DOCUMENT RESUME

ED 123 254 TM 005 313

AUTHOR Pugh, Richard C.; Gliessman, David
TITLE Measuring the Effects of a Protocol Film Series: Instrument Development and Use.
PUB DATE [Apr 76]
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS *Criterion Referenced Tests: Educational Psychology; Graduate Students; Higher Education; *Protocol Materials; *Questioning Techniques; Sound Films; Statistical Analysis; Teacher Behavior; Teacher Education; Test Bias; *Test Construction; Test Reliability; *Test Validity
IDENTIFIERS *Categorizing Teacher Behavior; Concepts and Patterns in Teacher Pupil Interaction

ABSTRACT

The effects of the protocol film series, Concepts and Patterns in Teacher-Pupil Interaction, were measured by a film-based test after the test had undergone validation. The test, Categorizing Teacher Behavior, evidenced a high level of reliability and validity in a graduate level course in educational psychology. The test was judged valid by a jury of experts and showed a significant increase in mean test performance from a pretest situation to a posttest situation. Further, there was a significant decrease in variances from pretest to posttest. There was evidence that the use of the Concept and Pattern films in an instructional setting (a course in educational psychology) had a significant effect on concept acquisition as measured by the film-based test. The evidence suggests that there was a gain on all of the concepts (probing, informing, approving, disapproving, productive questioning, and reproductive questioning) from a time period prior to instruction to a time period after instruction. (Author)

Measuring the Effects of a Protocol Film Series: Instrument Development and Use

by

Richard C. Pugh and David Gliessman

Indiana University

Annual Meeting
American Educational Research Association
San Francisco
April, 1976

Measuring the Effects of a Protocol Film Series: Instrument Development and Use

Broadly speaking, protocol materials in teacher education are designed to be used in the acquisition of concepts referring to behavioral events, particularly those occurring in classroom settings. The acquisition of such concepts, which may refer to behaviors in either the pedagogical or subject matter areas of instruction, should in turn lead to a greater degree of interpretive competence. In defining the need for materials to accomplish these ends, Smith (1969) hypothesized that such interpretive competence provided the conceptual background necessary for the development of skills in teaching.

At an operational level, concept acquisition implies the development of skillfulness in discriminating and categorizing ongoing behaviors in terms of a more or less complex set of concepts or categories. Practically speaking, this means that the use of protocol materials designed to "teach" a prescribed set of concepts should result in a high degree of accuracy in selecting and "sorting" the observed behaviors into those categories. In short, the minimal performance criterion for learning from protocol materials is skillfulness in categorizing complex behaviors in terms of specified concepts.

The evaluation task for the developers of any set of protocol materials, then, becomes two-fold: first, to construct and validate an instrument designed specifically to measure acquisition of those concepts upon which

that set of protocol materials is based; second, to evaluate the effects of training based upon the use of that set of protocol materials by means of this instrument. The present paper addresses these two evaluation tasks in that order. The most direct outcome of the evaluation studies reported in this paper is empirical evidence on the effectiveness of a set of protocol materials in producing gains in acquisition of a specified set of concepts. The more general contributions of this evaluation study are (1.) the design of a practical and objective format for assessing concept acquisition by means of a film-based (and thus, "observation-based") test and (2.) the development of a general strategy for assessing "mastery" in such learning tasks as concept acquisition. This strategy was, in fact, developed to resolve a major measurement problem posed in this evaluation study: the validation of a criterion-referenced instrument (Brown and Pugh, 1975).

Protocol Film Series

The protocol films that are the subject of this evaluation study are those in the Concepts and Patterns in Teacher-Pupil Interaction series developed through the Indiana University Protocol Materials Project. The purpose of the nine films in this series to to define, exemplify, and provide practice in the interpretive use of a specified set of concepts referring to teacher behavior in a classroom interactive setting. Six categories of teacher behavior were selected from the empirical literature on classroom behavior and adapted to the specific instructional purposes of this film series. Together with their definitions, they are:

REPRODUCTIVE QUESTION - a teacher question intended to directly elicit the recall of content specifically learned as part of a course or topic of study. In response to such a question, the student is expected to accurately reproduce such content or to recognize when it is accurately reproduced by someone else. Typical student responses are repetition, restatement or recognition of content.

PRODUCTIVE QUESTION - a teacher question that is intended to encourage the production of idess or combinations of ideas as opposed to simply the reproduction of specifically learned content. A student response to such a question may reflect the recall of specifically learned content but that content is used in such processes as interpretation, application, and evaluation.

