*THE EFFECTIVENESS OF CURRENTLY ENROLLED STUDENT
PROCTORS IN AN UNDERGRADUATE SPECIAL EDUCATION
COURSE¹*

JOHN F. GAYNOR AND WILLIAM D. WOLKING

UNIVERSITY OF DELAWARE AND UNIVERSITY OF FLORIDA

Potential users of Keller's personalized college classroom method might find it impractical to provide student proctors from the ranks of those who had previously mastered the course material. An alternative is to let currently enrolled students proctor and tutor each other. Such a method imposes the burden of showing that its effectiveness is not a function of (1) student collaboration or (2) the effects of "practice" received by students who listen to the performance of others before their own performance. The present study compared students proctored by previously trained students (Control) with students who received proctoring from classmates (Experimental). Experimental group performance was superior to Control group performance. The effects of "practice" and collaboration were ruled out as determinants of the difference between groups. It is hypothesized that the superior performance of the Experimental students was related to their activities as proctors.

While the adaptation of behavior principles to college instruction has taken a variety of forms in recent years (*e.g.*, Ferster, 1968; Lloyd and Knutzen, 1969), a particular genre has become associated with Keller's "personalized, proctorial, programmed" model (Keller, 1968, *unpublished*; Malott and Svinicki, 1969; McMichael and Corey, 1969; Sheppard and MacDermot, 1970). A modification of Keller's basic procedures has been tested by investigators at the University of Florida and described at length by Johnston and Pennypacker (1971). In the Florida model, several of the basic features of Keller's system are preserved: successive mastery of small units of the course material, the use of the constructed response in performance samples, immediate corrective or confirmative feedback on performance, statement of the terminal objective in behavioral terms, and individual proctoring by a previously trained student.

The use of student proctors who have previously mastered the course material has been treated as a cornerstone of both the Keller and Florida methods (Johnston and Pennypacker, 1971; Keller, 1968; Malott and Svinicki, 1969; McMichael and Corey, 1969), yet there is little evidence relating student achievement to the proctor's prior training. Although Farmer, Lachter, Blaustein, and Cole (1972) showed that proctored students performed better than those who received no proctoring, the question remains whether demonstrated mastery of the course material is a vital prerequisite to successful proctoring. This is a matter of some practical importance. Instructors of low enrollment or infrequently offered courses would have difficulty finding previously trained proctors. An alternate model, in which all the personnel needs of the course are met from current enrollment, might be more appropriate for some users of personalized instruction methods.

The present study compared the performance of two groups under different proctoring arrangements. The Control group was proctored by previously trained students; the Experimental group provided its own proctoring services, each student alternating as a listener and a speaker.

¹The authors wish to thank Mrs. Doris Davidson for her assistance in preparing the manuscript and Mr. Barker Bausell for his critical reading of the original manuscript. Reprints may be obtained from John F. Gaynor, Willard Hall Education Building, University of Delaware, Newark, Delaware 19711.

The introduction of the reciprocal proctoring arrangement raises secondary questions, which are also evaluated in the study: (1) will a "practice effect" occur on a similar performance test for students who monitor the performance of others? (2) will the reported performances of peer-proctored students be valid as measured by objectively administered summary tests, or will they be inflated by mutual interest or collaboration?

It is important to note that the focus of the present study is an *alternative* to the proctoring method of personalized instruction and not the method itself. Two critical components of the personalized approach, unit mastery and self-pacing, have been omitted because they would obscure the experimental effect. As a result, the unique benefits of previously trained proctors cannot be demonstrated, for it is only in the repeated tests leading to unit mastery that the proctor's tutorial expertise can be reflected. However, there is still reason to believe that the findings of the present study will generalize to other forms of personalized instruction. The argument can be made that the relationship between proctor expertise and student achievement is tenuous at best and is probably interactive with such factors as difficulty of course content, adequacy of study materials, and whether units are sequential and cumulative or relatively discrete. Given the complexity of the relationship, the margin of expertise of previously trained over currently enrolled proctors may not be sufficient to influence student achievement significantly.

The point is strengthened by reference to data provided by Johnston and Pennypacker (1971). They showed in three separate studies involving seven treatment groups that mean error percentages over multiple trials (where the expertise of proctors would be evident) were reduced from first-trial levels by negligible margins (1.2%, 5.2%, 0.1%, 1.2%, 1.5%, 0.3%, 0.7%). Assuming that some part of the error reductions were attributable to the student's restudying the material and some part

to the kind of corrective feedback that can be provided mechanically (*e.g.*, printed on an answer sheet), there would appear to be little remaining variance upon which the difference between the previously trained proctor and currently enrolled proctor could operate. The present study was not designed to shed much light on this issue. It has been introduced merely in support of the contention that the absence of unit mastery in the experimental design will probably not obviate the generality of the findings to other personalized instruction models.

