DOCUMENT RESUME

ED 384 672 TM 023 993

AUTHOR Martinez, Michael E.; Jenkins, Jeffrey B.
TITLE Figural-Response Assessment: System Development and Pilot Research in Cell and Molecular Biology. GRE Board Professional Report No. 89-02P.
INSTITUTION Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-92-50
PUB DATE Jan 93
NOTE 35p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; *Cytology; *Educational Assessment; Higher Education; *Molecular Biology; Pilot Projects; Scoring; Spatial Ability; *Test Construction; Test Items; Test Use; Verbal Ability
IDENTIFIERS *Figural Response Items; *Graduate Record Examinations; Open Ended Questions

ABSTRACT

The purpose of the Graduate Record Examinations (GRE) figural-response project was to design a prototype computer assessment system for delivering and scoring figural-response items in the domain of cell and molecular biology and to begin to investigate properties of the item format. This report describes progress to date in an effort that is intended to be continuous and lead to program implementation. Features of the delivery system are described, with sample items, and ancillary developments, such as a tutorial, are also noted. Findings from a pilot research study are described. The essence of the pilot study was to examine the relationships between two item types (figural-response and open-ended verbal questions) and measures of figural and verbal ability for undergraduates (n=17) and graduate students and professors (n=4). The data hint that whereas verbal items draw from verbal ability, figural-response items draw from figural and verbal ability. The report concludes with a discussion of possible new directions for research, development, and eventual program use of the item format. Two tables present some pilot study findings. An appendix contains sample computer test screen items. (Contains 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 384 672

GRE®

RESEARCH

Figural-Response Assessment: System Development and Pilot Research in Cell and Molecular Biology

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Coley
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Michael E. Martinez
and
Jeffrey B. Jenkins

TM 023 993

January 1993

GRE Board Professional Report No. 89-02P
ETS Research Report 92-50

Educational Testing Service, Princeton, New Jersey

ERIC

Figural-Response Assessment: System Development and Pilot Research in Cell and Molecular Biology

Michael E. Martinez
and
Jeffrey B. Jenkins

GRE Board Report No. 89-02P

January 1993

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Copyright © 1993 by Educational Testing Service. All rights reserved.

Abstract

The purpose of the GRE figural-response project was to design a prototype assessment system for delivering and scoring figural-response items in the domain of cell and molecular biology, and to begin to investigate properties of the item format. This report describes progress to date in an effort that is intended to be continuous and lead to program implementation. We first describe features of the delivery system and give sample items; ancillary developments, such as a tutorial, are also noted. Then, findings from a pilot research study are described. The essence of the pilot study was to examine the relationships between two item types (figural-response and open-ended verbal questions) and measures of figural and verbal ability. The data hint that whereas verbal items draw from verbal ability, figural-response items draw from figural and verbal ability. The report concludes with a discussion of possible new directions for research, development, and eventual program use of the item format.

Introduction

As a practice and as an industry, testing is simultaneously being viewed with criticism for its shortcomings and eyed expectantly for its potential to improve education. Perhaps in part because of this attention, important new developments in testing have begun to emerge. Among the most prominent of these developments are behavioral anchoring (proficiency scaling), incomplete block sampling designs (Messick, Beaton, & Lord, 1983), testlets (Wainer & Kiely, 1987), and diagnostic models that are compatible with item response theory (Tatsuoka, 1990). Further removed from traditional large-scale testing are portfolio and performance assessment. Another line of development looks to computer delivery of tests and compatible new technologies, such as computer adaptive testing (Reckase, 1989). Yet another line of research deals with constructed response items.

Two of these developments—technology and constructed response items—play a role in the project described here. The figural-response item format, the focus of this study, is defined by two features: constructed responses and the expression of proficiency through the manipulation of figural (pictorial) material. Computer delivery, though used in this project, is not a required feature of figural-response assessment. Figural-response items present an examinee with a picture or diagram and ask the respondent to carry out some task on the figure. In the domain of biology, these tasks might include labeling particular structures (such as a cell nucleus) or assembling structures from components (such as an organic molecule from atoms). The range of items possible and their potential value to assessment is open and amenable to research.

Constructed responses are often viewed as desirable, in part because they appear to reflect some target competencies much better than do multiple-choice questions. There is evidence that constructed-response items elicit cognitive processes that are qualitatively distinct from the kinds of thinking tapped by multiple-choice questions (Snow, 1980; Martinez & Katz, manuscript submitted for publication). The figural aspect is also important: Educators have argued that the dominant symbolic modes of formal education, including assessment, are verbal and logico-mathematical (Gross, 1974; Shavelson, Webb, & Lehman, 1986). These modes do not capture all possible ways of knowing, and in certain, visually oriented fields, communication of ideas in verbal form can distort their most direct and natural representation and hinder problem solving (Larkin & Simon, 1987).

The potential applicability of figural-response items is likely to vary from domain to domain. The item type is especially suited to content areas that are highly visual or graphical, and the format may enable the assessment of knowledge that cannot be tapped by verbal or quantitative representations or by more static means of testing. Biology, because it is so visual, invites this form of assessment, but assessment in other subject areas, such as engineering, might also be enhanced by the inclusion of figural-response items. Even in fields that are not predominantly graphical, it seems likely that figure-based assessment could draw upon understandings that are tapped poorly or not at all by other assessment forms. One can imagine asking a student to place key events on a timeline to demonstrate an understanding of event precedence and causality in history.

For the researchers involved, the motivation behind this project was a belief that items calling for constructed responses within a figural medium fill a gap in assessment—and also in instruction. What remained to be seen was, given the self-imposed constraints on the item type, whether tasks generated would have at least a face validity and appear to add value to assessment when combined with more typical kinds of questions. Apart from many technical challenges, the potential research issues, revolving mostly around validity, are many and of great practical importance. Finally, technology was an important aspect of the project because automated scoring was presumed to be virtually a prerequisite for large-scale use of the item format.

