A Systematic Analysis of Between and Within Grader Variability as Affected by Answer Key Condition, Grader Experience, and Test Set Assigned

Leonard Eugene Swistak
Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/masters_theses
Part of the Psychology Commons

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/2700

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Master’s Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
A SYSTEMATIC
ANALYSIS OF BETWEEN
AND WITHIN GRADER VARIABILITY
AS AFFECTED BY ANSWER KEY CONDITION,
GRADER EXPERIENCE, AND TEST SET ASSIGNED

by

Leonard Eugene Swistak

A Thesis
Submitted to the
Faculty of The Graduate College
in partial fulfillment
of the
Degree of Master of Arts

Western Michigan University
Kalamazoo, Michigan
December 1973
ACKNOWLEDGEMENTS

I cannot overacknowledge the contribution, in terms of advice and constructive appraisal, that was offered by my graduate advisor, Dr. E. W. Hitzing. I am deeply grateful for his unflattering assistance. I am also indebted to Dr. Jack Michael, and Dr. Herman Peine for their helpful suggestions while serving on my graduate committee. It is my particular wish to express my very deep gratitude to Dr. Michael Stoline for his expert advice and performance as a statistical consultant. Lastly, I wish to thank the many people who contributed their time and effort to serve as subjects in this study. Many thanks to you all.

Leonard Eugene Swistak
INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in “sectioning” the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from “photographs” if essential to the understanding of the dissertation. Silver prints of “photographs” may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
MASTERS THESIS

SWISTAK, Leonard Eugene, Jr.
A SYSTEMATIC ANALYSIS OF BETWEEN AND WITHIN
GRADER VARIABILITY AS AFFECTED BY ANSWER KEY
CONDITION, GRADER EXPERIENCE, AND TEST SET
ASSIGNED.

Western Michigan University, M.A., 1973
Psychology, general

University Microfilms, A XEROX Company, Ann Arbor, Michigan

THIS DISSERTATION HAS BEEN MICROFIEMED EXACTLY AS RECEIVED.
Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>METHOD</td>
<td>9</td>
</tr>
<tr>
<td>Subjects</td>
<td>9</td>
</tr>
<tr>
<td>Materials</td>
<td>9</td>
</tr>
<tr>
<td>Procedure</td>
<td>12</td>
</tr>
<tr>
<td>RESULTS</td>
<td>16</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>28</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>32</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>66</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
INTRODUCTION

The focus of research reports on operant behavior and behavior modification appears to be changing as the field continues to develop. The impressive segment of behavior modification literature devoted to the training of non-professionals in the principles of behavior modification is an indication of this expansion. Experimenters have found that changing the subject's inappropriate behavior in an artificial environment with the subsequent return of the subject to the unmodified, natural environment results in the restrengthening of the inappropriate behavior. A proposed solution to this problem would be to educate the significant, non-professional people within the subject's environment in the utilization of behavior modification techniques. It has been shown that the use of such a strategy will result in the modification of existing contingencies and the elimination of undesirable behavior.

The training methods have, for the most part, been inadequately documented. Several investigators have reported behavioral techniques for working with the parents of children exhibiting behavior problems. These include formal training in behavioral principles (Hall, Axelrod, Tyler, Grier, Jones, and Robertson, 1972; Hall Christler, Cranston, and Tucker, 1970), reprogramming the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
social environment (Patterson, McNeal, Hawkins, and Phelps, 1967), behavioral contracting (Stuart, 1972), and the training of parents in the home setting (Bernal, 1969; Hawkins, Peterson, Schweid, and Bijou, 1966; O'Leary, O'Leary and Becker, 1967; Tharp and Wetzel, 1969; Wahler, 1969; Wahler, Winkel, Peterson, and Morrison, 1965; Zeilberger, Sampen, and Sloane, 1968). School teachers have been instructed in the principles of behavior modification (Hall, Lund, and Jackson, 1968; Becker, Madsen, Arnold, and Thomas, 1967; Thomas, Becker, and Armstrong, 1968). However, this instruction has been limited to information relating to specific problem behaviors. Wetzel (1970) reported successful results with a training program designed to instruct teacher's aides in the principles of behavior modification. Peers have been instructed in the use of specific behavior management techniques to change the behavior of their classmates or siblings (Packard, 1970; Surrat, Ulrich, and Hawkins, 1969).

Techniques of training vary from instructions concerning what to say and do (Ayllon and Michael, 1959) to demonstrations (Wetzel, Baker, Roney, and Martin, 1966) to actual shaping of appropriate reinforcing behavior (Wahler, Winkel, Peterson, and Morrison, 1965). Some training programs provide formal instruction on the general principles of learning and techniques of behavior modification (Wetzel, 1966; Davison, 1965).
Although many training programs in behavior principles have been employed and reported, there is a paucity of tests or measures developed to evaluate the effectiveness of the various methods of instructional presentations. Thus, there remains the question of what the most effective method is for the instruction of behavior modification techniques.

Gardner, Brust and Watson (1970) have reported that there is generally a lack of objective evaluation of the effectiveness of the behavioral training programs. The evaluation of the various instructional methods requires that one have a reliable and valid means of measuring the trainee's behavior. According to the authors, there are "two basic ways to evaluate the behavior shaping proficiency of trainers (p. 633)." First, the trainer can be directly observed implementing a procedure with the patient and his effectiveness evaluated according to some pre-set criteria or the trainer may be evaluated indirectly in terms of the patient's progress. Their opinion was that neither measure is sufficient by itself to evaluate reliably the attendant's performance. The first measure yields an evaluation as to whether the trainee is correctly employing the different components of a specific type of procedure, but it does not measure his influence on the patient's behavior. The second approach evaluates the patient's behavior, but does not indicate where the deficiencies and strengths of the
trainers performance lie.

The Gardner study developed a scale called the Training Proficiency Scale (TPS). The scale attempted to measure effectiveness in applying behavior modification skills to the treatment of the mentally retarded. The TPS "was developed by breaking down behavior modification skills into their component parts and by constructing items corresponding to each of those components (p. 633)." The study consisted of 20 students employing the TPS to rate the role-played behaviors of 20 attendants. Inter-scorer, split-half, and test-retest reliability were calculated and all were high. The scores obtained from use of the TPS were found to be highly related to scores from a test on the principles of behavior modification and with global evaluations of training proficiency. It was the conclusion of this study that the TPS provided a "fairly short, efficient, reliable and valid method for assessing an individual's effectiveness in using behavior modification techniques (p. 636)."

Gardner and Biampa (1971) developed the Attendant Behavior Checklist (ABCL) designed to measure the on-the-job behavior of attendants upon completion of two different training programs; traditional and behavior modification. The study examined the reliability and validity of a time-sampling technique using the rating scale as their measure. The results showed that the ABCL provided a reliable measure of attendant behavior.
Furthermore, differences between test scores were shown to be highly related to differences in training conditions as reflected by the alternative approaches of the behavior modification group or the traditional training group programs. Direct measures of the changes in the patients behavior cannot be obtained with the ABCL, however, the percentage of attendant time spent engaged in behaviors purported to facilitate such "appropriate" patient behavior can be determined. It is relevant to note that the ABCL provides quantitative data and yields no data as to the qualitative nature of the attendant's interactions. It is the conclusion of this study that "on-the-ward behavior of attendant personnel is capable of reliable quantification (p. 622)."

Ashbaugh (1971) evaluated an attendant training program which was based on behavioral principles. The program had lasted for a total duration of three weeks with essentially two groups; an experimental group which was exposed to behavior modification training and a control group which received regular attendant training. The pre- and post-test consisted of six, two minute video-taped samples of commonly observed, inappropriate resident behaviors. The test question was simply, "How would you change this resident's behavior (p. 20)." Attendants were requested to employ a systematic analysis and procedure in an attempt to modify the deviant behavior. The attendants' responses were recorded on tape, and coded by
independent raters using a set of seven pre-determined criteria. Ashbaugh concluded that the experimental attendant training program was effective in training attendants to employ operant techniques.

