DOCUMENT RESUME

ED 382 649 TM 023 083

AUTHOR Martinez, Michael E.; And Others
TITLE Automatic Scoring of Paper-and-Pencil Figural Responses. Research Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-90-23
PUB DATE Oct 90
NOTE 23p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Automation; *Computer Assisted Testing; Educational Technology; *Multiple Choice Tests; Sciences; *Scoring; Test Construction; Test Format; *Testing; Test Items; Test Use
IDENTIFIERS *Figural Response Items; *Large Scale Programs; National Assessment of Educational Progress

ABSTRACT

Large-scale testing is dominated by the multiple-choice question format. Widespread use of the format is due, in part, to the ease with which multiple-choice items can be scored automatically. This paper examines automatic scoring procedures for an alternative item type: figural response. Figural response items call for the completion or modification of figural material, including illustrations, diagrams, and graphs. Twenty-five science items were written in cooperation with the National Assessment of Educational Progress (NAEP) and were printed with an ink that was invisible to scanning equipment. The items were answered with pencils; response sheets were scanned, and the resulting data were processed by computer-based scoring algorithms. The paper describes the technology that led to successful pilot scoring of 7 items for 10 subjects. Implications of this technology for the future of large-scale testing are discussed. Three tables and two figures illustrate the discussion. (Contains 10 references.) (Author/SLD)

RESEARCH REPORT

AUTOMATIC SCORING OF PAPER-AND-PENCIL FIGURAL RESPONSES

Michael E. Martinez
John J. Ferris
William Kraft
Winton H. Manning

Educational Testing Service
Princeton, New Jersey
October 1990

Automatic Scoring of Paper-and-Pencil Figural Responses

Michael E. Martinez

John J. Ferris

William Kraft

Winton H. Manning

Educational Testing Service, Princeton, NJ

Running head: AUTOMATIC SCORING

Abstract

Large-scale testing is dominated by the multiple-choice question format. Widespread use of the format is due, in part, to the ease with which multiple-choice items can be scored automatically. This paper examines automatic scoring procedures for an alternative item type: figural response. Figural response items call for the completion or modification of figural material, including illustrations, diagrams, and graphs. Twenty-five science items were written in cooperation with the National Assessment of Educational Progress (NAEP) and printed with an ink that was invisible to scanning equipment. The items were answered with pencils; response sheets were scanned and the resulting data were processed by computer-based scoring algorithms. The paper describes the technology that lead to successful pilot scoring of seven items for ten subjects. Implications of this technology for the future of large-scale testing are discussed.

Automatic Scoring of Paper-and-Pencil-Based Figural Responses

Large-scale testing is founded on test items that can be (a) answered quickly, on the order of one to two minutes per item, (b) scored dichotomously, right or wrong, and (c) scored automatically, using machines rather than human graders. This niche is dominated by the multiple-choice format. However, it is widely held that multiple-choice items fail to capture many of the desirable products of learning (Frederiksen, 1984; Haney & Madaus, 1989). In this paper, an alternative item type, figural response, is described. Figural response items share some features of multiple-choice items: they, too, can be answered quickly and scored dichotomously. But figural response items have properties of their own, such as dependence on figural material as the response medium and the construction of responses in the mind of the test-taker.

There have been other attempts to automate scoring of non-multiple-choice test items (see, for example, Braun, Bennett, Frye, & Soloway, 1990; Braswell, 1990). In a project that was a precursor to the work reported here, Manning (1987) developed a method for scoring an item format called cloze-elide. Using this format, test-takers are given a text passage in which extraneous words have been inserted. The task for the examinee is to strike out, or elide, the extraneous words. The test format is a reversal of the traditional cloze procedure, in which omitted words are supplied by the examinee. In cloze-elide, a count of the elided extraneous words provides a measure of text comprehension. In the automatic scoring of cloze-elide, lines of text are printed over a matrix of elliptical fields read by optical scanners. When the matrix is scanned, it becomes possible to know which words have

been elided, both original and extraneous. Manning and his colleagues developed computer-based procedures whereby elided words were noted and scored automatically. The procedures showed high scoring reliability.

For cloze-elide and other formats, automatic scoring is significant. Rapid machine scoring might render alternative item types useful for large-scale assessment programs, most of which depend on high speed and low cost associated with automation. Our interest in paper-and-pencil delivery grows out of the conviction that printed tests are likely to be used with large numbers of examinees for many years to come. While the potential of figural response and other item types in non-paper, computer-delivered tests is now being examined, interest in finding improvements to paper-and-pencil tests is likely to grow.

The purpose of this paper is (a) to describe the development of automatic scoring procedures for figural response items and (b) to present pilot data on the accuracy of the procedures.