METHOD

Thirty undergraduate students in an introductory special education course were randomly assigned to a Control group ($n = 16$) and an Experimental group ($n = 14$). The Experimental group was randomly subdivided into Team A ($n = 7$) and Team B ($n = 7$) for purposes of evaluating "practice effect". Materials consisted of 10 units of constructed response (fill-in-the-blank) items developed from the textbook *Educating Exceptional Children* by S. A. Kirk (1962) and six units based on class lectures. Each unit was randomly divided into alternate forms and typed on flash cards in the format described by Johnston and Pennypacker (1971). Alternate forms were color-coded to aid in controlling practice effect and number-coded for checking the accuracy of computations. Proctors were five undergraduate students who had assisted in developing course materials during a pilot project. The proctors had demonstrated mastery of the textbook materials but were unfamiliar with material covered in the lecture units. This arrangement was an unavoidable condition of field experimentation, but it provided a means of evaluating the effects of proctoring under different levels of proctor expertise.

Procedures

Conditions for both the Experimental and Control groups were identical except for the

proctoring variable. The Control group met with proctors in private offices with 30 min scheduled for each interview. The time required for a session seldom exceeded this but was freely given if needed. Proctors were instructed in procedures for reducing anxiety and generally managed to create a friendly, relaxed atmosphere. The Experimental group met *en masse* in the classroom. Seven pairs, each consisting of a speaker and a listener, were established by students arranging their chairs to face each other. Interviews were limited to 20 min to allow time for each student to act as both speaker and listener. With seven students speaking simultaneously during testing, noise levels were high compared to Control group values.

Performance sampling procedures followed the method described by Johnston and Pennypacker (1971) with the exceptions that unit tests were taken on a pre-arranged schedule and could not be repeated to achieve mastery criterion. During the interview, the proctor monitored the oral responses of the speaker, started and stopped him on time, counted correct and incorrect responses, held questionable items aside for discussion, and recorded response rates. Each item was read aloud by the speaker from a 3 by 5 in. (7.5 by 12.5 cm) index card with the answer written on the reverse side. Following the 5-min timed trial, the proctor reviewed missed items, asked for clarification of vague answers, and helped in any way that he could to provide corrective or confirmative feedback. Members of the Experimental group, serving as both proctors and speakers, met their dual assignment as follows: (1) Members of Team A proctored members of Team B and *vice versa*, but all members had different partners for speaking and listening; *e.g.*, if student #1 listened to student #2, a third student would listen to him when it was his turn to perform. This broadened the exchange of information across individuals and served as a foil to unwanted collaboration. (2) Both teams received an equal number of "listen first, perform second" assignments. The advantage of hearing the perform-

ance of another before performing oneself was held constant across both groups for both book and lecture units. (3) All members of the "listen first" team performed on an alternate set of test items; *e.g.*, if they listened to green cards, they performed on blue. Although there were similarities between alternate tests, the items were not the same. The purpose of this arrangement was to minimize the effects of "practice" on specific items.

Review Tests

In addition to the 16 basic unit tests, four review tests were administered. In each half of the course, students received a repeat test and a review examination. Repeat tests consisted of a second trial on the alternate form of the unit on which the student had previously obtained the lowest score. Review exams were administered as midterm and final examinations. Each was composed of 124 items randomly selected from the unit pools covered in the respective halves of the course. The review exams were administered and scored in the conventional manner as a means of validating performances reported by student proctors.

Dependent Measure

Following Johnston and Pennypacker (1971), student performance was measured in rates of correct and incorrect responding. For purposes of statistical analysis, a single performance measure has been derived by subtracting the rate of incorrect responding from the rate of correct responding.

RESULTS

The chief burden of the experiment was to answer the question of whether students in a self-contained classroom, monitoring each other's performance, would demonstrate levels of performance equal to or better than that of students who received monitoring services from previously trained proctors. The design yielded two kinds of information for evaluating this

question: (1) between-groups comparison on the 16 first-trial performances, and (2) between-groups comparison on four review tests.