It was Ashbaugh's belief that the most valid and reliable test of the effectiveness of the training programs would involve the observation of actual changes in the patient's behavior as a result of the attendant's behavior. However, it was not possible for Ashbaugh to design his experiment in this manner.

Schoonmaker (1971) reported on a workshop approach to training paraprofessionals in the use of basic behavior modification techniques. The short term method included lectures, audio-visual aids to illustrate the lecture topic, small group discussions, reading and writing assignments, behavior change projects, daily quizzes and pre- and post-test, which was designed to measure attitude change toward the treatment of the mentally retarded, showed a significant increase in the direction of a behavioral orientation as opposed to a medical model approach. However, this study made no attempt to systematically assess or evaluate the various methods and components involved in the workshop.

As mentioned previously, there is an increasing portion of behavior modification literature being devoted to the training of paraprofessionals in behavior modification. There are a variety of training programs with
respect to content and amount of material presented, methods of instruction, and objectives to be achieved. In light of such a movement it appears necessary to develop some type of reliable testing instrument to measure the effectiveness of such programs. Perhaps a solution to this problem would be to develop a standardized, "applied", written test. However at present there exists, in the opinion of this author, no adequate "applied" pencil-and-paper test that effectively measures an individual's comprehension of behavior modification's principles and techniques.

One of the problems in developing an "applied" paper-and-pencil test lies in the difficulty of making an objective and reliable answer key with which to judge the responses. For designers of such evaluation measures the question remains as to what type or style of answer key most reliably evaluates "applied" essay answers. The purpose of this study was to devise three different styles of answer sheets and to then test a grader's ability to measure reliably the qualitative content of the essay answers. The degree of consistency will be defined as the amount of variability in total test scores assigned to individual tests. Thus, the dependent variable may be considered the amount of variability of total test scores, both between and within graders. The independent variables are: (1) answer key condition, (2) grader group, (3) Test Set assigned to grade.
Different levels of subjects were administered the essay test, in an attempt to gain a wide range in the qualitative content of the answers. Initially this was also done so that the different levels of testee experience could be analyzed separately. However, this aspect was never carried out. Different levels of graders were used to increase the generality of the results. The study sought to compare the effect of graders who were deemed as experienced in the area of behavior modification with the effect of graders who had had only limited exposure to the area.
METHOD

Subjects

The subjects for this study consisted of 50 individuals who were, with respect to their exposure to behavior modification principles and techniques, of diversified backgrounds and experience. All test subjects were obtained from the following subpopulations: (1) naive subjects who had had no exposure to behavior modification, (2) high school seniors who were currently enrolled in the Behavior Science Institute (BSI) and whose exposure to behavior modification was limited to five weeks of intensive instruction, (3) graduate psychology students at Western Michigan University. The "testing" population consisted of 30 subjects of the following composition: 10 naive subjects, 10 BSI students, and 10 graduate psychology students. The "grading" population contained 10 subjects from the BSI subpopulation, and 10 graduate psychology students.

Materials

An abbreviated test consisting of four applied, procedural questions was constructed in order to generate written essay answers (See Appendix A). The questions
required the testee to formulate a descriptive operational design of the entire procedure that he would implement to correct the behavioral deficiencies or excesses presented in the questions. The "entire procedure", for purposes of this test, was defined as a description of all the steps preceding, subsequent to, and including the actual manipulation of the environmental contingencies. Questions 1 and 3 and questions 2 and 4 described an identical problem behavior and thus required similar answers. The distinguishing factor was that questions 1 and 2 did not state what type of procedure was to be used nor did it provide and structure with regard to a procedural outline. While the mate questions, questions 3 and 4 respectively, stated the proper procedure the testee was to describe; also providing him with a skeletal procedural outline.

The test format and the selected test population contributed to the generation of answers varying in quality. It was hypothesized that this variation in the quality of answers might have an effect upon the total test score assigned to a test and thus on the subsequent calculations of inter- and intra-grader variability.

There were three forms of answer keys: (1) open scale, (2) graduated scale, (3) graduated-defined scale (See Appendices B, C, D). The discriminating difference between them was in the amount of instruction and outline
structure provided the grader which was designed to instruct and assist him in the evaluation and assessment of the answers. The open scale (See Appendix B), instructed the grader as to the steps that an acceptable answer should incorporate and a model answer exemplifying the acceptable constituent elements. It instructed the grader to make a point award for each answer based on a 0-25 point scale; with the numerical appropriation depending upon the degree of quality as compared to the answer key. The graduated scale (Appendix C), provided the grader with the same answers as the open scale with respect to the requisite steps an acceptable answer should include and an example of such an answer. However, it provided additional instructions in that the grader was instructed as to how much he was to award for the inclusion of each step in the subject's answer. For example, the grader was instructed that a point award for a definition of the problem behavior is based on a 0 to 4 point range, the award being dependent upon the qualitative content of the answer, as compared to the answer key's answer. The graduated-defined scale (Appendix D) was not ostensibly different from the graduated scale with the exception that it further specified the assessment of the answer by stating the exact award for the inclusion of each step through the imposition of a 1 point gradation requirement. Each answer was concomitantly divided into 1 point components.
Thus, the grader was informed not only as to the point value for the inclusion of each step, but also as to the assignment of points to the integral parts of each step.

Data sheets were designed and distributed to the graders for the purpose of recording their assessments of the test answers. There were two types of data sheets, one for the open scale answer key condition and one for the graduated and graduated-defined scales (See Appendices E and F). The distinguishing difference between the two data sheets was that the data sheet for the open scale (Appendix E) required the grader to record his total score for each question of each test. While Data Sheet II (Appendix F) required the grader to assign and record a value to each of the constituent parts of an answer.

The initial reasoning behind the design of these data sheets was to obtain values for individual test items as well as a total test score. The variability of the individual test item scores was then to be calculated. However, due to the inordinate amount of data, the use of such results was deemed not feasible.

Procedure

All handwritten answers obtained from the subjects were typed on separate sheets of paper for purposes of grading. Inasmuch as there were four questions to a test, and 30 tests, the total number of these individual
answer slips was 120. The answers on the slips were identified by their corresponding question number, with the answer slip itself being identified by a code number on the back of the slip. This code number made it possible to establish which test the answer came from.

All 30 tests were randomly assigned to either Test Set A or Test Set B. It was not feasible, at any time, to have each grader grade all 30 tests because of the long time duration needed for such an activity. Thus, graders were randomly assigned to grade tests from Test Set A or B.

Test Graders A consisted of five subjects from the BSI group, and five subjects from the psychology graduate student group, each group being assigned to grade tests numbered 1-15. The remaining five members from each of these respective groups, a total of 10 subjects, composed Test Graders B and were responsible for grading tests numbered 16-30. Tests numbered 1-15 consisted of five tests taken by naive subjects, five tests taken by BSI subjects and five tests taken by psychology graduate students. Similarly, tests numbered 16-30 had the same distribution of test subject type.

The test answers were catalogued according to question number and were distributed to the grading subjects in that manner. All graders were given data sheets on which to record their assessment of each answer for all tests. That is, all graders received all the number one
questions for all 15 of their assigned tests, after completing their grading of those questions they were then given all the questions numbered two for their tests, and likewise for questions three and four. The test items were graded in this manner by all graders under all conditions. The order of the test numbers for each set of questions was randomly obtained and was the same for all graders within a Test Set. However, the order of the test numbers between questions 1, 2, 3, and 4 were randomly altered, but again remained the same for all graders. The test number order that was established for each question remained fixed for all graders across all answer key conditions. All graders began grading their assigned tests using the open scale answer key. Each grader evaluated his assigned 15 tests, after which he was then allowed a 10 minute break. He was then assigned three additional tests which were identified by different test numbers from his previously graded tests, but which were tests that he had in fact just graded. The subjects, upon completion of their assessments of these test, were finished for the day.

