DOCUMENT RESUME

ED 079 397 TM 002 999

TITLE	Development and Analysis of 18 Experimental Objectives-Based Measures Administered in 1972.
INSTITUTION	Educational Evaluation Associates, Los Angeles, Calif.
SPONS AGENCY	New Mexico State Dept. of Education, Santa Fe.
PUB DATE	Sep 72
NOTE	20p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Academic Performance; Accreditation (Institutions); Data Collection; Educational Accountability; Educational Objectives; *Evaluation Methods; Grade 12; *Program Evaluation; *State Programs; Tables (Data); Test Construction; *Testing
IDENTIFIERS	New Mexico

ABSTRACT

The development, field testing, and analysis of one component of New Mexico's statewide evaluation system, a set of 18 objectives-based tests administered to high school seniors in 56 districts, are summarized. The focus of this component is on providing: (1) information to school districts about the performance of their seniors on certain educational objectives, and (2) a data base to the New Mexico State Department of Education for the purposes of accrediting schools and evaluating State educational programs. Efforts to date are considered successful in view of the following findings: (1) a comprehensive catalogue of objectives has been developed and is ready for final field testing; (2) school personnel, students and community representatives were involved in selecting objectives with which each district is most concerned; (3) good tests were constructed to assess student performance on those objectives; (4) efficient procedures were used in administering these measures in 56 districts to a large, representative sample of seniors; (5) results of the testing indicated how the prototype measures should be modified for subsequent use; and (6) procedures were developed for reporting test results in terms of whether students are performing below, at, or above expected levels. (Author/KM)

A Report To The

New Mexico State Department of Education

on the

Development and Analysis of 18 Experimental Objectives-Based Measures Administered in 1972

prepared by

Educational Evaluation Associates
1072 Gayley Avenue, Suite 4
Los Angeles, California 90024

September, 1972

PREFACE

This report summarizes the development, field testing, and analysis of a set of 18 objectives-based tests administered to high school seniors in 56 districts in New Mexico. The purpose of these activities was to try out one component of a statewide evaluation system. The focus of this component is on providing information to school districts about the performance of their seniors on those educational objectives with which the districts are most concerned while at the same time providing a data base to the New Mexico State Department of Education for the purposes of accrediting schools and identifying the relative strengths and weaknesses of educational programs throughout the state.

The responsibilities for the conduct of this project were shared by the New Mexico State Department of Education and Educational Evaluation Associates (EEA) of Los Angeles, California. This project could not have been carried out, however, without the cooperation of numerous people throughout the state, especially local school administrators, teachers, parents, school board members and, of course, the thousands of students who participated in the testing. The strong support, encouragement, and suggestions by the New Mexico Legislative Finance Committee, School Study Committee, and State Board of Education was also most appreciated.

Table of Contents

Tables, Figures, and Appendices*

*Computer printout Appendices are in separate binders.

Introduction

The Public School Code of New Mexico requires that one-third of all the state's public schools be assessed and evaluated yearly for the purposes of accreditation.[1] The intent of this legislation is to improve educational programs throughout New Mexico and to provide school administrators, teachers, and the public with information regarding the quality of education in their schools and in the state as a whole. A statewide educational evaluation and improvement system is currently being developed to facilitate achievement of these ends.

One component of this system involves the use of a battery of measures based on the educational objectives considered to be most important throughout the state. The process of developing the objectives for this component involved curriculum advisors, school administrators, teachers, parents, community representatives, and students in 27 districts. This activity was concluded in the summer of 1971 and resulted in a set of 153 objectives covering the following four areas:

1. Mathematics
2. Communication Skills
3. Social Studies
4. Science

The relative importance of each of the objectives was then determined in each of the 27 districts that were involved in the development process and in 32 additional districts. Within each district, the kinds of people involved in this step were the same as those who constructed the objectives. The rationale underlying this approach is that school districts should establish a broad base of representation for determining those educational objectives with which they are most concerned. A detailed report of the procedures used for determining the relative importance of the objectives appears in Appendix A and a summary of the objectives chosen appears in Appendix B.

