THE ANALYSIS OF PERFORMANCE CRITERIA DEFINING
COURSE GRADES AS A DETERMINANT OF COLLEGE
STUDENT ACADEMIC PERFORMANCE¹

JAMES M. JOHNSTON² AND GEORGE O'NEILL

GEORGIA STATE UNIVERSITY

A series of five experimental conditions were designed to investigate the influence of minimum performance criteria and grade labels on college student academic performance. A college course in abnormal psychology was taught in an individualized manner so that each student could perform on each unit of subject matter in individual performance sessions whenever he wished. In each of the five experiments the minimum performance criteria that had to be attained before progressing to the next unit were varied during the quarter and the resulting changes in performance were recorded. In Experiment I there were no criteria; in Experiments II, III, and IV three levels of criteria (High, Medium, and Low) were varied but all of the criteria defined a course grade of "A". In Experiment V, the three criteria defined course grades of A, B, and C. The results showed that the criteria controlled performance to a high degree, so that regardless of what quality of performance had been demonstrated previously or was being produced currently, performance was immediately changed to attain new criteria put into effect. Students in Experiment I produced very poor performance compared to the other conditions.

The recent application of principles of operant behavior to the development of college-level teaching procedures based on the earlier work of B. F. Skinner has been initiated by a number of researchers (Postlethwait, 1964; Mahan, *unpublished*; Ferster and Perrott, 1968; Geller, 1968; Pennypacker, *unpublished*; Johnston and Pennypacker, 1971); the influence of Keller's work has been particularly important. Of the many efforts currently in operation (see Sherman, 1971), almost all are primarily utilitarian applications of a number of basic principles (Keller, 1968; Johnston and Pennypacker, 1971) to different subject matters, and the purpose is frequently to delineate the major characteristic effects of what usually are highly similar procedures from one institution to another.

Few educators have attempted to analyze in

detail college student academic performance in order to define precisely the relationships between performance and the variables of which it is a function (Malott and Janczarek, *unpublished*; Malott and Rollofson, *unpublished*; Johnston, *unpublished*). If teaching is defined as the arrangement of events in the academic environment so as to produce desired changes in student performance (Skinner, 1968), it would seem that ultimately it is from the empirical definition of those events and their relationships to the academic performance that instructional methods at the college level must evolve.

The present study originated from a program of research, conducted at the University of Florida and at Georgia State University, that has consistently suggested that a student's academic performance was strongly influenced by the teacher's definition of criteria for the various course grades. In particular, previous data suggested that the definition, or lack of definition, of specific criteria for academic performance and the association of grade labels with these criteria controlled student performance to a very large degree.

¹The authors wish to express their appreciation to the Student-teachers who contributed their energies to this project.

²Reprints may be obtained from James M. Johnston, Department of Psychology, Georgia State University, Atlanta, Georgia 30303.

Other empirical approaches to individualized instruction at the college level have varied widely in their definitions of minimum performance criteria for course grades. Keller (1968) referred to a unit perfection or 100% mastery requirement for frequent short quizzes. However, no such requirement existed for the final exam, and course grades were determined by 25% final exam scores and 75% quiz performances. Thus, the relationship of the unit quizzes to specific course grades was confounded by the exam, which did not occur until the end of the quarter. McMichael and Corey (1969) also required a perfect score on 10-item fill-in unit tests; however, again there was a final exam with no mastery requirement which contributed to the course grades. The exact grade definitions in terms of unit quizzes and the final exam were not reported. Sheppard and MacDermot (1970) used a criteria requiring a joint agreement between the student and the listener in brief interviews in combination with a "functionally correct" performance on 4- or 5-item short essay written tests to define letter grades. Grades of "A" through "D" were defined in terms of different quantities of interviews and tests taken. Meyer's (1970) criteria required the students to pass exams of different length and type with fewer than three errors, and a grade of "A" was earned by passing all 24 exams; a grade of "I" was given for any other performance. Born and Herbert (1971) set up a point system with all grades defined by some quantity of points. Seventy points were available from 12 unit tests and 30 points could be earned on a final exam. Born, Gledhill, and Davis (1972) required perfect scores on unit tests (of fill-in and short essay items) before taking midterm and final exams. However, the grading criteria were based on the curved class exam distributions only and were thus not announced during the quarter. Stalling (1971) also defined minimum performance criteria for course grades from frequent quizzes and major exams. Students could earn half of their grades from taking 1, 2, or 3 forms of each of 15 quizzes; only the best score

counted. The other half of their grade came from three large exams given during the course.

