RESEARCH
REPORTS

October, 1966 No. 15

CAN COMPUTERS WRITE COLLEGE ADMISSIONS TESTS?

James M. Richards, Jr.

Published by Research and Development Division · American College Testing Program · P. O. Box 168, Iowa City, Iowa 52240

Summary

For many years psychological tests have been scored by machines, and recently computers have assembled existing items into tests and have scored essay examinations. This study goes beyond these earlier techniques and explores the possibility of computer item writing. A computer procedure for writing *Verbal Comprehension* items was developed and used to write a 72-item test. This test, together with the Wide Range Vocabulary Test, was administered to University of Iowa freshmen. The test intercorrelations, reliabilities, and correlations with grades suggest that, in principle, computers can write college admissions tests. Possible objections to computer written tests are considered.

Can Computers Write College Admissions Tests?

James M. Richards, Jr.

For many years, objective tests and machine procedures have been reducing the role of human beings in measuring the aptitude and achievement of students. The first large scale application of machines in testing, of course, was to the scoring of tests. With subsequent advancement in the technology of scoring machines, it is now possible to score in two or three days the tests of several hundred thousand students who took a nationwide examination. In later applications of machines, Rock (1965) has shown that computers can simulate the behavior of human test developers in assembling a test with specified properties from an item file; Page (1966) has shown that computers can score some aspects of essay examinations; and Osburn (1966) has developed a computer procedure for writing statistics problems.

Recently, large, supplementary random access memory devices have been developed for computers. The characteristics of such devices also suggest the possibility of writing multiple choice test items on a computer. The purpose of the present study, therefore, is to determine whether, in principle, tests for screening college applicants can be written by computers.

The test factor most useful for predicting academic success in college, probably, is Verbal Comprehension. I attempted, therefore, to develop a procedure that a computer could use to write verbal comprehension items. Specifically, a procedure for writing synonyms items was developed. A synonyms test provides a good measure of verbal comprehension, and the use of the Wide Range Vocabulary Test from the ETS Factor Kit made it possible to compare a machine written synonyms test with a factorially pure synonyms test written by humans.

Computer Procedure

It is relatively easy to develop a procedure for choosing a stem and a correct alternative. One merely stores a dictionary of synonyms in a supplementary memory device, generates a random number, and uses this number to select a word to be the stem. One then generates another random number and uses it to select one of the synonyms for the stem as the correct alternative. If the stem word is more than one part of speech, it is necessary to choose one part at random for use in the item before choosing the correct alternative.

Just this procedure was used to pick stem words and correct alternatives in the present study. Since no computer with a large supplementary memory device was available, the actual operation of the computer was simulated. The simulation was rigorous, however, and the items correspond exactly to what would be written by a computer. This simulation used one of the oldest random access supplementary memory devices--namely, a book.

After a stem and correct alternative were chosen, obviously the next step was to choose "distractor" alternatives (i.e., wrong answers). It soon became clear that developing a sensible procedure for choosing distractors is the most difficult problem in writing tests on a computer. Indeed, a perusal of the literature, after this problem became obvious, suggested that distractor alternatives are little investigated and little understood and are perhaps the most important neglected problem in the construction of multiple choice tests. Most writers on test development merely suggest that the item writer must use his judgment and ingenuity in choosing distractors. Since computers have neither judgment nor ingenuity, such advice was useless for the present study.

The problem of distractors for synonyms items was solved in the following way. Roget's *Thesaurus* provides a classification scheme for word meanings in which each word is categorized into one or more categories. Numbers are assigned to these categories of meaning in accordance with an overall conceptual scheme. In the present study, the basic procedure for picking distractors was to choose randomly from words in adjacent categories in the *Thesaurus* scheme.

More specifically, the numerical code or, in other words, the category of meaning, shared by the stem and the correct alternative was determined. If more than one numerical code pertained, a random number was generated and used to select one of the codes. The procedure for modifying the numerical code as a step in picking distractors from adjacent categories consisted of adding *1* to or subtracting *1* from the code number for the stem and correct alternative. A random number determined the choice between addition or subtraction. Then the distractor was chosen at random from words with the modified code number and of the appropriate part of speech. A check was then made to determine whether the distractor had *any* code numbers in common with either the stem or the correct alternative. If so, a new distractor was chosen at random. Additional distractors were chosen by repeating the process of modifying the code number and choosing a distractor at random. Thus, the first distractor was chosen from words one category away from the stem, the second distractor from words two categories away, etc. In a few instances, of course, all the words in a category would share a common meaning with the stem. In such cases, the procedure was simply to modify the code number and choose a distractor from the next category. The locations of the correct alternative and the various distractor alternatives in the item were also determined by random numbers.

