EFFECT OF IMMEDIATE SELF-CHECK UPON RESPONSE TO OBJECTIVE TESTS

by

WAYNE WESLEY McINTOSH

B. S., Kansas State College
of Agriculture and Applied Science, 1938

A THESIS

submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

Department of Education

KANSAS STATE COLLEGE
OF AGRICULTURE AND APPLIED SCIENCE

1939

TABLE OF CONTENTS

INTRODUCTION

The purpose of this study was to determine what effect the use of Chemo-Score self-checking answer sheets has upon performance in objective tests. Magnitude of scores, reliability of the test, and validity of the test may be affected. Chemo-Score answer sheets may also have a different effect upon different levels of ability, and upon different levels of emotional stability.

RELATED LITERATURE

It has already been shown that self-checking answer sheets have a marked effect upon learning when used with guide questions.

Peterson (13) found that by using the Chemo-Score answer sheets in guided study in a course in general psychology, a marked increase in achievement was shown by the examination grades. On the average, groups who used this feature of the device in reading gained from 2.4 to 3 times as much in information as did those who used only questions as guides.

Marx (9) found a statistically significant difference in favor of Chemo-Score answer sheets as a study guide for grade school students.

Self-checking answer sheets have been used by Fleenor (4) as instruction devices in correspondence courses. By using Self-Instructor lessons instead of written lessons, students did better work in a correspondence course in general psychology.

Chemo-Score answer sheets serve as a study guide and motivate the learner's reading of assignments and his thinking about the subject.

In a recent experiment by McIntosh and McIntosh[1] carried on at the Manhattan High School, a small increase in reliability and a statistically significant increase in validity for a test was found by using Chemo-Score answer sheets.

In these previous experiments, the phenomenon of centripetal drift was not taken into account. Sir Francis Galton drew attention to the fact of regression a half century ago. He pointed out the tendency to regress toward the mean of a variable. In 1928, Peterson (12) pointed out the constant error in predicting for the extremes. Jung (6) states, "Whenever there is a perceptible degree of chance error, the centripetal drift must be at work." Upon reexamination of a group, both extremes are found to gravitate in the direction of the mean of the whole group. Johnson and Cobb (5) found, in a study of the value of

1. Effect of the Chemo-Score answer sheet. Unpublished paper. Kansas State College of Agr. and Appl. Science. 1939.

drivers' clinics, that any group selected according to their accident rates in one period will tend to regress toward the average of the population from which they were selected in the following period. It therefore appears that, if a group is divided into different performance levels and the performance of these sub-groups predicted, a correction should be made for centripetal drift.

Bingham (2, p.262) has devised a table to predict the most probable standard score on a second test from the standard score on a previous test when the correlation between the two tests is known. This table was used to predict the most probable standard scores on the second test in this experiment.

A test, to serve the purpose for which it was intended must have two qualities. It must have a high degree of reliability and it must be valid.

The reliability of a test may be defined as the accuracy with which it measures what it does measure. There are various ways of calculating the reliability of a test. It may be calculated by giving comparable forms of the test to the same individuals and correlating the scores made on the two forms. The objection to this method is that the forms may not measure exactly the same thing. The reliability of the test may be calculated by repeating the test

after some time and correlating the scores made. This is known as the "retest method" (Kelley, 7). The student will probably make the same mistakes he did the first time. Also, learning may take place between the times the test was taken first and second. The reliability may also be calculated by scoring the odd and even numbered questions separately and correlating the odd scores with the even scores. The reliability of the lengthened test may then be calculated by the Spearman-Brown prophecy formula. This formula will give a very close approximation to the reliability of the total form, as reliability of split halves will in general be approximately equal (Dunlap, 3). This method was used in this experiment.

"Validity is the extent to which a test does measure what it purports to measure" (Holzinger, 8). If a test is an intelligence test, it should measure intelligence and not some other trait or combination of traits. Validity is measured by correlating the test scores with scores on some criterion. Competent judgment and recognized tests in the field are also used as criteria. "In the case of recent school achievement tests, not only the tests as a whole, but every item separately in them has been selected because of its correlation with school records of achievement" (Kelley, 7).