---

[1] Public School Code of the State of New Mexico, 1971 edition, page 7, 77-2-2 W. State Board of Education, Article 2.

Test Construction

The results of the selection of objectives within each of the 59 participating high school districts indicated that there was a "common core" of objectives. In other words, of the total of 153 objectives considered, there was a small subset of objectives chosen as being very important by a vast majority of the districts involved (see Appendices A and B). The objectives in this common core also appeared to reflect current and local concerns, such as the need for greater consumer education. Since the time and funds available for test construction were limited, it was decided to construct a measure for each objective that was included in the common core rather than attempt to develop a measure for each of the 153 objectives reviewed in the selection process.

There were two exceptions to the foregoing procedure. The first exception involved constructing measures for objectives dealing with reading comprehension and writing skills. This was done because these two areas were considered to be of statewide concern (as opposed to just local concern) and were not among those objectives chosen most frequently by the local districts. The second exception involved combining certain objectives into a single test across areas. This was done for those objectives involving understanding graphs, tables, figures, and charts; and for those objectives involving reference skills.

The basic plan of constructing tests for the common core along with the two exceptions noted above resulted in a set of 18 measures covering 25 objectives. The appropriateness of this final set is evidenced by the fact that each school district had at least two of its most important objectives represented in the core of 18 tests (see Appendix B).

Workshops dealing with techniques for writing and editing achievement test items were conducted for SDE staff members. These staff members then worked with 32 districts in the development of test items to assess student performance on the important objectives. Concurrent with this activity, EEA also constructed test items. This simultaneous development was considered (and eventually proved) necessary in order to meet the schedule of a spring testing. The initial pool of items for each objective was synthesized into a 20 item prototype test. This was done by using items and ideas for items developed both by EEA and the districts involved in the test construction process. The prototype tests were reviewed and edited by SDE personnel; revisions were made on the basis of their comments (primarily in lengthening test time and changing emphasis on various measures) and final forms of the experimental tests prepared.

All tests consisted of 20 multiple choice items with four choices per item. A detailed examiner's manual accompanied each test to ensure standardized testing conditions. All students were told that there was no correction for guessing.

Sampling Plan

The purpose of using the 18 measures was to field test the basic procedures associated with the evaluation component emphasizing local concerns and to improve the tests developed for this component. These purposes were translated into the following guiding principles used to determine which tests were administered in which districts:

1. As many students as possible should be tested on each measure.
2. As many districts as possible should be involved in the field testing of each measure.
3. The districts and students involved in the field testing of each measure should be as representative as possible of high school seniors in New Mexico, especially with respect to ethnicity and geographical location.
4. The amount of testing time per student should not exceed one hour.
5. The test administered in a given district should deal with objectives that that district chose as being most important.

One of the chief methods used to achieve these ends was the random assignment of students to tests at the larger high schools. This meant that at some schools several tests could be administered without violating the self imposed restriction on total testing time per student. Table 1 contains a list of the number of districts and students who participated in the testing. Appendix C contains a list of the districts which administered each measure. An inspection of these data indicates the good success achieved in meeting the five criteria above despite the many compromises that had to be made. For example, approximately 1,000 students were involved in the field testing of each measure.

Test Administration

A member of the SDE delivered the test to each school no more than one or two days prior to the testing. All test booklets and answer sheets were collected immediately following the testing and returned to Santa Fe. The tests were generally administered by the classroom teacher according to instructions in a detailed examiner's manual. In some instances, however, the SDE staff person assigned to the district administered the tests.

Data Analysis

The answer sheets were marked with the appropriate district code numbers from which they came and then were sent to EEA for processing. EEA keypunched all the data and conducted the following item and test analysis:

1. An analysis was made of each item's average difficulty (i.e., the percentage of students who answered the item correctly), the percentage of students who chose each of the incorrect alternatives to the item, and the item's correlation with the total score on the test. An inspection of these data indicated that 24 of the 360 items (18 tests X 20 items per test) tried out should be deleted or modified in subsequent use of the measures. The reasons for these deletions and modifications ranged from confusing placement of an item in the test booklet to a general misconception regarding a technical term. A list of the item numbers deleted from each test appears in Table 3. The complete set of item analysis results appears in Appendix D.