Project Background and Overview

In its first instantiation, figural-response items were developed for the National Assessment of Educational Progress science assessment and printed on paper. From the beginning of the project, there was an interest in automated scoring. This technology was developed, but not to the point where it could be used with the reliability needed for program testing (Martinez, Ferris, Kraft, & Manning, 1992). When the current work was proposed, computer delivery of items was recommended for two main reasons. The first is that computers can collect the kinds of responses possible with paper and pencil, plus more (including assembly of structures from components). A second advantage is that some of the technical problems of paper-and-pencil scoring are no longer problems with the computer. A ready example is the problem of locating a response on the graphic. With paper-and-pencil delivery, this was not easy because variations in sheet feeding and paper imperfections and shrinking made the process less sure. Finding the location of a mouse click on a computer is trivial by comparison, as is determining the location of any object or the beginning and ending points of a line. A final reason for selecting computer delivery is that the GRE program was headed steadily in this direction for at least the verbal, quantitative, and analytical sections of the General Tests.

The project was primarily a development project; hence, development aspects of our work are emphasized in this report. This does not downplay the importance of future research; research is needed to shed light on the meaning of what is measured by any new item format. The report is organized around the products of development and the results of pilot research. The development portion is in a sense archival: The intent is to document the essential features of the delivery system. Another function is to provide something of a chronicle of our progress—even, and perhaps especially, our missteps. An understanding of our less-than-straight path might be of use to other researchers and reminders to ourselves of potential pitfalls in the development of a new assessment technology. Following a section on the path and products of development, research from a pilot study is presented. The pilot study focuses on one of a number of possible research perspectives, namely, connections between item format and aptitudes. The report concludes with a discussion of the potential contribution of figural-response assessment to GRE program testing.

System Development

Figural Response Authoring and Measurement Environment (FRAME)

The most significant product of the GRE Figural Response Project is a delivery system for figural-response items, which we call FRAME. The purpose of this section is to provide an overview of the most important features of the delivery system. In passing, it is worth noting that a functional delivery system was constructed fairly rapidly, within six months of the outset of the project. Refinements to the delivery system continued over the life of the project. These modifications were based on suggestions given by research subjects and professionals in the field of interface design. Feedback we have received makes us confident that the delivery vehicle is well designed and easy to use—even for someone who lacks significant experience with computers.

FRAME Version 2.0 presents a user with two types of displays: (a) a navigation screen, in which a list of items and their statuses are reported and (b) an item screen, which shows the item stem, the figure on which the response is made, and the tools needed for answering the item. Sample navigation and item screens are shown in Figures 1 and 2, respectively. All system input is made through a mouse. The only exception is when, as in the case of the pilot research, verbal (typed) responses are called for, in which case the keyboard is used. Incidentally, this illustrates that the figural-response delivery system can be used as a general assessment vehicle—for multiple-choice, verbal response, and figural-response questions.