In this way, a 72-item computer written (or at least writable) synonyms test with four-alternative items was developed. Such items and tests have several interesting properties. First of all, such items are definitely a random sample from a specified population of items. Thus, a test composed of them would conform rigorously to an assumption for several ways of estimating reliability. Second, while the population of items is very large, it is certainly finite. Similarly, while the number of alternative test forms of a given length that can be assembled from such items is very, very large, it also is finite. It appears, moreover, that the finite quality results primarily from the complete specification of the operations for writing items and may be common to many, if not most, item populations. Since the population of items and tests is very large, this has few practical implications. It may have theoretical implications, however. For example, true scores are usually defined in terms of an infinite number of tests.

Moreover, it appears that similar procedures can be used to write many other kinds of multiple choice aptitude test items. With only minor modifications the procedure used in the present study could be used to write analogies items. In addition, no insurmountable difficulties should be encountered in programming a computer to write mathematics items, although developing efficient distractors might be a problem.[1] The work of Osburn (1966) is a promising step toward computer written math items.

Evaluation of the Test

The computer written test, together with the Wide Range Vocabulary Test was administered in the fall of 1965 to entering freshmen at the University of Iowa. A generous time limit permitted all students to complete the computer test. At the end of the first semester, the grade point average was determined for these students.[2] The means, standard deviations, K-R 21 reliabilities, predictive validities, and intercorrelation of the two tests[3] for 599 male and 613 female freshmen are shown in Table 1.

These results are partly bad and partly good. The items in the computer written test, on the average, are easier than those in the Wide Range test (compare the mean to the length for both tests), and the reliabilities are lower for the computer test. This is disappointing because the greater length of the computer test should produce higher reliability.

[1] It is less clear that all possible types of items can be written by computers. John Holland (personal communications) has suggested a model for items in which alternatives are points equi-distant in a response space. It is most unlikely that existing computers can write such items.

[2] I want to thank Ted McCarrel, Charles B. Statler, and Willard L. Boyd for making it possible for me to obtain these data.

[3] Data analysis for this study carried out at the University of Utah computer center.

On the other hand, the validities are comparable to those for the Wide Range test, and the intercorrelation of the two tests is not too far from the limits set by reliability. On the whole, these results are encouraging enough to justify further analysis.

Table 1

Mean, Standard Deviation, Reliability, Validity, and Intercorrelation of Wide Range Vocabulary Test and the Computer Written Test

	Males (N=599)	Females (N=613)
Wide Range Vocabulary Test		
Number of Items	48	48
Mean	23.74	25.16
Standard Deviation	6.44	6.42
K-R 21 Reliability	.73	.73
Correlation with first semester GPA	.30	.30
Computer Written Test		
Number of Items	72	72
Mean	53.88	55.53
Standard Deviation	5.76	5.41
K-R 21 Reliability	.60	.57
Correlation with first semester GPA	.30	.32
Intercorrelation of two tests	.62	.64
Intercorrelation corrected for unreliability of both tests	.94	.98

Since the items in the Wide Range Vocabulary Test were carefully selected, while the computer test consisted of entirely unselected items the effect of item selection should be evaluated.[4] The Wide Range test consists of two parts with 24 items each. Accordingly, 24 item tests were selected from the computer test by three different procedures. The item selection was based only on data for males, so that the female sample would provide a cross-validation group. The first procedure, designed to maximize validity, was simply to pick the 24 items with the highest correlation with first semester grades. The second procedure, designed to maximize homogeneity, was to pick the 24 items with the highest correlation with the Computer Test total. The third procedure, designed for high discrimination among students tested, was to eliminate items answered correctly by more than 90% or fewer than 10% and then choose the 24 items with the highest correlation with the Computer Test total. Each of the procedures, of course, could easily be incorporated in a computer program for assembling test forms.