2. An analysis was made of each test's reliability by examining its internal consistency (coefficient alpha), i.e., the extent to which students were consistent in whether or not they got the items correct. The results of these analyses before and after each test was revised appear in Table 2 and 3. An inspection of these tables indicates that the tests had an average reliability of about .70, which is considered quite good for such short measures.

3. A set of mutivariate analyses of variance was made of the full 20 item tests as well as of the revised tests in order to examine the extent to which various ethnic-racial groups differed in their performance. A summary of the results of thse analyses appear in Table 4 and the complete set of results (including the average score on each item for each ethnic group) appear in Appendices E and F.

An inspection of these data indicate that Anglos performed consistently better than did the other groups in terms of total score on the tests. This trend was not consistent, however, across all the items in a given test. In other words, for certain items in a given test, the other groups did just as well and sometimes better than did the Anglos. This "interaction" (or discontinuity in performance) between group and test items had a very small but statistically significant impact on student scores (see Table 4). An investigation of the reason(s) for these interactions has indicated that they are not due to an item's sequential position in a test or its general level of difficulty (see Appendix G). With the assistance of EEA, the SDE is now conducting a study in order to determine what other factor or combination of factors may have produced these discontinuities. The results of these investigations will have a major bearing on issues dealing with the degree to which a test item or a total test is "biased" with respect to one or more groups.

Table 1. The number of students in different groups who took each test

Test #	Test Name	Group* Anglo	Indian	Negro	Spanish American	Other	Combi-nation	No Response	Total	Number of Districts Tested
1	MA-APP-01	609	51	10	251	22	25	45	1013	19
2	MA-APP-02	477	43	22	290	22	17	20	891	21
3	MA-#OP-01	466	82	14	253	26	9	36	886	17
4	MA-#OP-03	455	25	13	214	29	29	37	802	17
5	MA-#OP-05	559	68	15	302	49	25	39	1057	20
6	CS-GRA-11	514	161	20	262	29	29	29	1044	21
7	CS-ORA-01	556	36	14	275	38	19	27	965	18
8	CS-REA-01	481	54	15	286	22	12	20	890	14
9	CS-REF-04	574	19	15	243	21	21	23	916	18
10	CS-REF-08	475	26	7	275	47	34	25	889	14
11	SS-ECO-01	497	86	17	360	30	31	18	1039	19
12	SS-ECO-03	553	50	18	306	51	21	17	1016	18
13	SS-NMH-01	587	50	17	309	40	29	34	1066	18
14	SS-RSK-04	548	132	8	305	55	22	30	1100	18
15	SC-ATT-03	627	63	23	260	31	28	22	1054	17
16	SC-LIF-02	531	53	18	292	28	27	33	992	19
17	SC-LIF-03	506	57	10	283	28	23	31	938	20
18	SC-THE-01	549	26	23	293	25	22	10	948	20
Average Number		532.4	60.6	15.5	281.1	32.9	23.5	27.6	973	18
Percent of total		(55%)	(6%)	(1.5%)	(29%)	(3.4%)	(2.4%)	(2.8%)		

*The student was asked to indicate the group(s) to which he belonged. The category "combination" indicated that the student checked more than one group.

Table 2. Characteristics of experimental tests.

Test #	Test Name	Mean	Standard Deviation	Reliability*
1	MA-APP-01	10.47	3.74	.75
2	MA-APP-02	11.71	3.90	.76
3	MA-#OP-01	15.57	3.43	.78
4	MA-#OP-03	11.46	4.50	.83
5	MA-#OP-05	8.81	4.63	.83
6	CS-GRA-11	11.80	3.89	.77
7	CS-ORA-01	13.01	3.21	.68
8	CS-REA-01	13.61	2.97	.64
9	CS-REF-04	14.05	3.40	.72
10	CS-REF-08	9.81	2.88	.50
11	SS-ECO-01	8.47	3.02	.58
12	SS-ECO-03	11.23	3.06	.60
13	SS-NMH-01	9.75	2.85	.51
14	SS-RSK-04	12.87	3.14	.66
15	SC-ATT-03	15.71	2.54	.62
16	SC-LIF-02	12.99	3.41	.71
17	SC-LIF-03	12.35	3.26	.64
18	SC-THE-01	14.01	3.60	.75
	Average	12.09	3.41	.69