When self-checking answer sheets are used with objective tests, the person taking the test observes his mistakes immediately.

Thorndike and Woodyard (16) found that there was some impairment of ability to learn when a mind has been suffering from repeated frustrations. Sixteen tasks at the end of a hard series of problems were 76.2 per cent correct while the same tasks at the end of an easy series were 81.2 per cent correct. Thorndike and Woodyard state:

> It is a matter of common knowledge that a mind which for any reason becomes engaged in an activity and finds itself repeatedly and persistently failing therein, is impelled to intermit or abandon it. The person does abandon it unless this impulsion is counterbalanced by some contrary force, such as the hope of a turn of the tide toward success, or an inner sense of worth from maintaining the activity, or a fear that worse will befall him if he stop.

Peterson (11) found, in a study of rational learning with children, that slight errors caused confusions and consequent failure to avoid guessing answers the subject knows are wrong.

There is no literature available to show how students react to mistakes as shown them by self-checking answer sheets. One of the purposes of this experiment was to determine the reaction to mistakes.

Perfo-Score and Chemo-Score answer sheets were used in this experiment. The Perfo-Score answer sheets are uniform

answer sheets having a list of numbers to represent the questions or test items, every number being followed by a series of five numerals which represent the alternative answers to the questions, and every sheet having a hole punched at each of two corners so they may be stacked on a punchboard for scoring (Peterson, 10).

Chemo-Score answer sheets are similar in appearance to Perfo-Score answer sheets, but the answer spots are printed with moisture-sensitive inks so that the correct answer spots turn blue and the incorrect answer spots turn red when moistened (Peterson, 14).

MATERIAL AND METHOD

Two groups, totaling 154 students, were given Tests IV and V of the Group Test of Mental Performance by J. C. Peterson and H. J. Peterson, one -half the students being given Test IV first and then Test V, the other students being given the tests in reverse order. Chemo-Score (self-checking) answer sheets were used with Test IV, and Perfo-Score (not self-checking) answer sheets were used with Test V. There was no time limit and each student completed both tests, beginning the second immediately after completing the first. After completing each question, the student recorded on the margin of the answer sheet the time

in minutes and seconds as indicated by a laboratory clock. This became a cumulative time record, and the time for any one question could be calculated.

Most of the students taking the test were freshmen, but there were also sophomores, juniors, and seniors. These students were from four divisions of the college.

The odd and even questions on the tests were scored separately. The scores on the odd and even questions were correlated for each test and the reliability for the lengthened tests was calculated by the Spearman-Brown Prophecy formula.

The scores made on the two tests were also correlated, and although the tests are comparable, this correlation could not be used as a reliability coefficient because different procedures were used in giving the tests.

The validity of each test when given first, and when given second, was calculated by correlating the test scores with first semester grades. The validity was also calculated for each test with all scores included.

The mean and the standard deviation was calculated for each test when given first, when given second, and for the first and second groups combined.

The raw scores for each test were converted into standard scores. The most probable standard scores on the

second test were predicted from the standard scores on the first test by the use of Bingham's table (2, p.262). The predicted scores were then compared with the standard scores for the second test.

A study was made of the total mean time required for each test, the time required for the question on which the first mistake was made, each of the four questions immediately preceding the first mistake, and each of the four questions immediately following the first mistake. The percentage of mistakes for the four questions following the first mistake was also calculated for each test to show the reaction to mistakes as indicated by the Chemo-Score answer sheets. For this, only those papers that had no mistakes on the first four questions could be used. This same procedure was followed using the first mistake from the 81st to the 90th question, to find if the effect is different near the end of the test. This group was also divided into equal fifths to find the effect on different levels of ability, and the average time for each quintile was found for the nine questions studied.

RESULTS

The reliability of the tests as found by correlating the odd and even scores, and stepping up by the Spearman-Brown Prophecy formula was .930± .0074 (P.E.) for Test IV (Chemo-Score) and .915 ± .0090 (P.E.) for Test V (Perfo-Score). The critical ratio of the difference was 1.25. The correlation between the two tests was .842.

