DOCUMENT RESUME

ED 311 071 TM 013 908

AUTHOR	Giuli, Charles; And Others
TITLE	Instructional Effectiveness: Desperately Seeking Indicators of Quality.
PUB DATE	87
NOTE	17p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; *Educational Quality; Elementary School Students; Elementary School Teachers; Grade 1; Grade 2; Grade 3; *Hawaiians; *Instructional Effectiveness; *Language Arts; Outcomes of Education; Predictor Variables; Primary Education; *Program Effectiveness; Program Evaluation; Teacher Effectiveness; Teaching Experience
IDENTIFIERS	Hawaii; *Kamehameha Elementary Education Program; Metropolitan Achievement Tests; Otis Lennon School Ability Test; *Quality Indicators

ABSTRACT

The classrooms of 40 teachers in the Kamehameha (Honolulu) Elementary Education Program (KEEP) in grades 1 through 3 were used to investigate the influence on achievement of 12 important instructional variables. The KEEP is an elementary-level language arts program designed for Hawaiian learners. The independent variables concerned were, as follows: quality (program strategies, teaching strategies, program experience, teaching experience, and a composite of eight variables); task-ability match (work completion); motivation (participation); and time (daily instruction, attention, on task when not receiving direct instruction, management, and homework). The achievement of the 40 classrooms, with the median number of students in each classroom being 22, 21, and 25 for the 1st, 2nd, and 3rd grades, respectively, was measured using the Metropolitan Achievement Tests (MAT). The expected MAT mean for each classroom was estimated from that class's mean ability as measured by the Otis-Lennon School Ability Test. The measure of achievement used in the analysis was the standardized residual of the MAT score from the estimated mean. Only the number of years as a program teacher was related to achievement. This result was found in 2 consecutive years (1985 and 1986), but was puzzling because none of the 12 variables expected to be critical for instructional effectiveness was associated with achievement. Possible explanations are considered. Because the amount of teaching experience with this instructional program was associated with achievement, the results provide encouraging evidence for the effectiveness of KEEP, although they do not support the effectiveness of current models of instruction. (SLD)

Instructional Effectiveness: Desperately Seeking Indicators of Quality

Charles Giuli, Mark Troy, Roderick Calkins

Kamehameha Schools/Bishop Estate
Center for Development of Early Education

The classrooms of 40 KEEP teachers in grades 1-3 were used to investigate the influence on achievement of 12 important instructional variables (e.g., time on task, amount of instruction, quality of instruction). Achievement was measured by a standardized comprehension test. Only number of years as a program teacher was related to achievement. This result seemed robust because it occurred in two consecutive years, but was puzzling because none of the 12 instructional variables expected to be critical for instructional effectiveness—and for which increasing teacher proficiency over time seemed likely—associated with achievement. Possible explanations for these outcomes were discussed. It was concluded that study results were encouraging evidence for effectiveness of KEEP but were discouraging for the effectiveness of current models of instruction. Further effort to identify the effective instructional agent(s) apparently associated with implementation of KEEP were suggested.

The effort to identify effective instruction has, for roughly two decades, been the most active line of research on teaching (Shulman, 1986) and has often been labeled process-product or teaching effectiveness research (Dunkin & Biddle, 1974; Mitzel, 1960). This paradigm of educational research is based on the premise that outcomes of learning (i.e., the products of education) are determined by what happens in classrooms (i.e., the process of education). Much recent educational research is founded on the assumption that variations in teaching practices are related to and, because of their logical priority, likely cause differences in learner outcomes. Although, in principle, all outcomes of schooling (e.g., learners' thoughts, feelings, and attitudes) are relevant to the process-product research paradigm, the outcome most often studied has been academic learning as measured by achievement tests (Brophy & Good, 1986; Shulman, 1986).