PROBING - a teacher reaction in the form of a question or implied question that pursues some aspect of the substantive content of a preceding student response. Such probes typically seek further description, clsrification, explanation, or extension of that substantive content. By "preceding response" is meant any preceding response including, but not limited to, the student response that immediately precedes a teacher reaction. By "substantive content" is meant the formal content of classroom discussion as opposed to such procedural content as assignment making, the order of discussion, or disciplinary matters.

INFORMING - a teacher reaction in which information is introduced that is related to some aspect of the substantive content of a preceding student response. Such a reaction is often intended to produce some modification in the substantive content of that student response. By "preceding response" is meant any preceding response including, but not limited to, the student response that immediately precedes a teacher reaction. By "substantive content" is meant the formal content of classroom discussion as opposed to such procedural content as assignment making, the order of discussion, or disciplinary matters.

APPROVING - a verbal and/or nonverbal teacher reaction that is intended to encourage, or might reasonably be expected to encourage continued student responding or a continuation of student behavior.

DISAPPROVING - a verbal and/or nonverbal teacher reaction that is intended to discourage, or might reasonably be expected to discourage continued student responding or a continuation of student behavior.

As indicated above, the specific behaviors referred to by these concepts are commonly described and referenced in the literature on teacher behavior (see, for example, Dunkin and Biddle, 1974). In fact, it is in

part this commonality of use and the existence of some related empirical literature that led to the selection of these concepts. For purposes of clear definition and portrayal, these concepts were arranged in the following pairs: approving and disapproving, reproductive questioning and productive questioning, probing and informing. Minimally, use of the Concepts and Patterns series in instruction should result in skillfulness in using these concepts as interpretive categories. It is also entirely plausible to hypothesize that effective use of the series might also result in acquisition of the skills referenced by these concepts. Although, that hypothesis was not tested in the studies reported in this paper, some aspects of it are currently being investigated by the authors.

The films in the series are 16mm. motion pictures in color with synchronous sound. Each film is quite brief, ranging from five to eleven minutes in length. The series is arranged in two subsets: three Concept films, each of which introduces, defines and exemplifies a pair of concepts; six Pattern films, each portraying classroom behavior that can be interpreted in terms of varying combinations of these concepts. The series is designed to be used in conventional instructor-led discussion or presentation. The individual films may be used in a variety of sequences although the most common is probably that of a Concept film followed by one or more relevant Pattern films.

Test Validation

Test Development and Characteristics

The development of the filmed test, entitled Categorizing Teacher Behavior, was central to the evaluation plan for the entire film series.

The "logic" of developing an entirely film-based protocol series led quite reasonably to the specification that the principal evaluation instrument also be film-based. It was felt that the task of categorizing behavior portrayed on film would most closely approximate the interpretation of behavior in actual classrooms.

The test consists of 30 brief classroom episodes each less than one minute in length. These episodes were selected from "out-takes" from the films in the series (that is, from footage not included in the Concept and Pattern films themselves). As a consequence, the classroom settings, subject matter context, teachers and pupils are the same as in the Concept and Pattern films. However, the specific episodes appearing in the test film do not appear in the other films of the series.

Certain facts about the use of film as a medium in test development should be noted. Although all episodes for the test were drawn from scenes recorded on film and although the test is distributed as a film, the filmed footage was transferred to videotape for purposes of test development and preliminary trial. Consequently, in terms of numbers of subjects, most of the data obtained during validation was based on a videotaped representation of the test. The use of videotaped preliminary versions for the gathering of data is easily explained. As with printed tests, film-based tests must undergo a number of revisions, and as a medium, film is both expensive and difficult to revise. Videotape transfers, on the other hand, are both economical and convenient to modify. The authors' expectation was that since the changes introduced by this shift in medium were primarily

technical (for example, absence or presence of color and relative sharpness of image), the magnitude and pattern of quantitative results would not be substantially different for the two media forms. In fact, a comparison of the results obtained for the videotape and final film forms of the test confirms this expectation.

