J. OF THE NORTHEASTERN AGR. ECON. COUNCIL VOL. IX. NO. 1, APRIL, 1980

USE OF COMPUTER TESTING FEEDBACK FOR INSTRUCTIONAL IMPROVEMENT

Daymon W. Thatch

Abstract. This study reports the development of a computer assisted testing (CAT) system and the use of this system over the past five years to develop a normative testing tool in an introductory microeconomics theory course. The CAT system is more than a testing tool and can be used to aid in instructional improvement by pinpointing difficulties in technical material areas, levels of learning abilities, general lack of understanding and patterns of mathematical, graphic and written problems. It would appear that the techniques used to develop the normative tests in this report could also be used to develop individual standarized examinations tailored to the learning objectives in other courses.

The use of computers in instruction is not new—a number of references can be found in the AJAE and its predecessor, the Journal of Farm Economics, during the past 15 years or more (see C. French for a review of literature prior to 1973). Earlier references, as well as more recent ones, have mostly been concerned with the impact that computers have on teaching or the use of computers in games, teaching simulation and individualized instruction [Boehlje, Hammonds, Kendrick, Walker and White]. This article concentrates on the use of computers by the teacher as a tool in testing and instructional improvement. More specifically, three objectives are of primary concern. First to report on the development of a computer-assisted testing (CAT) system; second to present the results of using the (CAT) system and analyze the results of testing this system over the past five years in developing a normative testing tool; and third, to explore ways that the (CAT) system can be used to improve instruction.

Although the AJAE has been noticeably void of articles on computer-assisted teaching, a number of authors have documented programs that have been developed to grade and analyze the results of multiple-choice examinations [*McDonald, Oosterhof, Thatch, Wessel, Westcott*]. As one expects, all of the CAT programs are slightly different but all incorporate the ideas of using a true-false and/or multiple choice pencil-marked answer sheet that can be read by a special machine scanner, a way to feed the machine to read results into the developed program, and a computer program to analyze and print out the results. All programs also use a form of item analysis and several types of summary statistics and tables to report the results.

DEVELOPMENT AND USE OF COMPUTER STANDARDIZED TESTING

The original computer grading system was developed in the late 1960s [Westcott and Thatch]. The philosophy in developing the system was to use standardized acceptable statistics and yet incorporate as much flexibility for the user as possible. For the most part, the statistics reported in the Westcott article and the methods used to calculate them were the same as those used by the Educational Testing Service.[1]

After several program output format modifications, the computer grading system was pre-tested on a number of different classes of various sizes both within and outside the Agricultural Economics Department and within and outside of Rutgers—The State University of New Jersey. Since 1973, the program has been adapted so that the input could be taken directly from OPSCAN and more recent modifications allow the entire grading system to be handled with on-line equipment.[2]

In 1970, the author started to develop a test-bank of multiple-choice questions that could be used to develop a standardized microeconomic principle course final examination. After several years of test-question modifications and student and faculty feedback the examination was used in 1974 for the Principles of Microeconomics course. The examination was designed as a ninety (90) question, five-answer multiple-choice test that incorporated a number of questions on each major concept that the course was designed to cover. The design also covered a wide range of learning values[3] such as, recognition, understanding, simple application and complex applications. Graphic and numerical problems were used as well as various levels of written problems.[4]

Although the same microeconomic concepts and the total number of questions in the standardized examination are held constant each year, the wording in a number of questions has been changed over the years. Wording changes of both questions and/or possible responses were necessary to correspond to changes in textbook author's terminology, when data or facts became dated, or when experience showed that students were continually missing a question based on the wording in the question or its possible responses. As individual questions became better discriminators (that is, the students who did better on the exam versus the students who did poorer on the exam), the relaibility of the total examination increased. For example, for the microeconomic final examination developed above, over the last five years of use, the reliability increased from the mid '80s to the low '90s.

