DOCUMENT RESUME

ED 027 260 SP 002 238

By-Popham, W. James

Validation Results: Performance Tests of Teaching Proficiency in Vocational Education.

Pub Date 69

Note-22p.; Paper presented at the American Educational Research Assn. meeting, Los Angeles, Calif., Feb. 5-8, 1969.

EDRS Price MF-$0.25 HC-$1.20

Descriptors-*Academic Achievement, Achievement Gains, Achievement Tests, Behavioral Objectives, *Educational Experiments, *Effective Teaching, *Evaluation Techniques, Industrial Arts, *Performance Tests, Teacher Evaluation, Teacher Experience, Validity, Vocational Education

A project was undertaken to develop and validate a method of assessing teacher competence through the use of pupil performance tests. Teachers were given a list of specific, operationally defined objectives for a particular topic and directed to teach the objectives. Teacher competence was judged in relationship to the way their students performed on pre- and posttests of behaviors stated in the objectives. An attempt to validate this method of measuring teacher effectiveness involved contrasting the results produced by experienced teachers and nonteachers (28 pairs teaching a 10-hour auto mechanics unit and 16 pairs teaching a 10-hour electronics unit to high school industrial arts classes). Calculations of mean and standard deviation, internal consistency coefficients, intercorrelations between a number of variables, and analyses of covariance between pupil scores and interest revealed no significant differences between the teachers and nonteachers. Results (which confirmed those of an earlier study using social science classes) were interpreted as indicating that the experienced teacher is not more experienced than the nonteacher in modifying learner behavior in terms of previously established instructional objectives. Findings do not, however, refute the basic assumption that performance test measures are presently the most serviceable legitimate indices of teaching proficiency. (Related to ED 013 242, BR-5-0566.) (JS)

A Paper Presented at the Annual
American Educational Research Association Meeting
Los Angeles, California February 5-8, 1969

VALIDATION RESULTS: PERFORMANCE TESTS OF TEACHING PROFICIENCY IN VOCATIONAL EDUCATION

W. James Popham

University of California, Los Angeles
Southwest Regional Laboratory for Educational Research and Development

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education and Welfare.

INTRODUCTION

One of the most elusive targets in the history of educational research is a valid index of teacher effectiveness. Since the turn of the century literally hundreds of investigations have probed the question of teacher competence assessment and most of them have produced little, if any, significant progress.

In the last few years, however, evolving conceptions of the nature of instruction seem to offer promise to teacher effectiveness researchers. In most of the early investigations in which measures of teacher effectiveness were sought, there was an almost exclusive focus on the instructional means employed by teachers. Researcher after researcher attempted to identify "good teaching procedures" for, should such procedures be discovered, they would obviously have implications both for pre- and in-service teacher education, as well as for the evaluation of teachers on the job. Only recently have many educators come forward who are more inclined to accept the proposition that there are literally a variety of instructional means which can be used to bring about a single instructional end. This more recent thinking suggests that the focus in the evaluation of instruction should be on the results achieved by instructors, irrespective of the means they employ.

When Morsh and Wilder in their definitive review[1] of teacher effectiveness research during the first fifty years of this century indicated that no single teaching act had been discovered which was invariably associated with learner achievement, teacher competence researchers should have been more attentive. We should have first focused our efforts on identifying teachers who could produce superior growth in learners, leaving aside for the moment the question of how such improvements were brought about. If one can identify satisfactory measures of pupil attainment, then the next step is to identify the complicated procedures by which such achievements are attained. To repeat, the first task is to isolate satisfactory measures of quality instruction, the second is to use those measures to discover the means which contribute to teaching prowess. It is important to emphasize the complexity of this task, for the undoubted reason that reviewers such as Morsh and Wilder found few references to "good teaching procedures" is that effective instruction surely represents a series of subtle interactions among a given teacher, his particular students, the instructional goals he is attempting to achieve, the instructional environment, etc.

Of course, there have been researchers who have employed the criterion of learner growth as an index of a teacher's proficiency. Such

---

[1] Joseph E. Morsh and Eleanor Wilder, "Identifying the Effective Instructor: A Review of the Quantitative Studies, 1900-1952," Research Bulletin AFPTRC-TR-54-44, Lackland Air Force Base, Texas, 1954.

efforts would seem to represent proper attention to instructional ends rather than means. Unfortunately, most of these investigations relied upon the use of broad, standardized achievement tests which, while comprehensive, rarely took into consideration the idiosyncratic nature of an instructor's particular inclinations regarding what ought to be taught. Such gross measures of learner achievement invariably were reported as being too insensitive. Further, the standardized measures employed were invariably based on norm-referenced rather than criterion-referenced approaches to test construction and, as a consequence, were often inappropriate to measure group progress toward instructional goals.

Because of the methodological difficulties encountered to date by teacher effectiveness researchers, a heretofore untried procedure for assessing teaching competence was conceived at the University of California, Los Angeles during the year 1964. The heart of this procedure involved the use of so-called "performance tests of teaching proficiency." Two proposals were submitted to the U.S. Office of Education and support was secured from that agency for two related investigations, one of which was designed to develop and test a validity hypothesis regarding a performance test in the field of social science. Results of this investigation have been previously reported.[2]

The second study was to be conducted in the fields of vocational education and called for the development of two performance tests in that general area, more specifically, in the areas of (a) electronics and (b) auto mechanics. The present report describes results of the latter investigation.

