8-1977

A Point System for Maintaining Accurate Grading in a PSI Course

Terry E. McSween
Western Michigan University

A POINT SYSTEM FOR MAINTAINING ACCURATE GRADING IN A PSI COURSE

by

Terry E. McSween

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 1977

ACKNOWLEDGEMENTS

I wish to thank Shirley Cary, Gregg Beleuger, and Jim DeShane for the many hours of work each of them invested in this project. I also wish to thank R. W. Malott for arranging the contingencies that led to the original proposal; Jack Michael, Brian Iwata, and Wayne Fuqua for their invaluable comments on my original manuscript; Tom Welsh for his invaluable supervision; and also the entire SCEP staff of the Winter of 1977, Meg Dorsey, Tim Wysocki, Lonni Moffet, and Sue McSween for their roles in the completion of my research and this manuscript.

Terry E. McSween

TABLE OF CONTENTS

INTRODUCTION

Personalized systems of instruction (PSI) have consistently proved to be superior to traditional teaching procedures (see reviews by Hursh, 1976; Traveggia, 1976; and Williams, 1976). Hursh suggests that the consistency of these studies is tempered only by their lack of experimental control. He reviewed sixteen studies comparing PSI with traditional instruction and found only five provided any attempt to control for the reliability of the dependent measure while none report reliability procedures for the independent variable. Hursh also reports that seven other studies documenting the effectiveness of PSI all failed to report any reliability measures. Typically, when PSI researchers provide reliability data, they report levels of around 85 percent (Born, Gledhill, and Davis, 1972; Sheppard and MacDermot, 1970), while others use entirely multiple choice and fill in the blank items which minimize problems in grading reliability (Alba and Pennypacker, 1972).

With students' grades constituting one of the primary dependent variables of a majority of such studies, the Zeitgeist was set for the development of quality control systems to ensure reliable grading. Such procedures became increasingly important as researchers began investigating the components of PSI. Unfortunately very few researchers have addressed the problem. A computer search of both the psychological and educational literature

yielded no articles on the problems of monitoring and maintaining grading accuracy. Other sources provided only three references on the subject.

In the earliest paper to address the issue of grading accuracy, Coyne (1974) discussed the development of a monitoring and feedback mechanism for insuring accurate grading by student assistants. Subjects were student assistants responsible for grading weekly exams in an undergraduate course on Skinner's Verbal Behavior. Six assistants were paired on the basis of grading accuracy and one subject of each pair was randomly assigned to either a feedback or nonfeedback group. The experimenter randomly selected three of six tests graded each week and regraded 14 of the 30 items on each. During intervention the course instructor gave feedback on grading errors to subjects in the feedback group. The nonfeedback group served as a control and received no official feedback. While the feedback group showed a larger decrease in the percentage of grading errors than the nonfeedback group, the actual size of the difference was relatively small (about .2%) and his results did not show statistical significance and, in fact, grading errors increased after intervention. Even though Coyne failed to demonstrate experimental control, his study is significant as perhaps the first attempt to accurately monitor and improve the grading accuracy of student assistants.

In the first published study on grading accuracy Semb (1975) discussed grading accuracy and the use of student proctors in an

introductory child development class. Using a multiple baseline across groups, Semb provided proctors with feedback about their grading accuracy. At the beginning of each class, a student assistant gave each proctor written comments on items graded during the previous day. The percentage of agreement rose from 87 to 98 and from 89 to 98 for Groups 1 and 2 respectively. On the eight item quizzes the mean number of items graded correctly improved from 6.9 to 7.83 and from 7.12 to 7.82 for Groups 1 and 2 respectively. A third group received no feedback and showed no improvement in accuracy. Semb concluded that monitoring and feedback about grading errors were effective for producing and maintaining accurate grading by student assistants.

Semb had a problem with reliability, however. The student assistant agreed with the student proctors' grading of 98 percent of quiz items graded after intervention, but an independent observer agreed with the student assistant's grading on only 93 percent of the quiz items. It is not clear why student proctors were capable of 98 percent accuracy when the student assistants were only capable of 93 percent agreement. In addition, the reliability coefficient combines reliability data from the base-line and treatment conditions with reliability data from the control group. If reliability was high during baseline conditions and for the control group, it may have been low during treatment resulting in high reliability while the effect may have been due to changes in the assistant's grading criteria.