On the second day the graders were given their same assigned tests, in the same order as on the first day and were told to grade their tests using the graduated scale answer key. Once again the grader scored his 15 tests after which he took a short break. He was then, as before handed the same three additional tests which
were identified by different test numbers than his previous graded tests, but which were tests that he had already evaluated. Following the grading of these tests the "grading" subjects were dismissed for the day.

On the third day, the graders were given their same assigned tests in the same order as on the two previous days, but now told to grade their tests using the "graduated-defined" scale answer key. Once again the subject graded his 15 tests after which he was allowed a 10 minute break. He was then, as on previous days, assigned the same three additional tests to grade which were identified by different test numbers than his most recently graded tests, but were tests that he had already graded.

The data were then collected and compiled into total individual test scores, and placed in categories corresponding to the trader type (BSI or graduate student graders) and answer key condition. The single variance of the five total test scores assigned to a given test by the five members in a subpopulation under a given answer key condition was then obtained. Variances were calculated in this manner for each of the 30 tests across all three answer key conditions. The calculation of these variances will be explained further in the Results section.
RESULTS

The measure which was of interest was the amount of variability between graders and within graders, with respect to the total score assigned to an individual test. Since the variance data were to be subsequently analyzed by a three factor repeated measurement design, variances were calculated for each of the grading subpopulations in each Test Set. For example, the variance for the 15 total test scores assigned to an individual test by the five graduate psychology students in Test Set A were calculated. Thus, a single variance was obtained for that individual test, as graded by this subpopulation under a specific grading condition. The variance was calculated for each of the 15 individual tests the graduate psychology students graded for each of the three grading conditions. This then yielded 15 variance values for a grading subpopulation, for each answer key condition, or a total of 45 variances for each subpopulation across the three answer key conditions. Variances were also calculated in this manner for the five graduate psychology students in Test Set A and the BSI students in Test Set A and B. The total number of variances calculated for subpopulations, across all answer key conditions was 180.
In addition to the variability between graders within a subpopulation, this study was also interested in the within grader variability in terms of the regrading of individual tests. The above mentioned design was again used in assessing the variability. Briefly, it consisted of calculating the variability between the three total test scores assigned by a grader and that grader's regrading of those three tests. Thus, three variances were calculated for each grader, under each answer key condition. This then resulted in a total of nine variability values for each grader across answer key conditions or, considering that there are a total of 20 graders, a total of 180 variance values for all graders.

As previously mentioned, these variances were analyzed by Weiner's three factor analysis of variance (ANOVA) design. The independent variables consisted of: (1) Test Set (A and B), (2) grader type (graduate and BSI students), (3) answer key condition (open, graduated, and graduated-defined). It may be remembered that the 10 graders for Test Set A (replication 1) and the 10 different graders for Test Set B (replication 2) were grading different sets of answers. Thus, the set of tests assigned to be corrected by the grader would be a replication variable. It is reasonable to assume that without regard for the set of test answers, graders using the same answer key should assess their answers
with approximately the same variability as the other grading subpopulations grading both the same and different tests, but using the same grading key.

The between grader variability was initially analyzed with a three factor repeated measurement design with repeated measurements on one factor only. This analysis is essentially given as Winer's case two (Winer, 1962, pp. 559-571). The Test Set variable or replication (A) was considered a random variable while grader experience (B) and answer key condition (C) were both considered fixed variables. The traditional ANOVA is not appropriate when A is random and B and C are fixed, however, Winer gives a deprivation in Appendix C utilizing Table 7, 4-8 (p. 752), which allows one to derive the proper F-ratios.

The ANOVA computed for the between grader variability yielded inconclusive results, as shown in Table 1. The only independent variable that proved to be significant was the replication variable, Test Set. There was also significance in all cases in which the replication variable (A) interacted with the other variables. Thus, every term involving factor A (replications): A, AB, AC, and ABC was significant at the .01 level. It was the opinion of the consulting statistician that this statistical result was due to: (1) contamination of the analysis by the high degree of significance of the replication variable, (2) extremely low degree of freedom values.
TABLE 1

A three factor repeated measurement design with repeated measurements on factor C, for between grader variability.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F Obtained</th>
<th>F Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>416,449.8</td>
<td>1</td>
<td>416,449.8</td>
<td>21.3</td>
<td>**</td>
</tr>
<tr>
<td>B</td>
<td>368,289.8</td>
<td>1</td>
<td>368,289.8</td>
<td>1.11</td>
<td>--</td>
</tr>
<tr>
<td>AB</td>
<td>333,164.1</td>
<td>1</td>
<td>333,164.1</td>
<td>17.0</td>
<td>**</td>
</tr>
<tr>
<td>E₁</td>
<td>1,097,129.8</td>
<td>56</td>
<td>19,591.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>382,603.4</td>
<td>2</td>
<td>191,301.7</td>
<td>3.14</td>
<td>--</td>
</tr>
<tr>
<td>AC</td>
<td>121,977.3</td>
<td>2</td>
<td>60,988.6</td>
<td>25.75</td>
<td>**</td>
</tr>
<tr>
<td>BC</td>
<td>107,525.1</td>
<td>2</td>
<td>53,762.5</td>
<td>1.87</td>
<td>--</td>
</tr>
<tr>
<td>ABC</td>
<td>57,422.5</td>
<td>2</td>
<td>28,711.2</td>
<td>12.12</td>
<td>**</td>
</tr>
<tr>
<td>E₂</td>
<td>265,251</td>
<td>112</td>
<td>2368.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Significant at the .01 level  
-- Not significant at the .01 level

F.05,1,56  =  4.00  
F.05,1,1  =  161  
F.05,2,112  =  3.07  
F.01,1,56  =  7.08  
F.01,2,2  =  99.0  
F.01,2,112  =  4.79  
F.05,2,2  =  19.0

A = Test Set variable or replication  
B = Grader type  
C = Answer key type
throughout the analysis and (3) the high critical F values which, in order to obtain significance, necessitated high F obtained values. The conclusion was that the results of the between variability are very complex in that different conclusions may be reached for different replications.

In view of the inclusive outcome an additional analysis was performed on Test Set A (replication 1) alone and also for Test Set B (replication 2) data alone. The form of this analysis was a two-way repeated measurement design with repeated measures on one factor (Winer, 1962). As Table 2 shows, grader experience (B) was not significant at the .01 level, having an obtained F of .1234 and a critical F value of 7.60. However, the answer key variable was significant at the .01 level, with an obtained F of 42.748 and a critical F value of 5.45. Also significant at the .01 level was the interaction of B and C, the obtained F value being 11.924 and the critical F value was 5.45.

In Test Set B (replication 2) grader experience (B), answer key condition (C) and the interaction of B and C were all significant at the .01 level (See Table 2). The obtained F value for B was 19.059 the critical F value being 7.60, while the obtained F for C, the answer key variable was 52.29 with the critical F being 5.45. Finally, the interaction term, BC, had a F value of 19.84 significantly above the F value of 5.45.
**TABLE 2**

A two way repeated measurement design with repeated measures on one factor, for between grader variability.

**Test Set A - Replication 1**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>440.011</td>
<td>1</td>
<td>440.011</td>
<td>.1234 --</td>
</tr>
<tr>
<td>B within</td>
<td>99,791.57</td>
<td>28</td>
<td>3,563.98</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>45,113.87</td>
<td>2</td>
<td>22,556.93</td>
<td>42.75 **</td>
</tr>
<tr>
<td>BC</td>
<td>12,584.09</td>
<td>2</td>
<td>6,292.04</td>
<td>11.92 **</td>
</tr>
<tr>
<td>BC within</td>
<td>29,549.38</td>
<td>56</td>
<td>507.67</td>
<td></td>
</tr>
</tbody>
</table>

**Test Set B - Replication 2**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>704,548.54</td>
<td>1</td>
<td>704,548.55</td>
<td>19.86 **</td>
</tr>
<tr>
<td>B within</td>
<td>246,048.49</td>
<td>28</td>
<td>35,477.89</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>459,490.87</td>
<td>2</td>
<td>229,745.43</td>
<td>52.29 **</td>
</tr>
<tr>
<td>BC</td>
<td>174,379.49</td>
<td>2</td>
<td>87,189.74</td>
<td>19.844 **</td>
</tr>
<tr>
<td>BC within</td>
<td>246,048.31</td>
<td>56</td>
<td>4,393.72</td>
<td></td>
</tr>
</tbody>
</table>

** Significant at the .01 level  
-- Not significant at the .01 level

\[ F_{.05,1,23} = 4.20 \quad F_{.01,1,23} = 5.45 \]
\[ F_{.05,2,28} = 3.35 \quad F_{.01,2,28} = 7.60 \]

B = Grader type  
C = Answer key type

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The within grader analysis involved a three-way repeated measurement analysis with repeated measures on all three variables. The analysis showed that very little was significant, at least no main effects or double interactions. The triple interaction that was significant in Test Set B should be noted, but it was the opinion of the consulting statistician that accurate interpretation was not possible (see Table 3).