The test is divided into three parts, providing for all possible combinations of the pairs of concepts. Part I requires that the teacher behavior in each episode be categorized in terms of probing, informing, approving and disapproving. In Part II, the concepts involved are productive questioning, reproductive questioning, approving and disapproving. In Part III, the concepts are probing, informing, productive questioning and reproductive questioning. Ten episodes are included in each part; within certain practical limits imposed by editing possibilities, these episodes are representative of the content of the Pattern films themselves. Separate test items for each of four concepts are presented for each episode. In the case of each item, a "yes" or "no" option is provided for the question of whether or not a specified concept is illustrated in the episode. As a consequence, a total of 120 items (20 items for each concept) are contained in the test.

During development, the test was subjected to scrutiny by six members of a panel each of whom took the test independently and then deliberated together until unanimous agreement was reached on the correct response to each item. Only items for which unanimous agreement was reached were retained. From the set of items and episodes emerging from this refinement, episodes were selected for each of the major parts of the test.

Each episode is presented twice with a delay of from two to four seconds between presentations; the purpose of this repeated presentation is to minimize dependence on recall of the specific episode as a factor in the measurement of concept acquisition. Following the second presentation of each episode, the examinee is given a fifteen second period in which to respond by recording his answer on a separate answer sheet. The total time for the test (including repeated episodes, delays between episodes, and delays for responding) is approximately 35 minutes.

The restriction of items for each episode to four rather than six concepts was planned to achieve a balance between "stimulus simplicity" and the experience of "stimulus overload." The resulting "response task" was thought to be at an optimum level of difficulty. Overall, the examinee is required to indicate the presence or absence of each concept 20 times. The limitation of each form to a total of 120 items provides for a reasonably short test administration time as well as a reasonable opportunity to demonstrate concept acquisition.

In Table 1 is presented the frequency of occurrence of each concept for each part of the test. There are 20 episodes within the total test in which each concept might have occurred and 10 episodes within each part of the test in which each concept might have occurred. By comparing actual frequency of occurrence to the number of episodes (in each of which a given concept <u>might</u> occur) one can obtain the ratio of instance to non-instance items for each part and concept.

Insert Table 1 about here

In terms of a potential source of response bias, an ideal ratio of instance to non-instance items might have been 50:50. However, the method of item selection precluded obtaining an ideal ratio. In the case of this test, the representativeness of the items and episodes was considered to be very important even though the ratios of instance to non-instance correct answers for selected parts of the test deviated from the ideal. A "yes" or a "no" response on a given item has about the same probability of being correct for "approving" (12 yes to 8 no correct answers), "productive questioning" (10 yes to 10 no correct answers), and "disapproving" (9 yes to 11 no correct answers). For "probing," the ratio is 8 yes to 12 no; for "informing" it is 7 yes to 13 no; for "reproductive questioning" it is 6 yes to 14 no.

In Table 2 is presented the number of concepts instanced in each episode. From the Table, it can be seen that the typical number of concepts instanced for each episode is one or two. All episodes contain instances of one, two or three concepts.

Insert Table 2 about here

Definition of Strategy to Assess Test Characteristics

As a test, Categorizing Teacher Behavior is criterion-referenced because the content was derived from the meaning of the six underlying concepts (probing, informing, approving, disapproving, reproductive questioning and productive questioning) and a mastery performance level was sought. Since there are special problems in the appropriateness

of some of the conventional test characteristic indices, a strategy developed by Brown and Pugh (1975) was adopted to portray the characteristics of the test.

The strategy devised to assess the test characteristics assumed that the objective of the film-based instructional treatment was to bring the learner from a pre-treatment state below the criterion performance level to criterion level. It also assumed that if Categorizing Teacher Behavior detects this change of learning state consistently over the six concepts, over variations in the instructional setting, over separate groups of learners, and over half-tests within the test, then one has evidence both for (1.) the reliability and validity of the test and (2.) the reliability and validity of the instructional treatment.

The steps taken to establish the content validity of the test were (1.) developing a substantial number of episodes exemplifying the concepts, (2.) arranging for independent judging of the items by a jury of staff members who had acquired the concepts, (3.) balancing the number of items for each concept, and (4.) including a sufficient number of items so that each half-test is meaningful.

Pre-and post-treatment states. In the development of the test, it was most reasonable to assume a pre-treatment state reflecting some amount of prior learning. Since the test was developed to be used with both preservice and inservice teachers, it was plausible to conclude that the six concepts might be partially acquired prior to treatment. For the pre-treatment state, then, the mean of the test was expected to be above the

50 percent (or chance) level of difficulty; it was also expected that the variance would be greater than at the post-treatment state. The latter state was, consequently, expected to be characterized by a higher mean and attenuated variance.

