JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1984, 17, 521–538 NUMBER 4 (WINTER 1984)

TEACHER- VERSUS PEER-MEDIATED INSTRUCTION: AN ECOBEHAVIORAL ANALYSIS OF ACHIEVEMENT OUTCOMES

CHARLES R. GREENWOOD, GRANGER DINWIDDIE, BARBARA TERRY, LINDA WADE, SANDRA O. STANLEY, SUSAN THIBADEAU, AND JOSEPH C. DELQUADRI

UNIVERSITY OF KANSAS

In three experiments, we compared the effects of instructional arrangements that varied in: (a) teacher versus peer mediators, (b) methods used, (c) levels of student academic responding generated, and (d) content taught and tested. Instructional arrangements (i.e., tasks, structure, teacher position, teacher behavior) and students' levels of academic responding were measured by an observation system which served as an index of the independent variables. Students' accuracy on weekly spelling, arithmetic, and vocabulary tests and pre- and post-standardized achievement tests (Experiments 2 and 3 only) were the dependent variables. Results indicated that the classwide peer tutoring, compared to the teacher's procedure, produced more student academic responding and higher weekly test scores, regardless of treatment order or subject matter content (Experiment 1). The four lowest performing students in each class, in particular, benefited from peer tutoring, often performing as well as the other students. These findings were replicated in Experiments 2 and 3 wherein content taught/tested was also manipulated. Standardized test score gains were higher in those areas in which peer tutoring was used longest. Issues related to the functional analysis of instruction and achievement gain are discussed.

DESCRIPTORS: academic behavior, peer tutoring, classroom, elementary students

Only within the last 5 years have standardized achievement outcomes been causally attributed to teaching (Becker, 1977, 1978; Becker & Gersten, 1982; Brophy, 1979). Investigations of teaching practices that produce academic gains are beginning to yield exciting information on how to arrange lessons, how these arrangements affect student behavior, and in the long term, how they affect student achievement (Brophy, 1979). Recent educational research data have supported the notion that the opportunity to learn and students' academic engaged time are important correlates of achievement gain (Frederick & Walberg, 1980; Rosenshine, 1979; Stallings, 1975, 1977).

Experimental research on instruction has also indicated that procedures that create active student responding (e.g., Direct Instruction, peer tutoring, and Personalized Systems of Instruction [PSI]), have successfully affected academic outcomes (i.e., achievement, grades, test performance). For example, Carnine (1976) demonstrated that pacing in Direct Instruction (i.e., the rate of lesson presentation controlled by the teacher) produced high levels of student participation, correct responding, and student attention. Similarly, high levels of academic responding have been reported as characteristic of peer tutoring, due to the increased rate of task presentation, tutors calling for and prompting responses, and the use of immediate error correction (Delquadri, Greenwood, Stretton, & Hall,

This work was performed pursuant to Grant HD 0344 from the National Institute of Child Health and Human Development to Richard Schiefelbusch. Preparation of this manuscript was supported by Grants G007902271 and G007901332 from the Office of Special Education and Rehabilitation Services, U.S. Department of Education.

We are indebted to numerous individuals without whose help this work could not have been carried out: Gregorio Diaz, Becky Finney, and Janet Marquis of the Computer Applications Unit, Bureau of Child Research, for their help with software development and data analysis; Judith Carta and Frank Kohler, for their constructive comments on the manuscript; and Carmen Root and Alva Beasley, for their help preparing the manuscript.

Reprints may be obtained from Charles R. Greenwood, Juniper Gardens Children's Project, 1980 North 2nd Street, Kansas City, Kansas 66101.

1983; Dineen, Clark, & Risley, 1977; Trovato & Bucher, 1980). In PSI, study objectives, student access to course materials, frequent test taking, and frequent interactions with proctors have been related to increased study time and greater student engagement with course content (Born, Gledhill, & Davis, 1972; Born & Herbert, 1971; Kirigin, Braukmann, Atwater, & Wolf, 1982; Semb, Hopkins, & Hursh, 1973). The use of contingencies and rules has also been linked to gains in on-task and work behaviors during instruction and subsequent academic outcomes (Cobb & Hops, 1973; Greenwood et al., 1979; Hops & Cobb, 1974; Medland & Stachnik, 1972; Packard, 1970). The education literature also documents teacher behaviors that facilitate frequent student-teacher interactions and opportunities for students to respond as important features of effective instruction (Good & Grouws, 1977, 1979; Kounin & Gump, 1974; Leinhardt, Zigmond, & Cooley, 1981).

Of concern then, are recent reports that instruction in inner-city schools is often related to low levels of student engagement (Greenwood, Delquadri, & Hall, 1984; Hall, Delquadri, Greenwood, & Thurston, 1982; Stanley & Greenwood, 1983). These investigators reported that: (a) student academic behavior (i.e., writing, reading aloud, academic talk) occurred less often in inner-city 4th-grade classrooms than in suburban schools, and (b) the instructional methods used most in inner-city schools were those least associated with students' academic behavior and achievement.

Research to identify effective instructional methods, particularly in inner-city schools, requires information not only on the ecological variables arranged during instruction (e.g., materials, teacher behavior) and academic outcomes (i.e., achievement tests), but also on the rate and topography of student responding (Carnine, 1976; Sloane & Endo, 1981). Yet, few researchers have examined these three variables within the same experimental context (Bronfenbrenner, 1979; Brophy, 1979; Foster & Cone, 1980). In fact, in the majority of studies in both educational and applied behavior analysis research, assessment has been focused on outcome variables only, and context and behavioral treatment variables thought to produce changes in the outcome variables have been ignored (Brophy, 1979; Deitz & Baer, 1982; Dineen et al., 1977; Trovato & Bucher, 1980). This precludes the ability to explain why some procedures work and others do not.

In this study, we compared achievement outcomes using instructional procedures that differed in mediators, methods, contexts of instruction, and student behavior (i.e., teacher procedures versus classwide peer tutoring). We sought to account for and to analyze differences in: (a) ecological context variables, such as instructional materials used, grouping, teacher location and behavior, and (b) the quality and frequency of students' academic behavior during these conditions. Experiment 1 was designed to compare the achievement effects of teacher versus classwide peer tutoring methods with instructional time and order effects controlled. In Experiments 2 and 3, this analysis was replicated and extended by manipulating content taught/tested and instructional methods.

GENERAL METHOD

Participants and Settings

Five female teachers and 128 students (64 female, 64 male) participated. Classes were from two Title I elementary schools (3rd–6th grade) in inner-city Kansas City, Kansas. Student racial makeup was 106 black, 19 white, and 3 oriental.

On standardized achievement pretests, the students, as a group, were 5 months or more below grade level across subject areas. The four lowest students from each class were selected based on informal teacher nominations and pretest scores. These students were observed and assessed weekly. The other group in each class received weekly content tests but was not observed. The results of standardized test scores, summarized in Table 1, confirmed substantial differences between the low group and the other group in each class.

The three classes in Experiment 1 were taught in open-space classrooms (pods). The two classes in Experiments 2 and 3 were of traditional design

Table 1
Students' Preexperiment Achievement Levels

Experiment	Class	Grade	Groups	Spell		Math		Vocab	
				M	*SD*	*M*	*SD*	*M*	*SD*
1	1	4–5	Low ($n = 4$)	3.8	0.50	—[a]	—	—	—
			Other ($n = 21$)	5.7	0.71	—	—	—	—
1	2	3	Low ($n = 4$)	3.2	0.34	—	—	—	—
			Other ($n = 25$)	4.6	0.74	—	—	—	—
1	3	6	Low ($n = 4$)	—	—	—	—	2.7	0.33
			Other ($n = 30$)	—	—	—	—	4.8	1.47
2	4	3	Low ($n = 3$)	3.1	0.30	3.2	0.38	2.2	0.72
			Other ($n = 20$)	4.5	1.12	3.9	0.23	3.8	1.68
3	5	3	Low ($n = 3$)[b]	2.5	0.30	2.4	0.30	1.9	0.54
			Low ($n = 4$)	2.8	0.31	2.8	0.48	2.1	0.47
			Other ($n = 10$)	4.7	0.99	3.5	0.24	3.1	0.68

Note. Scores derived from the spelling and mathematics sections of the Wide Range Achievement Test and the vocabulary section of the Gates-MacGinitie Reading Test.