Table 1 shows the validity of the tests as found by correlating the test scores with first semester grades.

Table 1. Validity of Tests Used.

Test IV given first (Chemo-Score)	.461
Test IV given second	.380
Test IV groups one and two combined	.400
Test V given first (Perfo-Score)	.419
Test V given second	.383
Test V groups one and two combined	.407

The validity coefficients may seem rather low but it must be remembered that the students ranged from freshmen to seniors and were from four different divisions. In each case, the validity of the second test was lower than the validity of the first test.

Table 2 gives the means and the standard deviations of the tests.

Table 2. The Means, Standard Errors, and Standard Deviations of the Tests Used.

	Mean	Standard Error		Standard deviation
		Mean	Estimate	
Test IV*	66.62			12.72
Test IV**	70.64			12.52
Test IV***	68.62	.98	3.09	12.13
Test V*	65.42			12.94
Test V**	66.00			11.08
Test V***	65.60	.98	3.37	12.10

* Test given first
** Test given second
*** All scores on the test concerned

The difference between the means of the two tests was three points. However, there is an actual difference of 2.95 between the tests. There is therefore no significant change in the magnitude of the scores.

The raw scores for each test were converted into standard scores. The most probable standard scores for the second test were then predicted from the standard scores on the first test. In every instance, except the third quintile of the group taking the Chemo-Score test first, the actual scores were larger than the prediction. This is accounted for by the fact that the raw scores on the second test average about two points higher than the raw scores on the first test. The following figures show, by quintiles, the actual scores minus the predicted scores in terms of

points of raw score.

Table 3. Actual Scores Minus Predicted Scores.

Quintile	Perfo first (Chemo predicted)	Chemo first (Perfo predicted)
5	1.875	2.66
4	2.375	.53
3	3.133	- .20
2	1.625	.66
1	1.125	3.00

The predictions for the extremes were too high from the group taking the Perfo-Score test first and too low from the group taking the Chemo-Score test first. This indicated there was some other factor at work, presumably the Chemo-Score answer sheets.

To find the effect of the Chemo-Score answer sheets, the predicted scores were subtracted from the actual scores for the group taking the Perfo-Score test first, and the actual scores were subtracted from the predicted scores for the group taking the Chemo-Score test first, i.e., the Perfo-Score standard scores, actual or predicted were subtracted from the corresponding Chemo-Score standard scores. The results in terms of points of raw score are given in Figure 1.

This graph shows that the students in the middle ranges receive the most benefit from the Chemo-Score answer

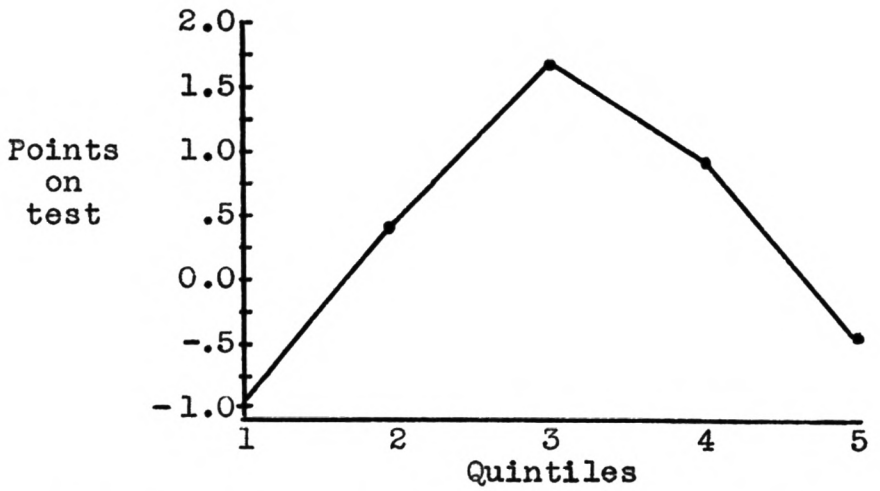

Fig. 1. Effect of Chemo-Score answer sheets.

sheets. This result seems reasonable. Those in the highest quintile have little room for improvement. The best students do not have a chance to show their ability if a test is too easy. Those in the middle ranges are motivated by knowing when they have the correct answers, and those in the lowest quintile may be discouraged by making so many mistakes.