The following procedures were designed to begin to analyze the influence of minimum criteria for course grades on academic performance.

METHOD

Subjects

The experiments were conducted with 65 undergraduate students routinely enrolled in a junior-level course in abnormal psychology at Georgia State University during spring quarter, 1971.

Procedure

The students participated in two kinds of activities. First, groups of 12 to 15 students met in weekly seminars with the instructor. These seminars involved various activities ranging from discussions to films. Attendance was not required, and the seminars played no part in the students' course grades and were not considered to be part of the study.

The second activity in which the students participated was a Performance Session. The procedures are described in detail in Johnston and Pennypacker (1971). When a student arrived at the Teaching Center at the time previously signed for on the Student-teacher's posted appointment sheet, the Student-teacher was usually ready with the student's data sheet, his cumulative graph, a pencil, and a watch. When the student announced which text or lecture unit he wished to work on, the Student-teacher shuffled the pool of items in that unit and randomly selected some portion of them appropriate to the period of time they had agreed to work. Each time a student came in to work on a certain unit, the items worked with were randomly selected from that same unit pool. Thus, on the second and third Performance Sessions of a unit, a student could see new items as well as those he might have performed on previously.

The student and his manager then sat down in a three-sided booth, and after the usual

friendly conversation, the student was handed the stack of item cards and the Student-teacher quietly noted the starting time on his watch. The student read each item aloud and filled in the blank orally with the correct answer or indicated that he did not know. After emitting either of these two specific alternatives, he turned the card over and read aloud the correct answer. He performed the same operation with all of the remaining items or until some previously agreed upon period of time (about 10 min) had expired. After the student was finished, the Student-teacher quickly counted the few missed blanks and gave the cards to the student to study. Then he counted the number of blanks correctly answered (there was sometimes more than one blank per item) and calculated the rate of reading and answering items correctly and incorrectly (the number of items was divided by the elapsed time).

This done, the Student-teacher and the student spent a variable period of time engaged in dialogue concerning the subject matter. This portion of the Performance Session was characterized by student defense of incorrect answers, discussion of errors, discussion of book and lecture material, review and suggestion of study techniques, and appropriate praise or curses. There was also discussion of personal and non-course matters.

Following this period of discussion, the Student-teacher plotted the rates of correct and incorrect responding with the student watching. His performance was compared to his past efforts and to clearly defined future goals as well as occasionally to the performance of other students. All performance graphs were displayed publicly on bulletin boards and students were free to come by and see their graph at almost any time.

The students could sign for and have Performance Sessions whenever they wished during a 50-hr week (although no more than once a day). In addition, they could perform on any unit as many times as they needed without penalty (except in Experiment I) to meet the stated rate

correct and rate incorrect criteria, although items were always randomly selected from the item pool on each attempt.

These procedures were designed to serve as a vehicle for research and did not necessarily represent in all respects procedures that would be most appropriate for strictly instructional purposes.

The above Performance Session procedures were identical for all nine units in all experimental conditions. The independent variable was the minimum performance criteria or mastery level that defined course grades and the course grades that were defined. Criteria were precisely stated in terms of the minimum number of items read and answered incorrectly per minute that could occur in the same session, and they were depicted along with actual performance on each student's graph. The reasons for the measurement of performance in terms of rate are described in Skinner, 1950; Johnston, *unpublished*; and Johnston and Pennypacker, 1971.

Experiment I. There were no teacher-defined rate criteria in this condition. Students were frequently told to do their best, that they would be graded "on the curve" at the end of the quarter with other students in their section. They were required to have exactly two Performance Sessions on each unit. This last requirement was designed to control for the amount of contact with the material that students in other experimental conditions would likely have. In the middle of the term, each student was told his rank order position among the classmates in his section. Giving this information was intended to be analogous to posting a frequency distribution of mid-term exam scores, which is often done in other courses.

Finally, the students in this experiment were not told that they had graphs or shown them until approximately one third of the way through the quarter, and the graphs were again removed during the last third of the quarter. When the graphs were absent the students were simply told their performance orally. When the graphs were present, they were not publicly

posted, but were shown only to their owners. This ABA manipulation was to permit an assessment of the possibility that the students might use the graphic depiction of their performance as a kind of criterion. The purpose of Experiment I was to compare the effects of having no teacher-defined criteria with the criteria used in other conditions.