*Coefficient alpha

Table 3. Characteristics of revised tests.*

Test #	Test Name	Mean	Standard Deviation	Reliability	Items Deleted
1	MA-APP-01	10.03	3.71	.76	1
2	MA-APP-02				
3	MA-#OP-01				
4	MA-#OP-03				
5	MA-#OP-05				
6	CS-GRA-11				
7	CS-ORA-01	12.86	3.16	.69	19
8	CS-REA-01				
9	CS-REF-04				
10	CS-REF-08	9.07	2.90	.58	3,5,8
11	SS-ECO-01	7.18	2.83	.64	3,8,17,19,20
12	SS-ECO-03	9.98	2.99	.64	3,6,19
13	SS-NMH-01	9.01	2.76	.55	2,12,13
14	SS-RSK-04	12.41	3.13	.69	1,19
15	SC-ATT-03	13.13	2.48	.66	2,3,7
16	SC-LIF-02				
17	SC-LIF-03	11.50	3.25	.69	5,14
18	SC-THE-01	13.22	3.55	.75	8

* In all instances, the reliability of the revised version of each test was the same or higher than the reliability of the original version of that test.

Table 4. Results of the multivariate analyses of variance.

Test Number	Test Name	Effects		
		Items	Groups	s X Groups
1	MA-APP-01	294.71	2.60	1.86
2	MA-APP-02	117.17	2.70	1.41
3	MA-#OP-01	110.13	2.21	1.76
4	MA-#OP-03	156.17	2.08	1.59
5	MA-#OP-05	92.53	3.71	2.35
6	CS-GRA-11	272.11	3.35	3.03
7	CS-ORA-01	346.30	2.72	1.77
8	CS-RE.-01	312.32	2.73	2.16
9	CS-RFF-04	146.08	2.96	1.85
10	CS-REF-08	169.32	2.24	1.98
11	SS-ECO-01	248.06	3.58	2.80
12	SS-ECO-03	246.80	2.99	2.38
13	SS-NMH-01	257.72	2.41	1.82
14	SS-RSK-04	300.62	2.43	1.58
15	SC-ATT-03	288.23	3.41	3.24
16	SC-LIF-02	216.28	3.58	2.45
17	SC-LIF-03	138.01	3.42	2.49
18	SC-THE-01	117.97	3.57	2.19

All F tests were statistically significant beyond the .01 level.

4. One of the most important sets of analyses of the data was an examination of the extent to which the seniors in a given district performed below, at, or above their expected levels on the experimental tests. The rationale underlying this type of analysis is that it is necessary to equate districts prior to making comparisons between them with respect to the performance of their seniors. The equating process is based on the general ability and achievement levels of the students entering the districts' schools so that an appropriate frame of reference is established for looking at the scores of seniors. For example, if the students entering district A's high school are less able than those entering district B's school, it would be expected that on the 12th grade tests the students graduating school A would score lower than those graduating school B. A fair determination of how well a school does with its "raw material" is, therefore, a function of how much better (or worse) it does than other schools relative to the typical relationship between entry skills and graduating performance.

The generally high positive relationship between entry skills and graduating performance is illustrated by Figure 1. The solid diagonal line running from the lower left hand corner to the upper right hand corner of this figure represents the typical relationship between entry skills and performance on one of the 12th grade tests. It can be seen from this figure that the higher the entry skills the higher the scores on the 12th grade tests.

