Gain vs. posttest scores as predictors of future performance
by Edward Roland Cimler

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE in Psychology
Montana State University

Abstract:
Predictors of future performance based on information derived from our educational system (such as college grades) have historically played an important role in the lives of individuals. One such predictor, the corrected gain score, which has received relatively little attention from psychologists, was compared for effectiveness to the posttest score. Students ($n = 90$) completed a unit of programmed instruction with pre- and posttests, and after one week were administered a retention test. Both posttest and gain scores correlated highly and significantly ($p <.01$) with retention test scores ($r = .84$ and $r = .65$ respectively), but posttest was shown to be significantly better than pretest as a predictor of retention ($p < .001$). Multiple regression techniques using SAT or ACT scores, pretests, and gain scores failed to significantly increase the predictability of retention test scores over posttest scores alone. Further, those students with the highest initial capability as measured by the SAT or ACT were shown to achieve the highest gains ($p < .05$). Results and implications for future research were discussed.

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at Montana State University, I agree that the Library shall make it freely available for inspection. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by my major professor, or, in his absence, by the Director of Libraries. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature *Edward Roland Amble*

Date *November 30, 1973*

GAIN VS. POSTTEST SCORES AS PREDICTORS OF FUTURE PERFORMANCE

by

EDWARD ROLAND CIMLER

A thesis submitted to the Graduate Faculty in partial fulfillment of the requirements for the degree

of

MASTER OF SCIENCE

in

Psychology

Approved:

Head, Major Department

Chairman, Examining Committee

Graduate Dean

MONTANA STATE UNIVERSITY
Bozeman, Montana

December, 1973

Acknowledgements

My sincere thanks are offered to my thesis advisor, Dr. Peter W. Dorfman, for his invaluable assistance in the preparation of this paper. Gratitude is also extended to Drs. M. Paul Willis and Douglas L. Herbster for their guidance and understanding.

Finally, my deep appreciation is extended to my wife, Lynne, for her assistance in data collection and typing of the manuscript.

Table of Contents

List of Tables

Abstract

Predictors of future performance based on information derived from our educational system (such as college grades) have historically played an important role in the lives of individuals. One such predictor, the corrected gain score, which has received relatively little attention from psychologists, was compared for effectiveness to the posttest score. Students (n = 90) completed a unit of programmed instruction with pre- and posttests, and after one week were administered a retention test. Both posttest and gain scores correlated highly and significantly ($p < .01$) with retention test scores (r = .84 and r = .65 respectively), but posttest was shown to be significantly better than pretest as a predictor of retention ($p < .001$). Multiple regression techniques using SAT or ACT scores, pretests, and gain scores failed to significantly increase the predictability of retention test scores over posttest scores alone. Further, those students with the highest initial capability as measured by the SAT or ACT were shown to achieve the highest gains ($p < .05$). Results and implications for future research were discussed.

Introduction

It is becoming increasingly difficult for individuals to cope with the daily requirements imposed by our highly complex technological society--requirements imposed by industry, education, and even mundane affairs such as the use of our mass transportation facilities. The primary means of training people for participation in this society is and has been our formal educational system. Other than general success or failure on the job at some future time, these training efforts have received little evaluation of the various components as they relate to future performance. Because our formal educational system is the main means of entry into the different strata of society, it seems we should be familiar with those aspects which contribute most to future success.

Our normal mode of prediction now is to test an indidivual after a training program in an effort to determine what knowledge he has gained from the course. This knowledge will presumably affect his future performance in that area. A series of grades obtained on tests measuring performance after instruction (posttests) are usually averaged, resulting in an assignment of competency which is also a prediction of future performance. For instance, a student receiving a "B" in an introductory chemistry course is considered to know more (and possess more skills) about the subject than a student receiving a "D" in the same course. The "B" student is also expected at a future time to perform better in other more advanced chemistry courses than the "D" student. More broadly, an accounting major receiving his bachelor's degree with a 3.8 overall average is expected to become a better accountant than one with a 2.5 overall average, and, on the basis of that predictor, will probably obtain a more desirable position if other factors are equal. Examples of other rewards based on educational predictors would be entrance to graduate school or even promotion on the job at some future time. Because of the tremendous importance of predictors used in the educational system, one which has received relatively little attention (for a number of reasons) will be examined in this thesis--the corrected gain score.

Historical Background of the Educational Process

A brief background of the evolution of our educational process seems in order. Most authorities (e.g., Cubberley, 1920) place the beginnings of our present system in early Greece. There education, along with birth into the proper class, was a means of training for entry into citizenship and acceptable religious standing.

Later, as Europe developed, education became primarily a means to spiritual enlightenment. These religious connotations of education spread to the first New England schools, but as the states increased in number, the additional concept of education for the good of the state evolved. This concept suggested that since education is largely financed by the state (or country) that training should be planned so as to maximize benefits to the state. The first laws for compulsory education in this country were passed in Massachusetts in 1642 and 1647. These laws made mandatory that children be taught to read in order to understand the principles of religious and capital law.