In order to obtain values which would reflect the absolute differences in grader variability between the various conditions, mean variances were calculated for each of the ANOVA cells. As Table 4 shows, the mean between grader variability of graduate graders assessing Set A tests was 69.6 using the open scale answer key, increasing to 76.7 using the graduated answer key and finally decreasing again to 18.3 while using the graduated-defined answer key. For the BSI graders assessing these same tests there was a continual decrease in the mean variability between graders, with the mean variability being 92.1 for the open scale answer key, 48.9 for the graduate scale key and 31.7 using the graduated-defined answer key.

The mean between grader variability of graduate students assessing Set B tests also continually decreased as the answer key became increasingly structured. While using the open scale key these graders had a mean variability of 95.2 decreasing to 63.4 under the graduated
TABLE 3

A three factor repeated measurement design with repeated measurements of all three factors for within grader variability.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>F.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>105,789</td>
<td>2</td>
<td>52,894.5</td>
<td>2.97</td>
<td>4.46</td>
</tr>
<tr>
<td>B</td>
<td>132</td>
<td>1</td>
<td>132</td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>88,857</td>
<td>2</td>
<td>44,428.5</td>
<td>1.79</td>
<td>3.98</td>
</tr>
<tr>
<td>AB</td>
<td>4,634</td>
<td>2</td>
<td>2,317</td>
<td>.39</td>
<td>4.46</td>
</tr>
<tr>
<td>AC</td>
<td>48,742</td>
<td>4</td>
<td>12,186</td>
<td>1.82</td>
<td>3.01</td>
</tr>
<tr>
<td>BC</td>
<td>39,521</td>
<td>2</td>
<td>19,760.5</td>
<td>.85</td>
<td>4.46</td>
</tr>
<tr>
<td>ABC</td>
<td>23,738</td>
<td>4</td>
<td>5,934</td>
<td>.36</td>
<td>3.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>F.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>274,859</td>
<td>2</td>
<td>137,430</td>
<td>4.251</td>
<td>4.46</td>
</tr>
<tr>
<td>B</td>
<td>5,800</td>
<td>1</td>
<td>5,800</td>
<td>.367</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>30,289</td>
<td>2</td>
<td>15,145</td>
<td>.61</td>
<td>4.46</td>
</tr>
<tr>
<td>AB</td>
<td>49,830</td>
<td>2</td>
<td>24,915</td>
<td>3.85</td>
<td>4.46</td>
</tr>
<tr>
<td>AC</td>
<td>54,470</td>
<td>4</td>
<td>13,617</td>
<td>.497</td>
<td>3.01</td>
</tr>
<tr>
<td>BC</td>
<td>92,073</td>
<td>2</td>
<td>46,036.5</td>
<td>1.51</td>
<td>4.46</td>
</tr>
<tr>
<td>ABC</td>
<td>97,234</td>
<td>4</td>
<td>24,308</td>
<td>4.58</td>
<td>3.01</td>
</tr>
</tbody>
</table>

A = Test Set variable or replication  
B = Grader type  
C = Answer key type

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
TABLE 4

The mean variances for between grader variability.

<table>
<thead>
<tr>
<th>Answer Key Type</th>
<th>Open</th>
<th>Graduated Defined</th>
<th>Row Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate Graders</td>
<td>$\bar{x} = 69.6$</td>
<td>$\bar{x} = 18.3$</td>
<td>54.9</td>
</tr>
<tr>
<td>BSI Graders</td>
<td>$\bar{x} = 92.1$</td>
<td>$\bar{x} = 31.7$</td>
<td>57.7</td>
</tr>
<tr>
<td>Graduate Graders</td>
<td>$\bar{x} = 95.2$</td>
<td>$\bar{x} = 24.6$</td>
<td>61.0</td>
</tr>
<tr>
<td>BSI Graders</td>
<td>$\bar{x} = 393.7$</td>
<td>$\bar{x} = 126.8$</td>
<td>240.3</td>
</tr>
<tr>
<td>Column Means</td>
<td>162.75</td>
<td>97.25</td>
<td>50.4</td>
</tr>
</tbody>
</table>
answer key condition and further decreasing to 24.6 using the graduate-defined scale. The BSI graders assessing these same tests had an extremely high mean variability between graders using the open answer key with the value being 393.7. However, this was followed by a large decrease to 199.7 when the BSI graders used the graduated answer key and to 126.8 using the graduated-defined scale.

Overall means were also calculated for between grader variability. Across answer key conditions the total overall mean was 162.75 for the open answer key, decreasing to 97.25 for the graduated key and further decreasing to 50.4 for the graduated-defined answer key. The overall mean for all graduate graders was 58.0, much less than the 149.0 overall mean of the BSI graders.

The means for within grader variability were also calculated, as shown in Table 5. The mean within grader variability for graduate graders assessing Set A tests was 94.7 using the open scale answer key, 46.5 using the graduated key, and 3.6 using the graduated-defined answer key. The mean within grader variability for BSI graders correcting these same tests was very similar with the mean variability being 98.9 for the open scale, 26.5 using the graduated key, and 5.6 using the graduated-defined key.

The mean within grader variability for graduate students assessing Set B tests continually decreased as the
TABLE 5
The mean variances for within grader variability.

<table>
<thead>
<tr>
<th>Answer Key Type</th>
<th>Open</th>
<th>Graduated</th>
<th>Graduated Defined</th>
<th>Row Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate Graders</td>
<td>$\bar{X}=94.7$</td>
<td>$\bar{X}=46.5$</td>
<td>$\bar{X}=3.6$</td>
<td>48.3</td>
</tr>
<tr>
<td>BSI Graders</td>
<td>$\bar{X}=98.9$</td>
<td>$\bar{X}=26.5$</td>
<td>$\bar{X}=5.6$</td>
<td>48.2</td>
</tr>
<tr>
<td>Graduate Graders</td>
<td>$\bar{X}=94.5$</td>
<td>$\bar{X}=46.5$</td>
<td>$\bar{X}=3.3$</td>
<td>48.1</td>
</tr>
<tr>
<td>BSI Graders</td>
<td>$\bar{X}=171.9$</td>
<td>$\bar{X}=9.1$</td>
<td>$\bar{X}=12.7$</td>
<td>54.2</td>
</tr>
<tr>
<td>Column Means</td>
<td>115</td>
<td>32.2</td>
<td>6.3</td>
<td>51.2 = Overall Average</td>
</tr>
</tbody>
</table>
answer key became increasingly structured. While using the open scale these graders had a mean variability of 94.5, decreasing to 46.5 under the graduated answer key condition and further decreasing to 3.3 using the graduated-defined scale. The BSI graders assessing these same tests had a mean variability of 171.9 using the open scale answer key. However, this was followed by a large decrease to 9.1 using the graduated scale and to 12.7 using the graduated scale key.