Process-product research typically has two shortcomings: (a) The instructional features studied are often so minute (i.e., decontextualized) their relevance for real classrooms can be questioned, and (b) the research conducted is often not based on theory. Many researchers have commented on the first shortcoming. Shulman (1986) points out the danger of synthesizing the micro-level results of process-product research into composite teaching styles which, in fact, may not have been studied as composites in the first place. Other researchers echo this sentiment:

> Researchers need to consider the whole picture of effectiveness and not necessarily separate the behaviors and analyze them as solely indicative of effective practice. (Stevens & Driscoll, 1986, p.10)

> Researchers will have to focus on the instructional unit rather than the lesson as their unit of analysis, and observe over several consecutive days rather than spread observations across the term. (Brophy & Good, 1986, p. 368)

> We risk finding that each variable is ineffective, failing to discover that combinations of the same variables are effective. (Slavin, 1986, p. 23)

Recently, Stallings and Krasavage (1986) addressed the problem of overreduction by including as part of their evaluation of program effectiveness high inference judgments of the appropriateness of the use of program components (i.e., whether a component was needed and, if needed, used effectively) rather than depending solely on low-inference ratings.

A second drawback of process-product research is the frequent absence of a theoretical basis for the research. Often a few variables of interest are used with no apparent guidance from theory. For instance, Bourke (1986) chose "a wide range of teaching practices" to investigate the relationship between class size and achievement. Such reliance on variables of convenience rather than those dictated by theory tends to make shallow interpretation of research results and to frustrate coherence of the research paradigm.

Description of Approach

In line with the process-product research paradigm, the research reported here attempted to demonstrate the effectiveness of major instructional components of an elementary-level language arts program designed for Hawaiian learners–the Kamehameha Elementary Education Program (KEEP) (Tharp, 1982).

In order to avoid the tendency of previous process-product research to decontextualize results, we used in the study high-inference variables; low-inference variables; variables that were global, synthetic, and at a macro level of analysis; and variables that were elemental, analytic, and at a micro level of analysis. Using two levels each for *Inference* (low inference, high inference) and *Globality* (micro, macro), we created a two-by-two matrix for classifying variables selected for this study (see Table 2). It was thereby possible to balance the variables used in this study with respect to level of inference and globality.

In order to address the second shortcoming of process-product research and to find theoretical guidance in the selection of variables for study, we reviewed the literature on models of educational effectiveness. Models of instructional effectiveness reviewed by Haertel, Walberg, and Weinstein (1983) (i.e., Bennett, 1978; Bloom, 1976; Bruner, 1966; Carroll, 1963; Gagne, 1977; & Walberg, 1980) include, as essential components of instruction, *amount of learning time, quality of instruction*, and *student motivation*. Although included in the foregoing models as one aspect of quality of instruction, *matching tasks to student abilities* is, in more recent formulations, generally treated as a separate component of instructional effectiveness (e.g., Beginning Teacher Evaluation Study, 1976; Cooley & Bickel, 1986; Slavin, 1986). Brophy, in fact, elevates matching task requirements and learner abilities to one of few "truly universal instructional principles" (Brophy, 1979, p.

735). We chose a model of instructional effectiveness propounded by Slavin (1986) as our archetype of classroom instruction because it is relatively recent and incorporates the four major elements of instruction that appear in all instructional models reviewed. Descriptions of Slavin's four instructional dimensions follow.

Quality of Instruction. Slavin defines *quality of instruction* as those procedures which make instruction comprehensible to students; such as, organization of material, clear specification of objectives, noting transitions, using examples, restating, reviewing, summarizing previously covered material, frequent assessment, and immediate feedback about performance.

This description of instructional quality covers a wide range of instructional behavior and, in so doing, is typical of the literature on instructional quality. For instance, Brophy & Good (1986) suggest three areas of instruction to which instructional quality pertains: structuring material for presentation, soliciting student responses, and responding to students. Suggesting that the criterion of instructional quality is the extent of learner attainment of integrated understanding and conceptual change, Brophy and Good go on to identify foci of instructional quality: teacher explanations, use of examples, questioning strategies, follow-up practice, and application activities. Walberg, Haertel, Pascarella, Junker, and Boulanger (1981) defined quality of instruction as "direct, didactic instruction" (p. 239), and others have given definitions such as clarity of instruction, provision for task difficulty, correct pacing, and appropriate feedback (Haertel, Walberg, & Weinstein, 1983). Such welter of definitions prompted Shulman (1986) to say "quality of instruction, so central to any research on teaching, remains frustratingly elusive" (p. 15).

