8-1974

The Effects of Informational Feedback on the Grading Accuracy of Undergraduate Assistants

Paul D. Coyne
Western Michigan University

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

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/2507
THE EFFECTS OF INFORMATIONAL FEEDBACK
ON THE GRADING ACCURACY OF
UNDERGRADUATE ASSISTANTS

by

Paul D. Coyne

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
August 1974
ACKNOWLEDGEMENTS

I wish to express my appreciation to Dr. Jack Michael for his encouragement and supervision during the course of this study and in the writing of this thesis. I express my gratitude to Dr. Paul Mountjoy, Dr. Galen Alissi and Dr. Bradely E. Huiteman for their criticism and aid in preparing this report. A special acknowledgement is presented to Alice Dybsky.

The financial aid of an assistantship and the intellectual training from the faculty of the Psychology Department have made my graduate study a pleasure and a privilege.

Paul Douglas Coyne
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 M-6122

COYNE, Paul Douglas
THE EFFECTS OF INFORMATIONAL FEEDBACK ON THE GRADING ACCURACY OF UNDERGRADUATE ASSISTANTS.

Western Michigan University, M.A., 1974
Psychology, general

Xerox University Microfilms, Ann Arbor, Michigan 48106
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>II</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>9</td>
</tr>
<tr>
<td>III</td>
<td>11</td>
</tr>
<tr>
<td>IV</td>
<td>25</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>30</td>
</tr>
</tbody>
</table>
INTRODUCTION

The application of behavioral principles to the goal of improving college instruction has been successfully demonstrated in a number of recent studies. Following Keller's (1968) lead a variety of similarly based contingency management techniques have been applied to the teaching of various psychology courses (Ferster, 1968; Malott, 1969; Hesse, 1971; McMichael and Corey, 1967; Lloyd, 1969) as well as to the teaching in areas including sociology (Witters and Kent, 1970) and introductory statistics (Myers, 1970).

The use of students as teaching assistants has correspondingly risen along with these programs of instruction.* For example, Malott and Whaley (undated) used sixty five student teaching apprentices

*The first system of instruction in which pupils, acting under the direction of a teacher, taught other less advanced pupils can be traced to Andrew Bell in 1797 and Joseph Lancaster in 1798. These attempts to offer some elementary schooling to large numbers of poor children was a result of a general movement of social reform characteristic of the age. Few trained teachers were available, but by using student teachers (monitors) a single instructor could manage 400 to 500 pupils. The instructor taught his monitors who taught what they had learned to groups of less advanced students. The monitors took full charge of the instruction, grading, and promotion, while the teacher worked out detailed plans and routines. The low cost of maintaining these schools promoted the growth of public education. The system lost its appeal, however, after more funds and teachers became available, and the fact that there was little technology developed to ensure a high quality of instruction. Contingency managed programs of instruction, on the other hand, have available a more advanced technology for ensuring quality performances. Teaching assistants are assigned specific duties in the instructional system. Some may be graders while others aid in the development of study aids. The technology used to develop stimulus materials is also more advanced so that the assistant does not become the sole source of instruction.
organized in a hierarchy of teaching apprentices, advanced teaching apprentices, undergraduate assistants, and graduate assistants to teach 1,000 introductory psychology students. The assistants programmed twelve hours of educational activity per week for the average student while providing two one-hour laboratories and two one-hour lecture and discussion sections each week.

The use of these paraprofessionals can provide the instructor with a "behavior amplifier", i.e., they can amplify the effects of the instructor's behavior toward the student in two areas. The first is in the presentation of adequate stimulus materials. The undergraduate assistant can be valuable in the development of educational materials such as objectives, quiz questions, and course manuals. However, as Michael (1969) points out, even excellent materials have little effect on the student if used improperly. This leads us to the second area, the provision of differential consequation that ensures the development of adequate study behaviors in the student and maintains his educational activity. To insure proper usage of educational material frequent sampling of the students' repertoire together with a vehicle for remediation is necessary. Often neither the instructor nor the student knows what the student knows until he has been tested. Frequent extensive tests can accurately assess the level of mastery the student possesses. A remediation procedure can be used to implement necessary revisions in the stimulus materials.