Predicted relationships between treatment states. Since the test was to be administered to subjects for whom some prior learning may have occurred, certain relationships between treatment states (i.e., pre-treatment and post-treatment states) were predicted. Repeated measurements across the two treatment states should show an increase in means and a decrease in variance. These directional predictions should hold for test composite scores and half-test composite scores. All of these composite comparisons should show statistical significance. The directional predictions should also hold for test concept scores and half-test concept scores; however, the actual directional comparisons may not be individually statistically significant using test concept and half-test concept scores.

All of these predictions were directional and assumed a comparison between groups of students who received effective training on the concepts through the use of the Concepts and Patterns film series and either (1.) the same groups in a pre-treatment state or (2.) an independent pre-treatment group. An additional set of predictions was stated involving comparisons between trained groups. In this case, if no consistent differences were found, additional evidence of test reliability and validity would be demonstrated.

Validation Studies

Instructional treatment. The Concepts and Patterns protocol films were used as part of regular classroom instruction in several intact sections of a graduate level course in educational psychology. The students enrolled in these sections were representative of those normally enrolled in the course: heterogeneous in age, experience and ability. The topics of teacher behavior and teacher-pupil interaction were a conventional part of the course. As indicated previously, all students may well have gained some general familiarity with one or more of the concepts from instruction in previous courses. All students in the first validation study were familiarized also with the specific concept definitions during the five-minute orientation prior to administration of the test. Thus, pretest performance in the first study was potentially influenced by both unplanned and planned prior learning; this influence probably was reflected in average pretest scores that were well above chance level. In the second study, the pretest orientation to the concept definitions was omitted; as a result, average pretest scores were closer to chance level.

Scoring procedure. In the case of the first validation study, responses were scored as correct (that is, an accurate categorization of the teacher's behavior); partially correct/partially incorrect (that is, an incomplete categorization of the teacher's behavior); or incorrect (an inaccurate categorization of the teacher's behavior). Two points were allowed for each correct response and one point for each partially correct/partially incorrect response. A perfect score was 240 points and a chance -

level score was 120 points. A perusal of examinee's answers indicated, however, that the partially correct/partially incorrect response was seldom used. As a result, this option was discontinued for the second validation study. In that study, all items were scored simply as correct or incorrect; thus, a perfect score was 120 points.

Data analysis. Designs were constructed which allowed for the statistical test of directional as well as nondirectional comparisons using the total composite score and odd/even split-half composite scores. For the directional comparisons, it was predicted that means would consistently increase and variances would consistently decrease from pre-instruction to post-instruction conditions. Further, concept scores would follow a similar directional pattern. The analysis is reported separately for each of the two studies.

First study design and results. The design used for the first study was a variation of the separate-sample pretest-posttest design with random assignment, within course sections, into two groups. This design was appropriate to the field setting and had superior external validity characteristics. The effects of pretesting and interaction of testing with treatment were controlled. The design is depicted as follows:

$$\begin{array}{ccc} O_A & X & O_B \\ & X & O_C \end{array}$$

From this design, two sets of directional contrasts of means were predicted, $B > A$ and $C > A$. Conversely, the directionality of contrasts of variances were predicted to be $B < A$ and $C < A$. A nondirectional contrast, $B = C$,

was also tested. A summary of the descriptive statistics is reported in Table 3.

Insert Table 3 about here

For the sets of directional contrasts of means and variances, the predicted trends using odd/even split-half composite scores and total composite scores were found. The B < A contrasts, which involved correlated data, showed significant differences in variance, [t (39) = 2.26 to 4.11, $p < .05$]. After using logarithmic transformations to obtain homogeneous variances, the B > A contrasts of mean differences were significant, [F (1,40) = 29.28 to 51.24, $p < .001$]. The C < A contrasts of differences in variance [F, (40,47) = 1.87 to 2.70, $p < .05$] and C > A contrasts of differences in means [t (64 to 72) = 2.17 to 2.64, $p < .05$] were significant. Since these contrasts involved independent data, t-tests using separate sample estimates of the population variance were used to test mean differences.

The nondirectional contrasts, B = C, resulted in nonsignificant [F (47,49) = 1.16 to 1.62, $p > .10$] variance differences but mean differences were consistently significant [t (87) = 2.91 to 3.97, $p < .01$] for odd/even split-half composite scores and total composite scores. For each contrast of means, B was greater than C, a difference that was attributed to the pretesting experience of the B Ss.