Experiments II, III, and IV. The criteria in these three conditions were sequential variations of three levels of the rate criteria. The High criteria were 3.8 items per minute correct and 0.4 items per minute incorrect, the Medium criteria were 3.1 correct and 1.1 incorrect, and the Low criteria were 2.5 correct and 1.7 incorrect. Note that the minimum total rate of reading and answering items was 4.2 per minute in each case. If the reading and answering of items is calculated independently of how much time is taken, the approximate percentages are: High, 90% correct/10% incorrect; Medium, 75% correct/25% incorrect; and Low, 60% correct/40% incorrect.

In all three experiments, the course grade defined by the High, Medium, and Low criteria was a grade of "A". In other words, if each student attained criteria on each unit, whatever the criteria were for him at that time, he received an "A" in the course. It should be noted that this was not a Pass-Fail system. Although only an "A" was defined in terms of criteria, a complete range of grades less than "A" would be given if earned. However, no criteria were announced for other grades, and the student had no way to figure out how other letter grades might be determined.

In Experiment II, the criteria changed for each student in a High, Medium, Low, High sequence during the course of the quarter. In Experiment III, the sequence was Low, Medium, High, Low for each student. In Experiment IV, the sequence was Medium, High, Low, Medium during the quarter. It was felt that of all possible arrangements of four phases of the three sets of criteria, these three would provide the best possible assessment of any effects on academic

performance due to the sequence of the different criteria.

Experiment V. This condition was identical to Experiment II with the following exceptions. There were only three phases: High, Medium, and Low. In addition, when under the High criteria, the students were told that meeting those criteria was "A"-level work. Approximately a third of the way through the quarter the Medium criteria were added to each student's graph, and he was told that meeting these criteria was "B" level work. Finally, when the Low criteria were added to the others already on the graph, each student was told that this was "C" level performance. Students were permitted to go on to the next unit after meeting any set of criteria currently stated on their graph. Students in this condition were not told exactly how their final grade would be determined except that it would take into account which criteria were attained on each unit. The purpose of this experiment was to assess the effects of the grade label itself as a form of criteria.

In all experiments, assignment of students to any condition was made by first having the students write their names in a column on a sheet of paper passed around the classroom on the first day of class. Then, using the sequence of names that this procedure generated, each name was placed in an experimental condition one at a time by repeatedly rotating through Experiments I through V until all students were assigned. Students joining the course at later dates were assigned similarly by adding one name to each experiment in rotating order through the five studies.

In Experiments II to V, the changes from one set of criteria to the next were made individually at different points for each student, depending on such factors as stability of his data under the present criteria, which unit he was on, the quality of his performance, and the remaining time in the quarter. The students were not told in advance what future criteria were or when a change might be made, but only that criteria would vary during the quarter. Thus, in

Fig. 1. Academic performance of a representative student for each of Experiments I to V. Triangles represent correct responses and circles represent incorrect responses. Data points are connected within units of subject matter. The top horizontal lines are the correct criteria, and the bottom horizontal lines are the incorrect criteria.

the most detailed sense this was an intrasubject design with 11 to 15 replications in each of five separate but related experiments.

RESULTS

The results of Experiments I to V are depicted in Figures 1 and 2. Each graph in Figure 1 shows the data of a student whose performance was highly representative of the effects of each treatment seen in the performance of all students under that condition. It is important to note here that these teaching methods consistently reduced variability among students, and in this study the treatment effects in each condition were extremely uniform from one student to another. All the students in every experiment showed similar patterns of responding under the different criteria, even though the absolute level of performance varied due to individual academic history, study skills, *etc.* The data points are connected only within units so that it is easy to see performance by units.

Fig. 2. Mean of final performances on units completed under each set of criteria for Experiments I to V. Diamond's represent correct responses and circles represent incorrect responses. H = high, M = medium, and L = low criteria.

The graph for Experiment I shows the typically poor performance of students under those conditions. Although there was slightly more variability in the absolute level of performance from one student to the next in this condition compared to the other treatments, even the best

student never maintained the High-level criteria. The first (and sometimes second) attempts on a unit were frequently inversions (rate correct lower than the rate incorrect), and the mean performance overall and final of the entire group just barely met the Low criteria used in the other experiments.

Table 1 shows the mean number of performance sessions taken to attain criteria under each set of criteria for all students in each experiment. The extreme right column shows the mean attempts across all units for each experiment. It may be noted that the number of attempts shows a fairly consistent positive relationship with criteria level, as might be expected.

Table 1

Mean attempts per unit under criteria manipulations.