The letters on the figure indicate the average score of the students in a given district. For example, the students in district A had on the average low entry skills and, as expected, a relatively low average score on the 12th grade test; while those in district E also had low entry skills, but had a much higher average score on the 12th grade test. This example illustrates the fact that although the solid diagonal line represents the typical relationship between entry skills and graduating performance, one frequently encounters the phenomena that some districts deviate markedly from this general trend. In order to note when this deviation is greater than one would reasonably encounter by chance, a band has been placed around the line of typical performance. This band is represented in the figure by a dotted line on either side of the solid one. If the performance of the students falls within this band, they can be said to be performing at the expected level (e.g., districts A and B in the figure). If, on the other hand, the students in a district perform outside of this band, then their performance is different than one would expect by chance. For example, the students in district E are performing above expected (i.e., their average 12th grade scores are higher than anticipated relative to their entry skills) while those in district K are performing below expected (i.e., their average 12th grade scores are lower than anticipated relative to their entry skills).

In interpreting the results of this kind of an analysis, it should be remembered that the solid line of typical relationship and the chance band around it was based on the actual average relationship between entry skills and the subsequent performance of the 12th graders. Since the precise nature of this relationship varies somewhat depending upon which entry and graduating skills are assessed, a separate figure must be constructed for each relationship that is examined.