By the late eighteenth century, individual rights were largely accepted by the leaders of the country. Thomas Jefferson voiced public sentiment in his preamble to the Declaration of Independence when he wrote "that all men are created equal...." A devotee of French enlightenment, he was probably echoing Rousseau's thought that all men should be considered morally and lawfully equal to counteract unequal strengths and intellects (Cubberley, 1920). In any event, the concept of mass equality came to be generally accepted by the nation and by the close of the nineteenth century, popular thought was that education should be for the state and individual rather than for religious purposes.

With industrialization came the necessity for large amounts of inexpensive manpower. These needs were filled to a great extent with children, and the demand on the schools became to teach technical and industrial skills. Because of child labor laws, compulsory education legislation, and the passage of the initial manpower demands of industrialization, the emphasis of schools returned to the more formal and traditional "education". However, our largely industrial society had demonstrated its power to influence the curricula of the schools.

Research Problem

Today, potential employers are increasingly asking what the product of our educational system can do, rather than how much or what type of education he received. Our schools have responded in part by recent emphasis on individualized instruction, the efficacy of which has been well documented (Alba & Pennypacker, 1972; Atkinson & Fletcher, 1972; Ferster, 1968; Keller, 1968; Witters & Kent, 1972). With individualized instruction the behavioral objectives for a course are specified and the student masters the material at his optimal pace.

Programmed instruction is one type individualized instruction. Subject matter is presented in small units called frames; all the frames together are called a program. In linear programs, all the necessary knowledge for proceeding through the programs is gained from the preceding frames (branching programs offer additional clarifying frames for those students having difficulty with certain concepts). In this manner, material may become more complex as the program progresses. After thinking or writing the answer to a frame (usually a short statement with a key word omitted), the correct answer is seen immediately afterwards. Advantages of this type of instruction are immediate feedback after each response and the ability to proceed at a comfortable pace without regard to the speed of others. The programs themselves are normally updated based on the comments of students having completed the program.

A great amount of research in the areas of individualized instruction appears to be the statistical comparison of one group trained with some form of individual instruction to another group trained with a different (usually traditional) teaching methodology (e.g., Johnston & Pennypacker, 1971; Suchett-Kaye, 1972). These studies generally show equality or superiority of individualized instruction techniques over traditional methods for a specific factor, such as speed of learning or student satisfaction. Although the major premise of individualized instruction is that individuals differ in capabilities and histories, it is not completely clear how such differences affect mastery of a training program. Moreover, individual differences frequently appear to be entirely ignored as legitimate factors contributing to success or failure of training (Bereiter, 1969; Newsome, Eischens, & Looft. 1965). Even as early as 1910, Carlton complained of lack of consideration of individual differences:

> "No matter whether the child is well or ill, over or under-worked, naturally quick or slow of comprehension, or whether he is or is not aided at home by the parents; the system operates like clock-work in the vain attempt to produce a ficticious, although much talked about, average child [p. 75]."

It is almost a certainty that any group of students will differ considerably in the amount of knowledge and preparation they bring to a given course. Normally, even with individualized instruction, final performance alone is used as a measure of success of the training, and is, presumably, a good predictor of future performance. That is, if a student receives a grade of "A" in a given course, it is assumed that his future performance dealing with the subject matter will be superior to that of a fellow student who received a "C" in the same course. It seems very possible that these differences in individual histories before instruction may affect future performance in the area, in that prior knowledge of a subject will determine how much material must be learned for mastery. For this reason it was felt that the gain score, which combines a pretest (measure of knowledge prior to instruction) with a posttest (measure of knowledge after instruction) and gives a measure of the increase in knowledge due to instruction, should be empirically compared with the posttest alone, which seems to be the dominant predictor in use.

Although the use of pretests is not uncommon in the literature (e.g., Atkinson & Fletcher, 1972), gain scores (derived from pre- and posttests) have received relatively little attention as potential predictors of future performance. According to Cronbach and Furby (1970), the attempted uses of gain scores fall into four categories:

1. To provide a dependent variable in an experiment on instruction, persuasion, or some other attempt to change behavior or beliefs.

2. To provide a measure of growth rate or learning rate which is to be predicted from other characteristics.

3. To provide an indicator of deviant development as a basis for identifying individuals to be given special treatment or to be studied clinically.

4. To provide an indicator of a construct that is thought to have significance in a certain theoretical network.

Stanley (1971) suggested the appropriateness of gain scores as a measure of group learning even though the instruments in use may be too unreliable for individual use. Davis (1964) felt that the simplest way to measure change in student performance could be estimated from his final test score minus his initial test score, if certain statistical precautions are exercised. Dressel and Mayhew (1954) cite the use of student gain scores as evaluators of teacher effectiveness. The scarity of utilization of gain scores is certainly due in part to the associated problems raised by Dressel and Mayhew (1954). They felt that those students with low pretest scores tended to show higher gains than those with high pretest scores for the following five reasons:

1. ceiling effect--high scores on the pretest have less potential for gain than do low scores. For example, a person scoring 20 correct of 100 could have a gain of 80, while a person scoring 90 correct of 100 could only have a maximum gain of 10.