[a] Data were not obtained.

[b] Non-English-speaking refugee students.

(self-contained). The study took place during instructional periods (i.e., spelling and vocabulary—Experiment 1; spelling, arithmetic, and vocabulary—Experiments 2 and 3).

Measures

Three measures were used: (a) direct observation, (b) weekly subject matter tests, and (c) standardized achievement tests. The first served as an index of the independent variable, whereas the last two were indices of the dependent variable. The weekly tests were experimentally controlled and served as the primary dependent measure. The standardized tests served as a social validation measure (Kazdin, 1977).

Direct observation. Observations were conducted in 20-min instructional sessions by trained observers using the Code for Instructional Structure and Student Academic Response—CISSAR (Stanley & Greenwood, 1981, 1983). (Copies of the CISSAR Observation Manual may be obtained from Charles R. Greenwood.) The code was used to assess eight event categories, including five categories of instructional context and three categories of student behavior. These categories and codes used within each category are summarized in Table 2.

Using momentary time sampling (Powell, Martindale, & Kulp, 1975), observers coded the activity, task, and structure in the first 10-sec interval, followed by six intervals in which teacher position, teacher behavior, and student behavior were coded. Intervals were signaled by auditory electronic timers mounted on clipboards. This 1:6 sampling pattern continued throughout the observation.

Observers were community persons who had completed high school. Five observers served in Experiment 1; 10 in Experiments 2 and 3. Applicants were screened using: (a) the Snellen Visual Acuity Test (Anastasi, 1961, p. 368), (b) the Wide Range Achievement Test, Level II Reading and Math (Jastak & Jastak, 1978), and (c) personal interview.

Selected trainees learned to use the CISSAR system in a 15-day workshop, with 4 hours of training each day. After observers passed mastery exams on definitions, they were taught to use CISSAR coding forms and practiced coding role-played and videotaped classroom events. When observers produced three reliable records in calibration with the observer coordinator (above 80% agreement), coding was initiated in classrooms. After 1 week, all observers obtained agreement scores of at least 80% over all code categories in calibration with the

Table 2
CISSAR Categories, Descriptions, and Codes

Ecological categories	Number of codes	Description	Codes
Activity	12	Subject of instruction	Reading, mathematics, spelling, handwriting, language, science, social studies, arts/crafts, free time, business management, transition, can't tell
Task	8	Curriculum materials or the stimuli set by the teacher to occasion responding	Readers, workbook, worksheet, paper/pencil, listen to lecture, other media, teacher/student discussion, fetch/put away
Structure	3	Grouping and peer proximity during instruction	Entire group, small group, individual
Teacher position	6	Teacher's position relative to student observed	In front, among students, out of room, at desk, side, behind
Teacher behavior	5	Teacher's behavior relative to student observed	Teaching, no response, approval, disapproval, other talk
Student behavior categories			
Academic response	7	Specific, active response	Writing, reading aloud, reading silent, asking questions, answering questions, academic talk, academic game play
Task management	5	Prerequisite or enabling response	Attention, raise hand, look for materials, move, play appropriate
Competing (inappropriate responses)	7	Responses that compete or are incompatible with academic or task management behavior	Disrupt, look around, inappropriate (locale, task, play) talk nonacademic, self-stimulation
	53 total codes		

trainer and with each other, and were permitted to collect data for the study.

Observers recorded one low performing student in each class during each 20-min observation. Observations occurred Monday–Thursday each week during spelling, math, and vocabulary sessions. Testing occurred each Friday, and there was no instruction in spelling, math, or vocabulary. Each low student was randomly selected for observation once each week, producing one data record for each student per week. In Experiment 1 (Phase 1 only), this resulted in five observations per student over 5 weeks or 20 total observations over the four students. The same procedure was used in Experiments 2 and 3. The total number of observations in these two experiments ranged from 2 per student during baseline to 15 per student during classwide peer tutoring.

Agreement was checked across phases, classrooms, and instructional periods, an average of four times per week. Each check lasted a standard 14 min. To control for observer drift, observers were randomly assigned to pairs for each check (Johnson & Bolstad, 1973). Calibration checks with the trainer occurred on 10% of all checks. Reliability checks were computed in two ways. Percent interval agreement methods were used during the study. Following the study, a Pearson r analysis was made on the percent scores reported herein (Hartmann, 1977).

The interval agreement method provided information on the equivalence of observers' records during the study. Percent agreement scores [(no. of agreements/total coded intervals) × 100] were computed separately for the six major code areas (e.g., activities, tasks, etc., and overall). In Experiment 1, the range in agreement scores was from 98.5% for structures (SD = 3.0) to 79.8% for student behavior (SD = 5.8). The overall score was 88.2% (SD = 4.2). Similar values were obtained in Experiments 2 and 3. Agreement percentages ranged from 99.8% for structures (SD = 1.9) to 88.3% for student behaviors (SD = 3.3). The overall percentage averaged 92% (SD = 2.62).

The Pearson r analysis was applied to the separate CISSAR code percentage scores. Forty-eight paired observer checks were randomly drawn from all checks completed and analyzed. The mean r was .77 (SD = 0.24) over the 53 separate CISSAR codes.

CISSAR variables were summarized as percentages of intervals in which each variable was noted to occur. Because the major changes in context categories occurred in teachers' use of tasks, scores were reported only for this context category. Further, because a major interest was students' academic behavior, only the seven academic response codes (writing, academic game play, reading aloud, reading silent, academic talk, asking questions, and answering questions) and their composite score were computed (see Table 2). The student behavior codes dealing with task management (e.g., raising hand, looking for materials) and competing, inappropriate behavior (e.g., disrupt), were not reported. Prior research indicated that the CISSAR academic response codes, particularly writing, reading silent, reading aloud, and the composite, were significant positive correlates of achievement. The remaining codes were either not significant correlates or were negative correlates (Greenwood et al., 1984).

Friday content tests. Weekly tests were administered to evaluate students' mastery of content taught during the investigation. Prior to the study, we helped teachers design weekly lists in spelling, arithmetic, and vocabulary that reflected their instructional objectives. These lists were prepared in sufficient numbers (one set of three per week) to span the entire investigation. The lists served as the exact material to be taught and the content to be tested each week.

To ensure the content validity of the tests, items were drawn from: (a) each school's curriculum for that grade level, (b) basal texts, and (c) other materials teachers planned to use. Our informal review of the lists confirmed that some items were also on the standardized achievement tests. No items on the content tests had yet been taught and all were considered of sufficient difficulty to challenge all students. To control for systematic variations in difficulty across content tests and to distribute the difficulty levels equally, item lists were randomly selected each week. Carryover effects, due to items repeating across several lists, were eliminated by removing all duplications.

The classroom teacher administered three weekly tests (i.e., spelling, math, and vocabulary) each Friday. Each test took about 15 min to complete, and was then passed to a peer for scoring. The teacher read the correct answer aloud and students marked the items as either correct or incorrect. The percent correct score was reported for each student.

Reliability on scoring was assessed by comparing: (a) teacher versus student (Experiments 1, 2, and 3), and (b) investigator versus student records (Experiments 2 and 3), to establish and maintain students as reliable correctors. The percent agreement, defined as [(smaller score/larger score) × 100] was used. Checks were distributed over each area and phase of the study. Teacher-student reliability averaged 98% and ranged from 97% to 100% over classrooms. Experimenter-student checks averaged 95% and ranged from 79% to 100%.