Rationale. In brief, the general approach used in the performance tests of teaching proficiency calls for a set of instructional objectives to be developed in extremely explicit terms to cover an instructional period of approximately nine to ten hours. Coupled with such objectives are examinations based explicitly on the objectives. In addition, a collection of possible instructional activities and references are provided in a form comparable to the "resource units" found in so many curriculum libraries. The procedure for using such performance tests requires that an instructor be given the objectives and resource materials well in advance of instruction, is told to devise a sequence of instruction suitable for accomplishing the objectives, and then allowed to teach to the objectives using whatever instructional procedures he wishes. In other words, only the ends are specified, the means are left free to the instructor. A participant in our investigation would be obliged, therefore, to attempt to accomplish the prespecified objectives, but would have complete freedom to choose instructional procedures which, to him, seemed likely to achieve those goals.

---

[2] W. James Popham, Development of a Performance Test of Teaching Proficiency, Department of Health, Education and Welfare, Project No. 5-0566-2-12-1, Contract No. OE-6-10-254, Final Report, 1967.

It is difficult, of course, to validate the merits of such an approach to the assessment of teaching competence. One does not have readily available the already established criterion measures which can be used to calculate concurrent validity coefficients. A construct approach to validation, therefore, appeared to be more appropriate. It seemed, considering the nature of the performance tests, that these measures ought to be able at least to distinguish between experienced teachers and those who have never previously taught. In other words, if one were to take a group of experienced teachers and ask them to teach to the objectives, in contrast to asking a group of "people off the street" to teach to the same objectives, the experienced teachers ought to out-perform their inexperienced counterparts. In order to test this validation hypothesis, it was proposed that performance tests be developed in the two fields previously mentioned and that the ability of the tests to discriminate between experienced teachers and nonteachers be determined. The project was initiated in May, 1965, and was concluded in June, 1968. Results of developmental work and field tests will be described in the following pages.

Related Results of the Social Science Performance Test Investigation. Because of its relevance to the present investigation, a somewhat extensive summary of the investigation dealing with the social science performance test should be presented. Accordingly, an abstract of that investigation is presented in full.

> A project was undertaken to develop and, hopefully, validate a heretofore untried method of assessing teacher competence, namely through the use of a performance test. The performance test was designed to function in the following way. Teachers were presented with a list of specific, operationally defined objectives for a particular topic and directed to teach the objectives. Following the instructional period, students were tested on the behaviors stated in the objectives. Teacher competence was judged in relationship to the way their students performed on the criterion test. An attempt to validate this method of measuring teacher effectiveness involved contrasting the results produced by experienced teachers and nonteachers on a performance test dealing with research methods in the social sciences.
>
> Two separate contrasts were conducted, the first involving six professionally trained, experienced student teachers versus six housewives for a six hour teaching period. The second involved thirteen regularly credentialed teachers and thirteen college students for a four hour teaching period. In neither contrast did the teachers perform significantly better than the nonteachers.
>
> The results were interpreted as indicating that the experienced teachers were not more experienced than the nonteachers in promoting learner achievement of previously established instructional objectives. An alternative approach

to validating the performance test strategy was discussed along with possible procedural modifications in the approach.

Because of the failure to support the validation hypothesis in connection with the social science performance test investigation, we were particularly anxious to learn whether tryouts of the two performance tests to be developed in the present investigation would confirm the validation prediction.

TEST DEVELOPMENT

Early months of the project were devoted to selection of tentative topics for the performance tests. The fields of electronics and auto mechanics had been selected because of the frequency with which such courses are taught in the public schools. But beyond the general topics, there was the specific question of what units would be appropriate for the performance tests. Ideally, it was judged that topics should be (1) sufficiently important so that teachers would be willing to include them in their curricula, and (2) sufficiently autonomous so that the units could be inserted rather freely at various points during the academic year. With these criteria in mind, the topic selected for the electronics unit was the "General Principles of Electronics Troubleshooting," and the topic selected for auto mechanics was "Carburetion."

Developmental work on the two units occurred in the following pattern: First, topics meeting the above criteria which might be covered in two or three weeks were selected. These were then submitted to several subject matter specialists who served as consultants during the project. From these tentative topics, the two mentioned above were selected and instructional objectives were prepared which were also screened by consultants. A preliminary set of these objectives was agreed on, and test items based directly on the objectives were developed. In addition, possible learning activities and reference materials were assembled. In some instances, these learning activities were designed to be particularly pertinent to the given objectives. In other cases, the activities were planned to be "flashy" but not germane to the objectives. It was thought that less experienced instructors might be attracted to the "flashy," irrelevant activities, but that the sophisticated teacher would tend to use the pertinent activities. These materials were revised several times prior to an initial trial. Of course, it was possible that the teacher might choose to develop his own activities and not use any of the materials provided in the unit.

The early forms of the post-tests were given to several teachers for administration to classes of students currently taking electronics or auto mechanics courses. Resulting data underwent item analysis procedures which resulted in improvement of many test items.

When ready for the first field trial, both the carburetion and electronics units consisted solely of objectives measurable by paper and pencil tests. The instructional time allotted to each was ten hours.

The carburetion materials (with 49 objectives) consisted of 22 pages while the electronics unit (with 23 objectives) consisted of 41 pages. Both units were considered incomplete, for it was planned during the second year to add objectives demanding performance on actual carburetors and electronics circuit boards.

For each unit a pre- and post-test were prepared. The post-test for electronics consisted of 52 multiple-choice items and the post-test for carburetion consisted of 97 multiple-choice items. Both of these tests were drawn specifically from the objectives. Items for the pre-tests were randomly selected from the post-test items in order to provide measures which could be completed in approximately 20 minutes. The pre-test for electronics contained 17 items while the carburetion pre-test had 20 items.