2. regression towards mean--the statistical concept that individuals scoring extremely high or low on a test will, on a future comparable test, tend to score nearer to the mean. Cambell and Stanley (1963) clarify this concept as follows:

> "The more deviant the score, the larger the error of measurement it probably contains, thus, in a sense, the typical extremely high scorer has had unusually good 'luck' (large positive error) and the extremely low scorer bad luck (large negative error). Luck is capricious, however, so on a posttest we expect the high scorers to decline somewhat on the average, and the low scorers to improve their relative standing [p. 11]."

3. focus of instruction effect--those students with the least ability often receive the most attention from the instructor.

4. If tests measure both ability and knowledge (some theorists feel that knowledge of a subject and the ability to use it are different phenomena and should be separated when measuring learning), scores could reflect predominately one or the other, resulting in incomparable scores.

5. differential motivation--often the "best" students are more highly motivated and work harder at the learning task.

After considering these difficulties many researchers (Lewis, 1968; Mitzel & Gross, 1956; Wheatley, 1969) suggested that gain scores be abandoned in favor of other techniques.

The Dressel and Mayhew (1954) problems with gain scores were not considered to be pertinent to this research in light of the following considerations:

1. The ceiling effect was negated by using a corrected gain formula (gain ratio developed by McGuigan and Peters, 1965) which takes into account that students scoring low on the pretest have greater potential for gain than those who score high on the pretest.

2. Regression toward the mean did not seem to be a problem in that the entire sample was correlated for individual data and no extreme groups of data were considered.

3. The focus of instruction effect did not occur because instruction was programmed with specified time limits, insuring equal attention to all students.

4. Separation of ability and knowledge was not attempted in that the tests were intended to measure only written behavior in response to test items and no other assumed capabilities.

5. There should have been no differential motivation in that all students were offered the same incentive for successful completion of the training program and participation itself was voluntary.

Lord (1963) and McNemar (1958) later stated that the use of gain scores is inappropriate if the pre- and posttests cannot be shown to be measuring the same variable. In this research, however, only one variable was under consideration; the written behavior of the students in response to test items dealing with operant techniques. Cronbach and Furby (1970) suggest the unfairness of using gain scores to make future decisions about an individual. They felt that categorization of people by college grades, I. Q., or any other test measure which could affect future job or training opportunities is an injustice which could lead to unnecessary discrimination. However, since this prediction of capabilities is common and an integral part of our society (for instance, course grades in college are used as predictors of future abilities in specific areas), it seems we have an obligation to use the most reliable methods of prediction, whatever they may be.

There has been considerable discussion of gain scores in the literature, (Bloom, 1963; Cronbach & Furby, 1970; Davis, 1964; Dressel & Mayhew, 1954; Dubois, 1957; Lewis, 1968; Lord, 1956, 1958, 1963; McNemar, 1958; Mitzel & Gross, 1956; Roebuck, 1972; Stanley, 1971; Wheatley, 1969) the bulk of which seems to be of a philosophical nature aimed at discouraging the use of these scores for various reasons. Because individualized instruction has developed from operant techniques, performance is normally the variable under consideration. Without reference to internal events, many of the arguments against the use of gain scores have no validity. Because gain scores have received relatively little attention from researchers, it was felt that a comparison between the most common predictor, posttest scores, and gain scores would determine their relative usefulness toward the end of effective and accurate prediction of future performance.

Research Objectives

The basic objective of this research was to determine the relative usefulness of gain scores and posttest scores as predictors of future performance. It was hypothesized that both gain and posttest scores would correlate highly with retention test scores. These correlations were then to be tested for significance of difference to determine the better predictor. By using multiple regression techniques, the best combination of variables for predicting future performance was to be determined.

A final hypothesis was that those students with the highest initial ability as shown by the Scholastic Achievement Test (SAT) or American College Test (ACT) would achieve the highest gains, indicating that the corrected gain score is, in part, influenced by initial ability.

Method

Subjects

One hundred seventeen students from two different sections of an introductory psychology course at Montana State University participated in the experiment. As incentive, they were offered two extra points to be added to their final grade, one point for participation and one point for successfully meeting criteria (criteria were not announced in advance; it was felt that this technique would insure some motivation of effort in the tasks.

Instruments

To test the hypotheses, a relatively short form of individualized instruction was necessary with pre-, post-, and retention measures which would allow for a high degree of variation. Because the accessible subjects were psychology students, the topic of operant techniques was chosen. The students had received a cursory presentation of this area, so variation on the pretest was expected. They had received little enough instruction that considerable gains in performance were possible after the training session.