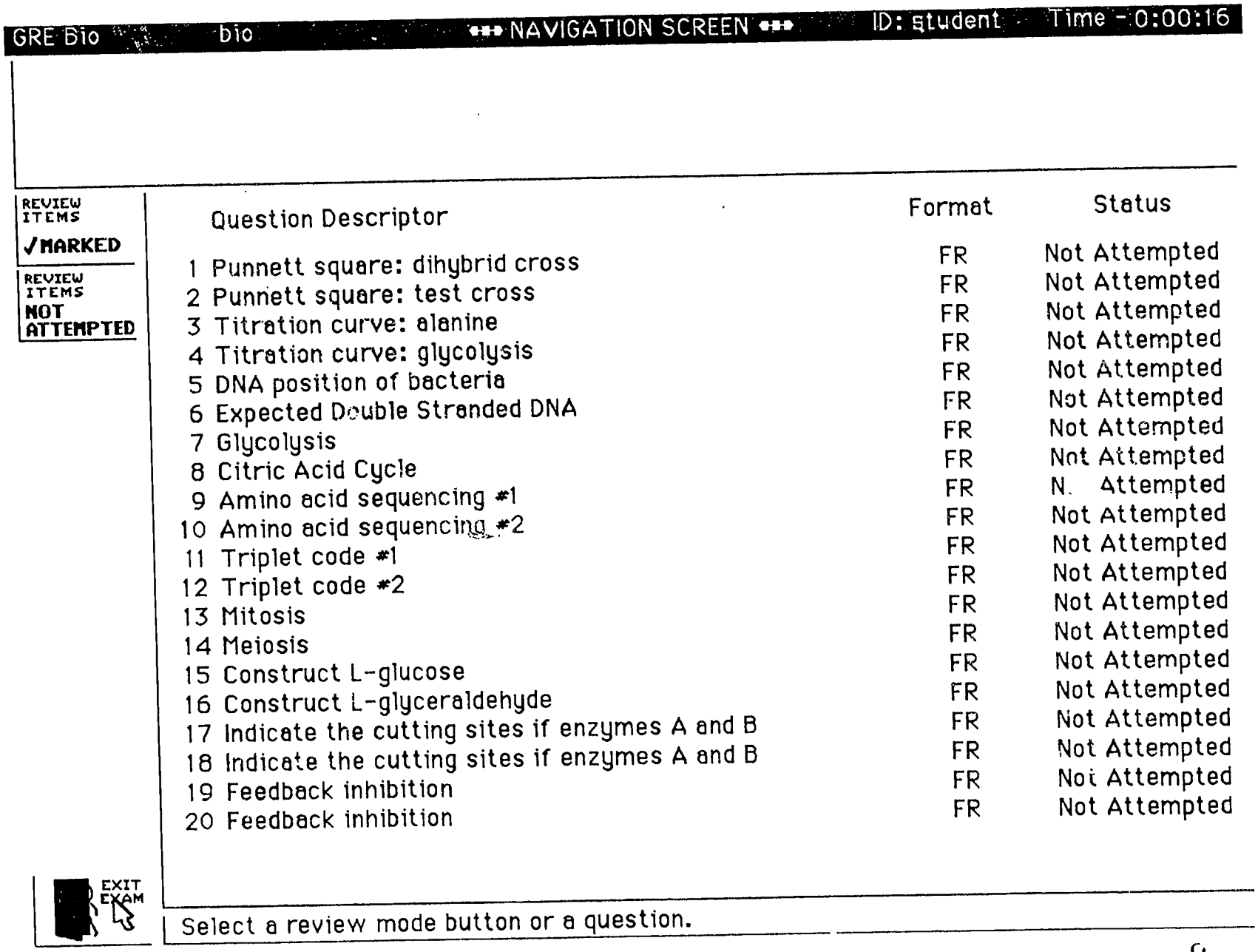
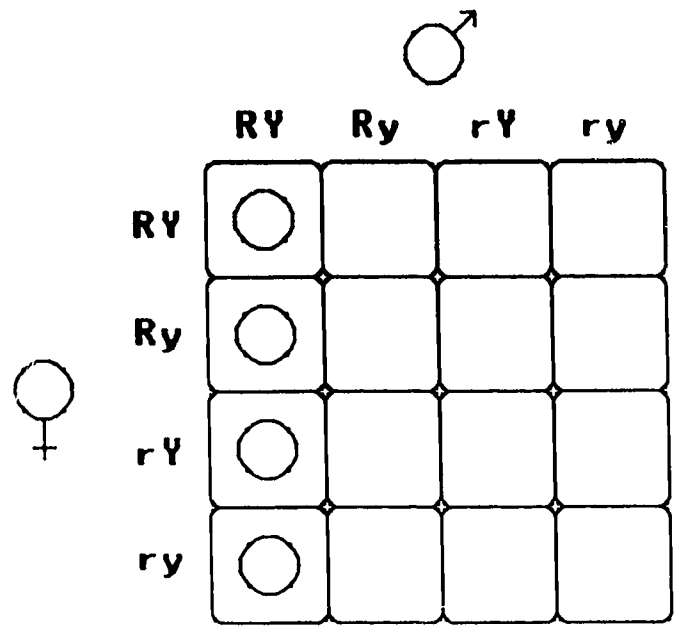


Figure 1. Sample navigation screen

Using pea plants, a dihybrid cross is performed that involves two independent alleles. Both parents are heterozygous for shape and color (genotype RrYy). Using the Punnett square and symbols provided, complete the expected phenotype of the F1 generation.



Pea Phenotypes

Yellow

Green

R = round seeds
r = wrinkled seeds
Y = yellow seeds
y = green seeds

ERASE

START OVER

MARK FOR REVIEW

PREV NEXT

NAVIGATE

To move an object, position the crosshairs on the object and click.

Figure 2. Sample item screen

Navigation screen. The navigation screen provides the user with a list of items that can be selected and attempted in any order. At the top is an administration line which lists the name of the test, the ID of the examinee, and the elapsed time. Below the administration line, and above the list of items is the stem area. When an item is selected, the verbal instructions to that item are displayed in the stem area.

The main window on the display shows, for each item, its number, a verbal descriptor for the item, the format of the item (such as FR for figural-response or MC for multiple-choice) and the status of the item (i.e., Attempted, Not Attempted, or Marked, which is explained below). If the number of items exceeds the display capacity of the screen, arrows for scrolling or paging are shown beneath the list of items. At the left are buttons that, when selected, automatically display only those items that have the characteristic shown on the button. For example, if the first button is chosen, only items that have the status Marked for Review will be displayed when items are viewed in sequence. Below the list of items is a context-sensitive help line that shows very simply and in general terms what the user is to do next, based upon the previous step. The button on the lower left, Exit Exam, permits the examinee to quit the test.

Item screen. At the top of the item screen display is an administration line that shows the name of the test, the name of the item, item number out of a total, the status of the item, the ID of the examinee, and the elapsed time. Below that is the item stem, where the verbal instructions to the item are given. The size of the box shown can accommodate five lines of text, which has been sufficient for all items we have created so far. Limitations on stem length are actually desirable because we wanted to keep the average response time per item to about 2-3 minutes. The stem area can accommodate long verbal instructions, if needed, and expand accordingly. The largest area of the screen we refer to as the work area. This contains a figure that is manipulated or modified. If objects are to be moved around the screen, these are usually placed on the right side of the work area for the sake of consistency. Figures are bit-mapped images stored in a file separate from the delivery system.

On the left-hand side of the screen is a row of buttons. The buttons at the top are tools used to respond to the item. The tools shown in Figure 2 are the Move Object and Erase tools; other tools are Draw Line (straight), Draw Line (free-form), Draw Arrow, Rotate, and Label. Only the tools needed to answer each question are provided with that item. At the onset of the project, we were not sure what set of tools we would ultimately have. This small set of tools is extremely flexible in the kinds of tasks it can facilitate.

The lower buttons handle administrative functions. For example, the Start Over button will redraw the current item, which is especially helpful if the examinee gets off to a bad start in answering the item. The Mark for Review button is useful if, after answering an item, the examinee wishes to return to that item at some later time. The new status, Marked, will then be shown in the column marked Status on the navigation screen. The Marked item is unmarked simply by clicking again on the button. Below that is a split button that will either advance to the next item or return to the previous item according to the order shown on the navigation screen. The last button, marked Navigate, returns the user to the navigation screen.