The overall means were also calculated for within grader variability. Across answer key conditions the total overall mean was 115 for the open answer key, decreasing to 32.2 for the graduated key and further decreasing to 6.3 for the graduated-defined answer key. The overall mean for all graduate graders was 48.2, slightly less than the 54.2 overall mean of the BSI graders.
DISCUSSION

The results of the three factor repeated measurement design was repeated measurements on factor C only yielded inconclusive results. On recommendation of the consulting statistician an additional ANOVA was performed in which each replication was treated separately. In terms of this ANOVA any conclusion is difficult to make in light of the high replication by factor B, C, BC interactions, but one possible partial explanation for the replication BC interaction is that B differences (grader experience) are not significant for Test Set A, but are significant for Test Set B. Thus, there was a "grader experience" difference for Test Set B, but not for Test Set A. However, it is even difficult to state this in light of the strong and persistent BC interaction. Thus, we must qualify the statement by stating that the differences between graduate "experienced", graders and BSI, "inexperienced," graders depend upon which level answer key condition (C) we are referring to. The BC interaction is highly significant for both replications.

The difficulty in the between grader analysis is also due to the significant ABC interaction found in the first analysis. Thus, the form or cause of BC
interaction depends upon which level of replication one is inferring about. As Table 5 shows, for Test Set A the graduate graders mean is lower than BSI graders for the open scale key and the graduated-defined scale, but this is reversed for the graduated answer key. This is an additional explanation of BC interaction for replication 1. For replication 2, Test Set B, the means decrease as you go from the open scale answer key to the more structured graduated-defined answer key for both experienced and inexperienced graders, but at different rates of decrease. This may be the cause of BC interaction for this replication. In light of this strong BC interaction it is difficult to interpret properly the significant answer key (C) effect differences for both replications. There are differences and the direction of the differences is the same for both replications, namely a decreasing sequence of means is observed as you go from the open scale key to the graduated scale to the graduated-defined answer key.

These results were most graphically shown through the calculation of mean variability values. As the answer key became increasingly structured, the overall mean variabilities for both between and within graders decreased. The difference in the overall mean variabilities between graders contrasting the graduated answer key and the graduated-defined key was greater than the difference between the open key and the graduated
scale. The overall mean variability decreased approximately 60% from the open key to the graduated key and decreased approximately 52% from the graduated scale condition to the graduated-defined key condition. Thus, in light of these results, if one wishes to eliminate between grader variability it is advisable to adopt an answer key similar in structure to the graduated-defined scale.

The difference in the overall mean variabilities within graders comparing the graduated answer key and the graduated-defined key was greater than the difference between the open key and the graduated scale. The overall mean variability decreased approximately 71% from the open key to the graduated scale condition and approximately 81% from the graduated scale condition to the graduated-defined scale condition. Again the design and implementation of an answer key similar to the graduated-defined scale is a profitable and effective method of eliminating grader variability.

The overall mean variability between graders of graduate students was 61% less than that of BSI students. However, the overall mean variability within graders was 115 for the open answer key, decreasing to 32.2 for the graduated key and further decreasing to 6.3 for the graduated-defined answer key. The overall mean for all graduate graders was 48.2, slightly less than the 54.2 overall mean of the BSI graders.
The results of this study, although somewhat inconclusive, should be of extreme interest to those individuals involved in developing evaluation measures. It is indeed a desirable feature, with respect to the standardization of a written test, to compose an answer key such that the assessments of several graders, grading a specific set of tests, would be highly similar. In reference to the results of this study, the test developer who desires to reduce the variability in the grading of his test, should design an answer key that closely resembles the structure of the graduated-defined answer key (Appendix D). The design should include a division of the answer into small point gradations, and an explanation of those gradations.
1. Mrs. Brown's social interactions are of inadequate duration. Her frequent conversations with fellow residents usually consists of simply saying, "Hi", or "How are you", and nothing else. Describe the entire procedure, including all the steps, that you would use to establish longer duration social interactions by Mrs. Brown.
2. Patti S., a fourth grade student, has recently been labeled a "problem student" by her teacher because of her high frequency of "out-of-seat" behavior. Several times an hour she leaves her seat and walks around the room. The teacher has attempted to eliminate this behavior by scolding Patti whenever she leaves her seat without permission. However, the result has been an increase in the frequency of "out-of-seat" behavior. Describe the entire procedure, including all the steps, that you would use to increase the frequency of appropriate "in-seat" behavior and to eliminate the "out-of-seat" behavior.
3. Mrs. Brown's social interactions are of inadequate duration. Her frequent conversations with fellow residents usually consists of simply saying, "Hi", or "How are you", and nothing else. Describe a "shaping" procedure that you could use to increase the duration of Mrs. Brown's social interaction. Be sure to include the following items in your answer:

a. Define the problem behavior.
b. Describe how you would obtain your baseline data.
c. Define the terminal behavior.
d. Identify the initial behavior or response to be shaped.
e. Describe briefly the general procedure.
f. Describe a strategy for maintaining the appropriate terminal behavior.
Patti S., a fourth grade student, has recently been labeled a "problem student" by her teacher because of her high frequency of "out-of-seat" behavior. Several times an hour she leaves her seat and walks around the room. The teacher has attempted to eliminate this behavior by scolding Patti whenever she leaves her seat without permission. However, the result has been an increase in the frequency of "out-of-seat" behavior. Describe a procedure which "differentially reinforces" appropriate "in-seat" behavior and extinguishes "out-of-seat" behavior. Be sure to include the following items:

a. Define the problem behavior.
b. Describe how you would obtain your baseline data.
c. Describe the behavior to be reinforced.
d. Describe briefly the extinction procedure.
e. Describe briefly the reinforcement procedure.
f. Describe a strategy for maintaining the appropriate terminal behavior.
Open Scale Answer Key

Grading Instructions: Grade each of your assigned four questions by using a 0-25 point scale, with 25 points being the equivalent of a superior response. Your actual point award is to be dependent upon the qualitative content of the answer as compared to the model answers given below.

Questions (1) and (3) The testee should include the following items in his procedure:

a. A Definition of the Problem Behavior. The behavior problem should be objectively defined in behavioral terms. Award no credit for definitions that use such "mentalistic" terms as "insecure", "emotionally upset", "immature", "wants attention", "seeks recognition", etc. The answer should include language to the effect that Mrs. Brown frequently engages in social interactions of short or inadequate duration.

Example: "Mrs. Brown engages in social interactions of short duration. These limited social engagements occur with a high frequency and generally consist of one word."

b. A Short Description of How They Would Obtain Their Baseline Data. In their description they should include how the problem behavior is going to be measured. The testee may state that he would obtain a baseline by timing the duration of verbal interactions or by counting the number of words emitted by Mrs. Brown during an interaction.

Example: "To obtain a baseline on the behavior I would observe Mrs. Brown at various times during the day and time the duration of her verbal interactions."

c. A Definition of the Appropriate, Terminal Behavior. The definition should acknowledge the fact that the terminal behavior to be reached is social interactions of a given, adequate duration.
d. An Identification of the Initial Behavior or Response. The initial response in the shaping procedure should be a behavior related toward the terminal behavior which occurs with some minimal frequency. The testee probably should initiate the shaping procedure by reinforcing Mrs. Brown for verbal responses consisting of at least one word or possibly a few words. A verbal response having an $X$ duration is not the best selection for initiating the procedure inasmuch as early approximations will be short and thus difficult to time. However, give full credit if they state that they would begin shaping by reinforcing verbal interactions of very short duration.

Example: "The initial response in my shaping procedure would consist of Mrs. Brown emitting a greeting consisting of at least one word, such as 'Hi'."

3. A Brief Description of the Treatment Procedure. The testee should give a short description of how he would proceed to implement his shaping procedure. The testee must begin his procedure by reinforcing a behavior that is only vaguely related to the terminal behavior, but that occurs with at least minimal frequency. From the initial behavior the procedure continues in a step-like manner with reinforcement being presented for durations of increasing length. The answer should include the idea that successive approximations are reinforced leading to the terminal behavior.