A short program (40 frames) was chosen as the means of training, both because of its recognition as a viable means of individualized instruction and its ease of administration to large groups. The first 30 frames were Set 7 (Introduction to Operant Conditioning) from The Analysis of Behavior by Holland and Skinner (1961). The last 10 frames were devised as a continuation of the program, but containing considerably more information per frame. It was reasoned that the inclusion of more difficult material would contribute to variation of the post- and retention test. The program is attached as Appendix I.

The pre-, post-, and retention tests were identical. The 50 questions on the test (Appendix II) were developed from the program, then randomized to avoid sequential clues. Test items were generally very similar or identical to the wording of the program. For example, the first item on the program read:

> Performing animals are sometimes trained with "rewards". The behavior of a hungry animal can be "rewarded" with ________________.
>
> food

The comparable test item read as follows:

> The behavior of a hungry animal can be "rewarded" with________________. (Answer: food)

Procedure

Initially, a group of nine students was administered the pretest, program, and posttest. From the data and comments of this pilot group, changes involving wording or clarification were made in the test. Additionally, times of 12 minutes for the tests and 20 minutes for the program were determined to be ample time for even the slower students to finish.

One 50 minute class period was scheduled for the volunteers of each section. Students were told that they would be given a 50 question fill-in-the-blank test covering the topic of operant techniques. They were asked to move quickly in that they had 12 minutes to complete the exam, skipping any question they did not know. If they had time and wanted to return to a previous question, they could. It was stated that the purpose of the test was to determine their level of knowledge in this area and that it was realized that they had not had extensive training in the subject matter. They were encouraged to do as well as possible.

After 12 minutes, the tests were collected and programs distributed. Instructions were to move a piece of paper down the front of the program until one frame was exposed. They were then to read the item and think the answer. The paper was then moved down to expose the answer. If it was not thought correctly, they were to redo the procedure for that frame and then continue. They were told that they could have 20 minutes to complete the program and to work quickly. Following completion of the program, the posttest was distributed with 12 minutes again allowed for completion. The students were thanked and asked to return in one week.

At the beginning of the class period one week later, students were admisistered the retention test with the same time limit (12 minutes).

Data Analysis

Ninety students constituted the data analysis sample in that of the 117 participants, 27 did not complete one of the examinations, almost all of the incomplete exams being the posttest. There were two criteria for incompletion:

(a) none of the last 15 items were attempted

(b) none of the items on the second page were attempted (this page was printed on the back of the first page and omission was probably by error rather than by time constraints)

For each student, five datum points were recorded:

1. SAT or ACT converted to stanines (because all the students had taken one of these exams, the scores were converted to a common scale, or stanine)

2. pretest score--number correct of 50 questions

3. posttest score--number correct of 50 questions

4. retention test score--number correct of 50 questions

5. gain score--the McGuigan gain ratio, observed gain divided by potential gain (for example, a student scoring 28 on the pretest and 39 on the posttest would have an observed gain of 11 and a potential gain of 22; 11 divided by 22 would yield his corrected gain score of 50%)

The means for each of the five variables were tested between the two sections to determine if any differences existed between the groups. A one way analysis of variance between the means of each test for the entire sample ($\underline{n}$ = 90) was performed to assess differences in performance before and after training and after the one week interval. Tukey's post hoc analysis utilizing the $\underline{q}$ statistic was then used to determine which treatment means differed significantly.

Correlation coefficients were computed between all combinations of the five variables to determine which variables were suitable predictors. The correlation of gain scores with retention test scores was compared to the correlation of posttest scores with retention test scores utilizing the procedure of Hotelling (1940) to decide which was the better predictor, if a difference existed.

Finally, multiple regression techniques were applied to determine if gain and posttest optimally combined would increase singificantly the predictability of retention test score from either gain or posttest alone.

TABLE A

Variable Means, Standard Deviations, and Possible Ranges

for Classes and Combined Sample

Statistic	Variables				
	SAT/ACT	Pretest	Posttest	Retention Test	Gain
			Section 01 (n = 57)		
Mean	5.12	20.91	31.84	33.22	38.84
S. D.	1.73	6.27	7.32	6.47	18.70
Poss. Range	1-9	0-50	0-50	0-50	0-100
			Section 03 (n = 33)		
Mean	5.48	20.34	32.40	32.61	42.33
S. D.	2.08	4.88	4.74	4.94	11.94
Poss. Range	1-9	0-50	0-50	0-50	0-100
			01 and 03 Combined (n = 90)		
Mean	5.25	20.59	32.07	33.02	40.13
S. D.	1.87	5.84	6.53	6.00	16.61
Poss. Range	1-9	0-50	0-50	0-50	0-100

TABLE B

Analysis of Variance of Means of Pre-, Post-, and Retention Test Scores

Source	df	MS	F
Between Groups	2	4308.62	114.70*
Within Groups	267	37.56	
Total	269		

