existed at a school rather than class level, significant differences in two MTS composite scores (Spelling/Writing and Oral/Silent Reading engaged time), and two ER variables (Spelling and Oral Reading rates) were hypothesized between programs. On the other

Table 1. Means and Standard Deviations (at a School Level) and Ranges (at a Class Level) for Five MTS and Two ER Variables in Breaking the Code and the Eclectic Program

	Breaking the Code		Eclectic Program	
School-Level Scores	*M*	*SD*	*M*	*SD*
MTS Measures				
Engaged Time$_1$	71.6	10.1	72.8	11.0
Active Engaged Time$_2$	29.2	9.4	39.8	10.6
Active Academic Engaged Time$_3$	17.2	5.9	28.7	8.5
Not Engaged Time$_4$	23.8	8.7	24.3	11.0
Not Active Engaged Time$_5$	66.2	8.5	57.3	12.0
ER Measures				
Oral Reading Rate	0.5	0.4	2.6	2.4
Spelling Rate	1.2	0.9	0.7	1.0
Individual Class Scores	low-to-high		low-to-high	
	M	*SD*	*M*	*SD*
MTS Measures				
Engaged Time$_1$	71.7-72.7	8.3 -11.2	66.3-87.9	4.4- 8.3
Active Engaged Time$_2$	26.8-30.9	7.9-10.5	35.3-50.3	7.1-12.5
Active Academic Engaged Time$_3$	16.4-18.2	4.4- 6.3	24.3-34.9	5.9- 9.5
Not Engaged Time$_4$	22.5-24.1	7.5- 9.3	7.6-32.1	5.8-10.1
Not Active Engaged Time$_5$	64.9-68.4	7.3- 9.3	45.4-63.2	9.1-14.0
ER Measures				
Spelling Rate	0.7-1.3	0.4-1.2	0.4-1.3	0.6-1.3
Oral Reading Rate	0.3-0.7	0.2-0.5	1.4-4.4	1.9-2.2

1 passive responding + oral reading + writing + copying + spelling + silent reading + other
2 oral reading + writing + copying + spelling + silent reading + other
3 oral reading + writing + spelling + silent reading
4 task management + off task
5 task management + off task + passive responding

hand, small differences on these MTS and ER variables were hypothesized among classes within each school. Within the Breaking the Code Program, between-class differences should be negligible; whereas within the more variable Eclectic Program, significant between-class differences should exist.

To test the hypotheses around differences among classes within schools, separate *F* tests were conducted for each school, with grade level (Grades 6, 7, and 8) as the between-subjects factor, and the same four observation variables as dependent measures. The two MTS measures consisted of (a) Spelling/ Writing engaged time, (b) Oral/Silent reading engaged time; the two ER measures included spelling rate and oral reading rate.

Table 2 presents mean score differences among the grade levels within each program. The Eclectic Program mean differences are more than eight times greater than the differences among the classes in the Breaking the Code. One-way ANOVAs were performed separately for each school to determine if differences among grade levels were significant. For the Breaking the Code Program, none of the *F*'s were significant. For the Eclectic Program, only one *F* value (Spelling/Writing) was clearly *not* significant, while the remaining three variables were significant or marginally so (p=.02, .05, .07).

To indicate the overall differences between the two programs, *t*-tests were conducted on time engaged in Spelling/Writing and Oral/Silent reading (MTS scores), as well Spelling and Oral Reading rates (ER variables). On three of the four variables, the mean difference between the programs was significant at the .0001 level: Spelling/ Writing engaged time, $t_{(53)}$=5.8; Oral/Silent read-

ing engaged time, $t_{(53)}$=13.3; and Oral Reading Rate, $t_{(53)}$ =5.78. On the fourth variable, the difference was not significant at the .05 level: $t_{(53)}$ =1.54.

Criterion-Related Validity

Engaged time scores and response rate scores obtained from the MTS and ER, respectively, were intercorrelated across the instruments. Because of the limited observation time, only two ER variables, oral reading rate and spelling rate, produced sufficient data for analysis. Therefore, the following three ER composite measures were employed in the intercorrelation: oral reading (correct and error), spelling (correct and error), and oral reading and spelling (correct and error). From the 15-second MTS, data were aggregated for oral reading, silent reading, oral and silent reading combined, spelling and writing combined, and active academic responding. The prediction was made that high reading or writing/spelling response rates as measured on the ER would correlate positively with high engagement rates on the MTS.