Example: "I would begin by reinforcing the initial behavior of saying, 'Hi', everytime it occurred. After this response had been adequately strengthened I would proceed to the next requirement or step. I would continue to differentially reinforce successive approximations"
until the terminal behavior, verbal interactions having a duration of two minutes, was reached."

f. A Strategy for Maintaining the Appropriate Terminal Behavior. The testee should recognize that following acquisition of the terminal behavior he would decrease the frequency of reinforcement presented to some intermittent schedule. He should also state that he would concurrently present social reinforcement to facilitate maintenance of the behavior. Lastly, the strategy should contain a statement concerning the generalization of the behavior to others. In order for the appropriate behavior to generalize in the presence of others the experimenter must reinforce Mrs. Brown for such behavior under those stimulus conditions.
Questions

(2) and (4) The testee should include the following items in his procedure:

a. A Definition of the Problem Behavior. The behavior problem should be objectively defined in behavioral terms. Award no credit for definitions that use such "mentalistic" terms as "insecure", "emotionally upset", "immature", "wants attention", etc. The testee should state something to the effect that the problem behavior consists of Patti leaving her seat and walking around the room at inappropriate times without permission.

Example: "The problem behavior to be eliminated is Patti's out-of-seat behavior. This inappropriate behavior can be defined as any response in which the subject's buttocks are not in contact with the seat of the chair under stimulus conditions in which out-of-seat behavior is inappropriate (e.g. teacher speaking, permission not granted, students to be engaging in school work, etc.)."

b. A Short Description of How They Would Obtain Their Baseline Data. In their description they should include how the problem behavior is going to be measured. The testee may state that the teacher or an independent observer would serve as a data collector and record the frequency or the number of times Patti emitted the out-of-seat behavior per hour, and/or per day. Another method of collecting the baseline data would be to use a time-sampling technique in which you would record the frequency or number of out-of-seat behavior for one hour three times a day.

Example: "I would obtain my baseline data by having the teacher record the frequency of out-of-seat behavior per hour."

c. A Definition of the Behavior to be Reinforced. The behavior to be reinforced should be objectively defined in behavioral terms. The testee should describe what in-seat behavior consists of.
Example: "The behavior to be reinforced is in-seat behavior which is defined as Patti sitting in her chair with buttocks in contact with the seat."

d. A Brief Description of the Extinction Procedure. The testee should objectively describe the mechanics of the extinction procedure. The testee's answer should state something to the effect that the teacher and the students should ignore Patti, contingent on the occurrence of her out-of-seat behavior. There should also be an indication that the teacher and students are to disregard Patti only when she is engaging in the inappropriate behavior.

Example: "The extinction procedure would consist of the teacher ignoring Patti whenever she engaged in out-of-seat behavior. The teacher should also instruct the students in the class to do likewise."

e. A Brief Description of the Reinforcement Procedure. The testee should objectively describe the mechanics of the reinforcement procedure. The answer should contain language to the effect that the testee would attend to Patti and provide social reinforcement contingent on her in-seat behavior. The teacher should reinforce Patti with tokens or points on a Fixed-Interval schedule with the X amount of points being awarded every hour, contingent on Patti having engaged in less than X frequency of out-of-seat behavior per hour. These points could be accumulated and later exchanged for various tangible goods or privileges.

Example: "I would present social reinforcement contingent on Patti's appropriate in-seat behavior. At the same time I would present points (token reinforcement) on Fixed-Interval one hour, contingent on Patti not engaging in out-of-seat responses on more than two occasions."

f. A Strategy for Maintaining the Appropriate Terminal Behavior. The testee should recognize that following acquisition of the terminal behavior he would decrease
the frequency of reinforcement by increasing the length of the fixed-interval required for reinforcement. He should also state that he would concurrently present social reinforcement to facilitate maintenance of the behavior.

Example: "I would maintain the appropriate in-seat behavior by gradually increasing the fixed interval requirement. Furthermore, I would concurrently facilitate continuance of the appropriate behavior through the presentation of contingent social reinforcement."
Graded Scale Answer Key

Grading Instructions: Grade each of your four assigned questions by using the given point scale following each part of the answer. Your actual point award is to be dependent upon the qualitative content of the answer as compared to the model answers given below.

Questions (1) and (3) The testee should include the following items in his procedure:

a. A Definition of the Problem Behavior. The behavior problem should be objectively defined in behavioral terms. Award no credit for definitions that use such "mentalistic" terms as "insecure", "emotionally upset", "immature", "wants attention", "seeks recognition", etc. The answer should include language to the effect that Mrs. Brown frequently engages in social interactions of short or inadequate duration.

Example: "Mrs. Brown engages in social interactions of short duration. These limited social engagements occur with a high frequency and generally consist of one world." Award 0-5 points.

b. A Short Description of How They Would Obtain Their Baseline Data. In their description they should include how the problem behavior is going to be measured. The testee may state that he would obtain a baseline by timing the duration of verbal interactions or by counting the number of words emitted by Mrs. Brown during an interaction.

Example: "To obtain a baseline on the behavior I would observe Mrs. Brown at various times during the day and time the duration of her verbal interactions." Award 0-4 points.

c. A Definition of the Appropriate, Terminal Behavior. The definition should acknowledge the fact that the terminal behavior to be reached is social interactions of a given, adequate duration.
Example: "The appropriate terminal behavior to be achieved is Mrs. Brown's engagement in social, verbal interactions having a duration of at least two (or X) minutes." Award 0-2 points.

d. An Identification of the Initial Behavior or Response. The initial response in the shaping procedure should be a behavior related toward the terminal behavior which occurs with some minimal frequency. The testee probably should initiate the shaping procedure by reinforcing Mrs. Brown for verbal responses consisting of at least one word or possibly a few words. A verbal response having an X duration is not the best selection for initiating the procedure inasmuch as early approximations will be short and thus difficult to time. However, give credit if they state that they would begin shaping by reinforcing verbal interactions of very short duration.

Example: "The initial response in my shaping procedure would consist of Mrs. Brown emitting a greeting consisting of at least one word, such as 'Hi'." Award 0-2 points.

e. A Brief Description of the Treatment Procedure. The testee should give a short description of how he would proceed to implement his shaping procedure. The testee must begin his procedure by reinforcing a behavior that is only vaguely related to the terminal behavior, but that occurs with at least minimal frequency. From the initial behavior the procedure continues in a step-like manner with reinforcement being presented for durations of increasing length. The answer should include the idea that successive approximations are reinforced leading to the terminal behavior.

Example: "I would begin by reinforcing the initial behavior of saying, 'Hi', everytime it occurred. After this response had been adequately strengthened I would proceed to the next requirement or step. I would continue to differentially reinforce successive approximations
until the terminal behavior, verbal interac-
tions having a duration of two minutes, was reached." Award 0-6 points.

f. **A Strategy for Maintaining the Appropriate Terminal Behavior.** The testee should recognize that following acquisition of the terminal behavior he would decrease the frequency of reinforcement presented to some intermittent schedule. He should also state that he would concurrently present social reinforcement to facilitate maintenance of the behavior. Lastly, the strategy should contain a statement concerning the generalization of the behavior to others. In order for the appropriate behavior to generalize in the presence of others the experimenter must reinforce Mrs. Brown for such behavior under those stimulus conditions. Award 0-6 points.
Questions (2) and (4) The testee should include the following items in his procedure:

a. **A Definition of the Problem Behavior.**
   The behavior problem should be objectively defined in behavioral terms. Award no credit for definitions that use such "mentalistic" terms as "insecure", "emotionally upset", "immature", "wants attention", etc. The testee should state something to the effect that the problem behavior consists of Patti leaving her seat and walking around the room at inappropriate times without permission.