Frequent sampling, however, involves a manpower problem. Traditionally the faculty member had two alternatives, either infrequent essay exams or frequent true-false, multiple-choice exams. Multiple-
choice exams have the advantage that they can be scored in a short period of time by a machine or clerk. They have three disadvantages: 1) they are difficult to construct without the alternative answers appearing too similar or too different from the correct choice; 2) the alternative answers act as prompts or clues for the correct answer. To recognize what is correct out of several choices is easier than producing the correct answer; and 3) the student's repertoire is difficult to sample. A student can study only those items which are transferable to these exams and consequently fail to acquire the complete repertoire the instructor wishes to develop.

Frequent essay exams, although they amply sample and help to develop the repertoire, cause a manpower problem because they are hard to grade. Careful scrutiny is needed to weed out superfluous information the student often includes in the answer. It is difficult to make up questions with unambiguous answers to avoid such superfluous material.

Teaching assistants can be an inexpensive means to solve this manpower problem. With the proper procedures and controls the undergraduate teaching assistant can facilitate the frequent use of essay exams even in the large classroom.

There are two classes of variables that control the accuracy of the assistants' grading. The first class can be labeled knowledge variables. They consist of those factors that ensure competency in the educational activity. Also considered in this class are those factors which control the behaviors commonly referred to as "emotional maturity" and "understanding". The second class is more appropriately labeled performance variables. Even though the assistant may have
competency in the stimulus materials, this does not guarantee that he will perform acceptably as a grader. This class of variables is responsible for the assistant participating in the grading activity itself and the quality of such participation.

Those factors included in the first type are most easily controlled for by appropriate screening and selection procedures. These variables have usually been established as controlling variables in the prior histories of the assistants. Consequently, they do not readily lend themselves to control by the instructor. An assistant's academic record or a screening exam, together with the personal observation of his previous interactions with students and teachers, should give the instructor a reasonable estimate of his competency.

Even under the best possible conditions, it is still the case that questions will arise that require a complex knowledge of the stimulus materials. The assistant grader's level of competency may not be extensive enough to ensure accurate grading at all times. It is probably necessary for the instructor to deal with test questions that are complex or answers that are difficult to grade. Consequently, it may be desirable to develop a filtering system that enables the instructor to spend time grading only those exam questions and answers with ambiguous features.

A verbal instruction given to the graders by the teacher to err in the direction of stringency when comparing the student's exam answer to the answer key, coupled with a re-grade period during which such errors could be corrected, facilitates the development of such a system. This procedure greatly reduces the number of exams that the
instructor must come in contact with. Those papers that are clearly of "A" or "F" quality can be expediently dealt with by the assistants. Even those papers that fall in between these two values can be expected to have some exam items that are obviously correct or incorrect. For those answers that have a vale of ambiguity surrounding them, this procedure would: 1) consequate the students' ambiguity, perhaps lowering the probability of such ambiguity occurring in the future, thus further reducing the number of exams the instructor must grade; and 2) appropriately bring them to the instructor's attention.

Performance variables are more difficult for the instructor to control. First, few prospective assistants have training in grading essay exams, and secondly, there are often available several competing and more reinforcing activities, e.g., sleep, social activities or studying other academic topics. Both of these factors may lead to the assistant's grading becoming inadequate in quality. The problem that remains is how can the instructor control these performance variables adequately.

In designing a set of procedures for the implementation of a behavioral system, it is also important to provide a systemic program to observe and consequate the staff's behavior. In the case of the assistant grader, it is important that a procedure be installed that encourages the grader for his compliance with the stringency bias but not overly so. In addition, it should compel the assistant to avoid the lure of becoming too lenient. After grading several essay exams, it may become clear to the grader that it is much easier to simply mark a particular answer correct than to make the sometimes difficult
discrimination between correct and not so correct. It is conceivable that the assistant could hand out "A's" to all students and never get "caught" since students generally never complain about injustices that improve their grade. Therefore, a viable method for monitoring and consequating should be employed.