System Development

The Items

Twenty-five figural response items were developed as part of the field test of the 1990 National Assessment of Educational Progress (1988). Questions were written in three science content areas: life sciences, physical sciences, and earth and space sciences. Three grade populations were targeted: fourth, eighth, and twelfth, with some overlap of questions across grade levels. The figural response questions were developed in parallel with comparable multiple-choice questions. Statistical comparisons between the two formats are reported in a separate document (Martinez, 1990).

The items were printed in an orange "drop-out" ink. The term "drop-out" is used because the ink is designed to be invisible (or nearly so) to the two types of scanning equipment we used. With such an ink, the scanners were sensitive only to the pencil markings made by the examinee. Automatic scoring procedures were then applied to these data.[1]

Data were collected for developing and testing the automatic scoring procedures. Approximately 500 examinees, about 170 each from grades 4, 8, and 12, marked their responses with a regular, No. 2 pencil. The response sheets were returned to Educational Testing Service (ETS), where the figural response questions were scanned.

Scanning and Data Transfer

The pencil responses were scanned with two kinds of equipment: (1) ETS optical mark scanners used in grading multiple-choice responses and (2) high-resolution image processing equipment. The two types differ in the density of information they can capture from a response.

The standard (MRC) scanner. Commercial test scanners read an array of circles or ovals (collectively, "bubbles"), noting the presence and darkness of pencil markings in each bubble. Only a fraction of the total array is used on a test answer sheet. One of the scanners ETS uses in its regular operations has a maximum resolution of approximately 8 bubbles to the horizontal inch and 6 bubbles to the vertical inch. Thus, one square inch of paper has 8 x 6 or 48 bubbles.

A benefit of using the MRC scanner is its place in the existing infrastructure for handling high-volume test scoring. A potential difficulty for figural response scoring, however, is that the 6 x 8 bubble array leaves some white space, or channels, separating the rows and columns of dots. It is

therefore possible that all or part of a drawn line would "fall between the cracks" in the bubble array. A second concern was the the MRC scanner would not offer the level of resolution needed to score hand-drawn responses.

Image processing. Image processing has been used for numerous applications, including map reading (Kasturi, Fernandez, Amlani, and Feng, 1989), form scanning (Nieberding, 1990), and photograph reproduction (Reid-Green, 1990). The resolution of our image processing system surpassed the MRC scanners by orders of magnitude. Our PC-based system had a maximum resolution of 300 x 300 dots, or pixels, per square inch (300 DPI). One square inch yielded 90,000 bits of information, and an 8 1/2 by 11 sheet of paper (allowing for margins) contained some 3.8 million pixels (Table 1). This

---

Insert Table 1 about here

---

wealth of information actually proved to be a stumbling block at some points in the project.

The solution to our "embarrassment of riches" was to have the image processing equipment read the images at 100 DPI and compress the data. Compression amounted to transforming the data so that the large amount of white space on each answer sheet was recorded parsimoniously. With this procedure, each response required approximately 1500 bytes, and all responses were stored on 15 high-density (1.2 megabyte) diskettes. Pixels were then aggregated in 4 x 4 squares to form 25 "superpixels" per inch. Each of these superpixels was assigned a weight in the range of zero (none of the 16 pixels marked) to one (all 16 pixels marked). This resolution seemed coarse enough to simplify data manipulation, yet fine-grained enough for reliable and accurate

scoring as well as retention of important response information. The ideal density of information is an empirical research question, and one that is still open, but 25 pixels per inch seemed to be a reasonable density for reading and recording pencil markings.

The data were transformed to x-y coordinate values and uploaded to a mainframe file. Printouts were made of the data on the mainframe file, and a visual inspection confirmed that the match between these data and the markings on the original answer sheets was good. However, across the sample we did experience considerable data loss and distortion during image processor scanning. The data loss was related to the partial visibility of the ink used to print the questions. As a result, most recordings of pencil responses were contaminated with elements of the original questions. This was compensated for, in part, by adjusting the sensitivity of the image processing equipment and using filters--but sometimes at the cost of losing some of the fainter pencil markings. The pilot evaluation of scoring accuracy was therefore confined to a subset of the total data set.

Scoring

Item types. The x-y coordinates representing responses were interpreted on the basis of scoring rubrics -- rules for classifying responses -- the same as used for hand-scoring. Scoring procedures were written for three kinds of items:

Type I: point-and-mark items, in which an examinee is asked to identify a structure on a figure by marking it. In the cell structure problem (Figure 1), the correct response is to mark a mitochondrion.