Task-ability match. Slavin's second element of effective classroom organization is *matching student capacities to the requirements of learning tasks*. Even though it seems obvious that learning progresses best when tasks are neither too easy nor too difficult, and even though Brophy virtually canonizes this instructional principle, a clear statement of its operationalization across a wide range of learning tasks is lacking in the literature on instructional effectiveness. For independent work in basic skills, there seems to be a general consensus that high success rates are necessary and that this is especially true for slower learners (e.g., Berliner, 1979). Brophy and Good (1986) cite a desired success rate of between 90% and 100%, and emphasize that this success rate should not result from mindless repetition but should include effort and thought on the part of learners.

Although guidelines for obtaining task-ability matching are available for simple school tasks (e.g., addition), complex tasks (e.g., composition) remain little studied. So little knowledge exists in how to match abilities to all but the simplest of tasks, that some researchers note with concern the severe limitations thus imposed on attempts to apply this knowledge to higher level learning (Peterson, 1979). To the extent that educational studies rely on this truncated understanding, then, attempts to evaluate success of task-ability match may, in some contexts, be fruitless.

Student motivation. *Motivation* is the interest students have in accomplishing academic tasks. Teachers can use extrinsic incentives such as praise, high expectations, and group contingencies to enhance student motivation. Subtle ways teachers can try to produce intrinsic motivation are by illustrating the relevance of content to things of interest to students and by matching the difficulty of academic tasks to student ability. Tasks that are too easy or too difficult usually hold little intrinsic value for learners (Csikszentmihalyi, 1975).

Learning time. Learning requires *time*. Time must be allocated, and allocated time must be productive. Though somewhat obvious, starting with Carroll (1963) amount and

quality of learning time has been much studied and continues to garner research interest (e.g., Rosenshine, 1979).

Research Interests

Having obtained a set of research variables balanced with respect to critical instructional dimensions and level of analysis, we were interested in the following issues: Which instructional variables were most strongly related to achievement? Which of the four instructional dimensions had the most impact on achievement? Was official participation in the educational program under study related to achievement and, if so, was this effect totally mediated by the instructional variables selected for study?

In addition to these program-specific interests, we were also interested in the following general measurement issues associated with the process-product paradigm: Were macro-level variables better predictors of outcome than were micro-level variables? Were high-inference variables better than low-inference variables for predicting outcomes?

Predictions

We made the following predictions about these research interests:

1. Macro-level, high-inference variables would better predict achievement than would micro-level, low-inference variables.

2. *Number of years teaching the program* would be positively associated with achievement, and this association would be mediated entirely by the instructional variables in the model.

3. *Number of years teaching the program* would have a stronger effect on achievement than would total years as a teacher.

Method

Sample

KEEP students in grades 1-3 from the classrooms of 40 teachers from four public schools in Hawai'i constituted the sample for this study and were tested during the 1985-86 school year. The number of classrooms from each of grades 1-3 were 16, 12, and 12 respectively. The median percentiles on the MAT for grades 1-3 were 58, 44, and 42, respectively; the median OLSAT scores for grades 1-3 were 93, 92, and 94, respectively; and the median number of students per class in grades 1-3 was 22, 21, and 24, respectively.

Dependent variables. The Metropolitan Achievement Tests (MAT), 1978 edition Form KI (published by The Psychological Corporation), were used as the outcome measure. Choice of test levels conformed to those recommended by the publisher for regular spring testing, i.e. *Primary I* at first grade, *Primary II* at second grade, and *Elementary* at third grade. The raw score number correct on the MAT Reading Comprehension Test was the dependent variable.

The Otis-Lennon School Ability Test (OLSAT), 1979 edition Form R (published by The Psychological Corporation), was administered to all children in the fall of the school year. Choice of test levels conformed to the publisher's recommended levels. Thus,

Primary I was administered at first grade, and *Primary II* was administered at second and third grades. Children's performance was expressed in terms of the School Ability Index (SAI), a scaled score based on age norms. The SAI has a mean of 100 and a standard deviation of 16. It is analogous to an IQ score.

Independent variables. Table 1 describes the twelve independent variables used for this study and classifies them in terms of Slavin's four dimensions of instruction. In addition to nine instructional variables, such as *quality of execution of teaching strategies*, we included in our analysis two variables that might be related to instructional quality and, therefore, to student achievement: *Total years as a program teacher* and *total years of teaching at the current grade level*[1]. We also included a composite variable which was an amalgam of all seven instructional variables in our study, that is, which comprised all variables except *amount of teaching experience, amount of program experience, whether instruction occurred daily* and *amount of homework*. This composite provided a very global level of measurement, thereby helping to round out our measurement package, to provide a way to increase the power of the analysis (Bourke, 1986), and to avoid possible problems of multicollinearity (Pedhazur, 1982).

