AN EXPERIMENT IN THE USE OF MACHINE-SCORED ANSWER SHEETS WITH FIFTH-GRADE PUPILS

H. George Loiselle
Dade County

Introduction

For a number of years there have been requests by teachers in Dade County to use separate machine-scored answer sheets in the achievement testing of elementary pupils. Small pilot studies, using the elementary battery of the California Achievement Tests and involving the test-retest method, first with the separate answer sheet edition and then with the consumable booklet edition at the fifth and sixth grade level, indicated a significant difference in student performance in favor of the expendable booklet in all areas measured.

In the spring of 1958, after several months of study and evaluation of various achievement tests at the elementary level, the elementary testing committee, composed of two members from each of the seven zones, recommended that the Stanford Achievement Tests, which measured more of our instructional objectives than other tests on the market, be administered in the fall of 1958 to all pupils in grades three through six. As these tests, at the fifth-and-sixth-grade level, utilized the method of recording answers by blackening spaces between dotted lines similar to that used on separate answer sheets, it was decided to experiment in the sixth grade in order to determine whether separate machine-scored answer sheets could be used at this grade level without penalizing pupils. In this experiment it was found that there was no significant difference between the performance of the control and the experimental groups and it was concluded that separate answer sheets could be used with sixth-grade pupils.

At the time of undertaking this experiment at the sixth-grade level, teachers were told that a similar experiment would be conducted the following year (1959) at the fifth-grade level in order to determine whether separate answer sheets could be used at this lower level.

Organization of the Experiment

By means of memoranda and a radio broadcast, the Supervisor of Testing instructed the principals and teachers in the method of administering the test and determining control and experimental groups within each school.

Each teacher was instructed to select his control and experimental group by checking every other name on an alphabetical list of his class starting with the first or second name. The names with the check were designated as the experimental group. These pupils were administered the Stanford Achievement Test, Intermediate Battery Partial Form KM, and used separate machine-scored answer sheets. The other pupils became the control and were administered the Stanford Achievement Test, Intermediate Battery Partial Form K, the consumable edition which had previously been selected as part of a county-wide testing program. Approximately 11,000 fifth-grade pupils were tested.

In order to eliminate any variations in the administration and timing of the test, it was decided that the Supervisor of Testing would personally administer the tests over the school system's FM radio band to the control as well as the experimental group. Each principal was to designate those teachers who would proctor the control groups and those who would proctor the experimental groups within his school. In order to orient the pupils in following directions given by radio as well as to familiarize them with the method of recording answers, each school was asked to utilize its public address system for a short orientation period. Materials on the proper manner of using separate answer sheets and recording answers were provided the schools for this purpose.

Each teacher scored the Form K test taken by his students; the separare answer sheets for those who took form KM were returned to the Testing Department where they were scored on IBM test scoring machines. These answer sheets were then returned to the teachers who converted raw scores to grade equivalent scores.

Each school prepared data sheets for control and experimental groups. One set was made out in alphabetical order for all control pupils within the school and another set made out in a similar manner for the experimental group. Data sheets, all scored separate answer sheets, and all scored consumable booklets were returned to the Testing Department. Only the scores of those pupils who had taken all sections of the test were used as a basis for sample selection. At the Testing Department all consumable booklets for each school were placed in alphabetical order for the entire county. Every tenth booklet was drawn in order to obtain the control sample for the statistical study. Each consumable booklet thus selected was then rescored by trained clerical workers who made out new data sheets for the county control sample. Data for the experimental group were obtained by selecting every tenth name on the alphabetical list of the experimental pupils within each school of the county; separate data sheets were also made out for this sampling of the experimental group. Raw scores were used in the statistical work.

Statistical Procedures and Findings

In the samples mentioned above, there were 553 cases in the control group and 534 in the experimental group. Means and standard deviations were computed for each of the groups in the six different areas of achievement measured by the test; paragraph meaning, word meaning, spelling, language, arithmetic reasoning, and arithmetic computation. The t for significant differences between groups was applied.

As shown in Table 1, the control group obtained a significantly higher mean than did the experimental group in paragraph meaning, spelling, arithmetic reasoning, and arithmetic computation. In word meaning and language the differences were not significant. However, in both instances, the direction of the differences is the same as that found in the areas for which significance is attained. When all the differences noted in the six areas studied are summarized into an over-all composite test of significance as shown in Table 2, they become significant at the .001 level in favor of the control group which used the expendable booklet edition of the test.

Table 1

Significance of the Differences in Six Areas of the Stanford Achievement Test, Intermediate Battery-Partial

Sub-Test	Control (N=553)		Experimental (N=534)		Difference in Means	t
	Mean	S.D.	Mean	S.D.		
Paragraph Meaning	22.99	9.72	20.73	9.57	+2.26[a]	3.90***
Word Meaning	24.05	10.99	23.02	11.11	+1.03	1.55
Spelling	34.32	12.22	30.79	13.30	+3.53	4.61***
Language	23.16	14.29	21.55	14.01	+1.61	1.90
Arithmetic Reasoning	21.96	8.34	20.84	8.52	+1.12	2.21*
Arithmetic Computation	15.90	6.09	14.81	6.40	+1.09	2.91**

[a]A plus value denotes a difference in favor of the control group.

*Significant at .05 level.

**Significant at .01 level.

***Significant at .001 level.

Table 2

Composite Test of the Significance of the Differences Obtained in the Six Areas Studied

	t	P	$-\log_e P$	df
t_1	3.90	.0000962	9.24908	2
t_2	1.55	.1212	2.11196	2
t_3	4.61	.0000068	11.89859	2
t_4	1.90	.0574	2.85771	2
t_5	2.21	.027200	3.60454	2
t_6	2.91	.003600	5.62682	2
			X^2= 35.34870*** with	12 df

*Significant at .05 level.
** Significant at .01 level.
***Significant at .001 level.

Results

1. In the areas of paragraph meaning, spelling, arithmetic reasoning, and arithmetic computation, there is a statistically significant difference between the means of the groups in favor of the control group which recorded its answers in the test booklet.

2. Although there is no statistically significant difference between the two groups in word meaning and language, the difference in each case approaches significance and is in the same direction as that found in the areas in which there is significance.

3. A composite test of significance, the chi-square test, indicates a significant difference in favor of the control group which recorded its answers in the test booklet.

Conclusions

1. Inasmuch as the samples were randomly selected, it seems to be logical to conclude that the use of separate machine-scored answer sheets with fifth-grade pupils apparently penalized the pupils by lowering their scores significantly (the average difference is from one to three-and-a-half months) even though the method of recording answers in the booklet is similar to that used in recording answers on the separate machine-scored answer sheet.

2. Further studies should be made to determine whether a period of orientation, which would involve several opportunities to record answers on separate machine-scored answer sheets prior to testing, might eliminate the differences noted in scores between the two groups.