In addition, concept scores and odd/even split-half concept scores were generated and the directionality of mean and variance differences were

compiled. These results are reported in Table 4.

Insert Table 4 about here

As reported in Tables 4 and 5, all the B - A contrasts of concept scores and odd/even split-half concept scores were in the predicted direction. That is, all variances were lower after instruction and all means were higher after instruction. With six concept scores and twelve split-half concept scores for both means and variances, 36 of 36 a priori predicted changes were found. For the C - A contrasts, 10 of 12 concept score predictions and 18 of 24 split-half concept score predictions were supported.

Insert Table 5 about here

Second study design and results. The final filmed version of the test was used in a graduate level educational psychology course following a pretest and posttest design. While it is acknowledged that the effects of pretesting and interaction of testing with treatment were not controlled, the absence of changes in means and variances would raise serious questions about the test and/or the effectiveness of the films. For this reason, the results are reported as evidence for the characteristics of the filmed version of the test.

The examinees who took the test as a pretest were simply introduced to the concept names as part of the test; no specific orientation on the

concepts was provided, as had been done in the first study. It was expected that the means would increase and the variances would decrease from pretest to posttest. A summary of the descriptive statistics is provided in Table 6.

Insert Table 6 about here

The predicted trends were found using odd/even split-half composite scores and total composite scores. The correlated variances on the posttest scores were significantly lower than the pretest scores [t (17) = 3.76 to 5.96, $p < .01$]. The means of the posttest scores were significantly higher than the pretest scores [F (1,18) = 22.58 to 35.17, $p < .001$].

The concept scores and odd/even split-half concept scores of the pretest were compared with the same scores on the posttest. The predicted directionalities of means and variances were found and are reported in Table 7. The probability of obtaining each number of predicted differences by chance alone is reported in Table 8.

Insert Table 7 about here

From Tables 7 and 8, it can be seen that all the B - A contrasts of total concept scores and odd/even split-half concept scores were in the predicted direction. The means from the posttest were higher than the means from the pretest and the variances from the posttest were lower than the variances from the pretest. Although several alternative explanations might

be offered, the test consistently gave indications of being sensitive to instruction through the use of the Concept and Pattern films.

Insert Table 8 about here

Summary

The following conclusion regarding the test characteristics of Categorizing Teacher Behavior seems warranted: by virtue of the proportion of predictions that were substantiated, the test evidenced a high degree of reliability (consistency) and validity. Directional predictions were confirmed using composite test scores and composite half-test scores in most instances. There clearly was a discernible and significant gain in mean test performance which can be accounted for by instruction through the use of the protocol films. Further positive evidence on the characteristics of the test was found using full test concept scores and half-test concept scores. The overwhelming majority of these comparisons substantiated a priori predictions.

Response Bias

As a supplementary analysis to the strategy of assessing the technical characteristics of this test, response bias was investigated. The test was constructed with the realization that, even without knowledge of the concepts, examinees could conceivably score higher than chance level by biasing their responses; that is, one could try to "second guess" the test. This possibility is due primarily to the fact that instancing of the concepts

was done with a view to approximating a realistic classroom setting rather than constructing an ideal correct response distribution.

It is quite reasonable to conclude that longer classroom episodes might evidence a greater number of concepts. In fact, even though all episodes are less than one minute in length, the longer episodes do tend to illustrate more concepts. There was a zero-order correlation coefficient of .37 ($df = 28$, $p < .05$) between the length of the episode (measured in seconds) and the number of concepts instanced. Since this relationship was found, the tendency for an examinee who has received concept training to respond "yes" in the longer episodes was investigated by computing the average number of "yes" responses for each episode in the first study. Further, first-order partial correlation coefficients were computed between the length of the episode (in seconds) and the average number of "yes" responses per episode with the correct "yes" responses partialled out. The first order partial correlation of .11 ($t(27) = .55$, $p > .05$) was insignificant. This finding indicated that a "yes" response bias in the case of longer episodes was unlikely. It was concluded that "score inflation" by such response behavior was not present in that study.

Another possibility for examinees in biasing responses was to respond to each item with "no." A person who used this response pattern could score somewhat higher than chance level (fifty percent correct) since concepts are more often absent than present. It was felt that the general and often demonstrated tendency to respond "yes" when in doubt made the possibility of this type of bias negligible.