Table 1

Definitions of Variables and Classification by Instructional Dimension

Dimension/ Variable	Definition	Scale
Quality		
Program Strategies	Quality of execution of program teaching strategies	Low/Adeq/High
Teaching Strategies	Quality of general teaching strategies	Low/Adeq/High
Program Experience	Amount of experience as a program teacher	Years
Teaching Experience	Total amount of teaching experience, both program and nonprogram, at this grade	Years
Composite	All variables except program experience, teaching experience, homework, and daily instruction.	—

table continues

[1] We used *years at grade level* as a proxy for years of teaching experience, which was not known for all teachers.

Dimension/ Variable	Definition	Scale
Task-Ability Match		
Work Completion	Quality of independent work (includes errors)	Low/Adeq/High
Motivation		
Participation	Extent to which children participated during small group instruction	Low/Adeq/High
Time		
Daily Instruction	Whether direct instruction in Language Arts occurred each day in small groups for 20 minutes	Yes/No
Attention	Level of children's attentiveness during instruction	Low/Adeq/High
On Task	Whether students were on task at least 75% of time when not receiving direct instruction	Yes/No
Management	Quality of teacher's classroom management	Low/Adeq/High
Homework	Amount of homework	Number of minutes per week

Table 2 classifies the twelve independent variables used in this study in terms of their levels of inference and analysis. Variables were classified as high inference if their measurement was based on evaluators' expertise and, therefore, involved judgment of skill quality (such as how well teaching strategies were executed) rather than simply noting occurrences of clearly defined behaviors (such as whether instruction occurred daily). The latter were classified as low inference variables. Variables were classified as macro-level if they embodied global, complex instructional repertoires (such as number of years of teaching) rather than unitary, elemental instructional phenomena (such as student attention). The latter were classified as micro-level variables.

Table 2

Classification of Variables According to Level of Inference and Complexity

		INFERENCE	
		LOW	HIGH
COMPLEXITY	MACRO	Program Experience Teaching Experience	Teaching Strategies Composite
	MICRO	Daily Instruction Homework	Attention Management On Task Participation Program Strategies Work Completion

Procedure

Implementation of the KEEP program relies on highly trained consultants, each of whom typically works with between four and six teachers at a given school. Barring changes in assignment, the same consultant and teacher continue to interact from year to year with meetings and observations occurring frequently. Thus, consultants become well informed about their teachers' instructional performance.

As part of a previous study investigating the effects of variables similar to those used in this study (Au & Blake, 1984) participants in the current study had been trained to use an instrument similar to the one used in the present study until they obtained 95-100% reliability with criterion. During the year-long study which ensued, each observer had spent about 60 hours observing in classrooms. Thus, observers for the present study had extensive training and experience in using an observation protocol to assess instructional practice.

Training was provided to 12 consultants to ensure that variables (e.g., number of years in program) were recorded in an identical manner. Consultants were instructed to base their ratings on how the classes were currently functioning. The presence or absence of features were recorded based on direct observations and experiences the consultants had with the classrooms under review. All ratings were made during the first two weeks of May 1986. Particular attention was paid to ensure that all raters used similar behavioral anchors in rating those categories requiring inference (low, adequate, high). An "adequate" rating was defined as that level of teaching skill which connoted mastery of the given component within the KEEP program. A "low" rating was given

when teacher and consultant were still working on improving that skill to a mastery level. A "high" rating denoted exceptional performance; in fact, a useful concept raters were asked to keep in mind was that of "master teacher" being the level of skill necessary to warrant a "high" rating. Six of the seven high-inference variables were measured by a single item. The seventh, *quality of teaching strategies*, was measured by summing for each teacher ratings for the teaching strategies observed (e.g., language experience-text, directed-reading-thinking-strategies).

Derivation of residualized scores. Implementation of program components and instructional practice occurs at the classroom level. Therefore, the appropriate unit of analysis is the classroom mean. Using the mean raw score, however, presented problems of interpretation and comparability between grade levels. A different problem arises with the use of derived score such as a percentile rank or the normal curve equivalent based on the percentile rank. The MAT provides no group norms so percentile ranks of means are not obtainable. Moreover, evaluating the classroom means using individual norms is inappropriate (Angoff, 1954).