*$p < .005$

TABLE C

Tukey Analysis of Differences Among Means

of Pre-, Post-, and Retention Test Scores

Statistic	Differences Between Means		
	A–B	A–C	B–C
q	17.77*	19.25*	1.48

Note.--$q_{critical}$ = 4.12

* $p < .01$

TABLE D

Correlation Matrix for Five Variables

	SAT/ACT (S)	Pretest (A)	Posttest (B)	Retention test (C)
Pretest (A)	.11			
Posttest (B)	.26*	.61**		
Retention test (C)	.24*	.52**	.84**	
Gain (G)	.26*	.10	.75**	.65**

$^*p < .05$

$^{**}p < .01$

Results

In order to determine if there were differences between the two classes with regard to a) initial ability, b) performance on the three tests, or c) gains, the means for each variable were compared between the classes by _t_ test. The differences between the means were not statistically significant ($p > .05$). No further analyses were made between the classes and future analyses used combined scores. The combined and class means and standard deviations are shown in Table A.

A one way analysis of variance (Table B) indicated a significant difference among the means of the pre-, post-, and retention test scores ($F = 114.70$, $df = 2, 267$, $p < .005$). Since treatment _n_'s were equal and normality and homogeneity of variance assumptions appeared reasonable, the Tukey multiple comparison method was used for _post hoc_ analysis. The Tukey procedure provides a statistically powerful and conservative test for comparisons between two treatment means. Analysis showed significant differences between the means of the pre- and posttests ($p < .01$) and the pre- and retention tests ($p < .01$), but not between the means of the post- and retention tests (Table C). These data indicated that, on the average, students performed better after the training session than before with no appreciable difference in performance immediately following training compared to performance one week later.

Pearson product-moment correlations were computed between all variables. The matrix appears as Table D. The data show that those students with the highest stanines achieved the highest gains ($r = .26$, $p < .05$). The gain score proved to be an excellent predictor of future performance on the retention test ($r = .65$, $p < .01$), but the posttest score alone proved to be better ($r = .84$, $p < .01$). Utilizing the procedure developed by Hotelling (1940), the correlation of gain score with retention test score was compared to the correlation of posttest score with retention test score. The difference between the two correlations was found to be statistically significant ($t_{d_r} = 3.67$, $p < .001$), indicating the superiority of the posttest score compared to the gain score for prediction of future performance.

Finally, multiple regression techniques were used to optimally combine gain with posttest scores to predict retention scores. The addition of gain scores to posttest scores for predicting retention added an insignificant amount to the variance accounted for by using posttest scores alone ($r^2 = .7058$ and $r^2 = .7064$ respectively). Moreover, multiple regression techniques optimally combining all four other variables added very little to the prediction of retention score ($r^2 = .7084$) over prediction from posttest scores alone.

Discussion

The significant increase in test scores from the pretest to the posttest as shown by the Tukey analysis ($p < .01$) was an expected result. This analysis indicated a definite increase in performance on the subject matter following completion of the training program and reiterated the usual finding that learning does occur using programmed instruction (e.g., Ferster, 1968; Suchett-Kaye, 1972; Witters & Kent, 1972).

The almost identical performance on the post- and retention tests was a surprising result. A significant decrease in performance was expected after a one week interval in light of the normal finding that a substantial portion of learned material is forgotten very soon after learning, especially when learning occurs in only one session, rather than over a period of time (Bass & Vaughan, 1966). There were two probable explanations for this result. The training program with pre- and posttests was a substantial amount of work for a fifty minute class period, and, by the time the pretest was administered, a phenomenon like ratio strain seems to have occurred--the students were just tired of performing in the absence of reinforcement. One week later the retention test was given at the beginning of the class period, and the students had not been working in the subject area immediately prior to the test. It appeared that the scores on the retention test were the result of a legitimate effort by relaxed and refreshed students, while the scores on the posttest may have suffered from the effort required by the preceding test and program (and possibly accompanying anxiety).

Secondly, it was determined by informal conversations that the students did discuss the subject matter among themselves after the first session. This factor would account for a slight increase in knowledge between the post- and retention tests, which, when negated by normal forgetting over the one week interval, may have resulted in the equivalent scores.

The correlation between SAT/ACT stanines and gains ($r = .26$, $p < .05$) confirmed the hypothesis that those students with the highest initial ability as measured by the SAT or ACT would achieve the highest gains. Actually, a higher correlation was expected between these two variables, but the resulting correlation appeared to indicate effectiveness of the program used. Although the more capable students (as measured by SAT/ACT performance) did achieve the higher gains, initial capability accounted for less than six percent of the total variance. It is possible, however, that this finding is not generalizable to all programmed instruction, in that there was a time limit to complete the program. Normally, a program is completed at the pace of the individual so that he will, hopefully, master the material. It was considered likely that, without a time constraint, there would have been lesser differences among the gains since the slower students would probably have utilized more time and felt under less pressure to finish.