Standardized achievement tests. Students took the Gates-MacGinitie Reaching Test, Level B or D, depending on grade level and ability (MacGinitie, 1978). The test manual reports Kuder-Richardson Formula 20 reliability figures above 0.90 for the vocabulary test. The Wide Range Achievement Tests, Level I spelling and mathe-

matics, were also used (Jastak & Jastak, 1978). The split-half reliabilities reported for these tests are 0.96 and 0.94, respectively.

Instructional Procedures/Arrangements

Procedures were based on prior research examining teacher-mediated procedures and classwide peer tutoring during instruction (Greenwood et al., 1984). These procedures were defined in terms of: (a) teacher versus peer mediation, (b) context arrangements, and students' behavior as measured by the CISSAR system, and (c) procedures implemented by teachers but not directly assessed. *Teacher-mediated procedures* were defined by use of tasks such as teacher-student discussion, media, readers, paper/pencil, and worksheets with the entire class. Teacher positions were: in front, among students, or at desk. Teacher behaviors were: teaching, no response to students, or disapproving behaviors. Students' behavior was largely passive attention or writing (composite academic responding was typically below 30% of session time) with students covering only a part of the week's material and emitting few individual academic responses. Occasionally groups competed, but neither point contingencies nor systematic error correction procedures were used. Little feedback was given for student performance.

In contrast, *classwide peer tutoring* was defined by: involving the entire group, peer mediation, and exclusive use of paper/pencil or worksheets for writing and practicing items. The teacher was among or to the side of students monitoring tutoring pairs. The teacher was typically engaged in answering students' questions, observing their performance, or awarding points for correct tutoring behavior. Tutors presented each item the tutee was to write and corrected errors. Students' composite academic behavior was high during tutoring (ranging from 45% to 75% of session time) and much more diverse than during the teacher-mediated procedure. Individual academic responses included writing, academic talk, reading aloud, and reading silently. All words/items were typically covered during a session and coverage of the entire list several times was dependent on each student's work rate. All students responded rapidly to the material. Competing teams and group and individual point contingencies were used. Group and individual point totals were also posted.

Baseline (2 to 4 weeks—Experiments 2 and 3 only). A baseline condition served to establish students' performance on the Friday tests in the absence of teaching. Teachers' instructional methods were not manipulated; however, the specific content tested these weeks was not the content taught during the teachers' lessons.

Teacher procedure (5 to 15 weeks). In this condition, teachers focused exclusively on the designated content in a 20-min instructional period. Teachers used a kitchen timer to signal the duration of this Monday through Thursday daily session each week. They were instructed to design a lesson to teach the content for that week, but to refrain from teaching this content at other times of the day. They were asked not to use peer tutoring but to use instead a procedure using teacher-student discussion, media, readers, and paper/pencil tasks, in rank order. Beyond these requirements, teachers were free to create this activity.

Classwide peer tutoring (3 to 15 weeks). We trained teachers in peer tutoring by having them read a short descriptive manual and roleplay and practice the procedures. We visited each classroom during the first week when the teacher presented the rules of the game, and students practiced tutoring. We provided feedback and praised teachers' correct use of the procedures.

Each Monday, students were randomly assigned to a tutoring partner for the entire week. Each pair was then randomly assigned by the teacher to a team. When an odd number of students was present (most often due to absences), a triad accommodated the extra student. During a 10-min tutoring session, the tutor dictated the item to the tutee who wrote a response. The tutor awarded points for a correct response or modeled the correct response if an error was made. Tutees earned 2 points for each correct response and 1 point for practicing the correct response three times following an error. The pair exchanged roles at the end of 10 min. At the end of this second period, each

Table 3
Phase 1 Comparison of Tasks Used and Associated Academic Responding Across Classrooms (Experiment 1)

Most used task summary	Intervals	Associated student academic composite %
Class 1—Peer tutoring in spelling		
Paper/pencil	58.9	70.1
Worksheet	17.6	76.0
All tasks/total observation	100.0	68.7
Class 2—Teacher procedure in spelling		
Paper/pencil	50.3	42.0
Readers	17.3	13.0
Teacher-student discussion	8.0	8.0
All tasks/total observation	100.0	29.4
Class 3—Peer tutoring in vocabulary		
Worksheets	68.9	42.6
Paper/pencil	17.7	89.9
All tasks/total observation	100.0	52.8

Note. This table portrays the two or three tasks most frequently used by teachers. Students' composite academic responding associated with each task is the conditional probability of academic responding [*p*(response/task type)] given the task specified or for the entire session (all tasks/total observation).

student reported his or her total points to the teacher, who summed and posted them for each team; she then announced the winning team.

Baseline (2 to 3 weeks—Experiment 3—math only). This reversal condition was used to recover the original Friday test baseline level. During the 20-min period, the teacher used her own teaching procedure to teach three lists of math facts. However, Friday tests covered three content lists not taught that week.

EXPERIMENT 1

Students

Students from three classrooms in an inner-city Title I school participated (see Table 1). All 88 students were black; 47% were male; 53% were female.

Table 4
Academic Responding Means by Class and Procedure (Experiment 1)

Academic response codes	Phase 1		
	Class 1 Peer tutoring	Class 2 Teacher procedure	Class 3 Peer tutoring
Writing	46.1	26.6	15.6
Academic game play	—	—	—[a]
Read aloud	2.5	—	9.2
Read silently	3.2	—	7.4
Talk academic	16.9	0.6	17.7
Answer question	—	—	1.1
Ask question	—	—	1.8
Academic response composite	68.7	29.4	52.8

[a] Occurrences of this behavior were not noted.

Design and Procedures

A counterbalanced reversal design with three phases was used across the three classrooms to compare instructional conditions and to control for order effects. In Class 1 and Class 2, conditions were implemented during daily 20-min spelling sessions; in Class 3, during 20-min vocabulary sessions. The conditions were classwide peer tutoring (A), the teacher-developed procedure (B), followed by peer tutoring (A). By random assignment, the order of conditions was ABA, in Classes 1 and 3 and BAB in Class 2.

RESULTS

Instructional Arrangements

During Phase 1, all teachers were observed teaching entire group lessons in front of and among students. Paper/pencil and worksheet tasks were also typical and common across all three classrooms. However, as Table 3 shows, only teacher 2 used reader and discussion tasks. These data confirmed that standard peer tutoring contexts (i.e., paper/pencil or worksheet tasks) were systematically used in Classes 1 and 3.

Students' academic responding. As shown in Table 4, composite academic response means for Classes 1 and 3 low students were both greater than the teacher-implemented procedure in Class