Effects of Protocol Based Instruction

Having gathered supporting evidence on its validity as an evaluation instrument, Categorizing Teacher Behavior was next used to assess the results of protocol based instruction in terms of concept acquisition. In a real sense, of course, the same type of evidence (data on change in concept acquisition attributable to training) is used to present information both on test validity and instructional effects. However, in this section, the data will be viewed from the perspective it provides on training effects for all concepts rather than from the perspective of the technical qualities of the test.

Instructional Treatment

Selected films from the Concepts and Patterns series were used in instruction with a graduate level class in educational psychology. The class was composed of both experienced and inexperienced teachers representing a wide variety of teaching or administrative levels and areas in professional education. The class met for one evening session of approximately two and one half hours each week. The instructor, one of the authors of this paper, consistently takes a "teaching skill or behavior" approach to this course; thus, the content of this film series fits naturally into the context of the course.

All three Concept films and a specifically selected four of the six Pattern films were used in instruction. The Concept films were used to present the concepts; analysis of the behavior portrayed in these films and in the Pattern films was conducted through instructor-led class

discussion. A few brief related readings were assigned as collateral work. The class was advised that performance on a posttest would be used to determine letter grades for the unit. A total of approximately eight hours of classroom time over a four week period was devoted to this instructional treatment (including both pretesting and posttesting).

Design and Results

As in the first validation study, a variation of the separate sample pretest and posttest design was used. As shown in the diagram below, a random half of the total group was pretested with Categorizing Teacher Behavior (O_A).[1] All students were posttested, following instructional treatment, with Categorizing Teacher Behavior. The results are reported separately, however, for those students who had taken the test as a pretest (O_B) and those students who had not been pretested (O_C). It is important to note that in neither the pretest nor the posttest did students have access, either immediately before or during testing, to concept definitions.

$$O_A \quad X \quad O_B$$

$$X \quad O_C$$

As in the case of the test validation studies, directional contrasts of means ($B > A$ and $C > A$) and of variances ($B < A$ and $C < A$) were predicted. A nondirectional contrast, $B \approx C$, was again tested.

Means and standard deviations for concept and composite scores are reported in Table 9 for the pretest (A) and the two posttest groups (B and C). As depicted in the design, the posttest information reported under B is for the random half of the class who were pretested. The

posttest information under C is for the random half of the class who were not pretested.

Insert Table 9 about here

A comparison of the B means to the A means using a repeated measures analysis of variance revealed a significant [F (1,22) = 4.47 to 61.49, p = .01 to .05] difference for all concept scores and for the composite score. For the pretested group there was a significant gain on all concepts.

Using a simple analysis of variance, the C means were contrasted with the A means. A statistically significant [F (1,51) = 9.06 to 34.06, $p < .01$] difference was found for the composite score and the probing, productive questioning, and reproductive questioning scores. For all concept scores, the C means were greater than the A means; however, the difference did not reach statistical significance for informing, approving and disapproving concept scores. In addition, the B means were contrasted to the C means and no difference was statistically significant for the two posttest groups.

The B < A comparisons of variances were tested using a t-test (for differences between dependent samples). With the exception of probing, the directionality for all concept scores was in the predicted direction--but none reached statistical significance ($p < .05$). The C < A comparisons of variances yielded similar results. With the exception of probing all differences were in the predicted direction but failed to reach statistical significance.

Summary

There is evidence that the use of the Concept and Pattern films in this instructional setting had a significant effect upon concept acquisition as measured by the film-based test. The evidence suggests that there was a gain on all concepts from a time period prior to instruction to a time period after instruction. This gain was reflected in the differences in means. The lack of a significant decrease in the variances, however, suggests that there was not a convergence of the instructional group toward a criterion. While it is clearly evident that several students reached a high level of concept acquisition, the absence of a decrease in variances indicates that other students did not reach a similarly high level.

It is quite likely that the failure to demonstrate a decrease in variance is attributable to certain limitations of the specific course conditions as a training setting. A protracted general training period, week long intervals between instructional sessions, and a large class enrollment, for example, are all inimical to the kind of intensive instruction that is probably involved in teaching towards a mastery criterion. The adverse effect of such conditions is probably more pronounced for some trainees than for others. Some degree of irregularity in attendance over the general training period probably also helps to account for the absence of a decrease in variance.