The high correlations for all combinations of the three tests (r_{AB} = .61, $p < .01$; r_{BC} = .84, $p < .01$; r_{AC} = .52, $p < .01$) indicated that students scoring high on any test tended to score high on all three. These data were partially explained by the correlation of SAT/ACT with performance on the post- and retention tests (r_{SB} = .26, $p < .05$; r_{SC} = .24, $p < .05$), which showed that the more capable students scored higher on exams following instruction (it was noted that initial capability did not significantly contribute to performance prior to instruction [r_{SA} = .11]). Probably, higher motivation among the more capable students (with a history of reinforcement for academic effort) contributed to the phenomenon of high on one test, high on all. Also, the more capable students may have been more likely to discuss new terms and concepts among their peers following instruction.

As hypothesized, both gain scores and posttest scores were shown to be good predictors of future performance (r_{GC} = .65, $p < .01$; r_{BC} = .84, $p < .01$). When these two correlations were compared, the difference was determined to be statistically significant (t_{d_r} = 3.67, $p < .001$), indicating the superiority of the posttest score compared to the gain score for prediction of future performance. One of the tenets of individualized instruction is that criteria are specified, and once a given performance level is achieved (in effect, a posttest), certification of competence for the subject area is in order. In consideration of the above data, it seemed that our usual means of prediction (based on posttest measures) are appropriate, if an estimate of future performance is desired.

It was not surprising that multiple regression of gain and posttest score failed to significantly increase the predictability of retention test score. Correlation between gain and posttest scores was .75 and, with so much common variance, the optimal combination was not expected to be much better than posttest alone for prediction of retention.

Based on the data obtained from this research, it appears that commonly used measures, such as college grades or grade point averages do, with high probability, predict retention of the course material. Measures of competence prior to training may predict success or failure of the student with the course, but final performance after training is the best predictor of future performance in the subject area.

Limitations of Research

One of the characteristics of programmed instruction is that each student may proceed at his own pace. In this research, it was not possible to allow unlimited time to complete either the tests or the training program because of scheduling considerations. Although, it was determined by a pilot group that time limits were adequate, on the basis of group performance of the sample, it seemed that more time would have been beneficial.

Because of the availability of the subjects, a one week interval prior to retesting was scheduled, rather than two retention tests as originally planned. Several retention tests over longer intervals may have been very instructive for comparisons with the predictors. It was assumed that, after an interval of three or four weeks, performance on the retention test would have decreased enough to show any effects of individual differences. Unfortunately, this was not possible due to constraints imposed by the schedules of the subjects.

Implications for Future Research

Gain scores and posttest scores both were shown to be excellent predictors of future performance after a one week time interval. Because predictors are usually concerned with longer periods of time, it would seem that testing after several longer periods of time would be in order. These retention test scores could then be correlated with other predictors to determine if existing relationships hold over longer periods.

Additionally, although posttest scores account for much of the variance as predictors of future performance ($r^2 = .71$), there is still a substantial amount of variance unaccounted for which is, presumably, attributable to other variables. The nature and extent to which these other variables contribute should be investigated.

Finally, since generalizability is often questioned, research employing a paradigm similar to the above (without the limiting factors) should be conducted using different topic areas, different methods of instruction, and with students of varying initial capabilities concerning the subject matter. If comparable results were then obtained, the conclusions based on present research could be applied to prediction of future performance following instruction in general.

References

Alba, E. & Pennypacker, H. S. A multiple change score comparison of traditional and behavioral college teaching procedures. Journal of Applied Behavior Analysis, 1972, 5, 121-124.

Atkinson, R. C. & Fletcher, J. D. Teaching children to read with a computer. The Reading Teacher, 1972, 25, 319-327.

Bass, B. M. & Vaughan, J. A. Training in industry; The management of learning. Belmont, Cal.: Wadsworth Publishing Company, 1966.

Bereiter, C. The future of individual differences. Harvard Educational Review, 1969, 39, 310-318.

Bloom, B. S. Testing cognitive ability and achievement. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand-McNally, 1963.

Cambell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand-McNally, 1963.

Carlton, F. C. Educational and industrial revolution. New York: MacMillan, 1910.

Cronbach, L. J. & Furby, L. How should we measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80.

Cubberley, E. P. The history of education. Boston: Houghton Mifflin, 1920.

Davis, F. B. Educational measurements and their interpretation. Belmont, Cal,: Wadsworth Publishing Company, 1964.

Dressel, P. L. & Mayhew, L. B. General education: Exploration in evaluation. Washington, D. C.: American Council on Education, 1954.

Dubois, P. H. Multivariate correlational analysis. New York: Harper, 1957.

Ferster, C. B. Individualized instruction in a large introductory psychology college course. The Psychological Record, 1968, 18. 521-532.

Holland, J. G. & Skinner, B. F. The analysis of behavior. New York: McGraw-Hill, 1961.

Hotelling, H. The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters. Annual of Mathematical Statistics, 1940, 11, 271-283.

Johnston, J. M. & Pennypacker, H. S. A behavioral approach to college teaching. American Psychologist, 1971, 36, 219-244.