Table 3 presents correlation coefficients between three ER variables and five MTS variables; correlations are presented for the two schools separately and together. The 15 correlations range from .32 to .84, and all but one are statistically significant at the .05 level.

As in Table 2, the sizes and patterns of correlations differ greatly between the two programs. The combined data, if presented alone, would obscure important differences in correlation patterns between the two instructional programs.

Another measure of criterion-related validity is the degree to which the 15-second momentary time sample and 2-minute event record are predictive of student achievement. Passage reading and word list reading tests were administered in the fall (mid-October), winter (mid-January), and spring (mid-May). Scores from the two types of tests were

Table 2. Variation Among Grade Levels* for Both Programs on Two MTS Composite Engaged Variables and Two ER Response Rate Variables

MTS Measures		Spell/Write				Oral/Silent Read			
Grade		6	7	8	M	6	7	8	M
Breaking the Code	*M*	15.0	15.0	15.4		4.9	5.6	6.6	
	M Diff				0.20				0.10
	$F_{(2,41)}$				0.01				0.84
	p				0.98				0.43
Eclectic Program	*M*	4.7	7.8	6.4		21.4	31.3	34.0	
	M Diff				2.00				8.30
	$F_{(2,12)}$				0.72				5.10
	p				0.50				0.02
ER Measures		Spelling Rate				Oral Reading Rate			
Grade		6	7	8	M	6	7	8	M
Breaking the Code	*M*	1.2	1.2	1.1		0.4	0.6	0.5	
	M Diff				0.05				0.14
	$F_{(2,41)}$				0.02				1.30
	p				0.97				0.27
Eclectic Program	*M*	0.4	2.1	1.0		1.5	3.4	4.9	
	M Diff				1.10				2.30
	$F_{(2,12)}$				3.24				4.10
	p				0.08				0.05

*Grade levels were selected because missing data would not permit an analysis by classroom. Students initially were placed in classrooms by grade level.

Table 3. Intercorrelations Between Academic Responding Measures from the Event Record (ER) and Momentary Time Sampling (MTS) Observations

ER Rates	MTS Percentages	*r*	*p* value
Breaking the Code (N=45)			
Oral reading correct & error	Oral reading	.43	.003
Oral reading correct & error	Silent reading	.84	.0001
Oral reading correct & error	Oral & silent reading	.84	.0001
Spelling correct & error	Spelling and writing	.34	.01
Oral rdg/splg correct & error	Active academic responses	.32	.02
Eclectic Program (N=15)			
Oral reading correct & error	Oral reading	.69	.0038
Oral reading correct & error	Silent reading	.38	.15
Oral reading correct & error	Oral and silent reading	.52	.04
Spelling correct & error	Spelling & writing	.46	.01
Oral rdg/splg correct & error	Active academic responses	.58	.022
Combined Programs (N=60)			
Oral Reading correct & error	Oral reading	.34	.006
Oral Reading correct & error	Silent reading	.67	.0001
Oral reading correct & error	Oral & silent reading	.74	.0001
Spelling correct & error	Spelling & writing	.40	.001
Oral rdg/splg correct & error	Active academic responses	.50	.0001

standardized and averaged for the purpose of this analysis. As was stated earlier, the focus of the Eclectic Program was oral and silent reading of text, while Breaking the Code focused on spelling and writing letter combinations and words. Because of this difference, it was hypothesized that the reading tests would correlate more highly with engagement rates from the Eclectic Program than from Breaking the Code. It was hypothesized also that the ER variable of proportion of correct oral reading responses (response accuracy) would predict reading test scores in Eclectic Program, but not in the Breaking the Code program.

Table 4 presents correlations between reading scores and MTS and ER variables. Correlations of engaged and not-engaged time with passage and word list reading scores for students within the Eclectic program reflect moderate and significant relationships. Within this program, correlations are strongest for the winter test, the time closest to the in-class observations. However, the correlations between the same process variables and reading achievement scores are low and nonsignificant for Breaking the Code students. Correlations between the ER variable, response accuracy, and passage and word list reading scores are also higher in every case for the Eclectic Program, but none of the coefficients is significant at the .05 level.