Experiment	Criteria and Mean Attempts Per Unit				Mean Attempts All Units
I(N-15)	no criteria-2.0				2.0
II(N-12)	H-2.3	M-1.8	L-1.4	H-2.0	1.9
III(N-13)	L-1.6	M-2.5	H-2.7	L-1.6	2.1
IV(N-14)	M-1.9	H-3.2	L-1.6	M-1.8	2.1
V(N-11)	H(A)-2.5	M(B)-2.5	L(C)-2.1		2.4

The graphs for Experiments II, III, and IV show a strong consistent functional relation between criteria and performance. In addition, there is no discernible sequence effect. Each set of criteria had the same degree of control over performance regardless of what level of criteria preceded them. However, all students in Experiments II to V tended to exceed both the correct and incorrect criteria by an unnecessarily large degree in the first phase, whatever the criteria were. This "wasted" performance was quickly cut to a minimum so that in the final phase, when the criteria reversed to those used at the beginning of the quarter, the same criteria were now attained by a much narrower margin.

In the graph for Experiment V, the same patterns of responding observed for Experiment II may be seen; however, they are now confounded by the different grade labels used in Experiment V for the same criteria. The result

is that performance in Experiment V follows the successively lower sets of criteria somewhat less closely than when the same criteria are defined as a course grade of "A" in Experiment II.

Figure 2 shows the mean correct and incorrect performance of all students under each set of criteria for each experiment. The same patterns of responding observed in the representative individual graphs are thus summarized here. It is easy here to observe the "overshoot" tendency on units early in the quarter for Experiments II to V, whatever the criteria at that time. It is important to note the consistency of the changes in performance in response to changed criteria in all conditions, as well as to remember that the data points represent considerable variations in the quality of academic work.

DISCUSSION

The performance data from Experiments II, III, and IV make clear the conclusion that variations in minimum performance criteria defining a course grade of "A" were functionally related to marked variations in the quality and quantity of student academic performance. The students' final performance on a unit tended to follow the current minimum criteria to a consistent and close degree. In other words, regardless of what quality of performance had been demonstrated previously or was being produced currently, performance was immediately and sharply changed to attain new minimum criteria put into effect, whatever the direction of change. Under the conditions of this study, performances early in the quarter also were related to initial criteria but exceeded the criteria by an unnecessarily large margin. This margin quickly decreased, so that by the end of the quarter the reversal to the original criteria produced much tighter control over performance, which now exceeded minimum criteria very narrowly.

The influence of the actual grade label itself is seen by comparing the results of Experiments II and V. The grade label of "A" applied to successively lower criteria produced a much

higher degree of control over performance than the grade labels of "B" and "C" applied to those same successively poorer quality criteria. This sharper control may be seen from the fact that in the Medium and Low phases, the performance was much closer to the minimum criteria in Experiment II than in Experiment V, where not all students dropped their performance to the lower criteria and where most did not drop as low as in Experiment II. In addition, the fact that in Experiment V the High—"A" criteria remained in effect throughout the entire quarter, and the fact that the Medium—"B" and Low—"C" criteria did lower the quality of the performance previously demonstrated under the High—"A" criteria, make it clear that the "B" and "C" grade labels did control performance to a considerable degree, although less well than a grade label of "A".

The effects of the absence of any teacher-defined criteria (except for the mid-term rank ordering of each student with the others in his section) may be seen by comparing the data from Experiment I with that from Experiments II, III, and IV. First, the announcement to individual students in Experiment I of his rank among his classmates in his section had no immediate or long-term effect on subsequent performance, in spite of their knowledge that they were going to be graded "on the curve" at the end of the quarter. Second, the overall quality of academic performance of these students was consistently poor throughout the quarter in comparison to the performance of students in Experiments II, III, and IV. They barely averaged performance meeting the Low criteria used in the other experiments. This was in spite of the facts that (1) they used the same procedures in all other respects, (2) they averaged the same amount of exposure to the items, and (3) they were in open competition with their peers (students in other experiments were graded only with respect to the criteria). In addition, the presence or absence of their performance depicted on a graph had no measured effect on their performance.

In Experiment I, the students were each left to define their own criteria (with the exception of the mid-term ranking). As a result, the instructor had no idea of what kind or level of criteria were being used by any one student. It is probable that each student drew upon his entire academic and personal history as well as from current variables to guide his study and performance. Given the uniqueness of individual histories, it is likely that when criteria were personally defined they were highly variable in both kind and quantity and were always peculiar to each student. The result of this practice was extremely inferior performance compared to that of students with teacher-defined criteria.