Heath, Jansen, Fortna, Bianchini, and Young (1967) used means on an ability test to evaluate means on an achievement test. They employed a simple regression procedure to estimate each group's expected achievement test mean based on the group's mean ability. Then they used the standardized difference (*z*) between the observed and expected means as a descriptive index of each group's achievement relative to it's ability and to the performance of other similar groups.

The same procedure was followed in this study. The ability test was the OLSAT administered to all pupils in Fall 1984. Pupils who entered school or transferred in after that time were given the OLSAT in Fall 1985. It was assumed that scores on the OLSAT are relatively stable so that scores from earlier and later administrations could be pooled.

Only students having both an MAT reading comprehension raw score and an OLSAT SAI were used in the analysis. At each grade level, classroom means were calculated on both variables, and a simple linear regression model was used to estimate the expected classroom mean achievement from classroom mean ability. The standardized residual (*z*) was taken as the standing of the observed classroom mean in the theoretical normal distribution. The classroom *z*-scores were then used as the dependent variables in subsequent analyses.

Analysis. Values for dichotomous variables were derived by assigning 1 to "no" responses and 2 to "yes" responses. Values for trichotomous variables were derived by assigning 1 to "low," 2 to "adequate," and 3 to "high" responses. Because the focus of this study was the effect on student achievement of teacher behavior, the appropriate unit of analysis was the classroom. The use of within-grade residualized scores made it possible to pool teachers in grades 1-3 as a single group for analysis. Residualized scores for teachers were regressed on the set of independent variables using the SAS STEPWISE regression procedure (SAS, 1985). This regression procedure is designed to find a final set of independent variables each of which has individual *F* values significant at $p \leq .15$.

Results

Means and standard deviations for independent variables appear in Table 3 along with *F* values for variables significant in the regression analysis at the .05 level. *Number of years as a program teacher* was the only statistically significant predictor of

achievement. The composite of seven instructional variables was, surprisingly, no better at predicting achievement than any of its constituents.

Table 3

Means, Standard Deviations, and Regression Analysis of Instructional Variables

Variable	Mean	*SD*	*F*	*p*
Program Experience	2.44	1.76	6.93	.01
Teaching Experience	5.72	5.29		
Program Strategies	2.31	.61		
Teaching Strategies	2.41	.58		
Work Completion	2.31	.57		
Participation	2.49	.56		
Daily Instruction	1.87	.34		
Attention	2.54	.55		
On Task	1.84	.36		
Management	2.24	.67		
Homework	109.87	68.15		
Composite	2.31	.46		

Note: $N = 40$

Discussion

These results partially confirmed our hypotheses for this study, though often in surprising ways. Our last hypothesis--that *number of years in the program* would be a stronger predictor of achievement than would *total number of teaching years*--was confirmed. However, *number of years as a program participant* was the only variable of all tested to reach significance. Therefore, its effect on achievement was not mediated by other instructional variables as we thought it would be. Even though *number of years as a program participant* was a macro-level variable, therefore partially confirming a third expectation that macro-level variables would be better predictors of achievement than would micro-level variables, we thought other variables in the model would also be related to achievement. None were.

Although not formally stated as a hypothesis for this study, one further result was somewhat surprising: The composite of seven instructional variables was no better at predicting achievement than were any of its constituents. This was surprising both because previous research (Bourke, 1986) had found a block of instructional variables to

be related to achievement when none of its members were, and because the power gained by combining the variance for several independent variables was not adequate in this study to cross the threshold of either practical or statistical significance.

These results seem especially robust because they replicated those of a previous study (Giuli, 1986), which employed the same procedures described here. This prior study used seven instructional variables. Four of these were identical to variables used in the present study, viz., *program experience, student attention during reading instruction, quality of execution of program strategies*, and *whether on-task rates were high*. For both studies, *amount of program experience* was the only statistically significant variable. Not only were results replicated for four identical variables in two studies, then, but expanding the set of independent variables in the second study from 7 to 11 variables did not result in increments in explained variance for achievement. These results are especially powerful because they derive from a within-group design. Compared to between-group designs, the total variance available for analysis in within-group designs is typically restricted. That is, any differences found must be detected among members of the same experimental group rather than between two different experimental groups such as "treatment" and "control."