Keller, F. S. Good-bye teacher....Journal of Applied Behavior Analysis, 1968, 1, 79-89.

Lewis, D. G. Experimental design in education. London: University of London Press, 1968.

Lord, F. M. The measurement of growth. Educational and Psychological Measurement, 1956, 16, 421-437.

Lord, F. M. Further problems in the measurement of growth. Educational and Psychological Measurement, 1958, 18, 437-451.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

McGuigan, F. J. & Peters, R. J. Assessing the effectiveness of programmed tests: Methodology and some findings. Journal of Programmed Instruction, 1965, 3, 23.

McNemar, Q. On growth measurement. Educational and Psychological Measurement. 1958, 18, 47-55.

Mitzel, H. E. & Gross, C. F. A critical review of the development of pupil growth criteria in studies of teacher effectiveness. New York: Division of Teacher Education, Office of Research and Evaluation, City of New York, 1956.

Newsome, R. S., Eischens, R., & Looft, W. R. Intrinsic individual differences: A basis for enhancing instructional programs. The Journal of Educational Research, 1972, 65, 387-392.

Roebuck, M. Floundering among measurements in educational technology. Programmed Learning & Educational Technology, 1972, 9, 87-97.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1971.

Suchett-Kaye, C. Personality factors and self-instruction: A survey. Programmed Learning & Educational Technology, 1972, 9, 206-215.

Sulzbacher, S. I. Effects of response mode and subject characteristics on learning in programmed instruction. National Society of Programmed Instruction Journal, 1967, 4, 10-11.

Wheatley, R. A. Measurement in programmed learning experiments. In A. P. Mann & C. K. Brunstrom (Eds.), Aspects of educational technology III. London: Pitman, 1969.

Witters, D. R. & Kent, G. W. Teaching without lecturing: Evidence in the case for individualized instruction. The Psychological Record, 1972, 22, 169-175.

Appendices

Appendix I

Operant Conditioning

Name: Soc. Sec.#: Section:

1. Performing animals are sometimes trained with "rewards". The behavior of a hungry animal can be "rewarded" with__________.

food

2. A technical term of "reward" is reinforcement. To "reward" an organism with food is to________________it with food.

reinforce

3. Technically speaking, a thirsty organism can be_____________ with water.

reinforced (NOT: rewarded)

4. The trainer reinforces the animal by giving it food___________ it has performed correctly.

when (if, after)

5. Reinforcement and behavior occur in the temporal order: (1)____________, (2)___________.

(1) behavior
(2) reinforcement

6. Food given to a hungry animal does not reinforce a particular response unless it is given almost immediately____________the response.

after

7. Unlike a stimulus in a reflex, a reinforcing stimulus________ act to elicit the response it reinforces.

does not (will not)

8. A reinforcement does not elicit a response; it simply makes it more__________that an animal will respond in the same way again.

probable (likely)

9. Food is probably not reinforcing if the animal is not_________.

deprived of food (hungry)

10. If an animal's response is not followed by reinforcement, similar responses will occur___________frequently in the future.

less (in-)

11. To make sure an animal will perform, the trainer provides_______ for the response frequently.

reinforcement(s)

12. A hungry pigeon in the park flicks dead leaves about with quick movements of its beak. This behavior is________whenever it uncovers bits of food.

reinforced

13. A pigeon is occasionally reinforced for flicking leaves about because of the common natural arrangement of leaves over_________.

food (seed, insects, reinforcers)

14. The reinforcement used by animal trainers is deliberately arranged, while the arrangement of leaves and food in the park is___________.

natural (non-deliberate)

15. Food is not reinforcing unless the animal has first been_________ food for some time.

deprived of (without, hungry for)

16. Reinforcing a response produces an increase in the_____________ that the response will occur again.

probability (likelihood, chances)

17. We do not observe "probability" directly. We say that a response has become more probable if it is observed to occur more_________ under controlled conditions.

frequently (often)

18. When a response has been reinforced, it will be emitted __________ frequently in the future.

more

19. To get an animal to emit a response more frequently, we __________ the response.

reinforce

20. In laboratory research, various devices are used to reinforce responses. Heat can be used to __________ the responses of a cold animal.

reinforce

21. An electrically operated food magazine which presents food can be used to reinforce a(n) ____________ of an organism deprived of food.

response

22. If the cold (or food deprived) organism turns on an electrically operated heat lamp (or food magazine), the response of turning on will be ______________.

reinforced

23. The response of turning on the electrically operated heat lamp or food magazine will be emitted more _________ in the future.

frequently (often)

24. In a typical apparatus, the depression of a horizontal bar automatically operates a food magazine. The apparatus selects "bar pressing" as the _________ to be reinforced.

response (behavior)

25. Since no eliciting stimuli are observed for such responses as flicking leaves or bar pressing, we cannot say that these responses are _______________ by stimuli, as are the responses in reflexes.