The average time for Test IV with the Chemo-Score answer sheets was 20 per cent longer than the average time for Test V with the Perfo-Score answer sheets. It takes somewhat longer to mark the Chemo-Score answer sheets; also the increase in time required may be caused by more careful consideration of the questions.

The time consumed for the question on which the first mistake was made, each of the four questions immediately

preceding, and each of the four questions following the first mistake, was found for each test. The average time was found for each question and the results tabulated. No significant differences were found for the two methods of giving the tests. However, on both tests more time was taken for the question on which the mistake was made. No difference was found in the number of errors following the first mistake. However, when similar tabulations were made for the first mistake from the 81st to the 90th question, in addition to a slowing up on the question on which the mistake was made for both groups, the Chemo-Score group used more than average time for their group in completing the question following the first mistake in the above mentioned range. On the same questions, 17 per cent more mistakes were made by the group taking the Perfo-Score test. Evidently when those in the Chemo-Score group found they had made a mistake, they would slow up and avoid making a number of consecutive mistakes.

Figure 2 shows the percentage of the average time for each group taken by each quintile for the nine questions near the end of the test.

The lowest quintile in the Chemo-Score test and the lowest two quintiles in the Perfo-Score test took less than the average time for their respective groups in completing

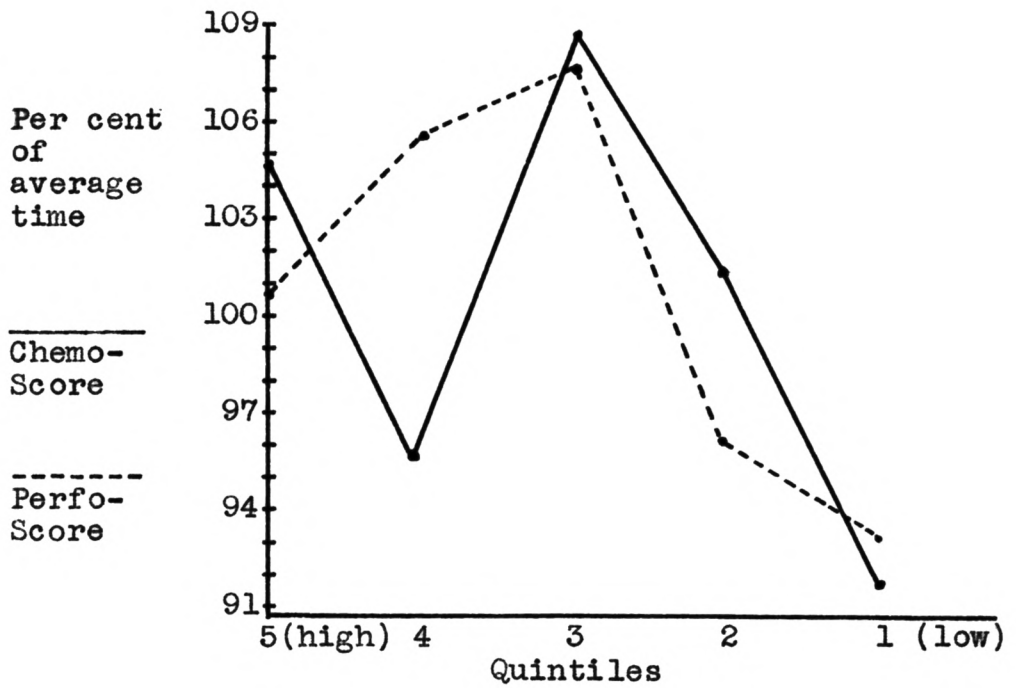

Fig. 2. Percentage of average time taken by each quintile.

the nine questions studied. This probably indicates that these students were reading the questions over hastily and guessing at the answers.