Finally, criterion validity was addressed by intercorrelating workbook accuracy scores with the following ER variables: (a) spelling error rate, (b) correct spelling rate, (c) total spelling rate (correct + error), and (d) response accuracy (correct rate/total rate). At the end of the academic year, workbooks from all students in Breaking the Code were analyzed across 24 written review quizzes for the number of correct responses, spelling errors, and number of omissions. The workbook quizzes shared a common format, all containing the same number of items on sound spelling, word spelling, and sentence dictation. Table 5 shows four weak to moderate significant relationships between ER observation data and workbook quiz scores. Spelling error rate, and response accuracy correlate significantly with the number of workbook errors (r=.5 and .39; p<.05). Alone, the workbook scores of number of problems correct and number of problems omitted did not significantly correlate with direct observation variables. When combined in a multiple regression formula, however, number

correct and number omitted did correlate significantly with both spelling error rate and proportion of correct spelling responses (R=.57 and .45; p<.05).

DISCUSSION

The purpose of this research was to report treatment and criterion-related validities of two classroom observation instruments, a momentary time sample (MTS) and an event record (ER). Keys to the validation process were the simultaneous application of two instruments across two disparate language arts programs. Application across programs allowed treatment validation, and simultaneous use of two instruments permitted exploration of criterion validation, with engagement and response rates related to achievement outcomes. The two metrics of engaged time and response rates were intercorrelated and were correlated with achievement criteria. Within this multiple validation approach, treatment validity, which is the sensitivity of an instrument to program content and process, was considered the most crucial. An important need within special education is the ability to apply program-sensitive process measurement to formative evaluation of program components.

In this study, two widely divergent reading programs were concurrently observed with two observational instruments. One program, Breaking the Code, was highly structured, with scripted teacher instructions, tightly organized activities, and a multi-modality approach to synthetic phonics. In the other, the Eclectic Program, teacher instructions were less scripted, little standard sequencing of activities within or between lessons existed, and more emphasis was placed on oral and silent reading of longer passages.

Treatment Validity

Much of the student performance data generated by the observation systems reflected differences between the programs, which could be validated by class schedule, lesson plan, and curriculum. Great variability in Eclectic Program content and opportunities for students to actively respond within the Eclectic Program were reflected in performance means and standard deviations of scores compared across classrooms. These results should caution researchers about too readily generalizing student process-behavior findings across different types of classrooms or types of curricula. The ability to generalize findings apparently depends also on what variables are measured and what sampling techniques and metrics are used.

Interestingly, the molar engaged time variable, which included both active and passive responding, revealed few differences between programs, while larger differences, as much as 1 *SD*, occurred

Table 4. Correlations Between Both MTS and ER Measures and Combined Passage/Wordlist (P/WL) Reading Scores for Fall, Winter, and Spring Testing

	Fall P/WL		Winter P/WL		Spring P/WL	
	Engaged	Not Engaged	Engaged	Not Engaged	Engaged	Not Engaged
MTS Measures						
Breaking the Code (N=40)	-.12	+.10	-.13	+.13	-.22	+.22
Eclectic Program (N=13)	+.66**	-.61*	+.71**	-.69**	+.57*	-.52
Combined (N=53)	-.04	+.05	+.01	-.03	-.06	+.05
ER Measures	Proportion of correct oral reading responses					
Breaking the Code (N=40)	+.18		+.20		+.05	
Eclectic Program (N=13)	+.46		+.36		+.40	
Combined (N=53)	+.25		+.15		+.14	

* p<.05
** p<.01

Table 5. Single and Multiple Correlations Between ER Variables (Spelling Error Rate, Correct Spelling Rate, Spelling Rate Proportion of Correct Spelling Responses) and Workbook Quiz Scores

	Workbook Scores			
	No. Correct	No. Errors	No. Omissions	(Multiple R) No. Correct & No. Omissions
ER subscores				
Spelling error rate	-.08	.50*	-.30	.57*
Spelling correct rate	.16	-.03	-.09	.16
Spelling total rate	.15	-.03	-.13	.15
Response accuracy (Proportion spelling correct)	.27	.39*	-.06	.45*

* $p<.05$

on the two composite measures of percentage of *active* engagement. This highlights the utility of obtaining data on relatively narrow response classes and combining them only where appropriate. The smaller active engagement scores and response rates in the Breaking the Code program reflect that program's large group instruction, in which the teacher paces virtually all student responses. A potential effect of excessive teacher pacing appears to be a reduction in active engagement and rates of responding.

Content differences between the two programs were reflected accurately by the MTS and ER process measures. Students in the Breaking the Code Program were engaged in spelling or writing at a rate nearly three times that found in the Eclectic Program. In contrast, the magnitude of difference for reading engagement was equally strong; students in the Breaking the Code classes spent only one-fifth the amount of engaged time as those in Eclectic Program classes.