Initial Tryout. Both performance tests were given their first test during January and February, 1966. The electronics test was tried out in eight electronics classes at Los Angeles Trade Technical College.[3] The carburetion performance test was tried out in two classes, one at the junior college level at Los Angeles Trade Technical College, and one at Fullerton Union High School.[4] Thirty-six students took part in the carburetion tryout and 108 students in the electronics tryout.

Each instructor was given a copy of the unit objectives and the resource materials approximately one week prior to the time he was to teach the unit. Each was instructed to attempt to accomplish the objectives stated in the unit, but to use any instructional techniques he wished. For purposes of this trial, participating instructors were also asked to make suggestions regarding ways in which the materials could be improved.

Arrangements were made with each teacher so that a member of the research staff administered (1) a twenty minute pre-test during the first day of the ten hours devoted to the unit and (2) a fifty minute post-test at its conclusion. In addition, a questionnaire was administered to the students at the close of the unit. A questionnaire was also given to the teacher at that time soliciting his suggestions regarding the unit. Finally, in two of the electronics classes and one carburetion class the Wonderlic Personnel Test, a twelve minute test of "problem-solving ability" was administered to the students at the time of the pre-test. In all, 25 different variables were represented by the two questionnaires and the Wonderlic. The trials were completed between the dates of January 17 and February 10, 1966.

---

[3]Appreciation is expressed to the administration and staff of Los Angeles Trade Technical College for their participation in this investigation.

[4]Appreciation is also expressed to the administration and staff of Fullerton Union High School for their participation in this investigation.

Analysis. Two different types of analyses were conducted on the preliminary data. The first was to compute item analyses and coefficients of internal consistency (Kuder-Richardson Formula 20) on the pre- and post-tests. The second was to compute intercorrelations among the several measures. Key interest in the latter analysis focused on the possibility of detecting variables which could be used, in part, to control for differences among the pupils due to such factors as "set" toward the unit's material, intelligence, etc. Further, of course, the responses of the instructors were carefully considered. The overall purpose of the initial analysis was essentially heuristic. We were attempting to find possible variables to be considered in subsequent trials of the materials.

It was fully expected that a great number of deficiencies would exist in the first experimental versions of both of the performance tests. Procedural defects regarding such details as test administration, relations with instructors, etc. were also anticipated.

Results. The performance of students in this first extensive field test is summarized in Table 1. Item analysis results revealed a considerable number of items, particularly in the electronics tests, which were in need of revision. The KR_{20} coefficients were markedly higher for the carburetion tests than for the electronics tests.

Table 1

Electronics and Carburetion Pre- and Post-Tests Results

	Electronics					Carburetion				
	n	items	$\bar{X}$	s	$KR_{20}r$	n	items	$\bar{X}$	s	$KR_{20}r$
Pre	108	17	9.68	2.71	.56	36	20	10.41	3.48	.71
Post	98	52	23.68	9.57	.44	33	97	51.64	17.86	.78

Of the 25 variables constituting the pupil and teacher questionnaires and the Wonderlic test, interest centered on those which might be of value in adjusting for initial differences among pupils and/or teachers. In the case of both carburetion and electronics, the variables which were most strongly related to post-test performance were: (a) pupils' overall grade point average, and (b) pupils' estimate of the pre-test's difficulty, i.e., the students who thought it easier tending to score higher on the post-test. Negligible correlations existed between post-test scores and (a) pre-test performance or (b) Wonderlic performance.

The chief suggestions from participating instructors concerned the addition and deletion of certain objectives for the units. A number of

criticisms were made of the technical terminology employed in the objectives and reference materials. Many instructors thought that the topics could not be adequately treated in ten instructional hours. Many minor deficiencies in the quality of the reference materials were also noted.

In the case of the carburetion test, several additional small-scale trials were conducted during the spring of 1966, resulting in a number of revisions of the instructional materials, but in the overall confirmation of the choice of carburetion as a suitable topic for the instructional unit. Revisions in three experimental versions resulted in a 27 page set of resource materials to be given to participants in the study. In these materials were included 29 specific instructional objectives (as opposed to the previous 49 objectives), a 99 item post-test and a 20 item pre-test based on the objectives. The student questionnaire was refined so that it consisted of a 17 item instrument. The instructor questionnaire of 16 items was also modified. All of the materials were surveyed numerous times by vocational education consultants, usually junior college and high school auto mechanics instructors.

The disappointing results with the unit on electronics trouble shooting led to a serious reappraisal of the value of that particular topic. After extensive consultation, particularly with a number of practicing electronics instructors, a complete shift of topic was undertaken and a new unit was developed on the topic of "Basic Power Supplies." After three revisions, an experimental version was completed in January, 1967, consisting of 30 pages and 23 specific instructional objectives. A 73 item post-test, a 20 item pre-test, and student and teacher questionnaires comparable to those used with the carburetion unit were prepared. These materials were examined by many electronics teachers and the response of consultants to the newly selected topic was most encouraging.

The tests for the new electronics performance measure were administered to almost 100 subjects in an early field trial. Thirty-three of these were advanced junior college students whose performance could be considered that of sophisticated learners, that is, those who should know the material treated in the test. An encouraging Kuder-Richardson (20) internal consistency coefficient of .91 was yielded on the basis of this tryout. The test took approximately 90 minutes to complete, however, and had to be shortened. This was done according to results of item analyses on the total tryout results.