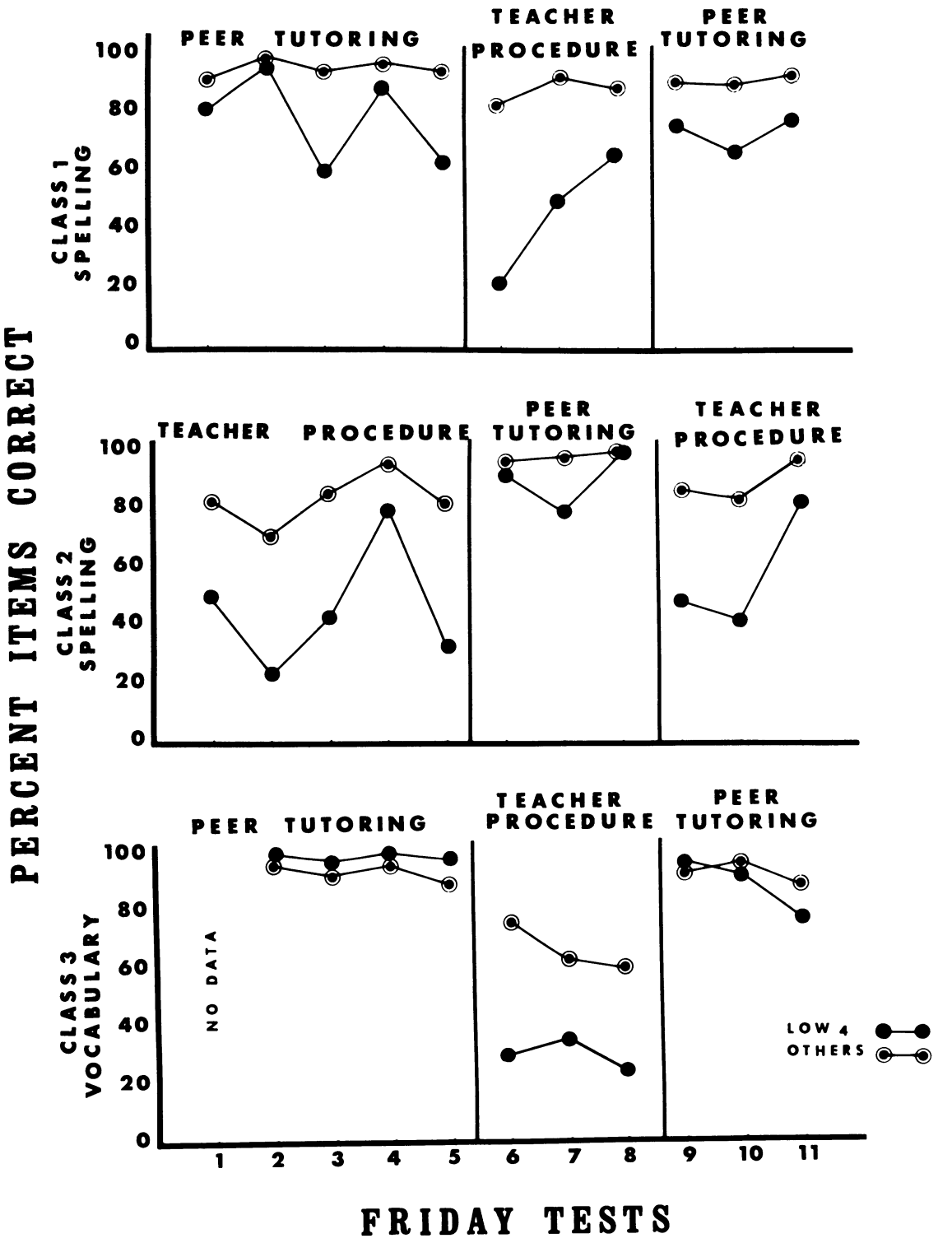

Figure 1. Percentage of items correct on Friday subject matter tests over three classrooms and two content areas (i.e., spelling and vocabulary) in Experiment 1.

2. Students in Classes 1 and 3 also exhibited greater variety in their academic responding.

Achievement outcome—weekly tests. As Figure 1 illustrates, Class 1 students scored higher on Friday tests during both tutoring phases than during the teacher procedure. Phase means were 81.0%, 46.0%, and 74.2% for the low group, and 96.5%, 90.3%, and 93.5% for the other group. This pattern was more evident in Class 3, which received the same order of conditions. The low group had phase means of 94.4%, 30.2%, and 91.0%; and the other students' means were 95.3%, 66.4%, and 93.1%. Higher Friday test scores were also produced during the peer tutoring condition in Class 2, even though the order of conditions was reversed. The phase means were 45.4%, 91.9%, and 53.5% for the low group; and 83.7%, 99.2%, and 89.2% for the other students.

Across the three classrooms, during 14 of the 18 peer tutoring weeks, low group students performed higher on Friday tests than they did during teacher procedure weeks. On two occasions, low students did equally as well, and on only two occasions did the teacher procedure result in better performance.

Discussion

Both low and other group students performed best on Friday tests under the classwide tutoring condition. Moreover, during tutoring the low group often performed as well as the other group students did. This was not the case during the teacher condition. These effects were demonstrated across three different classrooms, in different subject matter (Class 3), and regardless of treatment order. The data from Tables 3 and 4 suggest that: (a) the exclusive use of paper/pencil and worksheet tasks, and (b) the increased amount of academic responding, which was twice as high in the tutoring conditions, covaried directly with high Friday test gains. During tutoring, students engaged more frequently in writing task items, academic talk, reading aloud, and reading silently. In the teacher condition, the use of readers and teacher-student discussion for 25% of the time appeared to reduce students' academic responding (see Table 3, Class 2). These data confirm that during the first phase, teachers in each classroom and condition implemented the designated procedures. Because observers were available to collect data only during Phase 1, however, no conclusions can be made about subsequent phases.

The major purpose of Experiment 2 was to replicate both teacher and tutoring conditions under: (a) the continuous monitoring of context and student response variables, and (b) a baseline condition wherein teachers received no instructions concerning procedures to use, and students received no instruction on the content tested.

EXPERIMENT 2

Students

Of the 23 students in Class 4 who participated in Experiment 2, 11 were black, 12 were white; 13 were male, 10 were female. (See Table 1).

Design and Procedures

A multiple-baseline design across content areas was used. As in Experiment 1, the instructional context and student academic behaviors were directly observed and recorded. However, in Experiment 2, these observations were conducted throughout the study.

Three conditions were introduced in an ABC sequence. Baseline (A) reflected students' Friday test performance in the absence of any teaching of the content tested (other lists were taught that week) and without any effort to modify the teacher's instructional procedure. In the teacher-developed procedure (B), used in Experiment 1, teacher-student discussion was the primary means of teaching the content tested. Following this condition, classwide peer tutoring (C) was implemented to teach the content tested.

Results

Instructional Arrangements

As in Experiment 1, the teacher taught lessons to the entire group and was frequently observed in front of students or at the desk. The teacher mostly