Criterion Validity

The three critical comparisons for establishing criterion validity (between instruments, between observations and test scores, and between observations and workbook scores) produced inconsistent results. The MTS and the ER instruments appeared to correlate well together across a number of behavioral categories. In correlations with student test achievement, however, the momentary time sample scores related more strongly than did the event record scores. In addition, the stronger correlations between MTS and reading test achievement scores were highly program specific; they were obtained only in the program that focused mainly on oral/silent reading. One could argue that a portion of the Breaking the Code program held less relevance for student reading achievement.

Validation of the ER Program by comparison with workbook scores also produced inconsistent results; while spelling error rate correlated significantly, spelling correct rate did not. Individual correlations diverged widely, depending upon how workbooks were scored: by number of errors, number of correct responses, number of omissions, or a combination of these. Again, the case is made for caution against over-generalizing from simple, unidimensional metrics, whether it be in applying direct observation instruments or evaluating permanent student products.

Future Research in the Classroom

In summary, validation findings are consistent for treatment validity across different criteria. That is, differences between the two programs are reflected clearly in the relationships between the student responding and other criteria of successful student functioning (i.e., reading and workbook performance). The relationships between classroom responding and these external criteria are more apparent when data from the two programs are analyzed separately. This finding affirms the importance of developing process-product relationships specific to particular curricula content, lesson organization, and classroom procedures; generalizability cannot be assumed.

As Berliner (1983) noted, the classroom can be described in terms of activity structures (Doyle, 1977). Some behavior is seen as functional for any particular activity, while other behavior is essentially non-functional or dysfunctional. The classroom can be conceived as comprising many activity structures, each of which creates differential

opportunities to respond in a particular manner. "An activity structure perspective helps to decompose many of the classroom activities into easily discernible sub-units so that the study of teaching and learning in classrooms can be accomplished with more precision" (Berliner, 1983, p. 3).

Findings from this research are consistent with Berliner's perspective. In most analyses, differences between programs were substantial, with data aggregation across reading programs diluting the program-specific relationships. This is interesting, given the replicability of the treatment descriptions across classrooms. Unfortunately, it is more typical in classroom observation research to find little or no specification of the treatment or program.

Within Breaking the Code, several activity structures were present, requiring different responses from the student. Most of the teacher's cues called for students to answer academic questions, write, copy, or spell. In fact, reading was seldom a functional response, especially at the level of reading sentence-length units or longer passages. From this perspective, it is not surprising that the gains made in reading were not significantly different from those achieved in the relatively unstructured Eclectic Program. Future research should ascertain opportunity for different types of active student responding within different activity structures.

Unfortunately, no pre- and postest writing and spelling measures were administered, leaving open to question the relationship between observational data and achievement in these areas. It is apparent, however, that if gains in spelling and writing occurred, they did not transfer to reading. This fact again reinforces the notion that observational instruments must be sensitive to the curriculum content and activity structures in the classroom. Further research is needed on the development of stable process measures at the molecular level which can be adapted systematically to different curricula, content, and activity structures.

REFERENCES

Anderson, L.W. (1980). *New directions for research on instruction and time-on-task.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Berliner, D. (1983). Developing conceptions of classroom environments: Some light on the T in classroom studies of ATI. *Educational Psychologist, 18*(1), 1-13.

Bloom, B. S. (1980). The new direction in educational research: Alterable variables. *Phi Delta Kappan, 61*, 382-385.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

CTB/McGraw Hill (1985). *California Achievement Tests.* Monterey, CA: CTB/McGraw Hill.

Early, M., Cooper, E. K., & Sonteusonio, N. (1979). *Bookmark reading program.* New York: Harcourt-Brace-Jovanovich.

Doyle, W. (1977). Paradigms for research on teacher effectiveness. In L. S. Shulman (Ed.), *Review of research in education* (Vol. 5). Itasca, IL: Peacock.

Fisher, C., Berliner, D., Filby, N., Marliave, R., Cahen, L., & Dishaw, M., & Moore, J. (1978). Teaching and learning in the elementary school: A summary of the beginning teacher evaluation study [Technical Report VIII-1 Beginning Teacher Evaluation Study]. San Francisco: Far West Laboratory for Educational Research and Development.

Frederick, W. C., & Walberg, H. J. (1980). Learning as a function of time. *Journal of Educational Research, 73*(4), 183-194.