In a descriptive study on grading accuracy Fuqua and Heckler (1977) reported the development of a quality control system designed to develop and maintain accurate grading by 110 student assistants in a university wide instructional system serving 800 students. The procedures were not described in detail, but between 70 and 80 percent of the students' quizzes taken during the first two weeks were reevaluated for grading accuracy. This density of checks decreased as the semester progressed but was not allowed to drop below 30 percent. When an error was detected the grader received feedback and a short retraining session in grading procedures. Additional errors resulted in grade penalties. The graders made errors which increased a student's grade 12 times more frequently than errors which decreased a student's grade. Fuqua and Heckler reported that this ratio never dropped below 9 to 1 in two years of such monitoring. The percentage of grading errors dropped from more than 9 at the beginning of the semester to less than two, but as the authors pointed out, their study lacked the experimental manipulation of error contingencies needed to reliably demonstrate a functional relationship. Also, they provided no data to indicate whether or not the number of errors in the student's favor decreased in relation to errors against the student.

The purpose of this study was to experimentally validate a system designed to control the reliability of student graders in a personalized system of instruction. The study took place in

the Student Centered Education Project (SCEP) at Western Michigan University. SCEP is an accelerated two-semester program for psychology majors and minors. The two components are designed as SCEP I and SCEP II. In the SCEP I program, students earn seven credits for two freshman level psychology classes: an introduction to behavior analysis and analysis of children's behavior. SCEP II is a nine-credit course covering abnormal behavior, a more advanced course in behavior analysis, and an applied laboratory.

The program is administered by a hierarchy of student staff. After completing a semester of satisfactory work in the SCEP program, a student may apply to work as a teaching apprentice for either laboratory credit, if the student is enrolled in SCEP II, or for three hours of independent study credit. Teaching apprentices are responsible for giving quizzes, grading, and helping students with course materials. After completing a semester of satisfactory performance as a teaching apprentice, the student may become an advanced teaching assistant for additional credit. Advanced teaching assistants monitor and supervise the performance of teaching apprentices. Course assistants are paid students who worked a semester at each of the other positions. Course assistants supervise the other levels of staff, meet with graduate staff, and do research projects in the system.

Teaching apprentices worked for daily points that determined their final grade. Graders received five points for remaining "on task" during their scheduled time. Intervention in the present

study involved a change in this daily point contingency. After intervention, one point was given for each test that was scored accurately, as determined by an advanced teaching assistant. These consequences should theoretically offset the undesirable social contingencies existing between graders and students. The natural contingencies appear to encourage errors in favor of the student. This point consequence and feedback on grading errors constituted the independent variable for the present study.

METHOD

Subjects and Setting

Subjects were fifteen student teaching apprentices working in the SCEP program. Teaching apprentices are assigned responsibility for activities in one of three rooms: the quiz room, where students took quizzes; the study room, where students asked questions and reviewed course materials; and the grading room, where teaching apprentices graded the student quizzes. Teaching apprentices rotated to a new area and worked for two hours each day. There were three shifts, two morning shifts which began at 8:00 a.m. and 10:00 a.m. for SCEP II and the SCEP I shift which began at 12:00 p.m.

This study involved only the grading room and the teaching apprentices working in the grading room. The teaching apprentices grading on a given day were the subjects of that day. All teaching apprentices participated, though the number on any given day ranged from one to four.

Observation

One advanced teaching assistant from each setting monitored grading activities for research credit. Each day the advanced teaching assistant randomly selected five of the quizzes scored by each teaching apprentice and then regraded each quiz. The actual number of quizzes graded varied from day to day because

the SCEP courses were self paced and students could take quizzes at their own rate. In order to complete the course material however, students had to complete one unit per day. Generally, students had to retake about 20 percent of their quizzes in order to score above the 90 percent criteria that allowed progression to the next unit. Estimates based on these figures indicate that the advanced teaching assistants monitored about 27 percent of all SCEP I quizzes, about 30 percent of SCEP II (8 a.m.) quizzes, and about 33 percent of SCEP II (10 a.m.) quizzes.

On a data sheet the advanced teaching assistant recorded the name of the TA unit and form of the test he was monitoring. Each time he disagreed with the way the test was scored, he placed a "+" if the student was incorrectly penalized one point; "++" if the student was incorrectly penalized two points; and a "-" or "--" if the student received one or two points incorrectly. A "+" indicated the grader made a false positive identification of an error, while a "-" indicated that the grader made a false negative, or failed to identify an error.