After several revisions, particularly in the tests, the final set of electronics materials consisted of a 30 page unit which included 23 specific objectives, a 20 item pre-test plus a 46 item post-test based on those objectives, a 17 item student questionnaire, and a 16 item teacher questionnaire comparable to those in the carburetion unit. These revised materials were field tested on several small classes of students and results indicated suitability for use in the formal test of the validation hypothesis.

METHOD

The general approach to be employed in testing the validation hypothesis called for the identification of experienced teachers in the fields of electronics and auto mechanics and then comparing the achievement of their students in a performance test situation with the achievement of students taught by nonteachers. Accordingly, after developing the performance test materials, our next task was to locate a suitable number of teachers and nonteachers. This endeavor proved to be the most trying operation of the entire research project. At first we solicited the cooperation of large districts in which there would be a number of vocational education teachers in the desired fields. We also hoped to conduct our study in relatively large metropolitan areas where we might be able to locate numerous nonteachers who would have sufficient technical backgrounds to teach the units, but no prior teaching experience. In the case of the auto mechanics performance test, we anticipated using garage mechanics from service stations, auto agency service departments, etc. For the electronics performance test we hoped to recruit individuals such as television repairmen and workers in electronics industries.

The response from most large school districts in the state of California was both consistent and discouraging. For a variety of reasons, large districts were unwilling to participate. For example, the most convenient large district (to our research staff) had an unwritten board policy against evaluating personnel following employment. This policy precluded the conduct of such studies, even if teachers volunteered to participate. Other districts indicated an unwillingness to have a non-credentialed teacher in the classroom, although legal requirements could be satisfied by the regular teacher's remaining in the classroom. Of all the large districts contacted throughout the state, only the San Diego City Schools[5] agreed to participate in the project. In addition, a number of school districts in Orange County agreed to participate in the project.

The unanticipated difficulties in securing participants for the investigation actually delayed completion of the project by a full year. A request for a temporal extension of the contract from the U.S. Office of Education was graciously approved.

In addition to the location of districts which would participate, there was also the problem of locating teacher volunteers as well as nonteachers who would agree to teach in the schools. Because of their extra effort involved in this project, an honorarium of $50 was given to each participating teacher. Since the nonteachers would usually be obliged to take time away from their regular jobs, a $75 honorarium was given to each participating nonteacher.

---

[5] We are inordinately appreciative of the willingness of the San Diego City Schools and Orange County Schools to become participants in this project.

As an added inducement to participate in the project, teacher volunteers were promised certain equipment which could be used by their students (as well as the students taught by the nonteacher) during the unit. At the conclusion of the unit this equipment was to be left with the teacher for his use with other classes. In the auto mechanics class one carburetor was provided for approximately every four to six pupils. In the electronics classes a power supply kit consisting of one transformer, one capacitor, and four transistor diodes was provided for approximately every two or three pupils.

The difficulty of locating nonteachers was considerable. Research assistants assigned to the project had incredible difficulty in indentifying a suitable number of nonteachers who were both willing to participate in the project and could arrange their schedules in order to teach in the schools. We finally relied on a major amount of newspaper advertising in order to attract the attention of nonteachers. We tried to slant our call for participants so it would be particularly appealing to the nonteachers. We referred to them as "industrial specialists" and attempted to describe the value of the contribution they would make to the project. After untold hours of telephone conversations, letters, etc., a sufficient number of teachers and nonteachers were located so that we had 28 pairs (teacher and nonteacher) for the auto mechanics field test and 16 pairs for the electronics field test.

Because of the necessity of controlling the potential influence of "school effects" in the data analysis, we located a nonteacher "match" for every teacher who agreed to participate in the project. In all instances teachers were selected who had at least two sections of a class in which the unit could be taught for approximately nine hours. For electronics, the following kinds of classes were usually involved: first and second year electronics and introductory electricity. For auto mechanics the following kinds of classes were generally involved: first and second year auto mechanics and power mechanics.

In almost all instances one of the classes was randomly designated as that which would be taught by the teacher. We were anxious to avoid the possibility that teachers would select one of their "best" classes and, consciously or unconsciously, give the less able group to the nonteacher. In some instances, of course, the schedule of a particular tradesman dictated the selection of a certain class hour for him. In this instance, if more than one class was available to the regular teacher, the class to be taught by the teacher (for the project) was selected at random.

In general, the procedure involved giving the teacher and nonteacher sets of the instructional materials, that is, objectives and resource materials, approximately two weeks prior to the time when instruction was to commence. Then a member of the project research staff (or, in some cases, the teacher and nonteacher themselves) administered the pre-test to all students at the beginning of ten hours reserved for the project. The pre-test took approximately 15 to 20 minutes

to complete. The results of the pre-tests were immediately put into an envelope and returned by mail to the project staff. The regular teacher and nonteacher then, in their separate classes, taught for approximately nine hours. They attempted to achieve the objectives specified in the unit but, as indicated before, were free to use any methods they wished. For legal purposes, while the nonteacher instructed, the regular teacher remained at the rear of the classroom but was instructed not to interfere with the nonteacher's efforts. At the conclusion of the nine hours of instruction, the regular teacher or a member of the project staff opened a sealed envelope and administered the post-test to all students. These materials, along with a brief student questionnaire and instructor questionnaire, which were also filled out at the conclusion of the unit, were returned to the project staff.

ANALYSIS

Because of the interaction among pupils in given classes, it was considered appropriate to treat the data in terms of classroom units rather than individual pupils. Accordingly, the first step in the analysis called for the calculation of classroom means for each of the variables involved in the investigation. These means, then, constituted the data for subsequent analyses. The principal analysis concerned the test of the prediction that teachers would significantly out-perform nonteachers. Two analyses were conducted to test this hypothesis. The first was a correlated t test using the gross post-test score as the criterion. The second was an analysis of covariance in which pre-test scores and students' grade point averages served as covariates. Separate analyses, of course, were conducted for electronics and carburetion data.