STANDARD TESTING MICROECONOMIC RESULTS

Two different sets of results are reported using the above computer testing program and the developed standardized microeconomic principles examination (Tables 1 and 2). In Table 1, the results are from using the same 90-question microeconomics final over a six-semester, five-year basis. Table 2 reports the results of using the same standardized examination by a different professor who decided to use less than 90 questions (67 to 72) to

Daymon W. Thatch is Associate Professor, Department of Agricultural Economics and Marketing, Rutgers—The State University of New Jersey, New Brunswick, New Jersey.

[1]Educational Testing Service, Princeton, N.J., Berkeley, California and Evanston, Illinois.

[2]OPSCAN is an Optical Mark Reader (OMR) that permits students' answer sheets to be directly converted into punched data cards.

[3]A good discussion of learning levels can be found (Bloom).

[4]The five-choice multiple examination has also been used with only two choices for true and false examination and with five choices combined for ten possible responses. The method has also been used for weekly quizzes on hourly examinations as well as for finals. The computer matrix design can also be changed for various class sizes for a more economical operation. For our college needs we have used the program for class sizes from 15 to 600, but, in theory, there is no maximum to number of students that can be handled.

Table 1
Comparison of Professor X's Statistical Class Differences in Testing
Introductory Micro-Economic Theory[a]

	Years													
	Jan. 1974		Jan. 1975		Dec. 1975		Jan. 1977		Dec. 1977		May 1978		Overall Averages	Ranges
	Tests[b]													
	A	B	A	B	A	B	A	B	A	B	A	B		
Mean[c]	62.0	64.6	65.0	62.5	62.2	63.1	65.3	62.2	63.0	64.0	67.4	63.9	63.8	62.0-67.4
Standard Deviation[d]	11.08	10.34	10.03	10.43	12.65	11.31	10.96	10.62	11.51	10.97	9.26	9.76	10.74	9.26-12.65
High Test Score	85	81	79	79	81	80	86	80	87	85	80	78	81.8	79-87
Low Test Score	44	43	37	34	38	37	37	33	34	39	47	48	39.3	34-48
Students per Test	26	28	25	25	21	18	35	35	29	29	25	26	26.8	18-35
Class Size	54		50		39		70		58		51		53.7	39-70

[a]Based on a 90-question test.

[b]Tests A and B are the same questions but question order on tests and position of correct answer has been changed.

[c]None of the means in a given year or between years was found significantly different from any other means at a .05 level of significance except for Jan. 1974 A and May 1978 A, which was significant at .02 level.

[d]None of the standard deviations in a given year or between years was found significantly different from any other standard deviation at a .05 level of significance.

Table 2
Comparison of Professor Y's Statistical Class Differences in Testing
Introductory Microeconomic Theory

	Years						
	May 1975[a]		May 1977[b]		May 1978[c]		
	Test[d]					Overall	
	A	B	A	B	A	Averages	Ranges
Mean[e]	48.7	46.8	48.8	49.7	47.6	48.3	46.8 -49.7
Standard Deviation[f]	7.42	9.88	8.81	8.44	9.24	8.76	9.88- 7.42
High Test Score	64	61	66	64	64	63.8	61-66
Low Test Score	35	27	30	34	26	30.4	26-35
Students per Test	31	31	46	45	37	38.0	31-46
Class Size	62		91		37	63.3	37-91

[a]Based on 67 Questions.

[b]Based on 72 Questions.

[c]Based on 70 Questions. Due to small class size only one exam was given.

[d]Tests A and B are the same questions but question order on tests and position of correct answer has been changed.

[e]None of the means in a given year or between years was found significantly different from any other means at a .05 level of significance.

[f]None of the standard deviations in a given year or between years was found significantly different from any other standard deviations at a .05 level or significance.

correspond only to the areas that his course objectives were designed to cover.

Both professors administered all examinations in a similar fashion. The class was randomly divided in half—one group was given Exam A and the other, Exam B. Although the same questions appear on both examinations (A and B), the order of the questions, as well as the order of the responses, was different. This change between A and B, as well as different color coding, allowed for closer seating and a minimum of security-observation problems. The two examinations were then analyzed separately; comparisons of the results are shown in Tables 1 and 2.