Both the actual number of teaching apprentices and the number of tests monitored varied throughout the semester. The number of teaching apprentices scheduled to grade depended on the number of students taking quizzes and varied each day from one to four. The mean number of quizzes rechecked each day was six for SCEP II's 8:00 a.m. shift; thirteen for SCEP II's 10:00 a.m. shift, and eight for SCEP I.

About two weeks after implementation of quiz monitoring, teaching apprentices recorded the time when they began grading and the time they completed grading each quiz. Each time was rounded to the nearest five seconds and recorded on the right-hand corner of each quiz. Another teaching apprentice placed the quizzes in students' files and recorded grading time on a separate data sheet. This data sheet also provided a record of the number of quizzes graded each day.

Procedures

Baseline

The advanced teaching assistants monitored false negative and false positive identifications of errors without providing feedback of any kind during baseline. The teaching apprentices knew their grading was being monitored, but were accustomed to data being collected on various aspects of the system. If asked, the advanced teaching assistant explained that we wanted to determine the levels of grading accuracy in SCEP without mentioning potential consequences or specific data.

Consequences and Feedback

Advanced teaching assistants provided daily points that counted towards the teaching apprentices' grades and feedback as consequences for grading. They gave one point for each of the five quizzes without a grading error. Teaching apprentices earn 75 points each

week, so the five points earned for grading once a week comprised approximately seven percent of their final grade for the course.

Intervention began with a shift meeting during which each staff member received a handout explaining the new grading system and the rationale for the procedures (Figure 1). The staff then asked any questions they had and signed the handout to indicate that they read and understood the procedures.

During baseline, all subjects made substantially more false negative identifications of errors. This data prompted the development of a differential point system to encourage teaching apprentices to grade stringently, that is, they were encouraged to count unclear answers wrong. The advanced teaching assistant gave one-half point for a quiz which had one false positive identification of an error. In other words, the teaching apprentice lost one-half point if he made an error against the student but still received half credit for grading the quiz. If the teaching apprentice made more than one false positive, or one or more false negative identifications of an error, then he received no points for grading the quiz. On days when the advanced teaching assistant was too busy to recheck all five quizzes, he gave the teaching apprentice one point for each quiz that did not get monitored.

The point consequence made reassignment to the grading room less desirable than other areas of SCEP where teaching apprentices received daily points for their activities rather than outcomes.

Figure 1. Handout explaining the consequences for grading errors.

GRADING CONSISTENCY HANDOUT #1

Since the beginning of February, I have been collecting data on grading accuracy. An ATA is pulling five quizzes graded by each TA in the grading room and regrading all items. Frankly, your grading was much more accurate than I expected, but there are problems. We have been recording a mean of slightly more than five errors a day, but the problem is that in some cases the errors are running sixteen to one in favor of the students!

On page four of the student handbook we tell students that our graders are instructed to grade on the side of strictness in an effort to ensure the most accurate feedback possible with the regrade procedure and protect students from strict initial grading. Until now we have had no procedure to protect them from inaccurate feedback. In order to encourage you to grade accurately and on the side of strictness, the following contingencies will go into effect tomorrow. ATAs were giving five points for the activity of grading. Beginning tomorrow TA's will earn one point for each of the quizzes monitored without an error. If an error is made against the student, the TA earns one-half point (these errors are considered slightly more desirable than errors for the student because of the regrade option.) If the error is in the student's favor, the TA receives no point for that quiz. No more than one point may be lost per quiz. In the event that less than five quizzes are monitored, you may receive credit for monitored quizzes

plus the number of quizzes not monitored. The ATA will provide feedback on any items he believes are graded inaccurately. The point consequence is relatively small to keep this procedure as non-aversive as possible and still provide a source of motivation for careful grading. If you disagree with the feedback given by the ATA and if you sincerely need the points, you may appeal the point loss to the UGA. Please appeal only if the point loss is going to affect your grade. In most cases if an item is that debatable, it should have been counted wrong, then the student could clarify the answer through the regrade procedure.

The consequences are set up in a manner that should encourage you to count questionable answers wrong. I hope this will actually help decrease the time you spend grading. In any event, I am also collecting data on the time you spend grading and we will recycle if problems arise.

Advanced teaching assistants often have to reassign teaching apprentices to assure each area of SCEP has adequate staff. To counter the potential aversiveness of reassignment to the grading room, the advanced teaching assistant gave a bonus point to teaching apprentices reassigned to the grading room.