Software and hardware. The figural-response delivery vehicle was programmed in EASIS, a C-based programming language developed by ETS's Technology Research Group and used in building the National Council of Architectural Registration Boards (NCARB) simulations prototype and NCARB figural-response items. Because of the computational demands of scoring figural-response items real-time, Borland's C is being used to develop the scoring system. Object-oriented C++ is also being used to increase the efficiency of scoring and to improve the transportation of code between related projects. The hardware platform requirements consist of an IBM-compatible 286 microcomputer, a high-resolution VGA (640 x 480) graphics display, and a mouse. A 386-based micro is recommended. FRAME will take advantage of a math coprocessor if available.

Because the purpose of this document is to record the development process as well as the products, it is worth noting that the project inception was marked with several ambiguities regarding technical specifications. Most salient among these uncertainties was what hardware platform to use. To some this will seem obvious, but the rapid evolution of hardware made choosing a platform difficult. On one hand, certain levels of computational power and graphical resolution were essential for the tasks we wanted to pose. On the other hand, we feared that the platform might be either obsolete or out of production by the time program implementation took place.

The question of which operating system to use is another that has not yet been answered definitively. To this point, DOS has been used. The main problem here is that, normally, DOS does not allow access to more than 640K RAM. Many times during the project we reached this ceiling. Extending RAM or using a memory overlay scheme was itself problematic, and there were other concerns with switching to another operating system, OS2, which would have bypassed the 640K problems but could have introduced other difficulties. Even now, the use of object-oriented programming in scoring is being seriously considered. The point of these examples is to underscore the difficulties of working with technologies that evolve constantly, rapidly, and often unpredictably.

Figural Response Items

A second important product is a set of some 30 operational figural-response items in the domain of GRE cell/molecular biology (see Appendix for samples). The items were constructed by ETS test development staff who specialize in biology. The items were reviewed and revised iteratively before they were pilot tested.

Scoring

The development of the scoring system did not proceed as rapidly as we had hoped. This reflects an underestimation on our part of the magnitude of the work we were proposing. We have demonstrated scoring as a proof of concept with one item. In this, a Punnett square problem (Figure 2), peas that are either green or yellow and either round or wrinkled are placed into a 4 x 4 matrix to denote appearance of offspring, given parents of specified genetic identities. The program detected the shape and color of each placed pea and compared it with a record of the correct object in that cell. Thus, scores of between 0 and 16 were computed, depending on how many peas were placed correctly. In principle, this count could be used as a basis for partial-credit scoring on this item.

The problem with the scoring procedure is that it is "hard-coded." By this we mean that the programming needed to score the item is so bound to that particular item that it is unlikely to be transportable to others aside from fairly strict isomorphs. This seems unworkable for program testing, since development of each item using this method is likely to be time-consuming and expensive. Another problem with hard-coded scoring is that reporting is inflexible. In the case of the Punnett square items, a number correct is all that is given. A test developer might want to specify only one correct solution (the key) or a multitude of solutions that vary in value. Alternatively, many solutions might be equivalent with respect to some overall scale of proficiency, but would carry different diagnostic implications. Furthermore, the developer might want to expand or contract the number of these patterns over time. Also, the values and implications associated with these patterns might be better left independent of the pattern identification phase and modifiable in their own right. Our conclusion is that a more modular scoring system takes more work up front, but is potentially much easier to use and can provide a much richer report than is possible with a simpler but less flexible routine.

Feasibility requires rapid specification of scoring procedures using pre-formed algorithms that can be snapped together. Item construction should involve the same kind of modularity. Rapidity and ease of item construction and scoring can decide the difference between eventual program use or abandonment of the methodology. Besides making scoring components modular,

other practices might aid the item authoring process and contribute to feasibility. First, the use of object-oriented programming facilitates the re-use and sharing of code. We have begun to use an object-oriented approach to scoring items and will continue to do so. A second practice is the creation of variants (isomorphs) of existing items. Some of our items were actually variants of the same basic problem. The extent to which spinning off variants of items could be a standard procedure for item authoring is unknown but probably limited. More complex items, such as those one might find in a simulation, are probably more amenable to variation, but these items are likely to be hard coded at a profound level because they require significant domain knowledge.

Our goal then is modularity and flexibility in the design of a scoring system. The skeleton of a design that we think will meet these needs is as follows:

• Collect the raw response information (including temporal data) and "clean up" the examinee's response in order to provide to the scoring system a rich but interpretable protocol. Cleaning up means making low-level inferences about what the examinee meant and treating the responses accordingly. An example would be ignoring certain small, marginal stray marks.

• Describe the key features of an item. This might involve specifying which objects occupied particular response fields, or the angle, origin, and terminus of a drawn arrow. The routines used to describe key features would be separable and could be assembled independently according to the scoring requirements of the item.

• Compare features to a scoring rubric. This is essentially a pattern-matching step which in the limiting case would consist of a key pattern and all other responses. The judgment in this case would be dichotomous, as it is with virtually all multiple-choice items. A large number of patterns, presumably diagnostic, could be specified (but need not be). Decisions on the complexity of the interpretation would depend on the purpose of the assessment and the time resources available for item construction.

• Evaluate the examinee's performance quantitatively or qualitatively. Quantitative evaluation might mean the determination of a partial-credit score; qualitative evaluation might be a prescription for remediation. This information could lead naturally to placement and remediation recommendations.