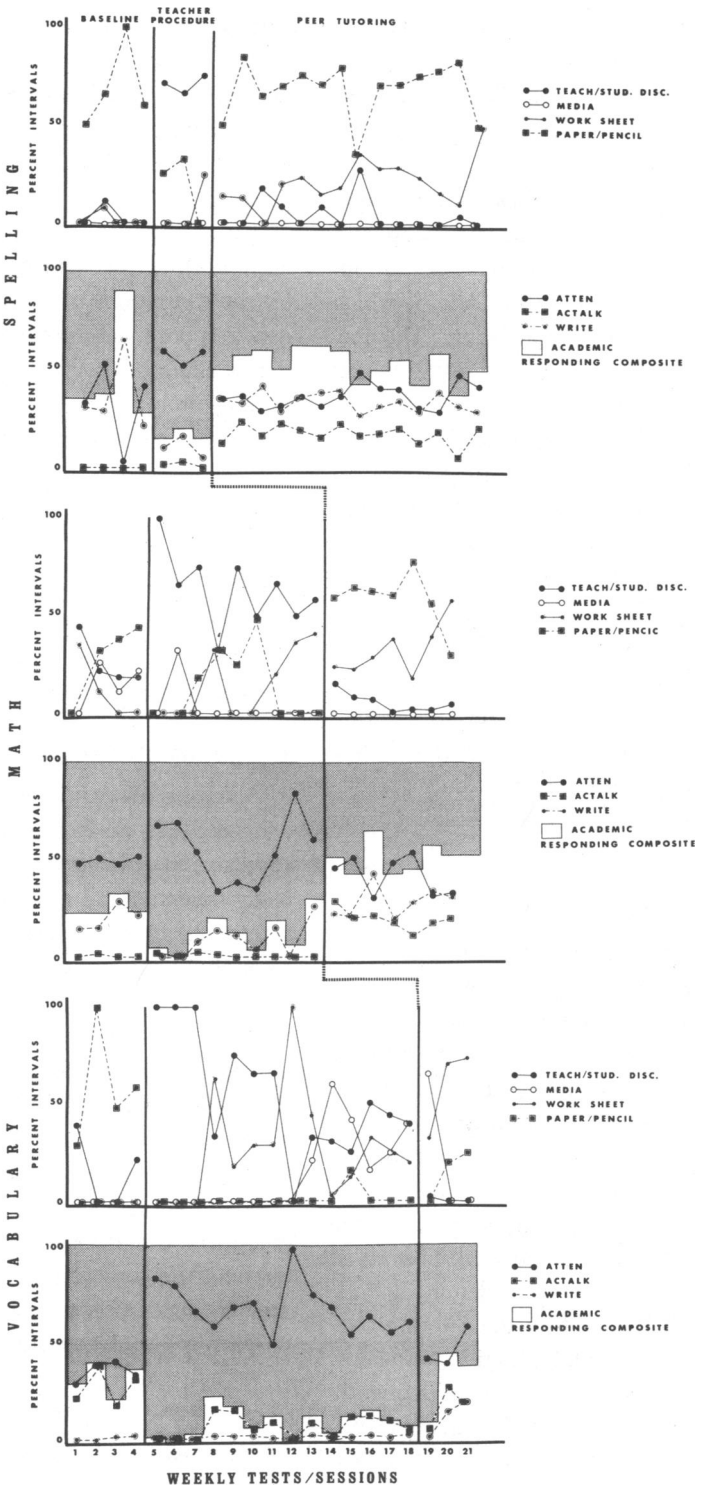

Figure 2. Percentage of intervals tasks were used (upper panels) and students' academic behavior occurred (lower panels) during weekly sessions for low group students in three content areas (i.e., spelling, mathematics, and vocabulary) (Experiment 2).

talked about the lesson or made no direct response to the students at all. Figure 2 (see upper panels for each content area) illustrates that during baseline, paper/pencil tasks were most often used. In math, however, tasks were diverse.

The change to the teacher-developed procedure in Phase 2 increased the use of discussion as the primary task. Over weeks, however, other tasks were increasingly used and discussion declined. In vocabulary, for example, worksheets or media were sometimes used more frequently than discussion. However, these data confirmed that the teacher procedure was implemented as in Experiment 1. The change to classwide peer tutoring in Phase 3 resulted in gains in the use of paper/pencil and worksheet tasks. Discussion was used much less often. This change replicated the use of peer tutoring in Experiment 1.

Students' academic responding. As shown in the lower panels of each content area in Figure 2, the students' academic response composite scores were highest during tutoring. During baseline, writing was the most frequent academic response in spelling and math, whereas academic talk was most frequent in vocabulary. In Phase 2, the academic response composite decreased dramatically, and passive attention increased in each area. During Phase 3, peer tutoring, the composite again increased. Specific gains were noted in writing and academic talk across the content areas, with a concomitant decline in the level of passive attention. These data indicate that students' performance patterns, characteristic of both procedures in the previous experiment, were also demonstrated in Experiment 2.

Achievement outcome—weekly tests. Students' weekly test means are shown in Figure 3. The weekly test means during the baseline phase were 54%, 44%, and 38% for the four lowest performing students in spelling, math, and vocabulary, respectively. Equivalent values for the other students in the class were 82%, 83%, and 57%. During Phase 2, the low students gained at least 20 percentage points or more to 75%, 73%, and 79%, respectively. The other group remained high but small gains were made by these students in each

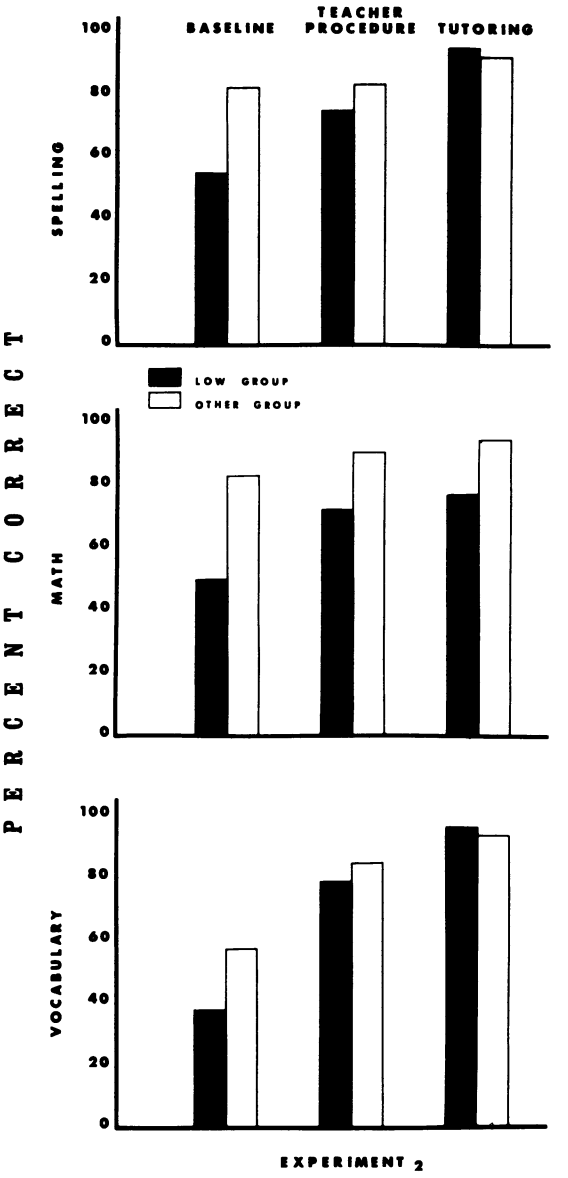

Figure 3. Mean percent correct on Friday tests by groups and content areas across conditions (Experiment 2).

area. Additional gains were made with the change to peer tutoring (Phase 3), particularly for the low students, who averaged 94%, 77%, and 96% during this phase in spelling, math, and vocabulary, respectively. The other group gained an additional 8%, 4%, and 9%.

Standardized tests. Students, as an entire group, made the largest gains in spelling and

mathematics. In spelling, students gained 5 months, increasing from 4.5 (pretest) to 5.0 (posttest), $t(11) = 3.53$, $p < .01$. (Degrees of freedom varied somewhat from the class N because only students with complete pretest and posttest data were used in the analysis. To reduce the effects of dependence in error components in these tests, due to mutual interaction between students, conservative tests were conducted. The usual $N - 1$ degrees of freedom were divided by 2 and the probability of t evaluated with conservative degrees of freedom [$(N - 1)/2$].) Students also gained 5 months in math, increasing from 3.8 to 4.3, $t(11) = 4.66$, $p < .001$. In vocabulary, students actually showed a decline, -0.2, scoring 4.0 (pretest) and 3.8 (posttest). These gains covaried with the length of time students used the tutoring program and maintained high levels of weekly mastery in each content area. Thus, 5-month gains in spelling and math were related to 14 and 7 weeks of tutoring, respectively; the loss in vocabulary was related to only 3 weeks in the tutoring program.

Discussion

Results of this study replicated those of Experiment 1 and again indicated that low performing students realized the largest gains in weekly achievement during tutoring. Although the other group students did not make such large gains, they did perform best during tutoring. For the low students particularly, it was demonstrated that both teacher and tutoring procedures were superior to a baseline condition in which students received no instruction on the content tested. It was also confirmed that: (a) teachers implemented the respective procedures correctly, and (b) largest gains in academic responding were evident during tutoring. Thus, systematic changes in context and student behavior patterns were again demonstrated to covary with accuracy on Friday tests.

The original purpose of Experiment 3 was to replicate the findings in Experiment 2. However, in the course of this study, the teacher appeared to deviate from standard tutoring procedures. The data in this study reflect the effect of procedural drift and demonstrate its impact on students' weekly achievement.

EXPERIMENT 3

Students

The 17 students in Class 5 (7 black, 7 white, 3 Laotian) participated in this experiment. As in Experiments 1 and 2, low and other groups were formed (see Table 1). Because three students did not speak English, a third group was formed. Ten of the students were male; 7 were female.

Design and Procedures

A combination multiple-baseline and ABA reversal design was used to replicate the effects of manipulating the content taught and prior instructional procedures. The multiple-baseline design was used in spelling and vocabulary baselines in an ABC format as in Experiment 2. The ABA design was used in math. During this reversal, as in the initial baseline, a list of content items was taught, but a different list was tested on Friday. The teacher was also instructed to use her original procedures to teach this list.