When the advanced teaching assistant finished monitoring quizzes, he would total the number of points earned by each teaching assistant involved in grading. He then recorded these points on the teaching apprentices' monitor sheets. If a teaching apprentice made an error, the advanced teaching assistant showed the quiz and error to the teaching apprentice and explained why he considered the item graded incorrectly.

Experimental Design

SCEP was composed of three shifts during the winter, 1977. Each shift implemented the grading point system at different times, constituting a multiple baseline across groups. Intervention began during a shift meeting on March 14th for SCEP II's 10:00 a.m. shift, on March 28th for SCEP II's 8:00 a.m. staff, and on April 4th for SCEP I.

Reliability

On each Tuesday and Thursday a course assistant collected all quizzes rechecked by advanced teaching assistants and scored the quiz for a third time. This data was used to assess the reliability

of the advanced teaching assistants.

Different reliability procedures yielded widely divergent results. The total number of agreements divided by the total number of items resulted in an overall Type II reliability coefficient of .97 (range .92 - .99) for the SCEP II advanced teaching assistant and .98 (range .94 - 1.0) for the SCEP I advanced teaching assistant. The number of agreements on the occurrence of a correctly scored answer divided by that number plus disagreements on the occurrence of a correctly scored answer yielded the same as the overall Type II calculations. The mean reliability coefficient for occurrence of correct scoring was .97 for SCEP II (range .93 - .99) and was .98 for SCEP I (range .94 - 1.0). However reliability on the occurrence of an incorrectly scored response, calculated by the last equation using grading errors rather than correctly scored items, was much lower. The SCEP II advanced teaching assistant had a mean reliability on the occurrence of an incorrectly scored answer of .34 (range 0 - .67) while the same coefficient for SCEP I was .36 (range 0 - 1). Table 1 presents reliability coefficients for each condition.

During the last three weeks of the study, beginning on March 31st, the reliability observer checked to insure that the advanced teaching assistant monitoring quizzes was applying the point consequence appropriately. A total of six such checks occurred and only one revealed any discrepancy between the number of points deducted. The SCEP II assistant correctly recorded 83 percent of

Table 1. Reliability data before and after intervention.

Table 1

Type II Reliability Data for Each Condition

Group	Baseline	Intervention
SCEP II 10 a.m.		
Overall	.96	.98
Occurrence of errors	.33	.36
SCEP II 8 a.m.		
Overall	.97	.98
Occurrence of errors	.30	.42
SCEP I		
Overall	.97	.99
Occurrence of errors	.36	.30

the grading points while the SCEP I assistant recorded 100 percent of the grading points correctly.

Due to limited staff, no reliability data was collected on other dependent variables.

RESULTS

The primary dependent variable was the percentage of items which teaching apprentices graded incorrectly (percent of errors). Table 2 presents the mean and range of errors for each subject, as well as the number of days per condition for each subject. False negatives were more common than false positive errors, and therefore of greater concern. Of the 15 subjects, 13 subjects averaged fewer false negatives after intervention, one subject averaged the same percentage of errors and one subject averaged a higher percentage of errors. Seven subjects averaged more false positives after intervention than during baseline. Six subjects made fewer false positives after intervention and two averaged the same number in both conditions. The mean percentage of errors for each group are in Table 3.

Figure 2 presents the daily data on the occurrence of errors. The largest effect occurred in SCEP I, the setting with the highest initial error rate. The percentages in Figure 2 are often based on different sample sizes because the number of teaching apprentices varied from one to four and because advanced teaching assistants often rechecked less than five quizzes.

The percentage of incorrectly scored quizzes plotted in Figure 3 give a better representation of the overall effect. This graph shows the reductions in the percentage of quizzes with one or more errors. The mean percentage of incorrectly scored quizzes

Table 2. Data summary for individual graders.