The description (pattern matching) step deserves some elaboration. One important function of the scoring system is the ability to determine the key features of any given solution. In order to allow the greatest flexibility in scoring different types of items, an interpretive language is being developed that will be used by the item developer to direct the scoring system toward relevant features and processing of those features. This capability will enable the item developer to specify the scoring procedures without requiring a programmer to develop a special scoring routine for each item type.

Features of the response that the scoring system could look for include the following:

• Does Response Field 1 contain all vertices of Object C? (Response fields and objects are associated with polygons fitted to their perimeters.)

• Is Object B in the correct orientation (e.g., 90 degrees)?

• Is Object A adjacent to Object B (ascertained by determining if Objects A and B are located in adjacent polygons)?

• Do Objects A and B overlap?

• Do Lines A and B cross ?

• How long is Line A?

- Where is Line A's origin/terminus?

- What label, if any, is in Response Field 1?

Tentatively, we would like the language to be able to specify logical operations (i.e., NOT, AND, OR). For example: Is Object A in response field C AND is Object A is rotated 90°. The language must also be able to interpret the amount of tolerance to determine a correct/incorrect response and tolerances must be specifiable and adjustable by the item developer. The basic structure of this interpretive language has been developed, as have some of the fundamental routines needed, such as INPOLY, the routine used to determine if an object is located within a response field.

Tutorial

Two ancillary project developments are worth noting. One is an on-line tutorial, mentioned above, which was used in the pilot test. The tutorial uses FRAME as the delivery vehicle. In the tutorial, the screen layout is described (Figure 3). Practice tasks, answered by using the appropriate system tool (Move Object, Draw Line, etc.), are given to subjects (Figure 4 for an example). This tutorial could easily be adapted for further research, and a later version might be used as a prelude to program testing.

In our experience, only a minimal exposure of about 10 minutes was needed to prepare the examinees for attempting the items. Our tendency was to elaborate the tutorial to the point where it took too much time without any corresponding increase in readiness for assessment. Controlling the mouse seemed as hard as mastering how to use the system. Some novices tend to orient the mouse the wrong way or to get stuck when the mouse reaches the end of the mouse pad. The tutorial gave explicit advice on these functions as well as practice tasks for developing control of the mouse. The problems did not persist for any subject. With mice being common on more hardware platforms, it is likely that this difficulty will become less important with time.

Human Scoring Utility

Even with fully automated scoring, some human scoring will probably be needed. This is so because with problems of any complexity, it is virtually impossible for automated scoring procedures to characterize 100% of the responses. Some small fraction of responses will not match any prespecified patterns or will contain an unusual array of features and will have to be examined by an expert grader. Even in routine human scoring, pre-specified rubrics have to be reworked while scoring is in progress because the original rubric fails to describe the universe of responses (M. Pearlman, personal communication). Another reason for human involvement is that automated scores will have to be validated and quality control maintained. This does not mean that human scoring is always the ultimate standard (human scoring is fallible), but it is one important way of examining the quality of the automated procedures.

To facilitate human grading, we have developed software that displays subjects' responses and that cues the user for a categorization of those responses. Another grader can later view the same answers and assign scores independently. Both sets of scores are recorded in a database. The utility then separates the items according to the difference between raters' scores and displays only those items for which the difference between the scores exceeds some specifiable criterion. A third score can be assigned and the data aggregated in a standard way (e.g., by eliminating the most discrepant score and averaging the other two). The utility can also accept the products of automated scoring as one set of scores.