   **Example:** "The problem behavior to be eliminated is Patti's out-of-seat behavior. This inappropriate behavior can be defined as any response in which the subject's buttocks are not in contact with the seat of the chair under stimulus conditions in which out-of-seat behavior is inappropriate (e.g. teacher speaking, permission not granted, students to be engaging in school work, etc.) ."  
   Award 0-3 points.

b. **A Short Description of How They Would Obtain Their Baseline Data.** In their description they should include how the problem behavior is going to be measured. The testee may state that the teacher or an independent observer would serve as a data collector and record the frequency or the number of times Patti emitted the out-of-seat behavior per hour, and/or per day. Another method of collecting the baseline data would be to use a time-sampling technique in which you would record the frequency or number of out-of-seat behavior for one hour three times a day.

   **Example:** "I would obtain my baseline data by having the teacher record the frequency of out-of-seat behavior per hour."  
   Award 0-4 points.
c. A Definition of the Behavior to be Reinforced. The behavior to be reinforced should be objectively defined in behavioral terms. The testee should describe what in-seat behavior consists of.

Example: "The behavior to be reinforced is in-seat behavior which is defined as Patti sitting in her chair with her buttocks in contact with the seat." Award 0-2 points.

d. A Brief Description of the Extinction Procedure. The testee should objectively describe the mechanics of the extinction procedure. The testee's answer should state something to the effect that the teacher and the students should ignore Patti, contingent on the occurrence of her out-of-seat behavior. There should also be an indication that the teacher and students are to disregard Patti only when she is engaging in the inappropriate behavior.

Example: "The extinction procedure would consist of the teacher ignoring Patti whenever she engaged in out-of-seat behavior. The teacher should also instruct the students in the class to do likewise." Award 0-5 points.

e. A Brief Description of the Reinforcement Procedure. The testee should objectively describe the mechanics of the reinforcement procedure. The answer should contain language to the effect that the testee would attend to Patti and provide social reinforcement contingent on her in-seat behavior. The teacher should reinforce Patti with tokens or points on a Fixed-Interval schedule with the X amount of points being awarded every hour contingent on Patti having engaged in less than X frequency of out-of-seat behavior per hour. These points could be accumulated and later exchanged for various tangible goods or privileges.

Example: "I would present social reinforcement contingent on Patti's appropriate in-seat behavior. At the same time I would present points (token reinforce-
ment) on Fixed-Interval one hour, contingent on Patti not engaging in out-of-seat responses on more than two occasions."
Award 0-8 points.

f. **A Strategy for Maintaining the Appropriate Terminal Behavior.** The testee should recognize that following acquisition of the terminal behavior he would decrease the frequency of reinforcement by increasing the length of the fixed-interval required for reinforcement. He should also state that he would concurrently present social reinforcement to facilitate maintenance of the behavior.

Example: "I would maintain the appropriate in-seat behavior by gradually increasing the fixed interval requirement. Furthermore, I would concurrently facilitate continuance of the appropriate behavior through the presentation of contingent social reinforcement." Award 0-3 points.
Graduated-Defined Scale Answer Key

Grading Instructions: Grade each of your four assigned questions by using the given point scale contained in each of the examples provided in the model answer. Your actual point award is to be dependent upon the qualitative content of the answer as compared to the model answers given below.

Questions
(1) and (3) The testee should include the following items in his procedure.

a. A Definition of the Problem Behavior. The behavior problem should be objectively defined in behavioral terms. Award no credit for definitions that use such "mentalistic" terms as "insecure", "emotionally upset", "immature", "wants attention", "seeks recognition", etc. The answer should include language to the effect that Mrs. Brown frequently engages in social interactions of short or inadequate duration.

Example: ["Mrs. Brown engages in social interactions (1 point) [of short duration (1 point)].] These limited social engagements [occur with a high frequency (1 point)] and generally [consist of one word (1 point).""] Award 0-4 points.

"Social interactions"—The testee should recognize that social interactions are the behavior associated with the problem behavior.

"Short duration"—The answer must mention that the actual problem with Mrs. Brown's social interactions are that they are of short, inadequate, limited, or insufficient duration.

"Occur with a high frequency"—The answer should have something to the effect that the short interactions occur frequently.

"Consist of one word"—The testee should briefly describe what the typical conversation consists of.
b. **A Short Description of How They Would Obtain Their Baseline Data.** In their description they should include how the problem behavior is going to be measured. The testee may state that he would obtain a baseline by timing the duration of verbal interactions or by counting the number of words emitted by Mrs. Brown during an interaction.

Example: ["To obtain a baseline on the behavior I would observe Mrs. Brown (1 point) [at various times during the day (1 point)] and [time the duration (1 point)] of her [verbal interactions (1 point)]."]

"Observe Mrs. Brown"--The answer should identify the subject to be observed during baseline.

"At various times"--The testee should recognize the fact that he would observe Mrs. Brown on a "time-sampling" basis.

"Time the duration"--The testee must state that his baseline measurement would consist of timing the duration.

"Verbal interactions"--The behavior to be observed is Mrs. Brown's social-verbal interactions.

c. **A Definition of the Appropriate, Terminal Behavior.** The definition should acknowledge the fact that the terminal behavior to be reached is social interactions of a given, adequate duration.

Example: ["The appropriate terminal behavior to be achieved is Mrs. Brown's engagement in social, verbal interactions (1 point) having a [duration of at least two (or X) minutes (1 point)]."] Award 0-2 points.

"Engagement in social, verbal interactions"--The subject of the terminal behavior is engagement in social, verbal interactions.

"Duration of at least two (or X) minutes"--The testee should further define the terminal behavior by stating a duration requirement.
d. **An Identification of the Initial Behavior or Response.** The initial response in the shaping procedure should be a behavior related toward the terminal behavior which occurs with some minimal frequency. The testee probably should initiate the shaping procedure by reinforcing Mrs. Brown for verbal responses consisting of at least one word or possibly a few words. A verbal response having an X duration is not the best selection for initiating the procedure inasmuch as early approximations will be short and thus difficult to time. However, give credit if they state that they would begin shaping by reinforcing verbal interactions of very short duration.

**Example:** "The initial response in my shaping procedure would consist of Mrs. Brown emitting a greeting (1 point) [consisting of at least one word, such as 'Hi' (1 point).]"

Award 0-2 points.

"**Consist of Mrs. Brown emitting a greeting**"—The answer should provide a general description of the initial response such as a greeting, short conversation, etc.

"**Of at least one word**"—The answer should objectively state the reinforcing criteria of the initial behavior.

e. **A Brief Description of the Treatment Procedure.** The testee should give a short description of how he would proceed to implement his shaping procedure. The testee must begin his procedure by reinforcing a behavior that is only vaguely related to the terminal behavior, but that occurs with at least minimal frequency. From the initial behavior the procedure continues in a step-like manner with reinforcement being presented for durations of increasing length. The answer should include the idea that successive approximations are reinforced leading to the terminal behavior.
Example: "I would [begin by reinforcing (1 point)] the [initial behavior of saying, 'Hi', (1 point)] [every time (1 point)] [it occurred (1 point)]. [After this response had been adequately strengthened (1 point)] I would proceed to the next requirement or step. I would [continue to differentially reinforce successive approximations (1 point)] [until the terminal behavior, verbal interactions having a duration of two (or X) minutes, was reached (1 point)]."

Award 0-7 points.

"By reinforcing"--The testee should make reference to the fact that the shaping procedure involves the use of reinforcement.

"Initial behavior"--The answer must begin with a statement of the procedure to be used in reinforcing the initial behavior.

"Every time"--The testee must acknowledge that the reinforcer is to be presented on a continuous reinforcement schedule.

"It occurred"--The answer should take into consideration the fact that the reinforcer is presented contingently or immediately following the occurrence of the behavior.

"Adequately strengthened"--The testee should say something to the effect that he would proceed to succeeding stages of his program only after the initial behavior occurred with a high frequency.

"Differentially reinforce successive approximations"--The answer must contain a reference to the fact that the testee would continue to reinforce the successive steps in his program procedure.