Table 2

Percent of Errors by Individual Student Grader

Group		Baseline Mean	Baseline Range	Baseline Days	Intervention Mean	Intervention Range	Intervention Days
SCEP II 10 a.m.							
Subject							
1.	False positive	.4	0-2	11	.2	0-2	11
	False negative	1.8	0-8		.2	0-2	
2.	False positive	.7	0-4	10	.3	0-2	12
	False negative	2.9	0-6		2.0	0-7	
3.	False positive	.7	0-4	6	.2	0-2	12
	False negative	1.7	0-4		1.5	0-10	
4.	False positive	.6	0-4	8	.6	0-2	6
	False negative	1.2	0-2		1.0	0-4	
5.	False positive	.1	0-2	13	0	0	7
	False negative	3.1	0-10		3.1	0-12	
6.	False positive	0	0	3	.4	0-2.5	6
	False negative	2.5	0-4		.4	0-2	
SCEP II 8 a.m.							
7.	False positive	0	0	8	1.0	0-2.0	4
	False negative	2.3	0-6.7		1.0	0-4.0	
8.	False positive	1.0	0-4.0	14	1.1	0-6.0	7
	False negative	1.9	0-6.0		.6	0-2.0	
9.	False positive	.6	0-3.3	10	0	0	3
	False negative	1.3	0-4		.7	0-2.0	
SCEP I							
10.	False positive	.4	0-5	11	0	0	4
	False negative	4.2	0-6		0	0	
11.	False negative	.3	0-2	8	.7	0-2	3
	False positive	3.8	0-10		.7	0-2	

Table 2 (Continued)

Percent of Errors by Individual Student Grader

Group		Baseline Mean	Baseline Range	Baseline Days	Intervention Mean	Intervention Range	Intervention Days
12.	False positive	0	0	5	.7	0-2	3
	False negative	5.9	0-20		.7	0-2	
13.	False positive	0	0	7	0	0	2
	False negative	4.9	0-8		1.0	0-2	
14.	False positive	0	0	8	5	5	1
	False negative	4.3	0-8		7.5	7.5	
15.	False positive	0	0	9	1	0-4	4
	False negative	1.9	0-10		0	0	4

Table 3. Mean percentage of errors before and after intervention.

Table 3

Mean Percent of Errors for Each Condition

Group	Baseline	Intervention
SCEP II 10 a.m.		
False positives	.5%	.2%
False negatives	2.3%	1.4%
Total errors	2.8%	1.6%
SCEP II 8 a.m.		
False positives	.9%	.8%
False negatives	1.7%	.5%
Total errors	2.6%	1.3%
SCEP I		
False positives	.1%	.7%
False negatives	4.1%	.5%
Total errors	4.2%	1.2%

Figure 2. Daily error rates for each shift.