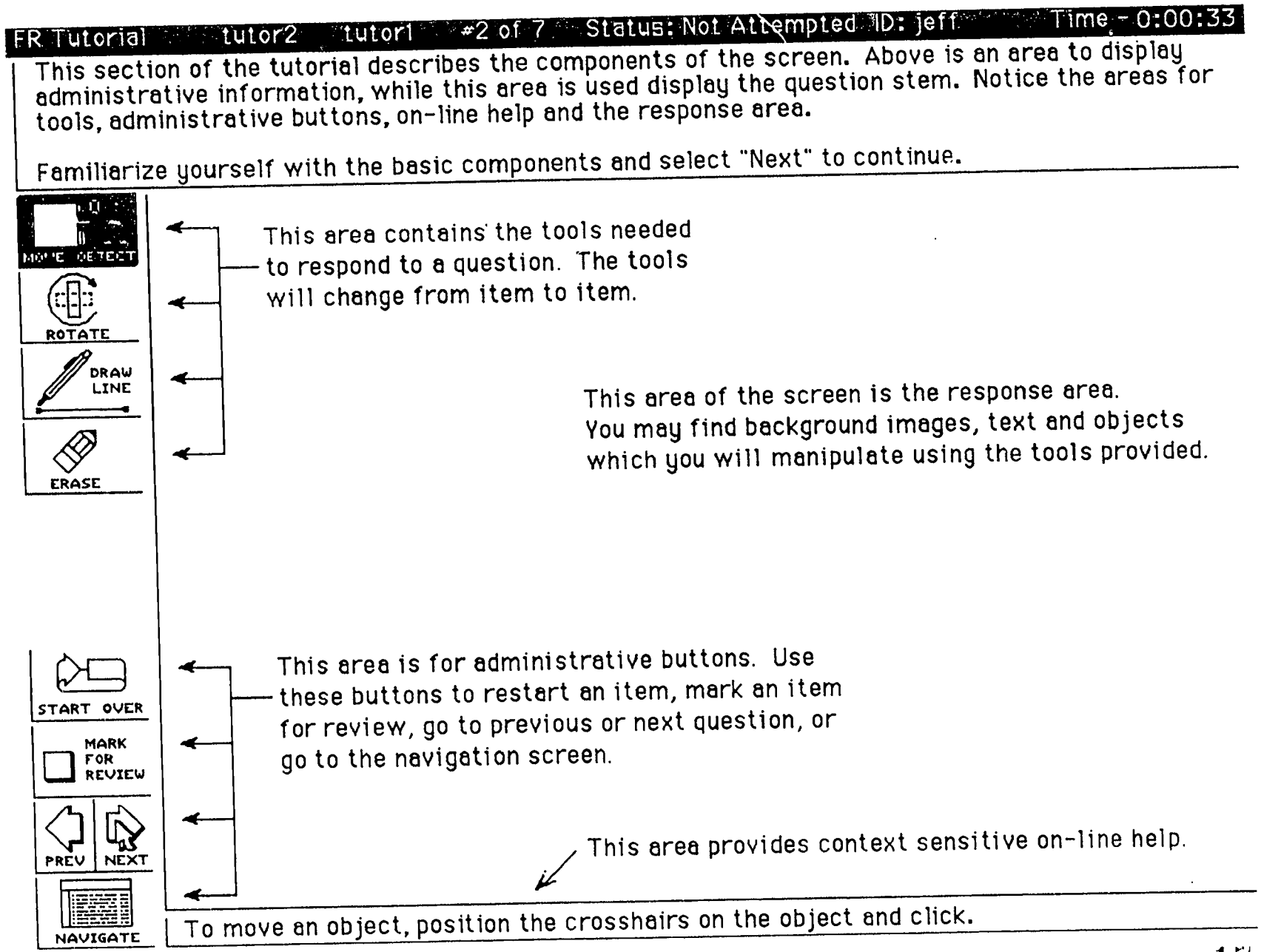


Figure 3. Sample tutorial screen

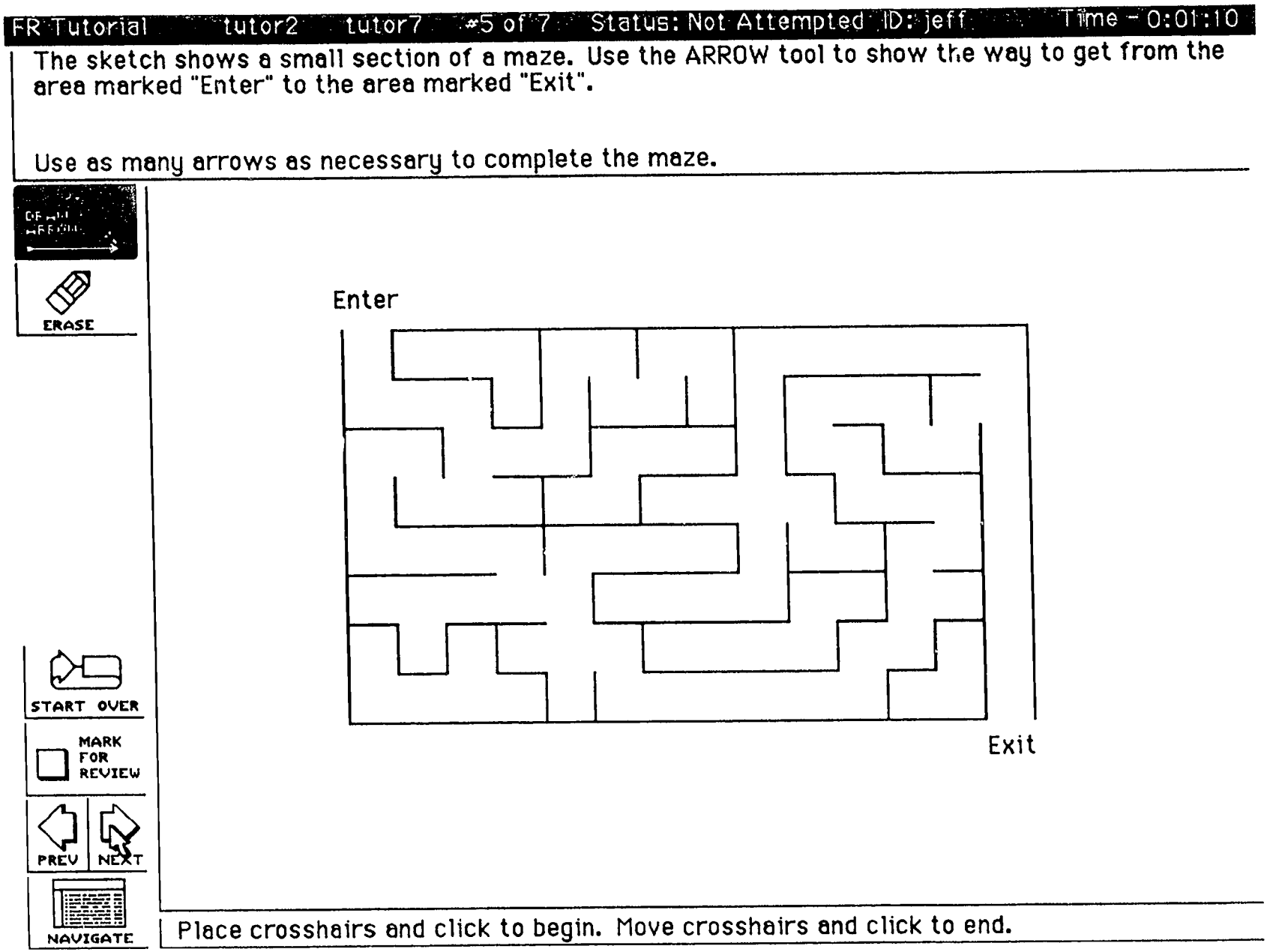


Figure 4. Sample practice task from tutorial

Pilot Research

The research questions that are germane to an experimental item type are numerous. For figural-response items, important research questions include:

(a) How does the constructed response/multiple-choice distinction influence the nature of what is measured?