We thought the results of the prior study were puzzling enough to warrant replication. The present study provides that replication. How is it, then, that none of the instructional variables–even when picked to represent both high and low inference as well as integrated and parceled variables–were themselves associated with achievement gain, while *number of years as a practicing teacher* in a program designed around these same instructional variables was relevant to student achievement gain? Additionally, neither *total years as a teacher*–regardless of number of years as a program member--nor a composite of various instructional variables was related to achievement. Clearly, then, there was something about being a program teacher that was associated with achievement but was not measured by our operationalization of any of the instructional elements of the program nor by teaching experience per se. Furthermore, shortcomings associated with previous research in this area had been addressed: Some independent variables were engineered to rely on expert judgment, and all variables were imbedded as a set in a living program.

The restriction of variance typically concomitant with within-group designs could explain the failure of any of the instructional variables to be associated with achievement. However, not only was a composite of instructional variables (which should result in increased power) not significant, but *number of years in the program* was significant. Assuming the program as a whole was a function of its parts, we expected any association of amount of program experience with achievement to be mediated by the effect of the several independent variables that we picked to represent individual program effects.

It is possible that these results are an artifact of a hypothetical selection mechanism: Good teachers stay with the program and bad teachers drop out, therefore confounding the effects of length of program membership with teacher quality. Though we cannot disprove this hypothesis, we can discount it. First, those who leave the program say they do so in large measure because they want to relocate to more convenient work places. (Giuli & Sloat, 1987). This reason for leaving seems justified and is applicable for both new and veteran program teachers. Second, assuming a major reason poor teachers would quit the program is because it require too much work on their part, it seems plausible that they would have more negative opinions about the program's value. We know, however, that current as well as former teachers of the program have strong

positive feelings about it (Giuli, 1985; Giuli & Sloat, 1987). Last, many of the teachers in this study were relatively new members of the program. Therefore, having not yet had much opportunity to decide to drop out, their number would have included both good and bad teachers, thus offsetting the hypothesized effects of selection due to teaching skill.

A second possible reason for the failure of the measures of teaching quality used in this study to be related to student achievement, while *amount of program experience* was, is a possible lack of reliability for these high-inference measures of instructional quality. For example, one rater's assessment of "mastery" might have been another rater's assessment of "needs more work." The association found here between *amount of program experience* and achievement is consistent with this thesis because *program experience* was a low-inference variable and most of the other variables were high-inference. Two reasons, however, argue against this interpretation of results: (a) If a relationship between instructional elements and achievement had been masked by unreliability of assessment, such masking would be less likely for low-inference variables. In fact, except for *program experience*, the low-inference variables in this study (i.e., *teaching experience, daily instruction,* and *homework*) were not related to achievement outcomes. (b) The raters used in this study had previously been trained to a rigorous standard of reliability for evaluating quality of program execution (Au and Blake, 1981), had used some of these identical assessment items in a study replication during the prior year, and had discussed the meaning of the anchors used for the high-inference items in this study just before classroom assessments were made. Furthermore, prior to entering service as practicing consultants, these raters received a year of standardized training in how to implement the program, a training based on acquiring the ability to make discriminations about program quality. Nevertheless, because the raters' assessments were not formally checked for reliability in this study, lack of reliability of assessment remains a possible reason for the failure of instructional variables to be associated with achievement here.

Another hypothesis for the apparent effect of *amount of program experience* on achievement is that the longer teachers stay in the program, the more they experience feedback associated with yearly standardized testing and, therefore, the more accurately they begin to match instruction to test objectives. All else equal, greater agreement between instructional and test content results in higher test scores. Although the data collected from this study were not designed to help sort out this hypothesis, some conjectures nevertheless seem pertinent. Like more experienced program teachers, inexperienced program teachers, prior to joining the program, had been tested with standardized tests and informed of results. This fact would argue against the hypothesis of differences between less and more experienced program teachers in the extent of match between instruction and test objectives. In favor of possible differences in test-instruction match, however, is the fact that the amount of testing was less and the kind of testing different for nonprogram teachers. Further, the test feedback received as nonprogram teachers was somewhat different than the feedback received by program teachers. Such differences in test type, frequency, and feedback may have been enough to cause a significantly greater convergence between test and instruction for experienced program teachers than for nonexperienced program teachers. In any case, the differential test-feedback loop hypothesized to account for the effect of program experience on achievement would have to be strong enough to override any effect derived from similar standardized testing for these teachers before they joined the program.