The results show that for the 12 observations in Table 1, there were no significant differences at a 5-percent level between the mean scores or standard deviations between any of the A's or between any of the B's, or between any of the A' and B's. The only exception was that in one year there was a significant difference between mean test score of two A's at the 5-percent level (but not at the 2-percent level). In a similar fashion, for the five observations (3 years) in Table 2, there were no significant differences at the 5-percent level between any of the means and standard deviation of any of the A's, B's or between any of the A's and B's.

In short, it appears that over the past five years of testing the developed standardized microeconomic examination, the examination has given consistent results in terms of class mean and standard deviation scores. As a result, the examination could be used to judge other students' performances in terms of the author's course in Microeconomic Principles (assuming constant course objectives). A second point that would follow is that, using the above general computer grading program and general question testing procedure, other teachers could develop individual

standardized examinations tailored to the learning objectives of their particular courses. This, in fact, is what happened both within the Department of Agricultural Economics and within a number of other departments at the College.

USING TEST RESULTS FOR INSTRUCTIONAL IMPROVEMENT

There are at least four potential areas where individual question analysis can be used to improve instruction. None of the areas are unique to this computer-testing program or to computer-testing in general. However, the effort to analyze each individual question on an examination without the computer would be excessive and probably few teachers would expend the effort. In each of the potential areas, the computer variables used are the same: the percentage of the students who obtained the correct answer, the discrimination index for the question, and the percent and/or number of students who responded to each choice on each question. In addition to individual question results, each student's overall score, the questions that each missed and how each did relative to the class can be part of the scoring process.

The first potential gain is for technical areas (or general knowledge) not well covered by the instructor or understood by the students. Consistently missing questions in one general or technical area with no one or two wrong choices being dominant usually indicates a general lack of understanding. If the 'best' students (as defined by the ones who do best overall on this test) miss the question along with the others, either the degree of difficulty is 'too hard' or the instructor has not adequately covered or stressed the point.

Levels of learning can be indicated by the types of questions missed and which students miss what types of questions. Most students who have studied, or at least attended most classes, can achieve reasonably good scores on recognition, understanding, and some simple application questions. On the other hand, for most students, complex application takes a better understanding and the ability to apply general concepts. By analyzing those questions missed, the instructor can determine the achievement level of the class and where additional emphasis is needed. The level of achievement also signals if the examinations are being over-or under-designed in terms of meeting the classes' progress and course objectives.

A third area for potential instructional improvement is through analyzing of the multiple choices (distractors). As indicated above, if all wrong choices are fairly evenly chosen there is usually a general lack of understanding. On the other hand, if one or two wrong choices are consistently chosen by the better students, that indicates something is wrong in the wording of the question or that something in the choices is giving the students a false indication. It can also indicate that the instructor gave false or misleading information on the subject.

A fourth area for instructional improvement is by analyzing how the class did by types of examination questions. By examining test results on mathematical, graphic or descriptive questions, patterns of student's problems often become apparent. For example, in the micro-principle course it was found that many non-math and non-science majors have had difficulty with quantitative and graphic problems.

OTHER ADVANTAGES OF USING COMPUTER TEACHING FEEDBACK

Although a complete paper could be written on the virtues of computer teaching and testing, perhaps a listing of some of these advantages may be of value, several of which are not unique to (CAT):

(1) Students get an analysis of the questions (concepts) they missed, why, and their relative position to the other members of the class. Rapid turnaround grading time (1 day or less) as well as the additional information has enhanced student interest in CAT.

(2) Teacher's confidence is enhanced in terms of what is and is not given students and the progress students are making in achieving the course objectives. Since computer grading is objective, once reliable and valid questions are selected, the teacher can feel comfortable with a particular grade.

(3) Given an amount of time to devote to a course, the teacher's scarce resource of time is utilized more efficiently.

(4) Examinations become more reliable in terms of meeting the course objectives.

(5) Norms of testing as well as standards are realistically set without use of subjective evaluations.

(6) The development of a test file of 'good' questions in terms of their ability to discriminate between students.

SOME FINAL THOUGHTS

In the final analysis what is really important in each course is the students learning the objectives as specified by the teacher. Evaluation testing should not be an end in itself but a means to see how well the course objectives have been achieved.