One alternative is to provide information to the assistants concerning their performances. The procedures employed to accomplish this can be defined as a feedback system. Feedback has been shown to be an effective technique for modifying behavior in a number of different settings and populations (Budzynski and Stayva, 1969; Kaess and Zeaman, 1960; Trowbridge and Cason, 1932; and O'Brien and Azrin, 1970). Feedback has been shown to be an effective consequence for nonprofessional hall attendants in a state institution for retarded children. Panyan et al. (1970) gave feedback to attendants as a reinforcer for applying operant conditioning techniques. After they had completed a four-week course in operant conditioning techniques, staff were instructed to carry out training sessions in self-help skills with retarded children and to keep daily performance records on each child in training. No contingencies on their performance were in effect during baseline, which was conducted for various lengths of time. Following baseline, a feedback program was initiated. Results showed an increase in the use of operant conditioning techniques, as measured by the number of sessions held, by staff personnel. The data further suggest that a consistent and systematic use of feedback can be an economical method of maintaining satisfactory performance of staff personnel in the place of daily supervision.
The present experiment was initiated to investigate the effects of feedback on the grading accuracy of undergraduate teaching assistants on tests taken by a population of undergraduate college students. Briefly described, the course involved in this study included: 1) weekly study objectives relevant to the lectures and reading assignments for each unit; 2) on Monday there was a short quiz over Part A objectives, a lecture, and a short quiz over the lecture; 3) on Tuesday there was a short quiz over Part B objectives, a lecture, and a short quiz over the lecture; 4) on Wednesday there was a fifteen minute study period, a one and one-half hour essay exam; 5) on Thursday there was a short quiz over exam answers, a re-grade period, a remedial lecture, and a short quiz over remedial lecture; and 6) on Friday there was a fifteen minute study period with an optional one and one-half hour re-exam.
METHOD

Subjects

The six subjects for this study were drawn from a population of fifteen students who assisted in grading exams for an undergraduate level course on Skinner's Verbal Behavior. All the students had received an "A" in the course the previous semester.

A four week baseline period was initiated during which all grading assistants were measured for their grading accuracy. Three matched pairs were then selected with the subjects of each pair randomly assigned to either a feedback or non-feedback group.

Setting

The investigations were conducted at Western Michigan University, Kalamazoo, Michigan. Every Wednesday morning, a two hour examination was administered to approximately 100 students enrolled in the Verbal Behavior class. After the exam the papers were collected and sorted alphabetically into fifteen folders. Each folder contained six papers. At the grader assistants' meeting each grader was given a folder which he had never previously graded. The graders' meeting served four main functions: 1) to review the material through a review lecture given by the professor; 2) to provide an opportunity to ask questions over the week's material; 3) to clarify any irregularities in the test or answer sheet; and 4) to present feedback to the appropriate graders.
Procedure

Baseline consisted of: 1) arbitrarily picking 14 points worth of test items from the 30 point exam; 2) randomly choosing three of six student exam papers in each of the grading assistants' folders; and 3) re-grading them and listing "points possible", "points given", "points erroneously given", and "points erroneously taken" on each item. A leniency ratio was computed by dividing "points erroneously given" by "total points given". A stringency ratio was computed by dividing "points erroneously taken" by "total points possible". This procedure was carried out during both phases of the study, however, during the experimental phase only the six selected subjects were assessed. During the graders' meeting the week before the experimental phase was to begin the instructor announced that: 1) three graders were randomly chosen to take part in a pilot study that might possibly evolve into a master's thesis; 2) out of their six graded exam papers, three papers would be re-graded with emphasis on one item; and 3) that personal feedback from the instructor to these three graders would be given each week after the graders' meeting. They were not told that the experimenter was attached to the class or a part of the procedure in any official sense. The remaining three subjects of the six selected subjects did not receive any official feedback. They were used for the purpose of control.

The personal feedback from the instructor to the three graders was initiated on the fifth week. The experimental phase lasted for eight weeks. Reliability was measured by an independent observer.
who randomly chose five test papers from the experimenter's set of re-graded test papers. The experimenter made no marks on the re-graded test papers; all re-graded adjustments and computations were tallied on a separate special re-grade sheet. The independent observer re-graded the selected items (14 points per paper) and assessed grading accuracy as described above. Reliability was computed by dividing agreements by agreements plus disagreements and then multiplying the quotient by 100. Reliability checks were taken twice during baseline and three times during the experimental phase. The names of the subjects in each group were never revealed to the independent observer.
RESULTS

The reliability assessment yielded the following scores: during baseline, the leniency error rates were 93% and 92%, for stringency errors they were 100% and 100%; during the experimental phase the figures were 86%, 95% and 95%, for leniency; 100%, 100% and 98% agreements on stringency error. These reliability scores were considered to reflect an accurate assessment of the assistant graders' scores by the experimenter.