Between-Groups Comparison on the 16 First-Trial Performances

Figure 1 shows that Experimental group performance exceeded Control group performance at all but three of the 16 data points ($p = 0.004$ Sign Test). Selection of a parametric test to analyze these results was governed by the need to test not only the differences at each point in the series of scores, but also the cumulative effect of these differences. A profile analysis procedure described by Morrison (1967) met these requirements. The first step was to test the two mean vectors for parallelism. In the absence of parallelism, a group interaction must be assumed, in which case the profile analysis would be inappropriate. The nonsignificant T^2 of 2.19 showed this not to be the case. A second test was conducted to determine whether the mean vectors differed. A significant difference was

found between the Experimental and Control groups ($T^2 = 3.69$, $df\ 16/13$, $p = 0.011$), indicating that the Experimental group was reliably superior to the Control group.

Between-Groups Comparison on Four Review Tests

A two-factor, repeated measures analysis of variance (Winer, 1962) was done. A tendency in favor of Experimental group superiority ($F = 4.06$, $df\ 1/28$, $p = 0.054$) and a significant interaction between the experimental treatment and the repeat tests ($F = 3.68$, $df\ 3/84$, $p = 0.015$) was found.

A highly significant effect of testing mode ($F = 16.45$, $df\ 3/84$, $p < 0.001$) was found, but may be assumed to be an artifact of the rate of responding in the written *versus* oral mode, and was not unexpected.

The Effects of Practice

If the effect of practice were the principal determinant of the performance differences re-

Fig. 1. Mean performance rate vectors of experimental and control groups on 16 first-trial performance samples.

flected in the 16 oral response sessions, it should be possible to demonstrate this by showing (1) that the seven students of the Practice group produced performance rates that were consistently and reliably superior to the rates produced by the seven students in the No-Practice group, and (2) that there was no reliable difference between the No-Practice group and the 16 students of the Control group, who also performed under conditions of no practice.

The mean vectors of the Practice and No-Practice groups were plotted and revealed no trend toward the superiority of either group. The Practice group was superior in nine cases; the No-Practice group was superior in seven cases. For comparison of the No-Practice ($n = 7$) and Control ($n = 16$) groups, profile analysis was suitable. The mean vectors were found to be parallel ($T^2 = 1.58$, $df = 15/7$, $n.s.$). The test of the difference between mean vectors showed that the seven Experimental group students who performed without practice were superior to the Control group at a high level of reliability ($T^2 = 10.49$, $df = 16/6$, $p = 0.004$).

The evidence from both comparisons (Practice *versus* No-Practice, No-Practice *versus* Control) supports the conclusion that the reliably higher achievement of the Experimental group cannot be attributed to the effects of practice. Practice effect in this case is defined as the benefit a student derives from proctoring the performance of another student before taking his own performance test on the same unit. The experiment was designed to reduce the effects of practice by having those students who listened first and performed second take an alternate form of the test. The evidence leads to the conclusion that this strategy was successful, and that an explanation for the superior achievement of the Experimental group must be sought elsewhere.

Validation of Student-Proctored Performance Sessions

Another factor that might reasonably be expected to account for the between-groups performance difference is the possibility that stu-

dents controlling each other's grades might overstate performance. Two written examinations were administered under standard conditions for purposes of testing this question. The assumption was made that written exam performance would provide a valid certification of prior performance records if the percentage of correct written items was approximately equal to the mean percentage correct of the preceding oral response units. Collaboration was possible in the latter case, but not possible in the former. If members of the Experimental group had reported difference rates that were based on something other than true performance, it should be seen in comparisons of the differences between student-administered and instructor-administered tests. For this purpose, it was necessary to make individual rather than group comparisons, since group validation might obscure the fact that one or two students enjoyed unintended advantages that the group as a whole did not.

Table I summarizes the differences in percentage of correct response between the eight oral response performance units on which the review tests were based and the review tests proper. The question of what criterion to accept for test validation can be safely ignored, because the differences in test performance are very slight, more often positive than negative, and of greater magnitude in the positive direction than the

Table I

Comparison of experimental and control groups on mean percentage of correct responses on eight first-trial tests with percentage correct on summary review tests.

	Number of Students Showing:		Mean Percentage of:	
	Loss	Gain	Loss	Gain
Midterm Review				
Experimental	1	13	-4.8	+8.8
Control	2	14	-8.2	+9.0
Final Review				
Experimental	6	8	-2.2	+6.1
Control	6	10	-6.8	+8.1

negative. Only one set of scores out of 60 showed a percentage loss in excess of 10%. This student was in the Control group, where the question of collaboration is not at issue. None of the Experimental group losses exceeded 5%. The data support the conclusion that the performance differences shown in previous analyses cannot be attributed to collaboration on the part of Experimental group students.