Results of pupil affective data on the questionnaires were compared by analyses of variance. As indicated previously, all of these analyses involved classroom means rather than data for individual pupils.

One case involving data for individual pupils, however, involved the calculation of Kuder-Richardson 20 internal consistency coefficients. These coefficients were based on the computations for the entire pool group of 710 students in the electronics classes and 1,248 students in the carburetion classes.

RESULTS

For simplicity of exposition, results of the electronics and auto mechanics analyses will be described separately.

Auto Mechanics. The initial analysis conducted was the determination of means and standard deviations for each of the classes taught by the 28 auto mechanics teachers and the 28 nonteachers. In a sense, of course, these represented pairs of instructors because a given teacher and nonteacher were instructing in a particular school. Results of these computations for pre- and post-test performance of the auto mechanics classes are presented in Table 2.

Table 2. Class Pre- and Post-Test Means and Standard Deviations of 28 Auto Mechanics Teacher and Nonteacher Pairs

	Pre-Test			Post-Test		
	n	$\bar{X}$	s	n	$\bar{X}$	s
Teacher	19	11.2	2.5	18	55.1	8.2
Nonteacher	22	11.1	2.1	20	51.5	9.7
Teacher	23	9.3	2.6	22	54.8	11.8
Nonteacher	22	10.0	2.3	19	47.6	10.8
Teacher	18	10.8	2.4	23	46.6	11.8
Nonteacher	22	11.5	1.7	19	46.3	9.4
Teacher	9	10.1	2.5	11	58.0	11.2
Nonteacher	19	10.4	2.3	23	65.7	10.5
Teacher	23	10.2	2.2	24	52.5	13.4
Nonteacher	22	10.8	2.3	24	50.4	10.7
Teacher	25	9.4	2.2	25	41.5	10.1
Nonteacher	24	9.4	2.6	20	45.7	12.2
Teacher	18	12.8	2.0	18	59.4	10.9
Nonteacher	19	9.4	2.7	19	40.8	8.8
Teacher	13	11.1	2.7	19	53.3	13.2
Nonteacher	20	10.8	2.4	20	56.6	8.8
Teacher	20	9.1	3.0	22	40.0	11.7
Nonteacher	21	7.8	3.6	21	47.3	11.9
Teacher	26	9.5	2.6	24	45.5	11.0
Nonteacher	26	9.7	2.7	22	43.3	9.1
Teacher	24	9.5	2.3	22	44.7	7.9
Nonteacher	22	9.5	2.5	19	37.8	11.0
Teacher	24	9.0	2.9	23	42.7	9.2
Nonteacher	26	11.6	2.3	22	48.3	7.0
Teacher	26	8.7	2.8	26	44.2	14.3
Nonteacher	19	8.6	2.7	16	42.3	10.8
Teacher	28	9.6	2.7	30	45.1	13.9
Nonteacher	27	10.5	3.0	28	50.8	14.3
Teacher	20	10.6	2.0	24	49.5	6.9
Nonteacher	20	9.6	2.2	21	50.1	9.9

Table 2. (Continued)

	Pre-Test			Post-Test		
	n	$\bar{X}$	s	n	$\bar{X}$	s
Teacher	15	6.3	2.9	17	37.0	7.7
Nonteacher	22	7.9	3.0	21	31.5	8.9
Teacher	18	10.0	2.5	18	48.3	12.4
Nonteacher	16	5.2	1.4	19	44.6	13.1
Teacher	17	6.5	1.7	18	39.1	10.9
Nonteacher	24	6.1	1.3	24	38.1	11.6
Teacher	28	9.9	2.7	28	46.6	8.6
Nonteacher	21	9.6	2.5	21	48.1	14.4
Teacher	17	9.5	2.2	19	50.1	12.7
Nonteacher	23	10.0	2.7	21	47.9	10.6
Teacher	21	10.1	2.5	21	52.6	10.5
Nonteacher	23	11.0	2.6	23	53.4	10.3
Teacher	9	9.6	3.2	6	47.5	8.0
Nonteacher	11	9.4	1.5	6	50.5	9.9
Teacher	15	11.2	2.4	17	60.8	9.1
Nonteacher	25	10.7	2.5	26	48.7	8.3
Teacher	18	8.6	2.4	20	42.2	9.9
Nonteacher	6	9.6	3.4	6	34.5	13.5
Teacher	17	9.9	2.3	18	52.0	9.9
Nonteacher	16	10.2	2.0	18	51.2	8.8
Teacher	25	9.8	3.1	22	50.2	12.4
Nonteacher	13	9.9	2.5	14	43.5	8.6
Teacher	19	9.1	1.9	19	33.4	11.1
Nonteacher	18	10.1	2.8	20	41.3	17.5
Teacher	26	10.3	3.1	28	48.0	10.7
Nonteacher	30	10.9	2.4	28	44.1	10.4

Several preliminary analyses were then conducted with the auto mechanics data. The first of these was a calculation of internal consistency (Kuder-Richardson 20) coefficients for the pre- and post-tests. These were calculated on the basis of the entire sample of pupils, considered irrespective of (a) class and (b) whether or not they were taught by a teacher or a nonteacher. The Kuder-Richardson 20 coefficient for the pre-test was .50 and for the post-test was .88.