Substantiating evidence for these hypotheses is in fact provided by the data reported earlier for the second validation study. In this case, there were significant decreases in variance as well as significant

increases in means. Instruction in this group was characterized by shorter instructional sessions occurring on a daily basis over a shorter general training period. Furthermore, class size was less than half that of the group described in this section.

In any event, it should be remembered that this protocol film series was explicitly designed for use in typically varied instructional settings. The fact that use of the oeries in two somewhat contrasting instructional settings resulted in consistent average gains in concept acquisition is an indication of the practical effectiveness of the series. It is clear, however, that some course arrangements can be identified that more closely approximate an effective training condition (as evidence by their "power" to move a group of trainees closer to a mastery level criterion).

Further Research

Having evidence available on (1.) the validity of Categorizing Teacher Behavior as a test and (2.) the effectiveness of the Concepts and Patterns film series as measured by that test opens the way for further investigation of the relationship between acquisition of this specific set of concepts and the teaching behaviors to which they refer. As suggested in the introduction to this paper, it is quite plausible to hypothesize that training in concepts referring to teaching behaviors might well result in acquisition of these behaviors themselves. As long as one accepts incidence of occurrence of behaviors as a performance criterion, considerable indirect evidence and at least one body of direct evidence (Kleucker, 1975) suggests that use of the Concepts and Patterns series can lead to behavioral

as well as conceptual change. Specifically, the authors of the present paper are presently investigating the relationship between level of concept acquisition, as evidenced by performance on Categorizing Teacher Behavior and incidence of use of the related skills in a simulated teaching setting. Briefly, the authors hypothesize that acquisition of the specific concepts measured by this film-based test is positively related to incidence of using the related skills in instruction. It may be that the effect of concept training on teaching performance has not been estimated accurately in some past studies because of failure to use a test of demonstrated and appropriate validity to measure concept acquisition.

References

Brown, L. D. and Pugh, R. C. A strategy for assessing the reliability of criterion-referenced tests. Proceedings of the 83rd Annual Convention of the American Psychological Association, 1975, 10, in press.

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

Kleucker, J. Effects of protocol and training materials. Acquiring Teaching Competencies: Reports and Studies, Report #6, 1974. Bloomington, Indiana: National Center for the Development of Training Materials in Teacher Education.

Smith, B. O. Teachers for the real world. Washington, D.C.: American Association of Colleges for Teacher Education, 1969.

Footnotes

[1] The criterion for randomization was not fully met because of absences, late arrivals, and other course administrative problems.

Table 1

NUMBER OF INSTANCES OF CONCEPTS
BY PART OF TEST

	Part of Test			
	I	II	III	TOTAL
Concept				
Probing	4	-	4	8
Informing	4	-	3	7
Approving	6	6	-	12
Disapproving	3	6	-	9
Productive Questioning	-	5	5	10
Reproductive Questioning	-	3	3	6
TOTAL	17	20	15	52

Table 2

NUMBER OF EPISODES WITH A GIVEN NUMBER OF CONCEPTS

Number of Concepts Per Episode	Number of Episodes			
	Part I	Part II	Part III	Total
0	0	0	0	0
1	4	3	6	13
2	5	4	3	12
3	1	3	1	5
4	0	0	0	0
TOTAL	10	10	10	30

Table 3

MEANS AND STANDARD DEVIATIONS FOR COMPOSITE SCORES

Composite Score	A			B			C		
	n	$\bar{x}$	sd	n	$\bar{x}$	sd	n	$\bar{x}$	sd
Odd	41	94.2	11.5	41	104.2	6.5	48	98.7	7.0
Even	41	99.2	10.2	41	108.5	5.9	48	104.3	7.5
Total	41	193.3	20.2	41	212.7	10.1	48	203.0	12.3