The computer is a valuable tool to increase labor efficiency in the classroom in the area of computer-assisted testing (CAT). On the other hand, it is not panacea for every class or every course. Yet, it does seem reasonable, as Theodore Schultz has noted that, "We have not really faced up to the fact that the services of the faculty have become more expensive relative to other instructional inputs. We have not really looked for substitutions." This statement is much truer today than when he made it in 1965.

The great value of CAT is that it helps both the teacher and the student; the teacher in terms of pointing out strengths and weaknesses of his course and the student, in terms of the strengths and weaknesses of his understanding of the courses concepts. The computer does not take the place of the teacher but it relieves him of many of the routine tasks he faces and thus frees him to interact in a more creative manner. Yes, we could do the same testing and question analysis without the aid of the computer but how many of us would devote the time and effort to the task? Furthermore, why should we devote the time to these routine tasks when with the aid of CAT it is so practical and economical?

In 1972, the Carnegie Commission on higher education published a report entitled, "The Fourth Revolution." This report refers to the fact that the fourth revolution in education will be the emergence of modern technology and the availability of electronic media devices for the use in education. Speaking on the broad concept of instructional technology, the Commission predicted that, "The widespread acceptance and application of this broad definition belongs to the future."

Although the future has not yet arrived for the Agricultural Economics teaching profession, there does appear to be signs of encouragement on the horizon.

REFERENCES

Bloom, Benjamin. "Classification of Educational Goals," *Taxonomy of Educational Objectives.* David McKay Co., New York, New York, 1956.

Boehlje, D. Michael and Vernon R. Eidman. "Simulation and Gaming Models: Application in Teaching and Extension Programs." *Amer. J. Agr. Econ. 60(1978):987-92.*

Carnegie Commission on Higher Education. The Fourth Revolution. New York: McGraw-Hill Book Co., 1972.

Edward, C. Richard and Arthur MacEwan. "The Teaching of Economics." *American Economic Review,* 60(1970):1-12.

French, E. Charles. "Selected Alternative Programs for Bringing the Real World to the Undergraduate Classroom." *Amer. J. Agr. Econ.* 56(1974):1163-75.

Hammonds, M. Timothy. "Guided Analysis: An Experiment in Educational Techniques." *Amer. J. Agr. Econ.* 58(1976):546-50.

Kendrick, G. James. "Techniques for Motivating Students." *Amer. J. Agr. Econ.* 55(1973):762-66.

McDonald, K. Sandra. *Computer Capabilities and Usage for Undergraduate Grading and Reporting.* Texas A & M University, Dept. of Agricultural Economics, 1976.

Mosher, T. A. "Higher Education in the Rural Social Sciences." *Amer. J. Agr. Econ.* 55(1973):711-19.

Oosterhof, C. Albert. *Preparation of Data Input and Interpretation of the Revised Kocher-Oosterhof Item Analysis.* University of Kansas, Dept. of Education, Research and Management, 1972.

Schultz, W. Theodore. "Reflections on Teaching and Learning in College of Agriculture." *Journal of Farm Economics,* 47(1965):17-22.

Thatch, W. Daymon. *Grading and Analysing of Multiple Choice Questions by Off-Line and Computer Equipment.* Cook College—Rutgers, The State University, Dept. of Agricultural Economics and Marketing, June 1976.

Walker, L. Odell. "Innovations in Undergratuate and Extension Teaching: Discussion." *Amer. J. Agr. Econ.* 60(1978):1003-1007.

Wessel, L. Kelso. *Computer Assisted Testing (CAT).* Ohio State University, Dept. of Agricultural Economics and Rural Sociology, March, 1975.

Westcott, R. Edwin and Daymon W. Thatch. *Computer Program for Grading and Analyzing Multiple Choice Questions.* Cook College, Rutgers—The State University, Dept. of Agricultural Economics and Marketing, May 1972.

White, C. Fred. "Simulation of Agricultural Policy Formation as a Teaching Approach." *Amer. J. Agr. Econ.* 59(1979):225-27.