(b) How does the technology affect the expression of proficiency?

(c) Can a process model be developed that will account for the solution steps of examinees in solving figural-response problems or subsets of them?

(d) To what extent does the ability to manipulate iconic and verbal symbols, and to translate between them, affect performance on figural-response items?

(e) Does performance on the format bear any relationship to certain kinds of mental abilities, such as figural memory?

The last question was the focus of the pilot study. It was chosen because observers frequently infer that figural-response questions are related to visual or figural ability of some kind. The question is empirical and seemed important enough to address at least on a small scale.

The chief research paradigm followed, that of a faceted test, was proposed by Guttman (1969) and more recently by Snow and his colleagues (Snow & Lohman, 1989; Snow & Peterson, 1985). The faceted test has experimental manipulations built into the test itself. Research on NCARB architectural items emphasized the constructed response/multiple-choice distinction (Martinez & Katz, manuscript submitted for publication); GRE pilot research highlighted symbol system differences, namely, the verbal/figural distinction in response mode. First, we sought to understand whether there was an aptitude-treatment interaction (ATI) of item format on mental ability (Cronbach & Snow, 1977; Snow & Lohman, 1989). Specifically, we tested the hypotheses that figural-response items would have a stronger relationship to figural ability than to verbal ability, and that verbal constructed-response items would have a stronger relationship with verbal aptitude. Second, we sought to determine whether the figural items or the verbal items would best separate the expert-novice status groups.

Method

Subjects. Three subjects were excluded from analysis because at least one of their scores was less than 1, and it was felt that there was either a motivational problem or a poor match between test instrument and ability. Data were analyzed from 24 subjects who formed three groups. The groups were (a) undergraduates, particularly sophomores, who had taken (or were taking) their first collegiate biology course (N=10); (b) undergraduates who had taken more than one college course (N=7); and (c) graduate students and professors of biology (N=4). We sought a range of ability in biology because greater variance would lend power to tests of our research questions. A subject pool limited to typical candidates for GRE Subject Tests would have restricted the range.

Procedure. Subjects began by taking a battery of paper-and-pencil aptitude tests in a group session. Ability measures were taken from the ETS Factor Kit (Ekstrom, French, & Harman, 1976). Scores from two of the measures, Advanced Vocabulary Test II and Completing Sentences, were aggregated to from a verbal aptitude score. The vocabulary test was chosen because vocabulary is generally a standard of verbal ability. Completing Sentences is a constructed response task, and so resembled the test items in that regard. Scores from two other measures, Form Board Test and Building Memory, were summed to form a figural score. The Form Board Test is a typical measure of visual-spatial ability; Building Memory was chosen because it seemed that proficiency in answering figural-response questions might depend on one's ability to

remember visual images, as is required in the Building Memory task.

Of the 30 or so figural-response items developed, 20 were selected for field testing. Twenty more counterpart items were created, and these required short verbal responses on the order of one to five sentences, which had to be typed in. The counterparts were not stem-equivalent items; since all subjects took each item, stem-equivalent pairs would have generated carry-over effects. Counterparts were items that tapped essentially the same concepts. For example, one figural question asked subjects to fill in a Punnett square by indicating the offspring of two heterogeneous pea plants (Figure 2). Its verbal counterpart was, "How is an autosomal (non-sex-linked) trait expressed?" Both questions involve understanding the nature of a recessive gene.

In individual sessions on a computer, subjects were shown figural-response questions and open-ended verbal questions. According to the faceted design, each subject attempted 10 each of the figural items and the verbal items. Both sets of items were scored according to scoring rubrics such that each subject had a maximum score of 10 on each item format, figural and verbal. Mean scores on figural and verbal formats were compared across different expert/novice status groups. Format scores were also compared with scores on figural and verbal aptitude measures. The most straightforward expectation was that proficiency in answering figural-response questions would have a strong relationship with "figural" aptitude, as operationalized by an aggregate score from aptitude tests. Likewise, a strong relationship between verbal performance and verbal aptitude was expected. Subjects were separated into high- and low-aptitude groups according to their ability relative to the sample medians. Mean scores across item formats were compared for these groups.

Results

Prediction of status groups. Table 1 shows the mean scores and standard deviations for three status groups. A simple analysis of variance was computed for each of two dependent variables: total figural-response scores and total verbal scores. The sample sizes are small even for an analysis of variance, so the findings should be regarded as suggestive.

Table 1
Simple ANOVA on Figural Response and Verbal Formats by Status Group

Status Group	N	Mean	S.D.	F	p
			Figural Response		
Undergrad; one biology course	10	1.77	0.80		
Undergrad; >one biology course	7	2.39	1.40	41.38	0.0000
Graduate or professor	4	7.23	0.79		
			Verbal Response		
Undergrad; one biology course	10	4.68	1.38		
Undergrad; >one biology course	7	5.82	1.63	9.03	0.0019
Graduate or professor	4	8.25	1.10		

Means for both measures, figural and verbal, showed a good spread in performance across the status groups, a difference in distributions that is statistically significant. An intriguing pattern is that the figural-response measure separated the status group distributions especially well. This is true of group means and, for each status group, the standard deviations are smaller on the figural-response measure than in the corresponding verbal measure. Separation of means and narrower distributions make for a rather large F value for figural-response (41.38) and a probability of Type I error less than .0001. If figural-response items are good indicators for separating status groups, a reasonable hypothesis is that they would also be effective predictors of successful graduate school study and professional achievement. Validation of such a claim would, of course, require additional research.