A fourth possibility for explaining why amount of program experience was related in this study to achievement is that the first year of program adoption may be detrimental to teaching effectiveness: The apparent effect of program experience, in this case, being

simply a return to normal levels of teaching efficiency by the second and following years of program teaching. This explanation for the effect of program experience on achievement is somewhat plausible because it is known that novice teachers in this program typically experience some anxiety, disorganization, and dependency during the first year of program teaching (Giuli, 1987). The effect on achievement of program experience, though, was strong and linear (rather than steplike as implied by this hypothesis), indicating that the effect permeated the range of values for program experience. This hypothesis of recovery from initial damage, then, did not seem especially tenable.

In finding amount of program experience but not specific instructional strategies to be relevant for student learning, these results seem to support the importance for understanding instruction of more aggregated, molar, and synthetic instructional variables. Possible psychometric reasons for this follow. Levels of variable inference and complexity relate to reliability and validity of measurement. As amount of inference required to measure a variable decreases, reliability of measurement tends to increase. Validity of the measurement of instructional effectiveness is likely to increase as the variables measured are more aggregated and thus complex. If so, then an optimum occurrence of both reliability and validity might be obtained with low inference, macro-level variables. The apparent success in capturing instructional quality of *amount of time with this program* may be due in part to the balance provided by this variable between the possibly high reliability stemming from it's low-inference measurement of number of years and from it's possibly high validity resulting from the complexity of meaning inherent in *program experience*.

If these results are to be taken seriously--and we think there is ample reason for doing so--length of time in this language arts program apparently brings with it a growing expertise on the part of teachers in how to produce learning (as reflected in standardized tests) in children. Maybe the teachers in this program learn how to produce in learners higher order learning strategies, which then mediate increased capacity for learning, the result of which is subsequently detected by achievement tests. At any rate, the conundrum needing solution is, What is it, if not instructional strategies or other program elements, that accounts for the effectiveness of these teachers?

Though we could not find it in this study, the mechanism by which the effect of being in the studied program is mediated is certainly knowable. To simply label this effect "instructional quality" without further understanding is to stop far short of the necessary goal. As Shulman said (1986):

> The continuing difficulty among both process-product investigators and the ALT [Academic Learning Time] proponents in dealing adequately with the issues of substantive instructional quality remains a nagging weakness in these research programs. In fact, it is the common flaw in all the extant programs of research on teaching. (p. 15)

Like others, we too have apparently not successfully identified the essence of instructional quality. Nevertheless, these results have shown important teaching effects to be associated with amount of teaching experience in a particular instructional program. Further, the critical variable detected in this study did not seem to be among those commonly studied in process-product research and did seem likely to be at a high level of instructional integration. Accordingly, further identification of the effective agent at work seems potentially helpful to the effort to improve instruction.

Because amount of teaching experience in this instructional program, but not teaching experience per se, was associated with learning, further study of this program seems advisable. Future research should be designed to focus on features of instruction which span days and even weeks such as: imparting higher order learning skills to students or, possibly, imparting to students a greater sense of confidence as a learner. In other words, future study of instructional effectiveness should broaden it's base from theories of instruction to theories of learning processes. Future efforts based on the program of instruction studied here, however, need to investigate carefully the possibility of *testing* and *damaged-recovery* artifacts, which were discussed as possible rival hypotheses for our results.

References

Angoff, W. H. (1984). *Scales, Norms, and Equivalent Scores.* Princeton, New Jersey: Educational Testing Service

Au, K. H., & Blake, K. M. (1984). *Implementation of the KEEP reading program, 1982-83; Results and methodological issues* (Tech. Rep. No. 112). Honolulu: Kamehameha Schools/Bishop Estate, Center for Development of Early Education.

Beginning Teacher Evaluation Study. (1976). *Proposal for Phase III-B of the Beginning Teacher Evaluation Study, July 1, 1976-June 30, 1977.* San Francisco: Far West Laboratory for Research and Development.

Bennett, S. N. (1978),. Recent research on teaching: A dream, a belief, and a model. *British Journal of Educational Psychology, 48,* 127-147.