elicited

26. The response of pressing a bar must be emitted at least once in order to be ______________.

reinforced

27. Responses such as bar pressing, flicking leaves, etc., are said to be emitted rather than elicited since there____________ (are or are no?) observed eliciting stimuli.

are no

28. If pressing the bar does not operate the food magazine, the response______________reinforced.

is not (will not be)

29. Reinforcement makes responses more frequent while failure to receive reinforcement makes responses___________.

less frequent (become extinguished)

30. No eliciting stimuli are observed for bar pressing, flicking leaves in the park, etc. Therefore, responses of this type __________classified as reflex behavior.

are not (cannot be, will not be)

31. The responses above are classified as operant behavior.__________ behavior is formed differently from reflex behavior by a process called shaping.

operant

32. Operant behavior is produced by a process called______________.

shaping

33. An animal is shaped to emit a new response by gradually changing the criteria for reinforcement, until only the new response is reinforced. This process is called__________.

shaping

34. Each time the criteria for reinforcement is changed during shaping, the acceptable behavior under the new criteria is called a successive approximation, because it__________more closely the behavior being shaped.

approximates

35. Because the criteria for reinforcement is changing during shaping, the reinforcement is said to be differential. Shaping is the__________reinforcement of successive approximations of the desired behavior.

differential

36. The end result of shaping is a new_____________.

behavior, repertoire

37. If the behavior of an organism is not reinforced, it will cease to be emitted. This phenomenon is called extinction. During extinction the behavior in question___________reinforced.

is not

38. When an experimenter stops reinforcing a response, the response is_________likely to be emitted in the future.

less

39. This process is technically called_______________.

extinction

40. Organisms are sometimes reinforced (and shaped) by chance in their natural environment. Such reinforcement is called accidental or adventitious reinforcement and leads to new_____________.

behavior

Appendix II

Operant Conditioning Test

Name: Soc. Sec.#: Section:

1. The behavior of a hungry animal can be "rewarded" with___________.

2. The technical term for "reward" is_______________.

3. Heat would be a____________________(TT) for a cold animal.

4. The end result of shaping is a new_______________________.

5. Differential reinforcement is used in the process of_____________.

6. _________________responses are physiologically determined.

7. When shaping or maintaining what type of behavior, does the stimulus (reinforcer) follow the response? ______________

8. Because no eliciting stimuli are observed for studying, it must not be________________behavior.

9. Failure to receive reinforcement makes future occurrences of the behavior__________________.

10. Reinforcement in an expreimental laboratory differs from reinforcement occurring in nature in that it is____________.

11. To be most effective, reinforcement should follow behavior________.

12. For a hungry animal, food is a ______________________.

13. With a reflex, the stimulus_________________the response.

14. Non-reinforcement is synonymous with__________________(TT).

15. Technically speaking, a thirsty organism can be__________with water.

16. Reinforcement makes future similar behavior___________likely.

17. To increase the probability of behavior in the future, it is_____.

18. In order to use food as reinforcement, the animal must be________.

19. An emitted, rather than elicited, response is called a(n)________.

20. Reflex behavior is determined by____________, in addition to stimuli.

21. Shaping is used to create what type of behavior?__________________.

22. A response must be___________by the organism if it is to be reinforced.

23. Technically speaking, operant responses are_____________.

24. If you fish in an area for awhile and catch no fish, your fishing behavior in that spot will be under___________________(TT).

25. In the natural environment, reinforcers may be arranged or occur ____________________.

26. When eliciting a_______________(TT)response, the stimulus precedes the response.

27. Reinforcement is given after__________________.

28. Because studying is maintained by certain reinforcers, it is a form of_________________behavior.

29. If the behavior of an organism is not reinforced, it will_________.

30. A trainer reinforces an animal by giving it food___________it has performed correctly.

31. The cessation of behavior due to lack of reinforcement is called _____________________(TT).

32. Accidental reinforcement is also called______________reinforcement.

33. To cause an animal to emit a response more frequently, one________ the response.

34. With operant behavior, the stimulus_____________the response.

35. Reinforcement changes the future____________of occurrence of a response.

36. During shaping, the________for reinforcement are gradually changed.

37. Shaping involves_______________reinforcement, as opposed to a normal schedule of reinforcment.

38. In shaping, the new behavior required is called a(n)____________(TT).

39. During extinction, reinforcement is_______________.

40. During shaping, successive approximations of the desired behavior are________________(TT) reinforced.

41. If responses are not reinforced, they will occur_________in the future.

42. Operant behavior is formed by a process called_________________.

43. Reinforcement occurring by chance in the natural environment is called_________________reinforcement.

44. Reinforcement changes the_____________of future similar behavior.

45. Food would be a reinforcer only if the animal is_____________.

46. Lack of reinforcement makes future similar behavior____________ likely.

47. Turning the key to your car is_________________when the engine starts.

49. In "trial and error" learning, reinforcement occurs_______________.

50. The stimulus-response paradigm deals with_______________behavior.