Another preliminary analysis involved the computation of intercorrelation matrices in which post-test performance was correlated with pre-test performance, as well as a number of variables reported by pupils in the student questionnaire. It was hoped that by inspecting the relationships among post-test performance and these variables that two or three measures could be identified which were (a) highly correlated with the post-test but (b) only moderately correlated with each other. Such variables would, of course, be suitable for the anticipated analysis of covariance which was to be performed. We were, in other words, looking for useful control variables for that analysis. Intercorrelation matrices were computed both on the basis of all pupil data, analyzed student by student rather than by class, and also based on class by class results. In the case of the auto mechanics data, the most useful covariates appeared to be (a) pre-test scores and (b) pupils' expressed interest in the field of auto mechanics. The correlations based on class means yielded an r of .59 between pre- and post-test scores, an r of .59 between post-test and expressed interest, and an r of .28 between pre-test and expressed interest.

The next analysis involved a comparison of the pre-test performance of the teacher and nonteacher classes. Because of the probable similarity of the performance of pairs of classes within a given school, a product-moment correlation coefficient was first calculated to determine the relationship between the teacher and nonteacher pre-test means on the basis of school. The correlation in this instance was .50. Because of this relationship, a correlated t test model was employed to test for significance or difference between the pre-test means of the teacher and nonteacher classes. In Table 3 the means, standard deviations, and correlated t test result for pre-test performance of the 28 auto mechanics teacher and 28 nonteacher classes are presented.

Table 3. Means, Standard Deviations, and Correlated t Test Result for Pre-Test Performance of Auto Mechanics Teacher and Nonteacher Classes

	n	$\bar{X}$	s	t
Teacher	28	9.8	1.3	.10
Nonteacher	28	9.7	1.4	

As can be seen from an inspection of Table 3, the pre-test means of the

two groups were almost identical. The correlated t test yielded a non-significant t value.

The principal hypothesis of the investigation concerned the predicted difference on post-test performance in favor of the classes taught by the teachers rather than the nonteachers. The next analysis involved a comparison of post-test performance of the teacher and nonteacher classes. Initially, the correlation between the pairs was calculated and because of its magnitude, that is, .60, again a correlated t model was used for contrastive purposes. In Table 4 the means, standard deviations, and correlated t test result for the post-test performance of the auto mechanics teacher and nonteacher classes are presented.

Table 4. Means, Standard Deviations, and Correlated t Test Result for Post-Test Performance of Auto Mechanics Teacher and Nonteacher Classes

	n	$\bar{X}$	s	t
Teacher	28	48.16	6.5	1.42
Nonteacher	28	46.53	6.8	

As can be seen, the predicted hypothesis was not confirmed. The difference between the teacher and nonteacher classes on the post-test, a post-test based on the pre-specified objectives which both groups were instructed to accomplish, yielded no significant difference in favor of the teacher group.

In addition to using the gross post-test results as the criterion for testing the validation hypothesis, several other potential ways of looking at the criterion performance were employed. For example, we calculated a gain score on the basis of post-test minus pre-test results. We also looked at a gain score based on the difference between only those items in the post-test which were present in the pre-test. (This "common items" gain score was, incidentally, used in a progress report[6] of this project at the 1968 meeting of the California Educational Research Association.) In addition, other adjusted gain scores were calculated based on an individual's pre-test potential to improve. These "efficiency" scores, along with all other ways of looking at criterion results, confirmed the general result presented in Table 4, namely, there was no significant difference in the performance of the two groups.

Because of the possibility that one of the two groups had been disadvantaged as a consequence of systematic favoritism on relevant variables, an analysis of covariance was computed in which the aforementioned

[6]Popham, W. J. Progress Report: Performance Tests of Teaching Proficiency in Vocational Education, 46th Annual Conference, California Educational Research Association, Oakland, California, March 15-16, 1968.

measures of (a) pupils' pre-test scores (on the 20 item test) and (b) pupils' expressed interest in auto mechanics (1 = low to 5 = high) were used as control covariates. Results of this analysis are presented in Table 5.

Table 5. Analysis of Covariance of Auto Mechanics Classes (Teachers Versus Nonteachers) Post-Test Performance, using Pre-Test Scores and Pupils' Expressed Interest in Auto Mechanics as Covariates.

Source	df	SS	MS	F
Between	1	22.7	22.7	.84
Within	52	1405.6	27.0	
Total	53	1428.3		

	Control Variables		Criterion Variable	
Group	Pre-Test $\bar{X}$	Interest $\bar{X}$	Unadjusted Post-Test $\bar{X}$	Adjusted Post-Test $\bar{X}$
Teachers	9.8	3.6	48.16	47.99
Nonteachers	9.7	3.6	46.53	46.71

As can be seen from Table 5, results of the analysis of covariance fail to reveal any significant difference between the teachers and nonteachers.

Analysis of variance tests of the difference between affective reactions of pupils, as reflected by responses to the student questionnaire, failed to reveal any significant differences between teacher classes and nonteacher classes. Measures involved in these analyses included responses to such questions as: "After this unit how would you rate your interest in the specific topic of carburetion?"

In summary, although a number of other analyses were conducted for heuristic purposes, the primary analyses reveal rather graphically that, contrary to prediction, the performance of the experienced auto mechanic teachers was not significantly superior to that of the nonteachers who were instructing comparable students.

Electronics. The initial analysis conducted was the determination of means and standard deviations for each of the classes taught by the 16 electronics teachers and the 16 nonteachers. Results of these computations for pre- and post-test performance of the electronics classes are presented in Table 6.