Effects of Proctor Expertise

With Control group proctors operating under two levels of expertise—mastery for book units, none for lecture units—it was hypothesized that the tutorial assistance given after first trials on unit tests may be reflected in differential performance on book *versus* lecture items appearing in the summary review tests. Chi square contingency tables showed a difference in the proportion of lecture items missed by the Control group as compared to the Experimental group on the midterm exam ($\chi^2 = 4.63$, $df = 1$, $p = 0.031$) but no difference on the final, indicating that proctor expertise was a factor of importance in the three lecture units covered in the first half of the course.

DISCUSSION

The clearest evidence in the study is related to performance on 16 first-trial unit tests. Students proctored by their classmates performed better than students of previously trained proctors. In attempting to account for the difference, two of the least salutary possible causes were ruled out: (1) inflation of performance reports by students in mutual collaboration, and (2) the effects of practice that students in the Experimental group might have received when they proctored classmates before taking tests of their own. The evidence against practice effect showed that with practice held constant, Experimental students still showed a highly reliable superiority. It was further shown that the method of proctoring made a difference before the tutorial component of proctoring had a chance to interact with the de-

pendent measure. For this reason, it can be stated that factors other than the proctor's demonstrated mastery of the subject matter account for the difference. This does not mean that proctor expertise is unimportant. The limited evidence on this point suggests that proctor expertise was a factor in at least one of the two review exams.

It is important to note that proctoring is a multivariate treatment, only one aspect of which is the proctor himself. Conditions between groups also differed with respect to the experimental space, total time on task, the presence of the instructor, and the presence of other students engaged in similar activity. Several factors, singly or in combination, might account for observed differences. Since the purpose of the present experiment was to demonstrate the effectiveness of the method as a whole, nothing in the data confirms the relative weights of components. However, one subtle difference in contingency arrangements between groups merits further consideration. On days when tests were taken by the Experimental group, students were observed to discuss unit material throughout the entire class period. They were often engaged in questioning each other's knowledge as they entered the classroom, continuing in that vein until formal testing started. This was not true of Control students. Meeting their proctors individually, they would seldom encounter others who were interested in discussing the material before the test. The proctors often asked them a preliminary question or two, but the purpose of this was to help the student relax, not to extend his level of preparation. It can be assumed that the discussions of the Experimental group were maintained in strength because everyone had an interest in maintaining them. For the Control group, good speaking led to reinforcement; for the Experimental group, there was an added incentive for good listening.

It has been understood from the beginning (Keller, 1968) that proctoring results in performance gains for the proctor as well as the proctee. However, the contingencies are not arranged to produce this result for the student who

has already passed the course at the mastery level. It is a by-product for the previously trained student, incidental to his main purpose as far as contingencies are concerned. By contrast, a currently enrolled student can apply any such performance increment to satisfying a grade criterion. The analysis of contingencies suggests that proctoring by currently enrolled students may possess unique didactic as well as practical advantages. With collaboration and practice adequately controlled, the method provides an effective alternative to proctoring by previously trained students.

REFERENCES

- Farmer, J., Lachter, G., Blaustein, J., and Cole, B. The role of proctoring in personalized instruction. *Journal of Applied Behavior Analysis*, 1972, 4, 401-404.
- Ferster, C. Individualized instruction in a large introductory psychology course. *The Psychological Record*, 1968, 18, 521-532.
- Johnston, J. and Pennypacker, H. A behavioral approach to college teaching. *American Psychologist*, 1971, 26, 219-244.
- Keller, F. "Good-bye, teacher . . ." *Journal of Applied Behavior Analysis*, 1968, 1, 78-89.
- Keller, F. *A programmed system of instruction*. Unpublished address given at Pacific Northwest Association for College Physics, University of Seattle, Washington, 1968.
- Kirk, S. *Educating exceptional children*. Boston: Houghton Mifflin Company, 1962.
- Lloyd, K. and Knutzen, N. A self-paced programmed undergraduate course in the experimental analysis of behavior. *Journal of Applied Behavior Analysis*, 1969, 2, 125-133.
- Malott, R. and Svinicki, J. Contingency management in an introductory psychology course for 1000 students. *The Psychological Record*, 1969, 19, 545-556.
- McMichael, J. and Corey, J. Contingency management in an introductory psychology course produces better learning. *Journal of Applied Behavior Analysis*, 1969, 2, 79-83.
- Morrison, D. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.
- 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.
- Winer, B. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Received 4 May 1972.

(Revision requested 2 August 1972.)

(Revision requested 12 December 1973.)

(Final acceptance 23 January 1974.)