Berliner, D. (1979). *Tempus educare.* In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching* (pp. 120-136). Berkeley, CA: McCutchan.

Bloom, B. S. (1976), *Human characteristics and school learning.* New York: McGraw-Hill.

Bourke, S. (1986). How smaller is better: Some relationships between class size, teaching practices, and student achievement. *American Educational Research Journal, 23,* 558-571.

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

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Whittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 328-374). New York: Macmillan.

Bruner, J. S. (1966), *Toward a theory of instruction.* New York: Norton.

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

Center for Development of Early Education (1981). *Essential features coding manual.* Honolulu: Kamehameha Schools/Bishop Estate.

Cooley, W., & Bickel, W. (1986). *Decision-oriented educational research.* Boston: Kluwer-Nijhoff Publishing.

Csikszentmihalyi, M. (1975), *Beyond boredom and anxiety.* San Francisco: Jossey-Bass.

Dunkin, M. J., & Biddle, B. J. (1974). *The study of teaching.* New York: Holt, Rinehart and Winston.

Gagne, R. M. (1977), *The conditions of learning* (3rd ed.). Chicago: Holt, Rinehart & Winston.

Giuli, C. (1985). *Results of a questionnaire about the KEEP program administered to all KEEP teachers* (Tech. Rep. No. 129). Honolulu: Kamehameha Schools/Bishop Estate, Center for Development of Early Education.

Giuli, C. (1986). *Summary of selected items from the Classroom Data Survey.* Unpublished manuscript. Kamehameha Schools/Bishop Estate, Center for Development of Early Education.

Giuli, C. (1987). *Four stages of KEEP implementation.* Technical Report in progress, Kamehameha Schools/Bishop Estate, Center for Development of Early Education.

Giuli, C., & Sloat, K. C. M. (1987). *KEEP teacher follow-up study* (Tech Rep. No. 142). Honolulu: Kamehameha Schools/Bishop Estate, Center for Development of Early Education.

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

Heath, R. W., Jansen, D. R., Fortna, R. O., Bianchini, J. C., & Young, M. R. (1967). The use of achievement and ability test averages. *Journal of Educational Measurement, 4,* 81-86.

Mitzel, H. E. (1960). Teacher effectiveness. In C. W. Harris (Ed.), *Encyclopedia of educational research* (3rd ed., pp. 1481-1486). New York: Macmillan.

Pedhazur, E. J. (198.). *Multiple regression in behavioral research* (2nd ed.). New York: CBS Publishing.

Peterson, P. L. (1979). Direct instruction reconsidered. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching.* Berkeley, CA: McCutchan.

Rosenshine, B. V. (1979). Content, time, and direct instruction. In P. L. Peterson & H. J. Walberg (Eds.), *Research on teaching.* Berkeley, CA: McCutchan.

SAS Institute Inc. (1985). SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: SAS Institute Inc.

Shulman, L. S. (1986). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. C. Whittrock (ed.), *Handbook of research on teaching* (3rd ed.) (pp. 3-37). New York: Macmillan.

Slavin, R. E. (April, 1936). *Quality, appropriateness, incentive, and time: Elements of effective instruction.* Paper presented at annual convention of the American Educational Research Association, San Francisco, CA.

Stallings, J., & Krasavage, E. (1986). *Peaks, valleys, and plateaus in program implementation: A longitudinal study of a Madeline Hunter Follow Through project.* Houston: University of Houston.

Stevens, D. D., & Driscoll, A. (1986). An intervention study of a staff development program on effective instructional strategies. *Journal of Classroom Interaction, 22* (1), 4-13.

Tharp, R. (1982). The effective instruction of comprehension: Results and description of the Kamehameha Early Education Program. *Reading Research Quarterly, 17,* 503-527.

Walberg, H. J. (1980), A psychological theory of educational productivity. In F. H. Farley & N. Gordon (Eds.), *Perspectives on educational psychology*. Chicago and Berkeley: National Society for the Study of Education and McCutchan Publishing.

Walberg, H. J., Haertel, G. D., Pascarella, E., Junker, L. K., & Boulanger, F. D. (1981). Probing a model of educational productivity in science with National Assessment samples of early adolescents. *American Educational Research Journal*, *18*, 233-249.