Table 6. Class Pre- and Post-Test Means and Standard Deviations of 16 Electronics Teacher and Nonteacher Pairs

	Pre-Test			Post-Test		
	n	$\bar{X}$	s	n	$\bar{X}$	s
Teacher	8	8.2	3.3	6	33.3	9.4
Nonteacher	7	5.8	2.2	7	28.4	5.1
Teacher	20	7.2	2.7	21	19.1	7.6
Nonteacher	18	7.6	2.9	16	20.7	4.9
Teacher	15	4.0	2.8	22	19.4	4.7
Nonteacher	17	5.4	2.4	18	20.0	3.8
Teacher	21	7.7	3.2	22	33.0	7.3
Nonteacher	19	6.7	2.4	23	28.9	7.2
Teacher	25	5.9	3.1	26	28.5	9.1
Nonteacher	26	6.4	2.7	25	24.0	8.5
Teacher	23	7.4	2.6	21	23.0	7.6
Nonteacher	21	7.3	2.4	22	20.0	4.4
Teacher	16	6.5	2.0	17	19.7	5.5
Nonteacher	16	5.4	2.1	14	17.5	3.0
Teacher	18	6.0	1.5	19	23.2	5.5
Nonteacher	23	5.6	2.9	21	23.6	7.8
Teacher	27	7.0	2.1	26	23.2	6.2
Nonteacher	27	8.3	2.9	27	27.6	7.6
Teacher	23	9.2	3.2	22	23.3	9.5
Nonteacher	22	7.7	2.7	19	22.2	5.8
Teacher	18	3.8	2.1	20	17.1	5.5
Nonteacher	24	5.0	2.0	26	17.6	3.9
Teacher	14	8.0	2.3	13	24.3	6.0
Nonteacher	7	9.4	3.8	9	25.2	7.8
Teacher	24	8.8	4.0	26	27.6	7.6
Nonteacher	23	7.8	2.7	20	26.9	9.6
Teacher	25	7.9	3.1	25	25.0	8.6
Nonteacher	27	7.5	2.4	26	20.4	6.8
Teacher	21	6.8	2.2	20	24.6	5.0
Nonteacher	23	8.5	3.5	23	21.1	7.5
Teacher	24	6.0	1.8	20	23.4	5.8
Nonteacher	24	5.2	2.2	25	17.6	4.3

Several preliminary analyses were then conducted. The first of these was a calculation of internal consistency (Kuder-Richardson 20) coefficient for the pre- and post-tests. These were calculated on the basis of the entire sample of pupils considered irrespective of (a) class and (b) whether or not they were taught by a teacher or a nonteacher. The actual number of such pupils was approximately 700. The Kuder-Richardson 20 coefficient for the pre-test was .58 and for the post-test was .62.

As in the analysis of the auto mechanics data, intercorrelation matrices were computed in which the post-test results were correlated with a number of variables potentially useful as covariates for the anticipated analysis of covariance. As in the case of the auto mechanics data, the most useful covariates appeared to be (a) pre-test scores and (b) pupils' expressed interest in the field of electronics. The correlations based on class means yielded an r of .47 between pre- and post-test scores, an r of .78 between post-test and expressed interest, and an r of .62 between pre-test and expressed interest.

The next analysis compared pre-test performance of the teacher and nonteacher classes. Because of the potential similarity of performance of pairs of classes, a product-moment correlation coefficient was first calculated to determine the degree of relationship between the teacher and nonteacher pre-test means on the basis of school. The correlation in this instance was .64. Because of this relationship, a correlated t test model was employed to test for significance or difference between the pre-test means of the teacher and nonteacher classes. In Table 7 the means, standard deviations, and correlated t test result for pre-test performance of the 16 electronics teacher and the 16 nonteacher classes are presented.

Table 7. Means, Standard Deviations, and Correlated t Test Result for Pre-Test Performance of Electronics Teacher and Nonteacher Classes

	n	$\bar{X}$	s	t
Teacher	16	6.9	1.5	.05
Nonteacher	16	6.8	1.3	

As can be seen from inspection of Table 7, the pre-test means of the two groups were almost identical. The correlated t test yielded a non-significant t value.

The main hypothesis of the investigation concerned the predicted difference on post-test performance in favor of the classes taught by the teachers rather than the nonteachers. The next analysis involved a comparison of post-test performance of the teacher and nonteacher classes. Initially, the correlation between the pairs was calculated and because it was of some magnitude, that is, .76, again a correlated t model was used for contrastive purposes. In Table 8 the means, standard deviations,

and correlated t test result for the post-test performance of the electronics teacher and nonteacher classes are presented.

Table 8. Means, Standard Deviations, and Correlated t Test Result for Post-Test Performance of Electronics Teacher and Nonteacher Classes

	n	$\bar{X}$	s	t
Teacher	16	24.26	4.5	2.0
Nonteacher	16	22.70	3.8	

As can be seen, the predicted hypothesis was confirmed. The difference between the teacher and nonteacher classes on the post-test, a post-test based on the pre-specified objectives which both groups were instructed to accomplish yielded a small, but significant difference in favor of the teacher group. The mean of the teachers was 24.26 and the mean of the nonteachers was 22.70. The difference was significant, using a one-tailed test, at the .05 level.

In addition to using the gross post-test results as the criterion for testing the validation hypothesis, several other potential ways of looking at the criterion performance were employed. These were identical to the schemes employed in the analysis of auto mechanics criterion performance. Most of these criterion performance estimates confirmed the general result depicted in Table 8, namely, a modest difference in favor of the teacher groups.