Green, S. B., & Alverson, L. G. (1978). A comparison of indirect measures for long duration behaviors. *Journal of Applied Behavior Analysis, 11*, 530.

Greenwood, C., Delquadri, J., & Hall, V. (1984). Opportunity to respond and student academic performance. In W. Heward, T. Heron, D. Hill, and J. Trap-Porter (Eds.), *Focus on behavior analysis in education* (pp. 58-88). Columbus, OH: Charles E. Merrill.

Haertel, G.D., Walberg, H.J., & Weinstein, T. (1983). Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research, 53*(1), 75-91.

Harnishfeger, A., & Wiley, D. E. (1985). Origins of active learning time. In C. W. Fisher & D. C. Berliner (Eds.), *Perspectives on instructional time* (pp. 133-155). New York: Longman.

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

Hoge, R. D. (1985). The validity of direct observation measures of pupil classroom behavior. *Review of Educational Research, 55*(4), 469-483.

Jenkins, L., & Stein, M. (undated). *Washington classroom observation schedule—experimental version.* Seattle: University of Washington.

Karweit, N. (1983). *Time-on-task: A research review* (Report No. 332). Baltimore: Center for Social Organization of Schools, The Johns Hopkins University.

Karweit, N., & Slavin, R.E. (1981). Measurement and modeling choices in studies of time and learning. *American Educational Research Journal, 18,* 157-171.

Lebo, J. D., Hughs, A., Thomas, N., & Gurren, L. (1975). *Breaking the Code.* LaSalle, IL: Open Court.

Leinhardt, G. (1980). Modeling and measuring educational treatment in evaluation. *Review of Educational Research, 50,* 393-420.

Powell, J., Martindale, B., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis, 8,* 463-469.

Powell, J., & Rockinson, R. (1978). On the inability of interval time sampling to reflect frequency of occurrence data. *Journal of Applied Behavior Analysis, 11,* 531-532.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of frequency, interval, and time-sampling methods of data collection. *Journal of Applied Behavior Analysis, 9,* 501-508.

Rosenshine, B.V., & Berliner (1978). Academic engaged time. *British Journal of Teacher Education, 4,* 3-15.

Stanley, S. O., & Greenwood, C. R. (1981). *CISSAR: Code for instructional structure and student academic response: Observer's manual.* Kansas City, KS: Juniper Gardens Children's Project, Bureau of Child Research, University of Kansas.

Stuck, G.B. (1980). *Time-on-task and school achievement: Classroom intervention research.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Test, D., & Heward, W. L. (1984). Accuracy of momentary time sampling: A comparison of fixed- and variable-interval observation schedules. In W. Heward, T. Heron, D. Hill, & J. Trap-Porter (Eds.), *Focus on behavior analysis in education* (pp. 177-196). Columbus, OH: Merrill.

Thurlow, M., & Ysseldyke, J. (1983). *Instructional intervention research: An integrative summary of findings.* (Research Report # 143). University of Minnesota Institute for Research on Learning Disabilities.

APPENDIX A

15 SECOND MOMENTARY TIME SAMPLE

Page #:

15 Sec. Momentary Time Sample

Date of Observation _____ Teacher _____
Time In_____ School_____
Time Out_____ Observer_____

A=Active/ P=Passive O=Off task/ D=Disruptive

	Time	Student Code #	Academic Activities: Reading	Math	Spelling	Written Comp.	Handwriting	Science	Social Studies	Health	Phys. Ed.	Art	Music	Non-Academic Activities
1														
2														
3														
4														
5														
6														
7														
8														
9														
10														
11														
12														
13														
14														
15														
16														
17														
18														
19														

APPENDIX B

TWO MINUTE ACTIVE RESPONDING EVENT RECORD

Two Minute Active Responding Event Record

Page #:

Date:_____ School:_____ Teacher:_____

Observer:_____ Period:_____

Start time:_____ Stop time:_____ Total Observ. time:_____

L=letters, W=words, Ph=phrases
S=sentences, Pa=passages/paragraphs

I=correct X=error F=feedback on errors C=self-corrects

	I.D.#	Activity Structure I/G/C	Oppor. to Resp. Pas/Act	Yes/No	Silent Reading Start	Silent Reading	Silent Reading Stop	Oral Reading Start	Oral Reading	Oral Reading Stop	Code	Writing Start	Writing	Writing Stop	Copy Start	Copy	Copy Stop
	●	●		●													
1																	
2																	
3																	
4																	
5																	
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	
24																	
25																	
26																	
27																	
28																	
29																	
30																	