Results

Instructional Arrangements

The results in Experiment 3 were generally similar to those noted in the prior experiments. However, as illustrated in the upper panels of Figure 4, the tasks used by this teacher were much more variable and diverse. During baseline, combinations of paper/pencil, worksheets, teacher-student discussion, and media were used. In Phase 2, teacher-student discussion was not always the major task used and tasks varied across content areas. In fact, some weeks, worksheets, paper/pencil, or media tasks were used more frequently than teacher-student discussion in all three content areas. Thus, this teacher was not as compliant with the instructions as prior teachers in this phase.

In Phase 3, the introduction of tutoring in spelling and vocabulary was associated with increased use of paper/pencil and worksheet tasks. This training tended to narrow and stabilize the range

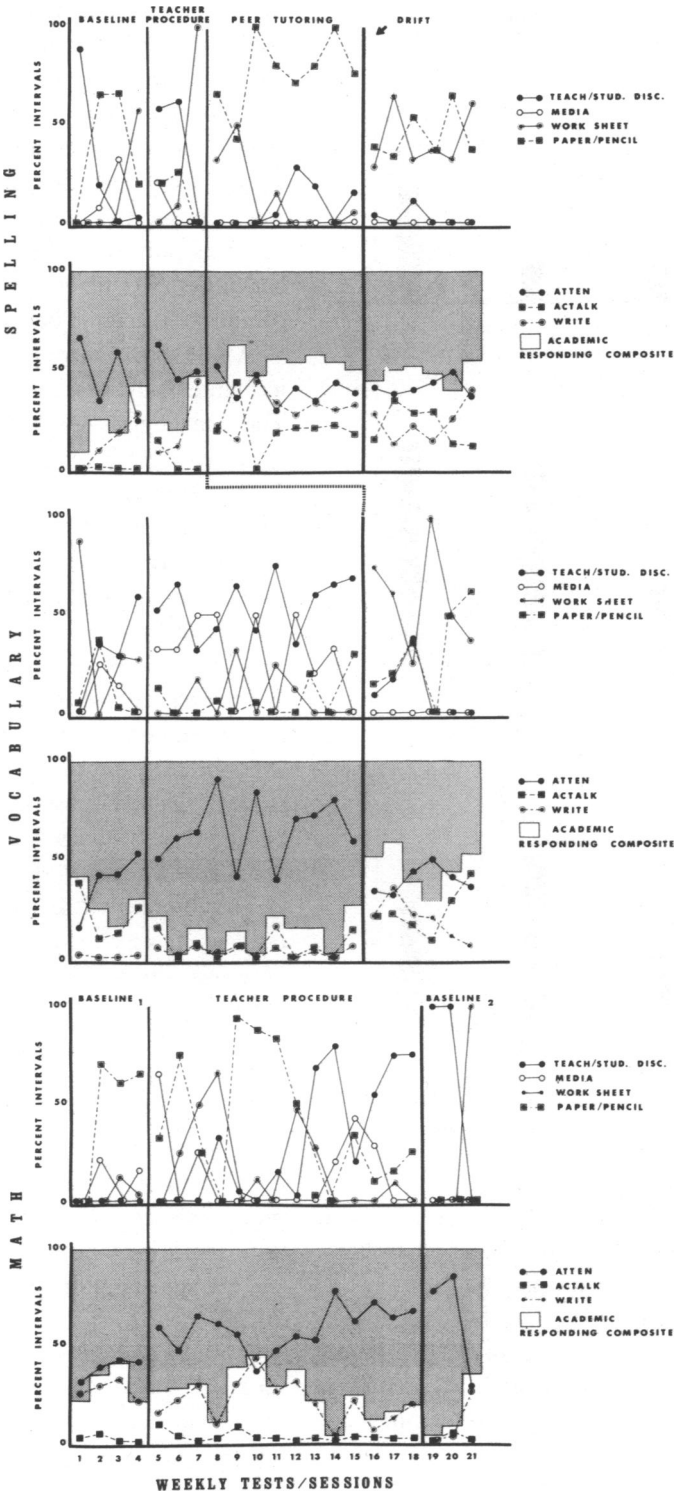

Figure 4. Percentage of intervals tasks were used (upper panels) and students' academic behavior occurred (lower panels) during weekly sessions for low group students (Experiment 3).

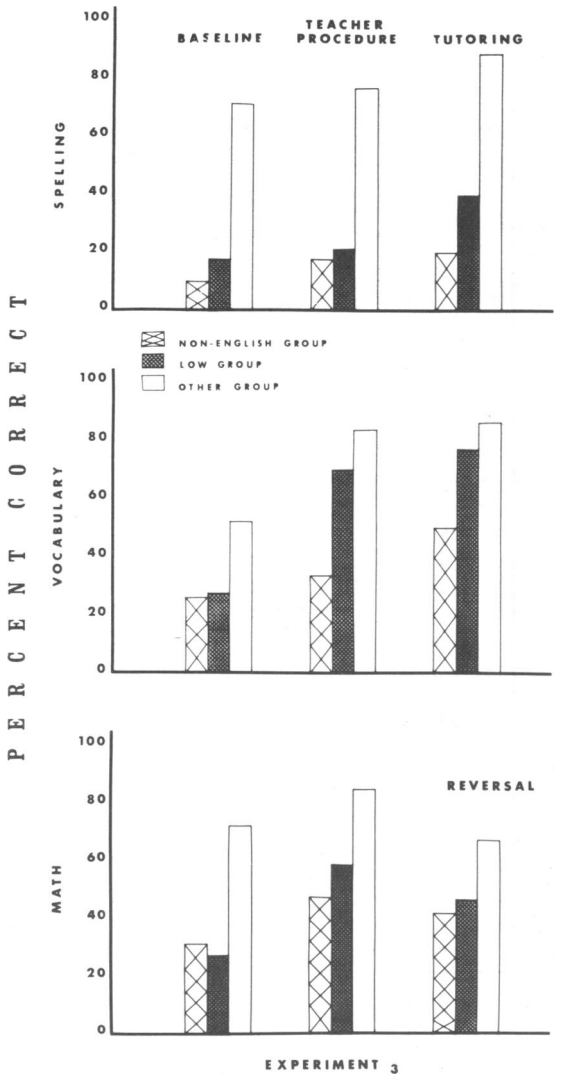

Figure 5. Mean percent correct on Friday tests by groups and content areas across conditions (Experiment 3).

of tasks used by this teacher to a greater extent than in prior phases. After Session 15 in spelling, however, it was apparent that the teacher had modified the tutoring procedure by using worksheets more than earlier in this phase (see Figure 4, spelling, upper panel). In math, the return to baseline resulted in the exclusive use of either discussion or worksheets.

Students' academic responding. As in the prior experiments, the students' academic response composite was highest during the peer tutoring condition (i.e., in spelling and vocabulary). Students' behavior, after drift in the tutoring procedure beginning Week 15, maintained the same overall level of composite academic responding but changed specific topography in spelling. That is, academic talk increased, and writing declined. This was opposite the pattern displayed by these two variables prior to Week 15. The reversal in math had little effect on students' already low levels of composite academic behavior.

Achievement outcome—weekly tests. As shown in Figure 5, the baseline means differed by groups. The non-English speaking and low groups averaged below 31% correct on tests and ranged from 11% to 31% across content areas. The other group's averages ranged from 53% in vocabulary to 72% in math. The teacher procedure (Phase 2) produced only minimal gains over baseline for groups in spelling, on the order of 4%–7%. In vocabulary and math, however, gains were more substantial and varied somewhat by group, ranging from 6% to 42% in vocabulary, and 13% to 32% in math, over the three groups. The non-English speaking students gained substantially less than the other two groups (6% in vocabulary) but gained somewhat more in math (16%).