Data for the two parts of the Wide Range Vocabulary Test and the three different tests composed of selected items from the Computer Written Test are summarized in Table 2. It should be remembered that some values for males are inflated because the correlations and item selection are based on the same group. There is also considerable item overlap in the three computer tests, which produces much inflated intercorrelations.

While the computer test is still somewhat easier, these data confirm without qualification that synonyms tests can be successfully written

[4] It should be noted, however, that selected items are no longer a random sample from the specified population.

Table 2

Comparison of Two Parts of Wide Range Test with Tests Composed of Selected Computer Written Items

	Males (N=599)				
	1	2	3	4	5
Wide Range Test					
1. Part 1	--				
2. Part 2	.51	--			
Computer Test					
3. Items selected for validity	.62	.43	--		
4. Items selected for homogeneity	.61	.41	.93	--	
5. Items selected for discrimination	.64	.45	.89	.88	--
K-R 21 Reliability	.56	.71	.61	.60	.63
Correlation with first semester GPA	.29	.21	.41	.35	.35
Number of items	24	24	24	24	24
Mean	12.32	11.54	18.63	19.03	14.94
Standard Deviation	3.59	4.34	3.17	3.08	3.78

	Females (N=613)				
	1	2	3	4	5
Wide Range Test					
1. Part 1	--				
2. Part 2	.50	--			
Computer Test					
3. Items selected for validity	.52	.56	--		
4. Items selected for homogeneity	.45	.56	.92	--	
5. Items selected for discrimination	.51	.58	.89	.88	--
K-R 21 Reliability	.68	.41	.55	.53	.61
Correlation with first semester GPA	.28	.22	.30	.29	.29
Number of items	24	24	24	24	24
Mean	13.57	11.58	19.30	19.66	16.46
Standard Deviation	4.14	3.14	2.82	2.69	3.52

Note--Some values for males are inflated because item selection and correlations are based on the same group. Intercorrelations of computer tests are inflated by substantial item overlap.

on computers. Each of the three computer written tests correlates about as highly with the two parts of the Wide Range Test as the two parts correlate with each other; the reliability of each computer test is somewhere between the reliabilities of the two parts of the Wide Range Test; and the predictive validities of the computer tests are, if anything, slightly higher than the validities of the Wide Range Test.

Discussion

The results clearly indicate that it is possible, in principle, for computers to write college admissions tests, which suggests that, with a thorough application of existing computer technology, it would be entirely feasible, in principle, to automate all aspects of college admissions testing. It would be possible to install in each high school an input-device to a central computer at a testing agency. On this device, the computer would display test items one by one on a screen; the student would push a button to indicate his response to each item; and his response would be transmitted back to the computer. Through time sharing, many students could be tested at once. The central computer would generate new items, include a few with each test administered, develop and equate alternate forms of the test, score each student's test, and even transmit the scores directly to another computer (or output device) at the college of the student's choice. Thus it would also be possible to eliminate answer sheets, scoring machines, and score report forms. All of this could take place entirely "untouched by human hands."

Let me emphasize that while this is possible with existing computer technology, a possibility is not an imperative, and there are many legitimate objections to such a dehumanized system. I hope, however, that humanists and others appalled by the 1984 overtones will make a more serious response to it than an implied Luddite computer smashing. For if computer written college admissions tests are objectionable, rejection of them on some basis other than predictive validity, while leaving unchanged the social system making high validity possible, would be a particularly dishonest and deplorable case of removing symptons rather than treating the underlying disease.

Any serious confrontation of the issues involved must recognize what is, on the evidence, an indisputable fact: namely, that multiple choice tests--even those written by computers--are the best and fairest currently available estimates of potential for academic success in colleges as they are now constituted. This is true in the sense that an able college applicant, if he is behaving rationally in his own self interest, will take such a test in preference to others ways of having his academic potential evaluated, such as taking an essay examination or being interviewed.

What convincing argument, then, can be advanced against having such tests written by a computer if the products are indistinguishable from, or superior to, tests written by people? I think any serious criticism of such tests must rather begin with a criticism of conventional measures of success in college, or, to put it bluntly, of grades given by college professors. If grades can be predicted quite well by a completely dehumanized test, just what is wrong with grades? That something may indeed be seriously wrong is indicated by the many studies showing, at best, a negligible relationship between grades and performance outside the classroom in important area of human endeavor (Hoyt, 1966; Richards, Holland, and Lutz, 1966).