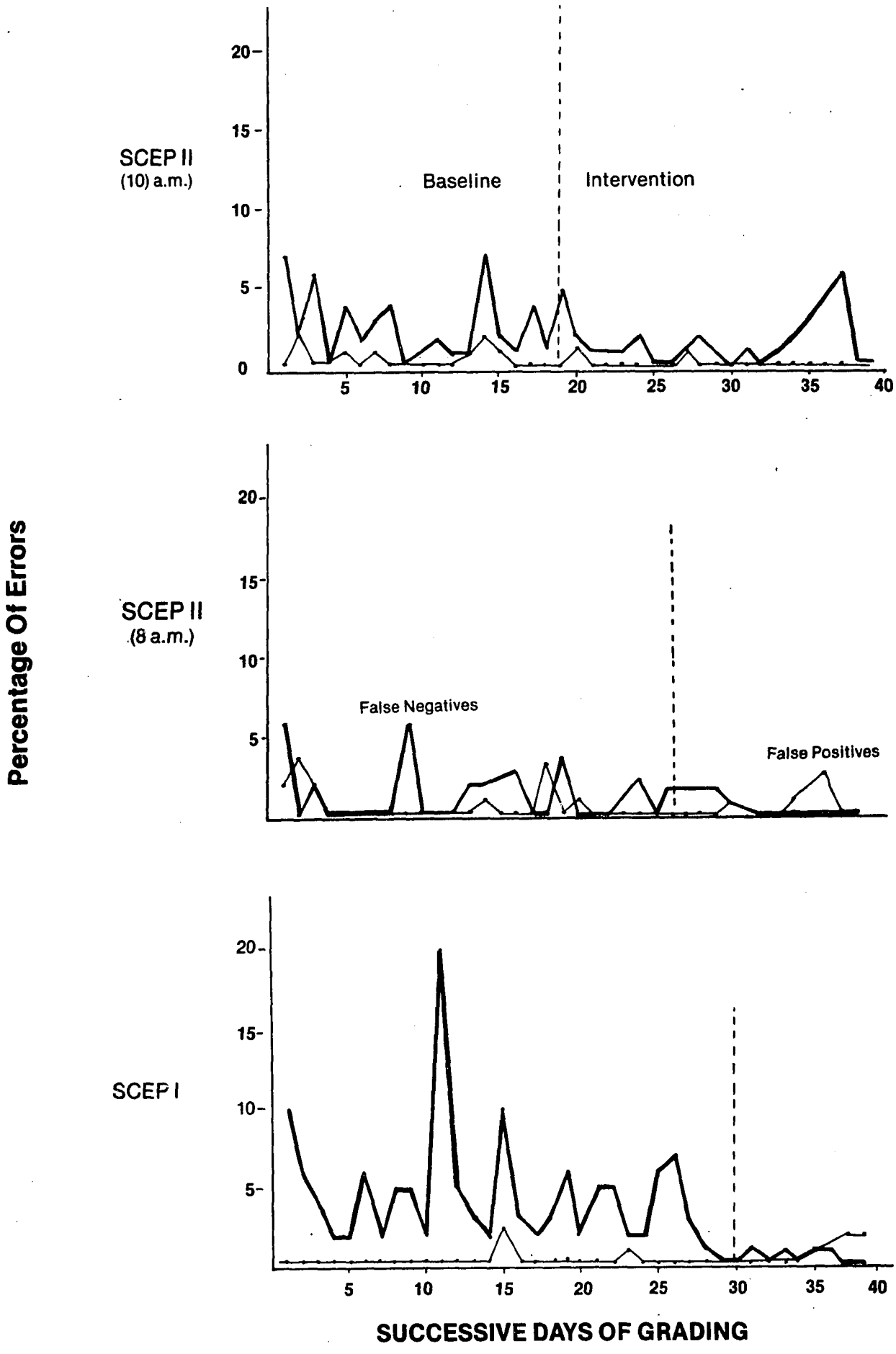
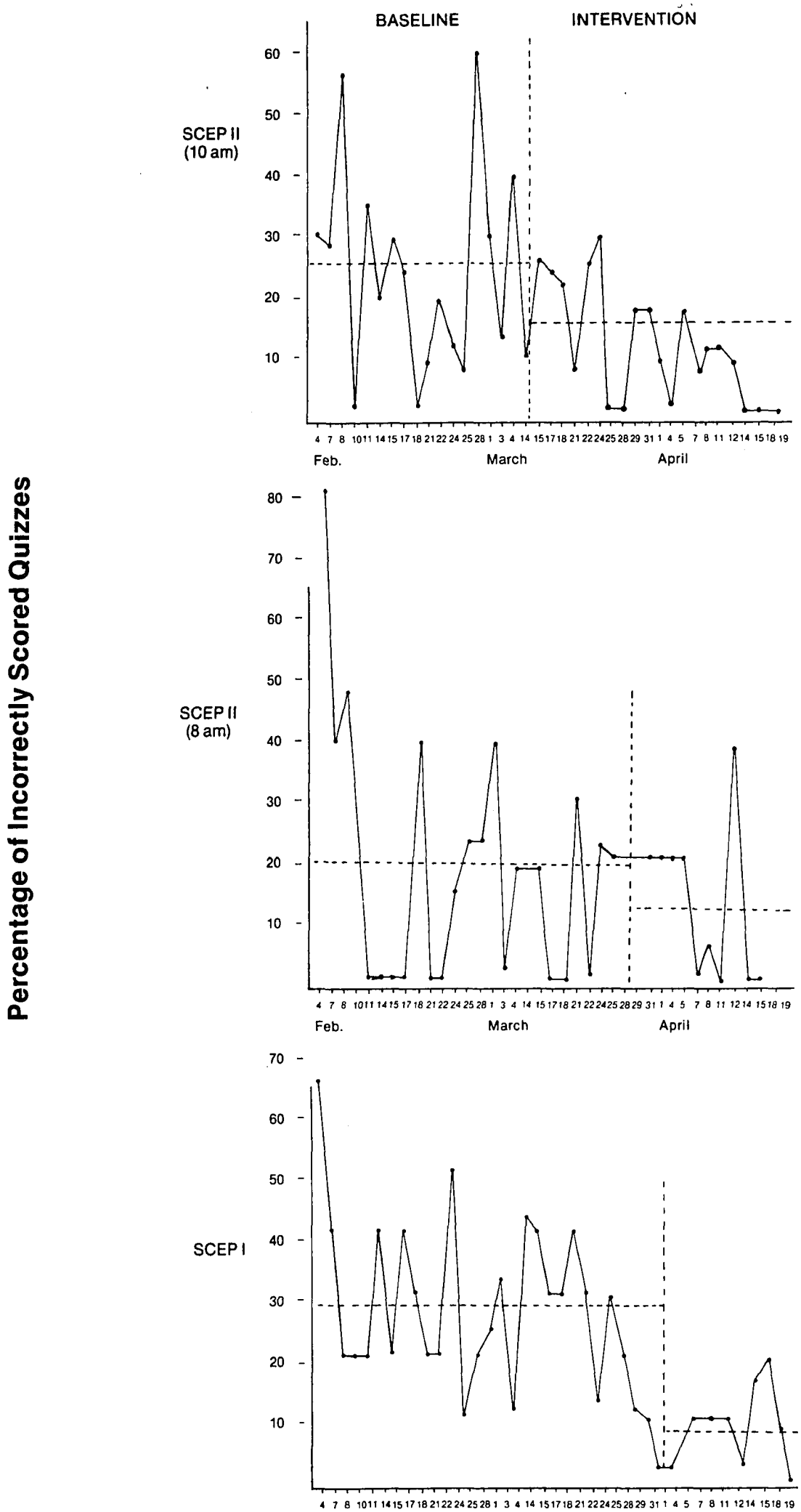