The comparison of Experiment I with the other experiments also suggests that the effectiveness of criteria in controlling performance may be related to the degree of specificity in their definition. In other words, the data suggest that regardless of the particular terms of the criteria the more precisely stated they are, the sharper will be the degree of control over academic performance. Conversely, the more vaguely described the criteria, the less clear will be their relation to performance, and the more variable that performance will become. These observations will require further investigation to determine fully the importance of specificity.

This study is only the first in a series that must be conducted to describe further the functional relationships between minimum performance criteria defining course grades and academic performance. It will be important, for example, to specify more clearly the differences between the control exerted by the grade label itself, holding criteria constant (*cf.* Experiments II and V) as opposed to the control of minimum performance criteria themselves, holding grade labels constant (*cf.* Experiments II, III, and IV).

However, some tentative implications of the present study are fully warranted by the data described here. The following tactics regarding the use of minimum criteria are suggested for the purpose of maximizing the control of criteria over performance in an educationally desirable

manner. (1) Define minimum criteria for academic performance. (2) Describe criteria as precisely as possible in terms of specifically defined student behaviors. (3) Define performance criteria at the beginning of the course. (4) Define criteria for only the highest standard of performance you have. If you must work in the traditional grading system, let that criteria define course grade of "A". Do not define successively lower criteria for lower course grades. If you work with a Pass-Fail system, define only the highest standard of performance as the minimum Pass criteria. If performance tends to attain criteria by the narrowest possible margin, the only way to determine if a student can produce better work is to raise the criteria. In summary, the data suggest that with respect to criteria the teacher should start high and go higher.

REFERENCES

- Born, D. G. and Herbert, E. A further study of Keller's personalized system of instruction. *Journal of Experimental Education*, 1971, **51**, 5-11.
- 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.
- Ferster, C. B. and Perrott, M. C. *Behavior principles*. New York: Appleton-Century-Crofts, 1968.
- Johnston, J. M. *An examination of critical variables in a behavioral approach to college teaching*. Unpublished doctoral dissertation. University of Florida, 1970.
- Johnston, J. M. ". . . to beat back the frontiers of ignorance." Unpublished manuscript. January, 1972.
- Johnston, J. M. and Pennypacker, H. S. A behavioral approach to college teaching. *American Psychologist*, March, 1971, **26**, 219-244.
- Johnston, J. M., Roberts, M., and O'Neill, G. The measurement and analysis of college student study behavior. In G. Semb (Ed.), *Behavior analysis and education—1972*. Lawrence, Kansas: Support and Development Center for Follow Through, Department of Human Development, University of Kansas, 1972. Pp. 377-392.
- Keller, F. S. "Good-bye Teacher . . ." *Journal of Applied Behavior Analysis*, 1968, **1**, 79-89.
- Mahan, H. C. *The use of Socratic type programmed instruction in college courses in psychology*. Unpublished paper read at Western Psychological Association. San Francisco, May 1967.
- Malott, R. W. and Janezarek, K. *The effect of daily quizzes on hour examination performance*. Unpublished master's thesis. Western Michigan University, 1970.
- Malott, R. W. and Rollofson, R. L. *An empirical evaluation of student-led discussions*. Unpublished master's thesis. Western Michigan University, 1970.
- McMichael, J. S. and Corey, J. R. Contingency management in an introductory psychology course produces better learning. *Journal of Applied Behavior Analysis*, 1969, **2**, 79-83.
- Myers, W. A. Operant learning principles applied to teaching introductory statistics. *Journal of Applied Behavior Analysis*, 1970, **3**, 191-197.
- Pennypacker, H. S. *Precision teaching of an undergraduate program in behavior principles*. Unpublished paper read at Midwestern Psychological Association. Chicago, February, 1969.
- Postelthwait, S. M., Novals, J., and Murray, H. *An integrated experience approach to learning—with emphasis on independent study*. Minneapolis, Minnesota: Burgess Publishing Co., 1964.
- 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.
- Sherman, J. G. *Personalized system of instruction newsletter*. Georgetown University, Washington, D.C. Issue #1, June, 1971.
- Skinner, B. F. Are theories of learning necessary? *Psychological Review*, 1950, **57**, 193-216.
- Skinner, B. F. Teaching science in high school. *Science*, 1968, **159**, 704-710.
- Stalling, R. B. A one-proctor programmed course procedure for introductory psychology. *Psychological Record*, 1971, **21**, 501-505.

Received 15 May 1972.

(Revision requested 1 September 1972.)

(Final acceptance 13 November 1972.)