References

Hoyt, D. P. College grades and adult accomplishment: A review of research. *Educational Record,* 1966, 47, 70-75.

Osburn, H. G. Computer aided item sampling for achievement testing. Paper read at American Psychological Association, New York, 1966.

Page, E. B. Automatic grading of student essays. Symposium presentation at American Educational Research Association, Chicago, 1966.

Richards, J. M. Jr., Holland, J. L, & Lutz, Sandra W. *The prediction of student accomplishment in college.* ACT Research Report No. 13. Iowa City, Iowa: American College Testing Program, 1966.

Rock, D. Assembly of tests by use of an automated item file. *Research Memorandum* 65-14. Princeton, New Jersey: Educational Testing Service, 1965.

ACT Research Reports

This report is the fifteenth in a series published by the Research and Development Division of American College Testing Program. Reports are published monthly and mailed free of charge to educators and other interested persons who have asked to be on the special research report mailing list.

The research reports have been deposited with the American Documentation Institute, ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. (ADI Document numbers and prices are given below.) Printed copies may be obtained free from the Research and Development Division, American College Testing Program. Photocopies and 35 mm. microfilms may be obtained at cost from ADI by citing ADI Document number. Advance payment is required. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress.

Reports in the series are:

*No. 1 _A Description of American College Freshmen_, by C. Abe, J. L. Holland, Sandra W. Lutz, & J. M. Richards, Jr. (ADI Doc. 8554; photo, $8.75; microfilm, $3.00)

*No. 2 _Academic and Non-academic Accomplishment: Correlated or Uncorrelated?_ by J. L. Holland, & J. M. Richards, Jr. (ADI Doc. 8555; photo, $3.75; microfilm, $2.00)

No. 3 _A Description of College Freshmen: I. Students with Different Choices of Major Field_, by C. Abe., & J. L. Holland (ADI Doc. 8556; photo, $7.50; microfilm, $2.75)

No. 4 _A Description of College Freshmen: II. Students with Different Vocational Choices_, by C. Abe., & J. L. Holland (ADI Doc. 8557; photo, $7.50; microfilm, $2.75)

*No. 5 _A Description of Junior Colleges_, by J. M. Richards, Jr., Lorraine M. Rand, & L. P. Rand (ADI Doc. 8558; photo, $3.75; microfilm, $2.00)

No. 6 _Comparative Predictive Validities of the American College Tests and Two Other Scholastic Aptitude Tests_, by L. Munday (ADI Doc. 8559; photo, $2.50; microfilm, $1.75)

*This report now available only from ADI.

ACT Research Reports (cont.)

No. 7 *The Relationship Between College Grades and Adult Achievement. A Review of the Literature,* by D. P. Hoyt
(ADI Doc. 8632; photo, $7.50; microfilm, $2.75)

No. 8 *A Factor Analysis of Student "Explanations" of Their Choice of a College,* by J. M. Richards, Jr., & J. L. Holland
(ADI Doc. 8633; photo, $3.75; microfilm, $2.00)

No. 9 *Regional Differences in Junior Colleges,* by J. M. Richards, Jr., L. P. Rand, & Lorraine M. Rand
(ADI Doc. 8743; photo, $2.50; microfilm, $1.75)

No. 10 *Academic Description and Prediction in Junior Colleges,* by D. P. Hoyt, & L. Munday
(ADI Doc. 8856; photo, $3.75; microfilm, $2.00)

No. 11 *The Assessment of Student Accomplishment in College,* by J. M. Richards, Jr., J. L. Holland, & Sandra W. Lutz
(ADI Doc. 8955; photo, $3.75; microfilm, $2.00)

No. 12 *Academic and Non-academic Accomplishment in a Representative Sample taken from a Population of 612,000,* by J. L. Holland, & J. M. Richards, Jr.
(ADI Doc. 8992; photo, $3.75; microfilm, $2.00)

No. 13 *The Prediction of Student Accomplishment in College,* by J. M. Richards, Jr., J. L. Holland, & Sandra W. Lutz
(ADI Doc. 9020; photo, $5.00; microfilm, $2.25)

No. 14 *Changes in Self-Ratings and Life Goals Among Students at Colleges with Different Characteristics,* by R. W. Skager, J. L. Holland, & L. A. Braskamp
(ADI Doc. 9069; photo, $3.75; microfilm, $2.00)