Insert Figure 1 about here

Type II: arrow shaft items, where the examinee must draw an arrowhead onto the proper end of shafts that are provided in the problem. In the plate tectonics problem, seven arrow shafts are provided on sections of the earth's crust. The arrowhead, drawn by the examinee, shows where each plate is believed to move over time.

Type III: free-form arrows in which examinees show direction (of, say, a projectile) by drawing an arrow shaft and head in the correct direction. The path can be curved or straight.

For both Type I and Type II items, the scoring procedures involved the specification of response fields in which the presence or absence of pencil marks is noted. In the cell diagram (Figure 2), each organelle constitutes a

Insert Figure 2 about here

response field. Each response field is framed by a close-fitting polygon; each polygon, in turn, is defined by a series of x-y coordinates denoting consecutive vertices. The entire figure is enclosed by a "window of interest" in which all relevant responses were expected to fall. In Figure 2, the entire cell constituted a window of interest. Data outside this window was registered by the image processing equipment, but was ignored by the scoring algorithms.

In Type I problems, the scoring algorithms searched for the presence of pencil marks in the response field(s) denoting the correct answer or "key," and the absence of marks in other fields and regions outside the fields.

Incorrect responses were classified according to a priori response categories, such as the various organelles depicted in the cell. For example, if an examinee marked a Golgi apparatus instead of a mitochondrion, this information was noted and would be available for reporting.

For Type II problems, the response fields were rectangles enclosing the ends of each arrow shaft in which the arrowheads could be drawn. The algorithm checked that only one arrowhead was drawn for each shaft and that the correct end was the one marked. Each problem contained a number of shafts so the number of correct responses was tallied.

In Type III problems, windows were established to find the heads and tails of the arrows. Orthogonal regression lines were fitted to each line to determine its angle, length, and curvature. Tolerance parameters reflecting the rubric categories as well as our own judgment were set for direction, length, and curvature. Each value and tolerance parameter could be set independently or shut off, if so desired. For example, one trajectory problem had a straight arrow as its solution. Parameters were initially set at 135 degrees, +/- 20. The length parameter was turned off, along with its tolerance, since the correctness of the response did not depend on the length of the arrow. The curvature of the arrow was set to zero with a tolerance of 15 degrees.

To prevent erasures and smudges from being registered as responses, the markings in a field had to exceed a weight threshold in order for that field to be considered "marked." Likewise, the sum of pixels turned on outside the response fields had to be below a certain threshold for a response to be marked correct.

Pilot Analysis of Scoring Accuracy

The described scoring procedures were tested on responses gathered in the field test. The results of hand scoring formed the standard by which judgments concerning the efficacy of automatic scoring were made. Responses were rated by two people, on researcher and one test developer, both of whom were familiar with the science content of the questions.

Each response was scored on the basis of scoring rubrics. Rubrics separated responses into between four and seven content-based categories, some of which separated different kind of conceptual errors, while others reflected ordinal degrees of correctness. Training on the scoring rubrics consisted of a trial period where scores were assigned, compared, and resolved if discrepancies were found. Most responses were scored once, while a sub-sample was scored twice so that inter-rater reliabilities could be established. The inter-rater (Cohen's Kappa) scoring reliabilities were 0.80, 0.77, and 0.08, for grades 4, 8, and 12, respectively. These reliabilities were based on scoring across all response categories, rather than on correct/incorrect judgments.

Because of the data loss described earlier, the pilot evaluation was limited to responses from ten 12th-grade subjects whose responses were consistently dark. Seven items for each of the ten subjects were chosen as representative of the item types we wished to score. Table 2 shows the

---

Insert Table 2 about here

---

agreement (a) between human raters and (b) between human raters and the computer scoring algorithms, for the responses. The table shows that agreement between human graders was greater than between human and

machine scoring. Beyond that, the table depicts promising reliability for some other, while less than satisfactory accuracy for others. We note, however, that the failure of the automatic scoring system was not always caused by the scoring procedures. Disagreement between the algorithms and human graders was more often attributable to a loss of data in the image processing phase than an inadequacy of the procedures.

Table 3 illustrates that the accuracy of the scoring algorithms varied

---

Insert Table 3 about here

---

greatly from one subject to another. For some examinees, the scoring algorithms were quite accurate, while for other examinees the procedures performed poorly. For one examinee, the algorithms marked the seven test questions in perfect accord with the human scorings. For two examinees, only one question in six was marked accurately. The other examinees were fairly evenly distributed with intermediate numbers of items scored correctly. Again, much of the variation in accuracy was a function of the consistency and darkness of marking, rather than the scoring procedures, themselves.