Figure 3. Percentage of quizzes with one or more errors.



dropped from 20 to 13 in SCEP II's 8:00 a.m. section. for SCEP II's 10:00 a.m. section the mean percentage fell from 20 to 13 and SCEP I the mean fell from 27 to 8.

A number of other dependent variables were monitored. Data were collected on the time teaching apprentices spent grading and the number of quizzes graded each day. These data were somewhat inconclusive due to the lack of reliability checks but should have revealed undesirable effects which might have resulted from the point system. The time spent grading quizzes decreased after intervention while the system processed more quizzes (Table 4).

Table 4. Changes in the mean grading time and mean number of quizzes graded before and after intervention.

Table 4

Secondary Data on Grader's Performance

Group	Baseline	Intervention
SCEP II 10 a.m.		
$\bar{X}$ Time grading each quiz	5.5 min.	4.0 min.
$\bar{X}$ # of quizzes per day	28	29
SCEP II 8 a.m.		
$\bar{X}$ time grading each quiz	3.9 min.	3.1 min
$\bar{X}$ # of quizzes per day	17	26
SCEP I		
$\bar{X}$ time grading per quiz	4.2 min.	3.3 min.
$\bar{X}$ # of quizzes per day	25	44

DISCUSSION

The purpose of this study was to experimentally validate a quality control sub-system for monitoring and maintaining accuracy of student graders. The procedures were successful at maintaining grading that was better than 98 percent accurate for all three groups. This finding replicates the grading levels obtained after intervention by Semb (1975) and is well above reliability figures in most PSI research. Further, the relatively small improvement in error rates resulted in a 13 percent increase in the number of tests with accurate scores.

Although the overall reductions in error rates were relatively small, the effect was replicated across all three groups thereby demonstrating experimental control. Data on individual teaching assistants show reductions in the error rates of fourteen of the fifteen subjects and strengthens the claim for experimental control.

The only subject who failed to show improvement only graded once after intervention. Therefore the grading accuracy cannot reflect contact with the point consequence and may simply be random variance resulting from chance factors such as peculiarities with an answer key, tests taken that day, or other variables.

Three problems arise in interpreting the results of this study. First the advanced teaching assistants knew when intervention occurred. This was necessary so that the observer could provide the grader with feedback on grading errors. The reliability

observer, however, did not know when intervention occurred and reliability coefficients did not change after intervention as would occur if the primary observer changed his criterion (Table 1).

The low reliability for identifying incorrectly scored items is also a problem. The low reliability coefficients are probably not surprising since each advanced teaching assistant and reliability monitor should count the item wrong if he has a question about whether it is correct. Advanced teaching assistants and reliability monitors have sufficiently diverse histories that an ambiguous item will be considered right by one reader and wrong or questionable by another. About one-third of the items considered an error by the advanced teaching assistant were also identified by the reliability observer. Hopkins and Herman (1977) reported a procedure for determining reliability coefficients occuring from observers recording instances of a response purely on the basis of chance. Chance reliability coefficients for the occurrence of errors are less than .002 for all three of the SCEP components, thus the reliability coefficients reported for the occurrence of errors in Table 4 are all well above chance. The point should be made that the advanced teaching assistant was grading correctly if he counted the item wrong when he considered it questionable.