The graph shows that the fourth quintile of the Chemo-Score group takes more time relatively than does the fourth quintile of the Perfo-Score group. The Chemo-Score group for this quintile made comparatively fewer errors. It therefore appears that the lowest two quintiles could improve their scores by taking more time on the test and giving the questions more careful consideration.

CONCLUSIONS

A slightly higher reliability was found for the test with which Chemo-Score answer sheets were used.

No change in validity was found in this experiment. This, however, may be masked by different effects at different levels of performance which were found in this study. A previous experiment showed a statistically significant increase in validity by using Chemo-Score answer sheets.

For this group, and with this test, Chemo-Score answer sheets had no effect on the magnitude of the scores for the group as a whole. It does not necessarily follow, however, for a different test, or for a group on a different level, that the same results would be found.

In each case, the standard deviation on the second test was smaller than the standard deviation on the first test. This would indicate that there is a narrowing of the range for the group on the second test.

Chemo-Score answer sheets appear to have a motivating influence, especially for students in the middle ranges. These answer sheets would be especially helpful in objective tests of subject matter, where it is desirable to have the student know the correct answer, so that he will not form mistaken ideas from the test.

ACKNOWLEDGMENTS

To Doctor J. C. Peterson, the author wishes to express his appreciation for encouragement, advice, and helpful criticism, and also for the use of the tests and answer sheets.

He also wishes to thank Doctor J. C. Peterson, Doctor O. W. Alm, and Doctor R. C. Langford for their cooperation in securing subjects for this study, and the Department of Physics for the use of the laboratory clock.

REFERENCES

(1) Adkins, Dorothy C. and Toops, Herbert A.
Simplified formulas for item selection and construction. Psychometrika, September, 1937.

(2) Bingham, Walter Van Dyke.
Aptitudes and aptitude testing. New York. Harper & Brothers. 1937.

(3) Dunlap, Jack W.
Comparable test and reliability. Jour. Educ. Psychol. 24:442-453. September, 1933.

(4) Fleenor, H. C.
The self instructor method in correspondence courses. Kans. Acad. Sci., Trans. 36:166-171. 1933.

(5) Johnson, H. M. and Cobb, Percy W.
The educational value of drivers' clinics. Psychological Bul. 35:758-766. 1938.

(6) Jung, F. T.
Centripetal drift: a fallacy in the evaluation of therapeutic results. Science, NS 87:461-462. 1938.

(7) Kelley, Truman Lee.
Interpretation of educational measurements. New York. World Book Company. 1927.

(8) Holzinger, Karl J.
Statistical methods for students in education. Boston. Ginn and Company. 1928.

(9) Marx, Lawrence Norbert.
Immediate check-up devices and their value in learning. Unpublished Thesis. Kansas State College of Agr. and Appl. Science, 1933.

(10) Peterson, H. J. and Peterson, J. C.
The self-instructor and tester; a learner's guide. Published by the Authors. Manhattan, Kansas. 1931.

(11) Peterson, Joseph.
The twenty-third meeting of the southern society for philosophy and psychology. Amer. Jour. Psychol. 40:515-520. 1928.

(12) _____
Tentative norms for a simplified rational learning test for children eight, nine, and ten years of age. The Va. Teacher. September-October, 1922.

(13) Peterson, J. C.
The value of guidance in reading for information. Kans. Acad. Sci., Trans. 34:291-296. 1931.

(14) Peterson, J. C., Peterson, H. J., and Higginbottom, H. H.
Further economies and new values in objective testing. Kans. Acad. Sci., Trans. 36:176-180. 1933.

(15) Richardson, M. W. and Adkins, Dorothy C.
A rapid method of selecting test items. Jour. Educ. Psychol. October, 1938.

(16) Thorndike, Edward L. and Woodyard, Ella.
The influence of the relative frequency of successes and frustrations upon intellectual achievement. Jour. Educ. Psychol. 25:241-250. 1934.