The performances of the three matched pairs of assistant graders are illustrated in Figures 1 through 6. Figures 1, 2, and 3 illustrate leniency errors whereas 3, 4, and 5 illustrate stringency errors. The figures indicate no marked difference in grading accuracy between the treatment and control subjects in any of the three matched pairs. An analysis of variance with repeated measures (Kirk, 1968) was performed on the data to test for significant differences. The test was run separately for both the leniency and stringency error scores. Accordingly, the results were nonsignificant at the .05 level. This suggests that the sample feedback used had little or no effect on the assistant graders' accuracy.

The mean leniency and stringency scores for the treatment group were 3.5% and 4.7%, respectively; for the non-treatment group they were 6.5% and 2.1%, a difference of 3% for leniency and 2.6% for stringency. These values were in the predicted direction; the treatment group had a lower leniency score but a higher stringency score than the control group. This possibly reflects the emphasis placed on the graders...
Figure 1

Distribution of leniency error scores for "R.D." and "D.B."
Figure 2

Distribution of leniency error scores for "A.R." and "B.B."
Figure 3

Distribution of leniency error scores for "S.S." and "G.L."
Figure 4

Distribution of stringency error scores for "R.D." and "D.B."
Figure 5

Distribution of stringency error scores for "A.R." and "B.B."
Baseline

Experimental Phase

Announcement

% of Error

Unit Number

Feedback (AR)

Control (BB)

Figure #5
Figure 6

Distribution of stringency error scores for "S.S." and "G.L."
incorporating a stringency bias in their grading. The total mean error score for the treatment group was 4.16%. The accompanying score for the control group was 4.3%. These scores were acquired by summing the error scores, both leniency and stringency, for the treatment group and dividing by the number of scores (i.e., 48). This score was computed for the control group in the same manner. These scores further reflect the insignificant difference in the percentage error rates between the two groups.

The experimenter found few problems in initiating the baseline and experimental procedures. The instructor was helpful as were his three graduate assistants. Some minor problems did arise, however, in the reacquisition of student papers to be re-graded. Students were instructed to hand in those test papers that had "PLEASE RETURN" marked on them in bold colored ink. Invariably, some students would hand their papers in late or not at all. This resulted in some of the weekly leniency and stringency scores being based on the number of papers the experimenter received for a given grader, rather than from a random sample of three, selected from the grader's total number of papers graded that week, as proposed. Although this problem probably affected the data slightly, it culminated during unit 10, when not enough papers were retrieved from the students to adequately derive error scores on any of the graders. This was the reason for the elimination of data for unit 10 on Figures 1, 2, and 3. The other missing data points (Figure 3, unit 8 for "B.B." and Figure 3, unit 13 for "D.B.") were due to the fact that the grader did not grade during that week.
DISCUSSION

The experiment was initiated within a college level course whose structure and orientation had evolved from Keller's (1968) program of instruction and is described by Michael (1971) and Williams (1973). The professor had incorporated the use of undergraduate assistants for the purpose of grading student essay exams. Procedures had previously been established to minimize the possible adverse effects of using these paraprofessionals. They consisted of: 1) a selection procedure which included an "A" in the course, and/or a personal interview; 2) the weekly graders' meeting, previously described; 3) instructions to the grader to grade with a stringency bias, i.e., taking off points when a student's answer is ambiguous as compared to an answer key; 4) a student re-grade period where the student was offered a chance to have various items re-graded by the professor and his graduate assistants. This was designed to offset the effects of the stringency bias and general grading errors. Nevertheless, it was felt that there were limitations with these controls.

A systemic method for monitoring and consequating the graders was not developed. Other than the negative comments from students and the periodic observation during the student re-grade period, effective procedures for assessing grading accuracy were nonexistent. Therefore, it was difficult to adequately evaluate the above "established procedures", either separately or collectively.

When developing a system of behavioral instruction, it is often the case that the instructor will put together several components of
a model from research that has individually shown significant effects, or that intuitively suggest an improvement over a more traditional counterpart. The entire package is then placed under scientific scrutiny. Unfortunately, however, it is still not known which component(s) of the system contribute to its success and which do not. It may be that some elements are essential to the success of the program while others offer little or nothing. Therefore, a long term evaluation of the instructional system should consider the contribution of each component of that system.