Discussion

The central purpose of this research was to test the feasibility of automatic scoring of figural response items, delivered on paper. Our conclusion, based upon a limited number of items and subjects, is that the technology is indeed feasible. Efforts to score the three item types described in the paper were successful at a level of accuracy lower than the inter-rater agreement of human graders. However, we believe that human automatic scoring agreement could be raised substantially with refined procedures.

Given a better scanning mechanism, including relatively dark responses and the printing of items in inks that would be fully invisible to the scanner, it is plausible that the reliability of automatic scoring could approach or even surpass that of human scoring. We have further concluded that older scanning technologies (i.e. OMR) are probably not sufficiently fine-grained to score drawn responses of the kind described here. However, image processing offers more than enough resolution to read and record faithfully the most intricate of drawn responses.

Type II problems illustrate an important feature of figural response problems: It is often possible to make judgments about degrees of correctness, as in open-ended problems. Almost always, some, but not all of the arrowheads were placed correctly. In the plate tectonics problem, seven shafts were provided. Examinees could score between 0 and 7 arrows correct. We marked each response dichotomously correct or incorrect, assigning "correct" status only when all arrowheads were placed correctly. Partial credit could be awarded on Type II and other items; if this were done, it is likely tnat much more information could be obtained for each item than was gathered in this study.

Conclusion

To the extent that paper continues to be a widespread medium for testing, the technology described here might be used to interpret, categorize, troubleshoot, and perhaps offer feedback on responses to test items. Responses other than the ones examined here are possible. Simple drawings, for example, might be evaluated for the presence of specific features or for spatial relationships among other. Ongoing work involves extension of the scoring methods presented here to other kinds of figural responses.

The larger implication of this study is that multiple-choice need not dominate large-scale paper-based testing. Furthermore, the project makes it clear that innovations in test item types, though perhaps more common in computer environments, are not limited to computer delivery. Operational figural response, delivered via paper and answered with a pencil--but using technology to automate scoring--might measure knowledge and skills neglected by more traditional test item formats.

References

Braun, H. I., Bennett, R. E., Frye, D., Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27, 93-108.

Braswell, J. (1990, April). An alternative to multiple-choice testing in mathematics for high-volume examination programs. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Frederiksen, N. (1984). The real test bias. American Psychologist, 39, 193-202.

Haney, W., & Madaus, G. (1989). Searching for alternatives to multiple-choice testing. Educational Researcher, 18 (9), 27-32.

Kasturi, R., Fernandez, R. Amlani, M. L., & Feng, W. (1989, December). Map data processing in geographic information systems. Computer, 10-21.

Manning, W. H. (1987). Development of cloze-elide tests of English as a second language. ETS Research Report RR-87-18. Princeton, NJ: Educational Testing Service.

Martinez, M. E. (1990, April). A comparison of multiple-choice and constructed figural response items. Paper presented at the meeting of the American Educational Research Association, Boston.

National Assessment of Educational Progress (1988). Science objectives: 1990 assessment. Princeton, NJ: Educational Testing Service.

Nieberding, M. (1990). Image processing of health insurance claims. IMC Journal, 26 (2), 16-18.

Reid-Green, K. S. (1990). A high-speed image processing system. IMC Journal, 26 (2), 12-14.

Footnote

[1] Another methodology we considered, but did not use, is image subtraction. In this procedure, both item and response images are recorded, but the item image is removed (subtracted) electronically to leave the examinee's response.

Table 1

Specifications for Scanning Equipment

	MRC Scanner	Image Processing
Maximum resolution		
Pixels per square inch	48	90,000
Pixels per page	4488	8,415,000
Resolution adopted		
Pixels per square inch	---	625
Pixels per page	---	58,437
Channels between pixels	yes	no

Table 2

Percentage Agreement Between Human and Automatic Scoring of Figural Responses, by Item (N=10 subjects)

Item Name	Type	Percentage Agreement	
		Human/Human	Human/Automatic
Half-moon	I	80	40
Mitochondrion	I	100	90
Nucleus	I	80	80
Eclipse	I	100	40
Plate Tectonics	II	80	50
Blood Flow	II	80	60
Weight & String	III	90	30
Mean Percentage	---	87	56

Table 3

Distribution of Agreement Between Human and Automatic Scoring of Figural Responses, by Item (N=10 subjects)