The third problem is the downward trend evident in the SCEP I and perhaps SCEP II (8:00 a.m.) graphs in Figure 3. The downward

trend in the data could be a function of several variables. The graders may have improved simply as a function of practice. Another factor may have been continued refinement of answer keys. When a student filed a regrade request with sufficient documentation of a novel answer, the student's answer was added to the answer keys. Also, teaching assistants could earn bonus points for correcting or clarifying the answer keys. The improved accuracy might also relate to the self-paced format of the course. As the semester approached, students took quizzes from within a narrower range of units. Though more students were taking quizzes, the range of units was smaller. Therefore, as contrasted with earlier in the semester, teaching apprentices graded more quizzes over fewer units.

A number of points can be made about the problem of downward trends. First, given the variability present during the baseline conditions, it is unlikely that the variability would have decreased so consistently after intervention. Further, the increase rate of errors occurring on the last few days of the semester is probably a function of the high rates of quiz taking on those days. The reduced variability after intervention and the reduced error rates for fourteen of the fifteen graders suggest that the feedback and point consequences were effective for improving grading accuracy. This conclusion must be accepted with caution, however, because of the downward trends in the data.

Such a system for improving grading accuracy has several

advantages. Data collected by this monitoring system provided reliability information for three other studies occurring in the SCEP program. Without such a system, educational research using grades as a dependent variable runs the risk of reporting results that are simply a function of unreliable grading.

The primary undesirable aspect of the system is the staff time required to monitor the graders. Making a conservative estimate on the basis of time data from Table 4, the monitor should spend about five minutes checking each quiz for grading errors. Monitoring five quizzes from each of three graders would require one hour and fifteen minutes of staff time. Estimating a total of three errors on the basis of a two percent error estimate (from Table 3), the monitor might spend fifteen minutes providing feedback. This yields an estimated total of one and a half hours staff time on a shift requiring three graders. It should be noted that the point system is probably cost effective even if the overall effect is small. The staff time involved in recording the points and providing feedback is relatively small as compared with the time required to monitor grading reliability.

A quality control system such as this one is desirable for programs like SCEP where nonpaid staff can do the monitoring. Other systems that utilize paid staff or those with difficulties recruiting adequate staff will have to carefully weigh the reliability of their present grading system against potential benefits and the cost of additional quality control.

Semb (1975) suggests the possibility of fading out feedback without sacrificing grading accuracy. Further research might pursue this concept and determine the frequency and the number of items that need to be monitored to maintain accuracy. Semb also targeted only the worst graders in his system. A more economical system might monitor grading long enough to determine which graders aren't meeting the desired criteria. Once these graders were identified, a monitor would give them feedback on grading and remove points for grading until each was meeting criteria regularly. Those meeting the criteria during the initial monitoring might be checked at less frequent intervals to insure accuracy. At any rate further research should attempt to develop a system which is more economical.

In addition, further research might try to determine whether or not monitoring alone results in improved grading accuracy. Grading accuracy in this study was much higher during baseline than levels reported in other studies. It would also be important to determine whether increased grading stringency or more accurate grading produces any significant effects in the performance of students. The answer to this question might well be an important determinant in the use of such a system in nonresearch oriented programs.

REFERENCES

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

Born, D. G., Gledhill, S. M., and Davis, M. L. Examination performance in lecture-discussion and personalized instruction courses. Journal of Applied Behavior Analysis, 1972, 5, 33-43.

Coyne, P. D. The effects of informational feedback on the grading accuracy of undergraduate assistants. Unpublished master's thesis, Western Michigan University, Kalamazoo, Michigan, 1974.

Fuqua, W. R. and Heckler, J. B. New directions in educational technology research: Some data and suggestions. Paper presented at the Midwestern Association of Behavior Analysis Convention, Chicago, Illinois, May, 1977.

Hursh, D. E. Personalized systems of instruction: What do the data indicate: Journal of Personalized Instruction, 1976, 1, 91-105.

Semb, G. Proctor selection, training and quality control in personalized instruction in Behavior Research and Technology in Higher Education. Edited by J. M. Johnston. Springfield, Illinois: Charles C. Thomas, 1975, 139-150.

Sheppard, W. C. and MacDermot, H. G. Design and evaluation of a programmed course in introductory psychology. Journal of Applied Behavior Analysis, 1970, 3, 5-11.

Traveggia, T. C. Personalized instruction: A summary of comparative research, 1967-1974. American Journal of Physics, 1976, 44, 1028-1033.

Williams, R. L. Personalized systems of instruction: Future research areas. Journal of Personalized Instruction, 1976, 1, 106-112.