In keeping with these guidelines this research presents three major developments: 1) it has developed a mechanism by which the assistant graders could be accurately monitored and evaluated; 2) it established a new component in the instructional system, i.e., a procedure that consequential grading accuracy by the feedback; and 3) it isolated the component to evaluate its contribution to the instructional system.

The contribution of the feedback component to the educational system, in terms of its effects on improving grader error, was shown to be negligible. One hypothesis is that the "established procedures" described above are sufficient in maintaining accurate grading. This hypothesis is supported by looking at the total mean grading error of 4.16% for the feedback group and 4.3% for the control group; also, in the low error scores of the treatment and control subjects in the three matched pairs, illustrated in Figures 1 through 6.

Another alternative hypothesis concerns a procedure initiated into the instructional system by the professor during the second week.
of the treatment phase (unit 7). Two hours before the graded exam papers were to be delivered back to the students, the professor and his three graduate assistants scanned those exams that were only a few points away from an "A" grade. Their purpose was to correct blatant stringency errors or make other necessary adjustments in a student's grade. This filtered out students who would otherwise have to participate in the official student re-grade period held the next day. The assistant graders were "aware" of this procedure, although no attempt was made to describe it to them in detail. It is quite possible that this "awareness" may have caused the graders to grade more accurately than if this component was nonexistent. In other words, the graders' performances were being monitored by the professor; therefore, he had the opportunity to observe the proficiency of the graders. Since a professor's esteem is held valuable by a number of students for a variety of reasons, e.g., letters of recommendation, future assistantships, and so on, the students' "awareness" of the procedure may have positively affected their grading accuracy. Consequently, it may have masked a significant difference between the error scores of the treatment and control groups. This hypothesis is important if the reader considers this new component separate from that previously labeled as the "established procedures". In any case, the previously mentioned suggestion of further research to evaluate the contribution of the various components seems reasonable since there is insufficient data to determine the degree to which this hypothesis is valid.
It can be argued that the "established procedures" are not effective enough in controlling grader accuracy because of the occurrence of individual grader scores above 10%. However, it should be pointed out that the criterion for acceptable grading error is quite arbitrary. It seems reasonable to base such a criterion on the practical costs of the mechanisms that ensure the desired level of grading error. What follows is a discussion of possible alternatives toward reducing grader error.

A possibility exists that the feedback systems developed there for monitoring and consequating student graders may offer beneficial results when incorporated into a program of instruction that does not have the number of components devised for the function of controlling assistant grading accuracy as employed here. It is reasonable to assume that the number of components mentioned (the selection procedure, the re-grade period, the monitoring and feedback procedure, the graders' meeting, the stringency bias, and the scanning procedure used by the instructor for correctional purposes) may be interchangeable in terms of what component(s) could be excluded due to practical considerations and still maintain an acceptable standard of grading error.

However, some modification of this procedure can be suggested. For example, the instructor could utilize an intensive monitoring and consequation feedback phase at the beginning of the semester to develop appropriate stimulus control over assistants' grading. The second phase might consist of a significantly smaller sample of the graders' work to be monitored with feedback given primarily to maintain
the acceptable degree of accuracy. This modification procedure is based on the hypothesis that: 1) there is some skill to be developed before an individual can become a grader of acceptable quality; and 2) that the prospective grader, having an academic history of reinforcement, does not already come equipped with this skill in his repertoire if selected by traditional screening procedures as described earlier in this paper.

Another suggestion lies in the selection procedures used in choosing prospective graders. The instructor might include typical exam questions and answers that the prospective assistant would be grading. A specific or curved criterion could be drawn with acceptances being contingent on those who met the criterion level of accuracy. Of course, instructors are often in the position where they must go out and recruit rather than sit back and select. Still, the accuracy scores obtained with this procedure could be used to compare with performances sampled later in the semester.

In short, this research has presented three developments: 1) a monitoring and evaluation mechanism for assessing assistant graders; 2) a new component in the ongoing instructional system; and 3) an isolation of the component to evaluate its effect on the instructional system vis-a-vis grading accuracy in undergraduate assistant graders. Furthermore, it has offered possible hypotheses to account for this effect together with suggestions for further research in this area. Finally, alternatives for practical applications of the procedures described in this paper were discussed.
REFERENCES