Number of Items Agreeing	Number of Respondents
1	2
2	2
3	0
4	1
5	2
6	2
7	1

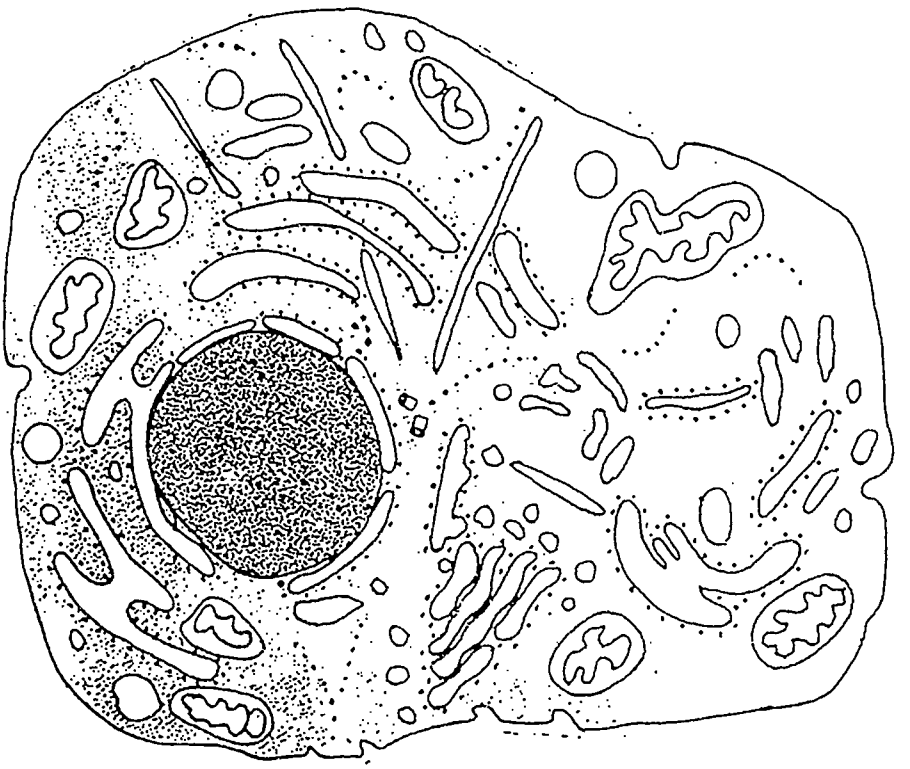

Figure 1. Example of a type I item. The question reads, "In the diagram of the cell shown below, mark an X on the part of the cell that produces most of the cell's energy as ATP."

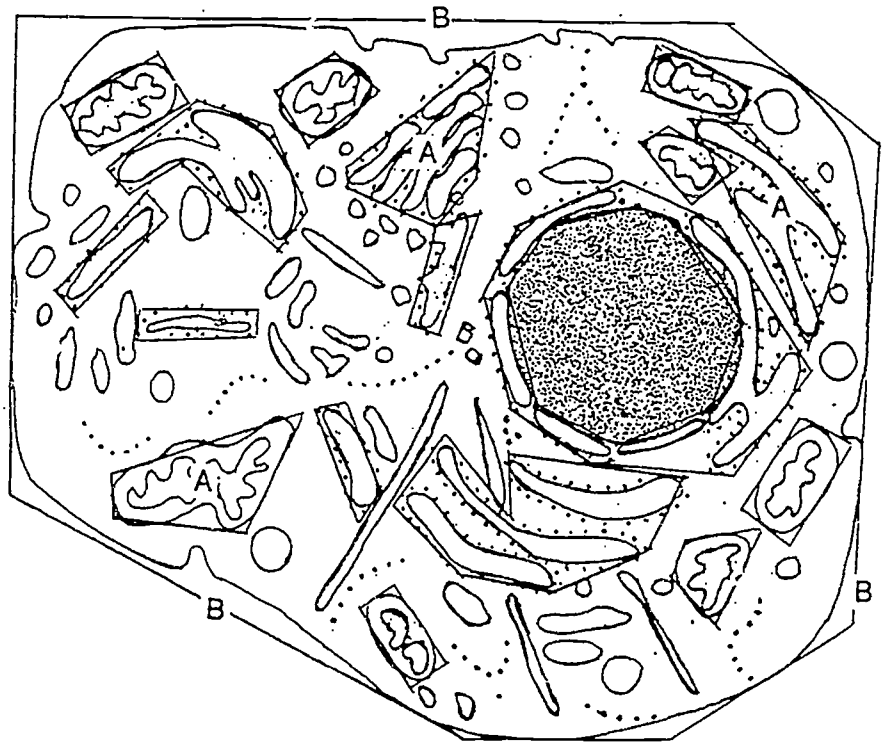

Figure 2. Sample demarcation of scoring zones on a type I item.

Note. "A" indicates examples of response fields defined by polygons. "B" indicates the "window of interest," in which all relevant marks are expected to fall.