The tutoring phase, as in the prior experiments, resulted in superior academic achievement gains. However, the non-English speaking students benefited least (See Figure 5). The low group gained 19% and 7% in spelling and vocabulary, beyond their gains in Phase 2. The other group made gains of 14% and 3%. The tutoring gains in spelling were temporary, however, and they declined when drift occurred. As indicated in Figure 6, the other group averaged 93.3% before, and 78.0% after, drift; the low group averaged 44.5% and 32.0%; and the non-English speaking group, 22.8% and 20.2%.

In math, the reversal produced declines in performance for all three groups as they were tested on content not taught these weeks. These declines ranged from −4% to 18% across groups compared to their Phase 2 performance. (See Figure 5.)

Standardized tests. Students made the largest

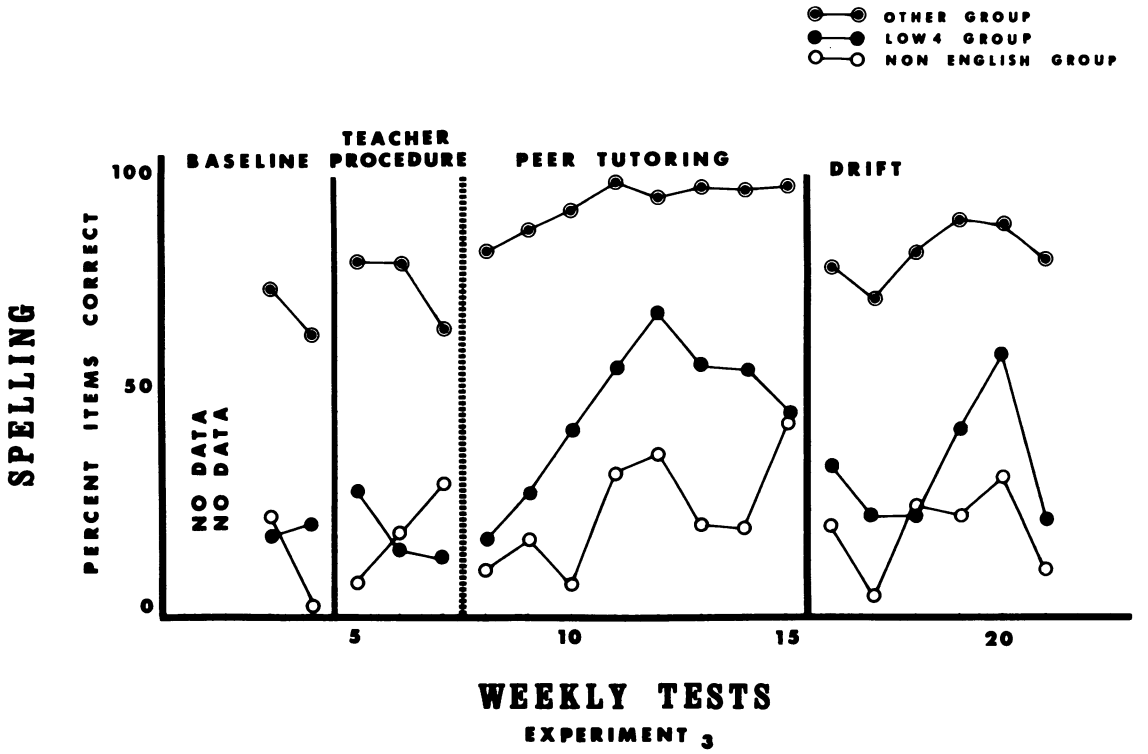

Figure 6. Weekly spelling test means by groups with an indication of procedural drift (Experiment 3).

gain in vocabulary, increasing from 2.9 (pretest) to 3.4 (posttest), t (7.5) = 3.78, $p < .01$. In this case, tutoring was in place for 6 weeks and weekly test scores were highest (at or above 80%). In math, students increased from 3.3 (pretest) to 3.7 (posttest). Spelling, where drift occurred, showed the smallest gain, from 4.1 to 4.2 months.

Discussion

This study demonstrated what can happen when the teacher alters or drifts from the standard intervention procedure. In this case, an unplanned change in tasks assigned to students and students' behavior during tutoring resulted in decreased spelling test performance. This procedural drift resulted in the students verbalizing the words while tutoring rather than writing them. One reason this affected spelling scores is that practice in writing the words is topographically similar to the required response on the Friday test. Also, increased oral spelling probably precluded tutors from correcting errors because tutors had been trained to focus correction on tutees' written answers. Thus, it appears that particular academic behaviors during instruction were more related to test performance than others and deserve further research.

Researchers must also examine how standard implementation can be maintained and how conditions may affect implementation over time. For example, we recently reported that teachers, even under optimal conditions, only implemented 76% to 80% of possible tutoring sessions over a 6-month period (Greenwood, Dinwiddie, & Delquadri, 1983). Method calibration, as discussed by Peterson, Homer, and Wonderlich (1982) and procedural reliability (Billingsley, White, & Munson, 1980) offer means for monitoring the quality of implementation. Systematic observation of the independent variable, allowed us to monitor the teacher's implementation. Such information could be used as teacher feedback on implementation (e.g., on such variables as tasks used and student

behavior occurring in relationship to implementation criteria), to limit intervention drift and undesirable teacher innovation. The implications of this problem for applied behavior analysis are profound because innovation with the independent variable may go undetected. Although procedural drift has been increasingly discussed in recent years as a quality control problem in mediators' use of behavioral procedures (Becker, 1977; Paine & Bellamy, 1982; Peterson et al., 1982), we are not aware of any investigations of this phenomenon. The use of an ecobehavioral assessment model as demonstrated in this investigation appears to offer the means for investigating such drift, the integrity of independent variables generally, and other factors related to outcome relationships within behavioral programs.

It was also evident in this study that the procedures were of limited effectiveness for non-English speaking students. It quickly became apparent that the 20-item list was too long for this group and the amount of practice per item under these conditions was insufficient. The amount of practice required by non-English speaking students to establish item mastery within peer tutoring needs to be examined in additional research.

GENERAL DISCUSSION

Results of this investigation confirmed with minor exception (the non-English speaking group in Experiment 3) that peer tutoring, compared to instructional procedures typically developed by teachers, produced superior weekly achievement effects for inner-city students. The experimental designs used in each experiment established these findings as functional relationships for the lowest students. Because of the measurement of both context and student behavior variables, it was also possible to conclude in relation to the lowest group students that: (a) procedures were implemented as designed, (b) procedures, in fact, differed in important context/behavior dimensions, and (c) these dimensions included differences in the quality and quantity of low students' academic responding during instruction. Levels of composite academic responding averaging above 45% were typical of peer tutoring and were most related to weekly achievement gain. Tutoring also produced a greater variety of academic responding (i.e., writing, reading aloud, reading silent, and academic talk), compared to teacher procedures. Thus, it was possible to compare and to replicate the achievement effects produced by different procedures and their associated student response patterns. These findings extend our prior knowledge of these concurrent effects, which were previously based on static designs (Greenwood et al., 1984) or on experimental studies of peer tutoring in which context and behavior data were not obtained (Delquadri et al., 1983). Moreover, these dimensions of the independent variables were helpful in explaining observed effects on the low students' weekly test performance.

Because these instructional methods differed on many additional variables, it was not possible in this investigation to attribute achievement gains to other than the molar package of procedures used in each method. These experiments were not intended to analyze the effects of separate components or the contribution of these components to achieve outcomes. Rather, the objective was to examine empirically the covariation of ecobehavioral and outcome changes.

Test gains were more dramatic for low group students because they did not encounter a ceiling on content lists and because peer tutoring enabled them to master this material. A pretest on words for each student was not completed, nor were lists individualized for each student, thus accounting for the generally higher baseline performance of the other group. Although individualized content lists would be desirable, the cost to teachers in preparing these for this study was too high. Procedures for individualizing content within the peer tutoring format warrant additional study and could enhance the weekly achievement of all students (cf. Slavin, Madden, & Leavey, 1984). There was also an indication that the largest gains in pretest–posttest grade equivalent scores occurred in content areas in which tutoring was used longest. This was consistent with other findings in which high levels

of mastery and content coverage have been related to standardized achievement test gains (cf. Brophy, 1979). In this study, however, we note only an instance of this covariation.