Table 4

DESCRIPTIVE SUMMARY OF CONCEPT SCORES

Concept Scores	Number of items	A (n = 41) $\bar{x}$	A sd	B (n = 41) $\bar{x}$	B sd	C (n = 48) $\bar{x}$	C sd
Total							
Probing	20	30.4	3.6	35.2	2.8	34.3	3.6*
Informing	20	30.5	5.1	34.7	2.4	33.5	3.5
Approving	20	32.3	4.0	33.6	2.7	32.0*	3.5
Disapproving	20	33.5	4.6	36.1	3.0	33.6	3.2
Productive Questioning	20	33.3	6.4	36.3	2.8	34.4	3.1
Reproductive Questioning	20	33.3	6.1	36.8	2.8	35.1	3.3
Odd							
Probing	10	15.6	2.3	18.6	1.8	17.8	2.5*
Informing	10	14.2	3.0	16.3	2.0	15.9	2.2
Approving	10	14.8	2.0	15.4	1.5	14.4*	2.3*
Disapproving	10	16.9	2.5	18.1	1.9	17.3	1.8
Productive Questioning	10	16.9	4.1	18.0	1.9	16.7*	2.2
Reproductive Questioning	10	15.8	4.1	17.8	2.3	16.6	2.4
Even							
Probing	10	14.8	2.0	16.6	1.7	16.5	2.2*
Informing	10	16.3	2.8	18.4	1.7	17.6	2.3
Approving	10	17.5	2.6	18.2	2.1	17.6	2.0
Disapproving	10	16.5	2.6	18.0	1.9	16.3*	2.4
Productive Questioning	10	16.5	3.0	18.3	1.9	17.7	2.2
Reproductive Questioning	10	17.5	2.3	19.0	1.4	18.5	1.8

* Reversal of prediction relative to A

Table 5

COMPARISONS OF TOTAL CONCEPT SCORES AND ODD/EVEN CONCEPT SCORES

	Total Concept Scores		Odd/Even Concept Scores	
	n_H	p_H	n_H	p_H
		B-A Comparisons		
Means	6	.02	12	.0002
sds	6	.02	12	.0002
		C-A Comparisons		
Means	5	.11	9	.07
sds	5	.11	9	.07

Table 6

MEANS AND STANDARD DEVIATIONS FOR COMPOSITE SCORES

Composite Score	A (Pre)			B (Post)		
	n	$\bar{x}$	sd	n	$\bar{x}$	sd
Odd	19	42.5	8.2	19	51.5	4.2
Even	19	41.0	11.9	19	51.3	4.8
Total	19	83.5	19.4	19	102.8	8.1

Table 7

DESCRIPTIVE SUMMARY OF CONCEPT SCORES

Concept Scores	Number of items	A (Pre) (n = 19) $\bar{x}$	A (Pre) (n = 19) sd	B (Post) (n = 19) $\bar{x}$	B (Post) (n = 19) sd
Total					
Probing	20	13.3	2.7	17.9	1.9
Informing	20	13.4	3.9	16.7	1.6
Approving	20	14.6	3.3	15.8	2.2
Disapproving	20	15.9	4.2	17.6	1.4
Productive Questioning	20	14.5	3.9	17.9	2.2
Reproductive Questioning	20	11.8	5.5	17.1	3.7
Odd					
Probing	10	6.3	1.5	9.3	1.0
Informing	10	6.2	2.1	7.5	1.0
Approving	10	7.2	1.5	7.6	1.3
Disapproving	10	8.1	2.3	9.1	0.6
Productive Questioning	10	7.5	1.7	8.8	1.2
Reproductive Questioning	10	4.8	3.6	8.2	2.4
Even					
Probing	10	7.0	1.7	8.6	1.2
Informing	10	7.3	2.1	9.2	1.0
Approving	10	7.5	2.3	8.2	1.3
Disapproving	10	7.7	2.1	8.5	1.0
Productive Questioning	10	7.0	2.5	8.9	1.6
Reproductive Questioniug	10	7.0	2.3	8.8	1.5

Table 8

COMPARISONS OF TOTAL CONCEPT SCORES AND ODD/EVEN CONCEPT SCORES

	Total Concept Scores		Odd/Even Concept Scores	
	n_H	p_H	n_H	p_H
Means	6	.02	12	.0002
sds	6	.02	12	.0002

Table 9

DESCRIPTIVE SUMMARY OF TOTAL CONCEPT SCORES AND COMPOSITE SCORE

Score	Number of Items	A (n=23)		B (n=23)		C (n=30)	
		$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
Probing	20	14.0	1.8	17.9	2.2*	17.5	2.3*
Informing	20	15.3	2.7	17.0	1.5	16.6	2.7
Approving	20	15.2	2.3	16.6	2.1	16.0	1.5
Disapproving	20	16.3	2.7	17.1	2.2	16.7	1.9
Productive Questioning	20	15.1	3.2	17.4	1.7	17.4	2.2
Reproductive Questioning	20	13.6	3.8	17.7	2.0	17.1	2.8
Composite	120	89.5	10.8	103.7	8.9	101.2	10.1

* Reversal of predicted direction.