"Terminal behavior"--The answer must recognize that the shaping procedure is to be continued in the above manner until the terminal behavior is realized.
f. A Strategy for Maintaining the Appropriate Terminal Behavior. The testee should recognize that following acquisition of the terminal behavior he would [decrease the frequency of reinforcement (1 point)] presented to some [intermittent schedule (1 point)]. He should also state that [he would concurrently present social reinforcement (1 point)] to facilitate maintenance of the behavior. Lastly, the [strategy should contain a statement concerning the generalization of the behavior to others (1 point).] In order for the appropriate behavior to generalize in the presence of others the [experimenter must reinforce Mrs. Brown for such behavior under those stimulus conditions (1 point).] Award 0-6 points.

"Decrease the frequency of reinforcement" -- The testee's answer should mention that he would decrease the frequency of reinforcement.

"Intermittent schedule" -- The answer should state that this decrease in reinforcement frequency would be accomplished by presenting the reinforcer on an intermittent schedule.

"Present social reinforcement" -- The testee should acknowledge the fact that the social reinforcers (praise, approval, etc.) would be presented.

"Generalization of the behavior to others" -- The testee's answer should recognize the fact that in order for Mrs. Brown's appropriate behavior to be maintained a generalization procedure must be used.

"Reinforce Mrs. Brown" -- The answer must state that a generalization strategy would involve reinforcement of appropriate behavior.

"Stimulus conditions" -- The answer should state that the generalization strategy would require Mrs. Brown to engage with others in social interactions of adequate duration and reinforcement to be contingent on this behavior.
Questions (2) and (4) The testee should include the following items in his procedure:

a. **A Definition of the Problem Behavior.** The behavior problem should be objectively defined in behavioral terms. Award no credit for definitions that use such "mentalistic" terms as "insecure", "emotionally upset", "immature", "wants attention", etc. The testee should state something to the effect that the problem behavior consists of Patti leaving her seat and walking around the room at inappropriate times without permission.

Example: ["The problem behavior to be eliminated is Patti's out-of-seat behavior (1 point).] This [inappropriate behavior can be defined as any response in which the subject's buttocks are not in contact with the seat of the chair (1 point)] [under stimulus conditions in which out-of-seat behavior is inappropriate (1 point)] (e.g. teacher speaking, permission not granted, students to be engaging in school work, etc.)." Award 0-3 points.

"Patti's out-of-seat behavior"--The answer should recognize out-of-seat behavior as the problem behavior.

"Behavior can be defined as"--The answer should contain a more detailed, objective definition. It need not be like the definition given above, as long as it is objective and accurate.

"Under stimulus conditions"--The answer should include a reference to the fact that the out-of-seat behavior is the problem because it is emitted under stimulus conditions when such behavior is considered inappropriate.

b. **A Short Description of How They Would Obtain Their Baseline Data.** In their description they should include how the problem behavior is going to be measured. The testee may state that the teacher or an independent observer would serve
as a data collector and record the frequency or number of times Patti emitted the out-of-seat behavior per hour, and/or per day. Another method of collecting the baseline data would be to use a time-sampling technique in which you would record the frequency or number of out-of-seat for one hour three times a day.

Example: "I would [obtain my baseline data by having the teacher (1 point)] [record the frequency] [record the frequency (1 point)] [of out-of-seat behavior (1 point)] [per hour (1 point)]." Award 0-4 points.

"Having the teacher"--The answer must state who will record baseline.

"Record the frequency"--The answer should state the particular measure the testee would use to record baseline data.

"Out-of-seat behavior"--The answer should mention what behavior is to be measured.

"Per hour"--The answer should include a statement as to what time scale is to be used in measuring the frequency of the problem behavior.

c. A Definition of the Behavior to be Reinforced. The behavior to be reinforced should be objectively defined in behavioral terms. The testee should describe what in-seat behavior consists of.

Example: "The [behavior to be reinforced is in-seat behavior (1 point)] which is [defined as Patti sitting in her chair (1 point)] [with her buttocks in contact with the seat (1 point)]." Award 0-2 points.

"In-seat behavior"--The answer should state that the in-seat behavior is the response to be strengthened.
"Which is defined as . . . " — The answer must objectively define what in-seat behavior consists of.

d. A Brief Description of the Extinction Procedure. The testee should objectively describe the mechanics of the extinction procedure. The testee's answer should state something to the effect that the teacher and the students should ignore Patti, contingent on the occurrence of her out-of-seat behavior. There should also be an indication that the teacher and student are to disregard Patti only when she is engaging in the inappropriate behavior.

Example: "The extinction procedure would consist of [the teacher (1 point)] [ignoring Patti (1 point)] [whenever she engaged in (1 point)] [out-of-seat behavior (1 point)].] The [teacher should also instruct the students in the class to do likewise (1 point).] Award 0-5 points.

"The teacher" — The testee should state that the individual implementing the extinction procedure would be the teacher.

"Ignoring Patti" — The testee should state that included in the extinction procedure to be used is the teacher's deliberate act of ignoring Patti's out-of-seat behavior.

"Whenever she engaged in" — The testee must in some manner give recognition to the fact that the teacher's disregard of Patti's out-of-seat behavior is either contingent, following, whenever it occurs, or every time the inappropriate response is emitted.

"Out-of-seat behavior" — The testee must state what behavior is to be ignored.

"Instruct the students" — The answer include language to the effect that the students would also participate in the extinction procedure.
A Brief Description of the Reinforcement Procedure. The testee should objectively describe the mechanics of the reinforcement procedure. The answer should contain language to the effect that the testee would attend to Patti and provide social reinforcement contingent on her in-seat behavior. The teacher should reinforce Patti with tokens or points on a Fixed-Interval schedule with the X amount of points being awarded every hour contingent on Patti having engaged in less than X frequency of out-of-seat behavior per hour. These points could be accumulated and later exchanged for various tangible goods or privileges.

Example: "I would [present social reinforcement (1 point)] [contingent on (1 point)] [Patti's appropriate in-seat behavior (1 point).] [At the same time I would present points (1 point)] [token reinforcement] on [Fixed-Interval one hour, (1 point)] [contingent on Patti (1 point)] [not engaging in out-of-seat responses (1 point)] [on more than two occasions (1 point).]"

Award 0-8 points.

"Present social reinforcement"--The answer should state that social reinforcement would be used.

"Contingent"--The answer must state that reinforcement is presented following, upon occurrence of, whenever Patti engages in, etc.

"Appropriate in-seat behavior"--The answer must mention that reinforcement is to be presented for in-seat behavior.

"Present points"--The testee must also include the presentation of token reinforcers as part of his proposed program.

"Fixed-Interval schedule one hour"--The answer should state the schedule on which the token reinforcement will be presented on.

"Contingent"--The answer must state that token reinforcement is to be presented following, upon occurrence of, whenever Patti engages in, etc.
"Engaging in out-of-seat responses"—The answer should identify the gross behavior to be reinforced.

"More than two occasions"—The answer should state the frequency required for reinforcement.

f. A Strategy for Maintaining the Appropriate Terminal Behavior. The testee should recognize that following acquisition of the terminal behavior he would decrease the frequency of reinforcement by increasing the length of the fixed-interval required for reinforcement. He should also state that he would concurrently present social reinforcement to facilitate maintenance of the behavior.

Example: "I would maintain the appropriate in-seat behavior by [gradually increasing the fixed-interval requirement (1 point).] Furthermore, I would concurrently [facilitate continuance of the appropriate behavior through the presentation of contingent (1 point)] [social reinforcement (1 point).]"
Award 0-3 points.

"Increasing the fixed-interval"—The testee's answer should mention in his procedure that he would decrease the frequency of reinforcement.

"Contingent"—The answer must state that social reinforcement is presented either following, immediately after, or for in-seat behavior.

"Social reinforcement"—The testee should state in his maintenance strategy that he would present social reinforcement, e.g. praise, approval, etc., contingent on in-seat behavior being exhibited.
APPENDIX E
<table>
<thead>
<tr>
<th>Test Number</th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
<th>Total Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Data Sheet II**

<table>
<thead>
<tr>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>C</td>
<td>B</td>
<td>A</td>
<td>D</td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td>F</td>
<td>A</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
REFERENCES