Because of the possibility that one of the two groups had been disadvantaged as a consequence of systematic favoritism on relevant variables, an analysis of covariance was calculated in which the aforementioned measures of (a) pupils' pre-test scores (on the 17 item test) and (b) pupils' expressed interest in electronics (1 = low to 5 = high) were used as control covariates. Results of this analysis are presented in Table 9.

Table 9. Analysis of Covariance of Electronics Classes (Teachers Versus Nonteachers) Post-Test Performance, using Pre-Test Scores and Pupils' Expressed Interest in Electronics as Covariates

Source	df	SS	MS	F
Between	1	5.3	5.3	.72
Within	28	205.9	7.3	
Total	29	211.2		

	Control Variables		Criterion Variable	
Group	Pre-Test $\bar{X}$	Interest $\bar{X}$	Unadjusted Post-Test $\bar{X}$	Adjusted Post-Test $\bar{X}$
Teachers	6.9	3.6	24.26	23.89
Nonteachers	6.8	3.5	22.70	23.07

As can be seen from Table 9, the analysis of covariance "washed out" the significant difference between the groups. The adjusted means are extremely similar and the resulting F value of .72 was not statistically significant.

Analysis of variance tests of the difference between affective reactions of pupils, as reflected by responses to the student questionnaire, failed to reveal any significant differences between teacher classes and nonteacher classes. Measures involved in these analyses included responses to such questions as: "After this unit how would you rate your interest in the specific topic of electronics?"

In summary, although a number of other analyses were conducted for heuristic purposes, the primary analyses yielded mixed results regarding criterion performance of the teacher and nonteacher groups. While the correlated t test yielded a one-tailed statistically significant result, the analysis of covariance failed to confirm this difference. The disparity in results is due, of course, to the initial differences on the covariates favoring the teacher group.

DISCUSSION

In looking over the results of two efforts to confirm the initial validation prediction that experienced teachers would be able to outperform nonteachers with respect to the attainment of pre-specified objectives, one might easily become discouraged. The prediction was not borne out by the data. Yet, the same results had been encountered once before with respect to tryout of the previously described Social Science Performance Test. Thus, the investigator is more prepared for the findings. Indeed, having already completed the Social Science

Performance Test Project and interpreted its results in a particular way, it would have been perplexing to have the validation prediction, made several years ago, actually confirmed. Because of the earlier results of the Social Science study, the investigator was in a rather unusual position of "pulling" for nonsignificant differences. It is hoped the reasons for this can be made clear.

The position was taken in the discussion of the Social Science Performance Test results that it was probably naïve to anticipate that experienced teachers would be better able to accomplish behavior changes in learners than nonteachers. The reasons for this were several. In the first place, teachers are not systematically trained to be changers of pupil behavior. Their teacher preparation experiences focus on a variety of other kinds of activities, but rarely address themselves to the question of how pupil growth can be systematically promoted. Further, few teachers are consistently reinforced by their administrators, school system, or community for being particularly skilled in modifying pupil behavior. In other words, the whole educational system, as it is currently set up, does little to foster the experienced teacher's skill in promoting pupil behavior changes. In essence, the experienced teacher is simply not more experienced at modifying learner behavior. This is an unfortunate state of affairs.

One might quickly conclude, and quite erroneously, that performance test approaches to the assessment of teacher competence cannot be validated through approaches such as that described herein and, therefore, should not be employed as indices of teacher competence. This could not be farther from the truth. There has been enough written through the years to support the general conception that the ultimate index of a teacher's proficiency must be his ability to modify his learners. This is his raison d'etre in the classroom. If pupils do not leave a teacher's classroom markedly modified in important ways, the teacher has been unsuccessful, no matter how rhapsodic his lectures, no matter how insightful his discussions, no matter what merit his administrator believes his classroom procedures to possess. This leads to the conclusion that a performance test measure of teacher effectiveness should be, indeed, the only legitimate index of teaching proficiency.

This may suggest that the next step in validating this general assessment approach is to provide both training and, subsequently, significant reinforcers for a group of teachers who would become proficient in bringing about pupil behavior changes. These teachers can be pitted against a group of nonteachers, or for that matter, even teachers who have not been so trained and who have not been, over a period of time, reinforced for modifying learner behavior. The prediction then would be similar to that made several years ago, that the teacher who is demonstrably skilled in modifying behavior of learners should be able to manifest that superiority in other performance test situations.

Another alternative exists, however. Perhaps we should establish the legitimacy of performance tests measures of teacher proficiency by fiat. Maybe we should simply assert that this is the only acceptable

measure of instructional effectiveness, and then move immediately to the manipulation of treatment variables which we hope can produce better results on such performance test measures. In other words, the original position of the writer was that we should, for the moment, circumvent the identification of powerful treatment variables in order that we could demonstrate graphically the validity of performance test measures of teaching skill. It may be more suitable to actually reverse this strategy and skip the validation phase of the research program. In other words, it may not be worth the trouble to validate these performance tests. The writer certainly believes it is possible to demonstrate their validity, but perhaps under the press of today's educational needs, there may be some merit in jumping ahead to the identification of powerful treatment variables.[7]

Having lived with performance tests and related considerations for several years now, the writer concludes the report with a reaffirmation of commitment to these measures as desirable ways of assessing teaching skills. The general strategy employed represents the epitomy of an ends-oriented rather than a means-oriented approach to instruction. As such, performance test measures seem to be the most servicable of those currently available to educators who require legitimate indices of teaching proficiency.

---

[7]The writer is indebted to Richard Schutz, Director of the Southwest Regional Laboratory for Educational Research and Development, for suggesting this alternative.