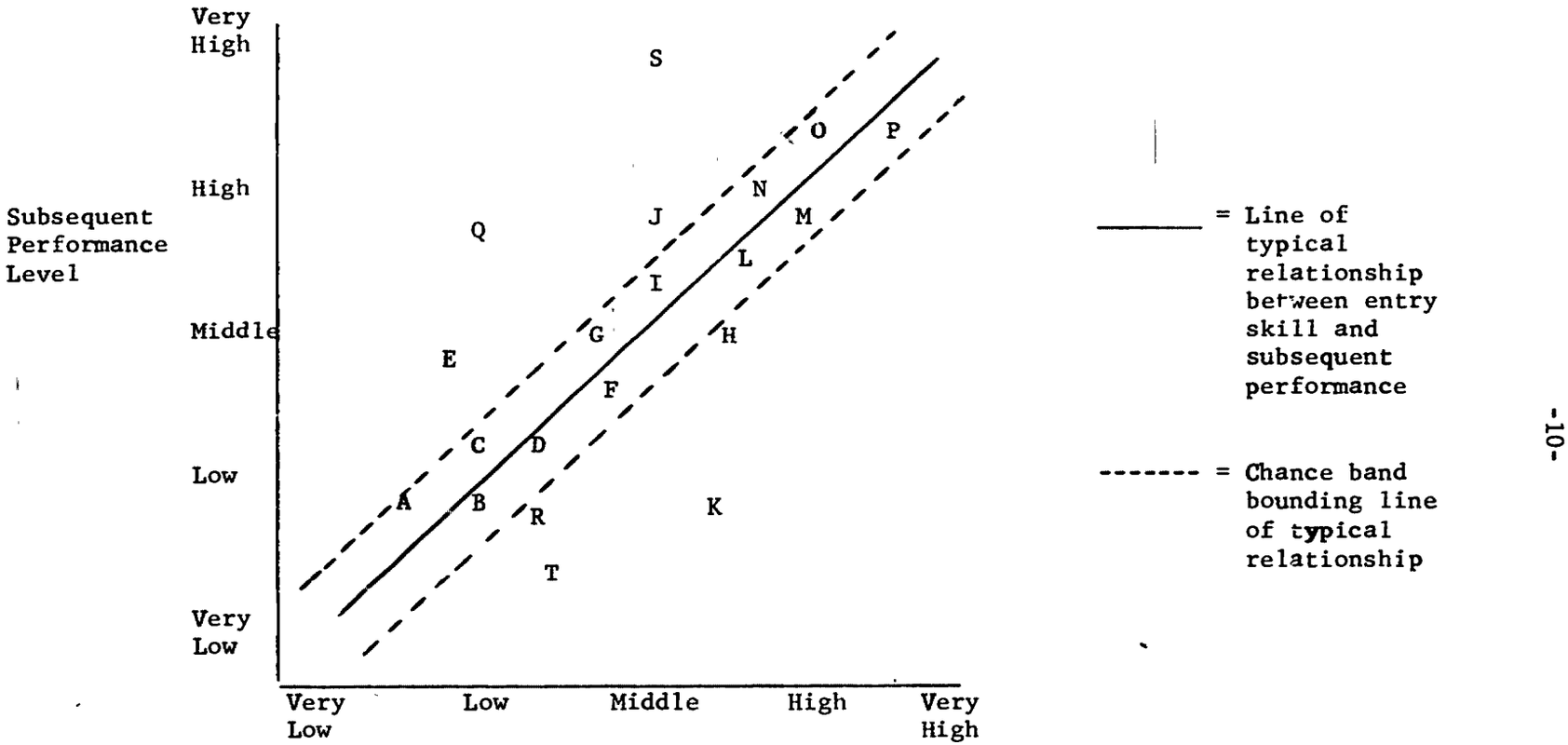

Figure 1. Illustration of the relationship between entry skill and subsequent performance.

In the case of tha analysis of the 18 experimental tests given in 1972, the only good and available indicant of the entry skills of the current 12th graders in a district was the average "total battery" score of 8th graders in that same district on the Comprehensive Tests of Basic Skills (CTBS). The typical relationship between this index of entry skill and the average score of seniors in each district was computed for each of the 18 experimental measures. The correlation coefficients in Table 5 indicate the strength of the relationship between the average CTBS total battery score and the average score on the experimental measure (note: the higher the coefficient, the stronger the relationship; the maximum possible is 1.00).

It should be noted that there are a number of limitations in the use of this kind of analysis for equating schools. Some of these are as follows:

1. Students at one level of entry ability may on the average learn faster than students at another level, e.g., students who start out higher progress faster than those who start lower. This kind of discontinuity is not taken into account well by the use of a straight line of typical performance. On the other hand, there is little evidence that such discontinuities exist in any significant way.

2. The 8th and 12th grade scores were not based on the same students. This was a practical constraint resulting from the fact that very few of the 12th graders tested had taken the CTBS in the 8th grade. This problem will be rectified when the statewide evaluation system becomes operational in that data on entry skills will become available for each class of graduating seniors.

3. The analyses were run on district rather than school averages. This was done for illustration purposes. For operational use of the system, analyses would be conducted on school rather than district averages or on the basis of individual student performance (e.g., each student's 8th grade score relative to his 12th grade score and then report results in terms of school averages).

4. The total battery score on the CTBS was used as the index of entry skills for all the analysis of the 12th grade performance. It may be argued that it would have been more appropriate to use 8th grade mathematics scores for basing expectations of 12th grade mathematics performance, 8th grade science for 12th grade science, etc. This type of differential prediction system should be considered for the operational use of the analysis. From a statistical point of view it probably is not worth the extra effort to do this since it is unlikely that it would lead to different conclusions. The reason for this is the high correlations between all the predictors and measures of subsequent performance, e.g., 8th grade reading scores are generally highly correlated with 12th grade mathematics scores.

Table 5. Correlations between the average score on the experimental 12th grade tests and the average 8th grade total battery score on the Comprehensive Tests of Basic Skills.*

Test #	Test Name	20 Item Test	Revised Test
1	MA-APP-01	.70	.72
2	MA-APP-02	.68	
3	MA-#OP-01	.67	
4	MA-#OP-03	.79	
5	MA-#OP-05	.57	
6	CS-GRA-11	.67	
7	CS-ORA-01	.53	.53
8	CS-REA-01	.86	
9	CS-REF-04	.73	
10	CS-REF-08	.58	.60
11	SS-ECO-01	.79	.78
12	SS-ECO-03	.40	.42
13	SS-NMH-01	.37	.31
14	SS-RSK-04	.57	.56
15	SC-ATT-03	.83	.86
16	SC-LIF-02	.60	
17	SC-LIF-03	.46	.45
18	SC-THE-01	.82	.81

* These correlations were based on the mean district scores. The average number of districts that took each experimental test was 18 (see Table 1). All districts administered the CTBS to their 8th graders.