Researchers must continue to examine the functional relationships between the stimulus controls arranged by teachers, student academic behavior, and academic outcome variables. It appears that a complete technology of instruction will depend on studies that both assess and manipulate these variables. The identification of procedures that maximize students' academic behavior and that can be systematically implemented by teachers over the school year is of primary importance to the academic success of inner-city students.

REFERENCES

Anastasi, A. (1961). *Psychological testing,* 2nd edition. New York: MacMillan.

Becker, W. C. (1977). Teaching reading and language to the disadvantaged—What we have learned from field research. *Harvard Educational Review,* **47,** 518–543.

Becker, W. C. (1978). The national evaluation of Follow Through: Behavior-theory-based programs come out on top. *Education and Urban Society,* **10,** 431–458.

Becker, W. C., & Gersten, R. (1982). A follow-up of Follow Through: The later effects of the direct instruction model on children in fifth and sixth grades. *American Educational Research Journal,* **19,** 75–92.

Billingsley, F., White, O. R., & Munson, R. (1980). Procedural reliability: A rationale and an example. *Behavioral Assessment,* **2,** 229–242.

Born, D. G., Gledhill, S. M., & Davis, M. L. (1972). Examination performance in lecture-discussion and personalized instruction courses. *Journal of Applied Behavior Analysis,* **5,** 33–43.

Born, D. G., & Herbert, W. A. (1971). A further study of personalized instruction in large university classes. *Journal of Experimental Education,* **40,** 6–11.

Bronfenbrenner, U. (1979). Contexts of child rearing: Problems and prospects. *American Psychologist,* **34,** 84–89.

Brophy, J. E. (1979). Teacher behavior and its effects. *Journal of Educational Psychology,* **71,** 733–750.

Carnine, D. W. (1976). Effects of two teacher-presentation rates on off-task behavior, answering correctly, and participation. *Journal of Applied Behavior Analysis,* **9,** 199–206.

Cobb, J. A., & Hops, H. (1973). Effects of academic survival skill training on low achieving first graders. *Journal of Educational Research,* **67,** 108–113.

Deitz, S. M., & Baer, D. M. (1982, May). *Is technology a dirty word?* Paper presented at the meeting of the Association for Behavior Analysis, Milwaukee, WI.

Delquadri, J., Greenwood, C. R., Stretton, K., & Hall, R. V. (1983). The peer tutoring spelling game: A classroom procedure for increasing opportunity to respond and spelling performance. *Education and Treatment of Children,* **6,** 225–239.

Dineen, J. P., Clark, H. B., & Risley, T. R. (1977). Peer tutoring among elementary students: Educational benefits to the tutor. *Journal of Applied Behavior Analysis,* **10,** 231–238.

Foster, S. L., & Cone, J. D. (1980). Current issues in direct observation. *Behavioral Assessment,* **2,** 313–338.

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

Good, T. L., & Grouws, D. A. (1977). Teaching effects: A process-product study of fourth-grade mathematics classrooms. *Journal of Teacher Education,* **28,** 49–54.

Good, T. L., & Grouws, D. A. (1979). The Missouri mathematics effectiveness project: An experimental study in fourth-grade classrooms. *Journal of Educational Psychology,* **71,** 355–362.

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

Greenwood, C. R., Dinwiddie, G., & Delquadri, J. (1983, May). Evaluating teaching success: Relationships between student academic responding and criterion achievement test gains. In M. Monteiro & S. Paine (Co-Chair), *The pursuit of clarity: Current research in direct instruction.* Symposium presented at the meeting of the Association for Behavior Analysis, Milwaukee, WI.

Greenwood, C. R., Hops, H., Walker, H. M., Guild, J. J., Stokes, J., Young, K. R., Keleman, K. S., & Willardson, M. (1979). Standardized classroom management program: Social validation and replication studies in Utah and Oregon. *Journal of Applied Behavior Analysis,* **12,** 235–253.

Hartmann, D. P. (1977). Consideration in the choice of an interobserver reliability estimate. *Journal of Applied Behavior Analysis,* **10,** 103–116.

Hall, R. V., Delquadri, J., Greenwood, C. R., & Thurston, L. (1982). The importance of opportunity to respond in children's academic success. In E. B. Edgar, N. G. Haring, J. R. Jenkins, & C. G. Pious (Eds.), *Mentally handicapped children: Education and training* (pp. 107–140). Baltimore: University Park Press.

Hops, H., & Cobb, J. A. (1974). Initial investigations into academic survival skills training, direct instruction, and first-grade achievement. *Journal of Educational Psychology,* **66,** 548–553.

Jastak, J. F., & Jastak, S. R. (1978). *The Wide Range Achievement Test.* Wilmington, DE: Jastak Associates.

Johnson, S. M., & Bolstad, O. D. (1973). Methological issues in naturalistic observation: Some problems and

solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice* (pp. 7–68). Champaign, IL: Research Press.

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification,* **1,** 427–452.

Kirigin, K. A., Braukmann, C. J., Atwater, J. D., & Wolf, M. M. (1982). An evaluation of teaching-family (Achievement Place) group homes for juvenile offenders. *Journal of Applied Behavior Analysis,* **15,** 1–16.

Kounin, J. S., & Gump, P. V. (1974). Signal systems of lesson settings and the task-related behavior of preschool children. *Journal of Educational Psychology,* **66,** 554–562.

Leinhardt, G., Zigmond, N., & Cooley, W. W. (1981). Reading instruction and its effects. *American Educational Research Journal,* **18,** 343–361.

MacGinitie, W. H. (1978). *Gates-MacGinitie reading tests (2nd ed., level D).* Chicago, IL: Riverside Publishing.

Medland, M. B., & Stachnik, T. J. (1972). Good-behavior game: A replication and systematic analysis. *Journal of Applied Behavior Analysis,* **5,** 45–51.

Packard, R. G. (1970). The control of "classroom attention": A group contingency for complex behavior. *Journal of Applied Behavior Analysis,* **3,** 13–28.

Paine, S. C., & Bellamy, G. T. (1982). From innovation to standard practice: Developing and disseminating behavioral procedures. *The Behavior Analyst,* **5,** 29–44.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis,* **15,** 477–492.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis,* **8,** 463–470.

Rosenshine, B. (1979). Content, time, and direct instruction. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 28–56). Berkeley, CA.: McCutchan Publishing.

Semb, G., Hopkins, B. L., & Hursh, D. E. (1973). The effects of study questions and grades on student test performance in college class. *Journal of Applied Behavior Analysis,* **6,** 631–643.

Slavin, R. E., Madden, N. A., & Leavey, M. (1984). Effects of cooperative learning and individualized instruction on mainstreamed students. *Exceptional Children,* **50,** 434–443.

Sloane, H. N., & Endo, G. T. (1981). The need for behavioral research on three topics related to classroom management. *Behavioral Counseling Quarterly,* **1,** 46–58.

Stallings, J. (1975). Implementation and child effects of teaching practices in follow-through classrooms. *Monographs of the Society for Research in Child Development,* **40,** (7–8, Serial No. 163).

Stallings, J. (1977). How instructional processes relate to child outcomes. In G. D. Borich (Ed.), *The appraisal of teaching: Concepts and process* (pp. 104–113). Reading, MA: Addison-Wesley.

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.

Stanley, S. O., & Greenwood, C. R. (1983). Assessing opportunity to respond in classrrom environments through direct observation: How much opportunity to respond does the minority, disadvantaged student receive in school? *Exceptional Children,* **49,** 370–373.

Trovato, J., & Bucher, B. (1980). Peer tutoring with or without home-based reinforcement, for reading remediation. *Journal of Applied Behavior Analysis,* **13,** 129–141.

Received October 17, 1983
Final acceptance July 2, 1984