5. The measures of entry skills and subsequent performance must have sufficient "cellars" and "ceilings" to register performance changes. For example, students with high initial skills should be challenged enough by the 12th grade tests in order for them to show how much they have really learned. If all the students at a given school get a perfect score on the 12th grade test, then one cannot be sure that they were sufficiently challenged. On the other hand, the tests were designed to assess what students should know upon graduation and this kind of directive implies a type of minimal standard which is inconsistent with the notion of a test that "really spreads the students out across the distribution of scores". It is necessary, therefore, to develop measures that provide a balance between these two conflicting needs of wide score range and adherence to the intent of the measure to focus on minimal standards.

It is evident from this list of concerns that the system for equating schools is not perfect. It is, however, much better than no system at all or just the simple comparison of average scores between schools. Further, the system can be modified in the ways indicated above in order to improve the quality of the results obtained. It is suggested, therefore, that this kind of analysis system be incorporated into the statewide evaluation program along with the reporting of actual average scores in each school. In this way, the school and the state will know both the level of performance obtained in a school as well as have some index of the extent to which this performance level is below, at, or above that which should be expected. Appendices H and I contain a listing by test of which districts fell below, at, or above expectancy on the expermental and revised tests. Appendix J contains a listing by districts of their actual performance on each test. These two ways of presenting the data correspond to their subsequent use. In other words, the listing by tests will be used by the instructional services division to identify districts with good or poor programs in certain areas while the listing by districts will be used for the purposes of school accreditation. In addition, Appendix C contains the average score for each district on each experimental and revised test administered in that district. This listing is presented only by test for this report, but in the future, the data in Appendix C would be incorporated into the total score report for each district.

Plans

The activities described in this report regarding the districts' selection of important objectives, and test construction, administration, scoring, and analysis took one year to complete. The next steps in the process of developing this component are as follows:

1. Review and revise the entire system of objectives to ensure its comprehensiveness. This activity was anticipated previously and, thus, provision was made during the process of selecting of objectives for districts to add to the initial total set of 153 objectives.

2. Add objectives for a fifth area, career education, to those already in the system. This area was chosen because of its high importance to the state.

3. Have the 30 districts who have not participated up to this point go through the selection process to determine which objectives they feel are important.

4. Revise experimental measures on the basis of the field test results. This will entail detailed review of the item analysis in order to delete poor items, modify good ones (such as improving the quality of distractors), and write new items in order to have measures of sufficient length (and reliability) to warrant confidence in the results. The time limits for the measures must also be shortened in many instances since there were several reports of students completing the tests long before the prescribed limit was reached.

5. Select or construct additional measures for objectives that should be added to the "common core". It is anticipated that this will be about 5-10 new measures at the 12th grade level.

6. Construct or select measures for the common core of important objectives at other checkpoints, such as grades 6 and 9.

7. Field test and revise all measures and procedures via testing in all districts in the state.

8. Finalize the procedures for including the results of the testing in the school accreditation process.

9. Finalize procedures for making the test administration, scoring, and reporting procedures as cost effective as possible, as well as including them in the overall plans for the statewide evaluation system.

10. Provide reports of results to the instructional services division to ensure the appropriate use of the data for the purposes of improving instructional programs, e.g., as part of the bases for the assignment of MAP personnel to districts.

All of the foregoing activities are planned to be completed by September, 1973. At that point, the evaluation component emphasizing local concerns will be considered fully operational.

Summary and Conclusions

This report summarizes the initial development and field testing of one component of New Mexico's educational evaluation and improvement system. The focus of this component is on providing information to the state and to local districts regarding student performance on those objectives with which each district is most concerned. The results of the efforts to date to develop this component have been quite successful. This conclusion is supported by the following findings:

1. A comprehensive catalogue of objectives has been generated, reviewed, revised, and expanded. It is now ready for final field testing.

2. School personnel as well as students and community representatives were involved fully in selecting objectives with which each district is most concerned.

3. Good tests were constructed with the help of New Mexico teachers to assess student performance on those objectives chosen most often as being important.

4. Efficient procedures were used in the administration of these measures in 56 districts involving a large representative sample of the state's seniors.

5. The results of the testing indicated how the prototype measures should be modified for subsequent use.

6. Procedures were developed for reporting test results in terms of whether students were performing below, at, or above their expected levels.

These findings indicate that the procedures used form a basis for a practical, legitimate, and realistic approach to educational accountability.