Relationship to aptitude measures. The relationships between verbal and figural aptitudes and the item formats are summarized in Table 2. Simple patterns of relationships were not found among the pilot data—there was no unmistakable interaction between performance on the figural-response items and figural aptitude, as derived from the figural tests of mental ability. The most salient pattern is that scores on the figural-response items were related to both figural and verbal ability, whereas verbal item scores were related to verbal ability only. This suggests an aptitude-treatment interaction (ATI), where the treatments consist of different item formats (Cronbach & Snow, 1977). The data hint that the figural-response format draws from different abilities, whereas verbal (written) responses do not. Written responses to our questions apparently drew from verbal ability, but not from figural ability.

Table 2
Means for Figural and Verbal Items, by Aptitude Group

	Figural Ability				t	d.f.
	Low (N=12)		High (N=12)			
	M	SD	M	SD		
Figural Response Score	1.61	(0.89)	3.93	(2.70)	-2.82*	13.4
Verbal Response Score	4.96	(1.22)	5.88	(2.77)	-1.05	15.1
	Verbal Ability				t	d.f.
	Low (N=12)		High (N=12)			
	M	SD	M	SD		
Figural Response Score	1.68	(0.61)	3.86	(2.81)	-2.61*	12.0
Verbal Response Score	4.44	(1.87)	6.40	(2.01)	-2.47*	21.9

* $p<.05$

Explanations for these interactions are far from clear, but some hypotheses are tenable. One is that the figural-response item format requires cognitive interplay between figural and verbal symbol systems. The stem (or problem instructions) are presented verbally, but often figural perception is required before the goals and rules of the problem are understood. This appears to require at least some matching between entities and relationships presented in both codes. A second possible explanation is that much of the cognitive processing of figural items is actually mediated verbally. Counterintuitively, the presence of pictorial representations might actually aid those who are relatively low in visual/pictorial kinds of ability (Cronbach & Snow, 1977). Also, the relationship between figural aptitude and the figural-response format may not be linear. Several of the undergraduate subjects had relatively high figural aptitude scores but low figural-response totals. Some sort of figural ability may be necessary, but not sufficient, to encode and express figural understandings of a domain.

Discussion and Conclusions

The figural-response delivery vehicle (FRAME) is an accessible platform for administering figural-response items and other item types, including multiple-choice items. A scoring system is being designed and built in a way that will allow a test developer to specify scoring parameters rapidly, just as items can be assembled rapidly now. The scoring system will be based on pattern recognition so diagnostic evaluation will be possible. The patterns identified in responses will lead to quantitative or qualitative valuations, which might include partial credit scoring or suggestions for remediation.

The pilot test findings, though tentative, are intriguing. Figural response items were better able than open-ended verbal response items to distinguish between experts and novices. If replicated, this discrimination might serve well the first-order requirement of GRE testing, namely, to predict which candidates will be the most successful graduate students. A second finding is that there was no simple correspondence between performance on the figural-response items and figural aptitude. Figural response scores were related to both figural and verbal aptitudes. Verbal scores were more clearly related to verbal aptitude alone. Possible explanations were given, including a mechanism in which figural-response items might draw from both figural and verbal aptitudes.

The delivery system will require periodic modification, but the most important direction for new work is on scoring. This work has begun and its conceptual base has been largely worked out. A second essential area for research is on the psychometric modeling of response data. The best psychometric model will draw from a cognitive characterization of examinee responses, and will allow those responses to be weighted along a continuum rather than assigned dichotomous scores. Some sort of partial credit scheme is essential, especially as items become more complex and time-consuming. From the point of view of GRE program direction, it is also crucial to determine the domain in which development is best focused. One promising area is engineering, which is likely to be undergoing some changes in the not-too-distant future. Figural response items seem to be a natural supplement to other item types that might be used in engineering. Assembly and error-detection items might be especially germane to proficiency in engineering.

Validity issues underlie the research and development effort. Ultimately, one must know what is being measured in a test or a test item. For measures like figural-response, it is important to know how what is measured differs from the normal test methodology, and how the two might complement each other. A variety of methods are useful here, including protocol analysis for cognitive modeling and faceted test design to relate measures to human abilities, as reported in the pilot study. Prediction of first-year graduate GPA is an important criterion for examining validity—but not the only one.

The GRE program has for some time committed to using technology in its program testing. Alternative item types like figural-response can capitalize on the power and flexibility of computers to offer forms of test authoring, administration, scoring, and reporting that cannot be achieved with paper and pencil. The figural-response research and development effort is intended to be open-ended in its ultimate form. Possible directions include the use of photographic and dynamic images, response data in which time (latency and order) plays a role, and detailed diagnostic reporting to students, teachers, and institutions. The more near-term goal, however, is the development of a robust authoring, testing, and scoring system that can be used on a large-scale basis.

References

Cronbach, L. J., & Snow, R. E. (1977). Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Gross, L. (1974). Modes of communication and the acquisition of symbolic competence. In D. R. Olson (Ed.), Media and symbols: The forms of expression, communication, and education (Seventy-Third Yearbook of the National Society for the Study of Education). Chicago: University of Chicago Press.

Guttman, L. (1969). Integration of test design and analysis. Proceedings of the 1969 invitational conference on testing problems. Princeton, NJ: Educational Testing Service.

Larkin, J. H. & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11, 65-99.

Martinez, M. E. (1992). An initial taxonomy of item types for demonstration of proficiency in a figural medium. Proceedings of the 33rd Annual Conference of the Military Testing Association.

Martinez, M. E. (1991). A comparison of multiple-choice and constructed figural response items. Journal of Educational Measurement, 28, 131-145.

Martinez, M. E., Ferris, J. J., Kraft, W., & Manning, W. H. (1992). Automated scoring of paper-and-pencil figural responses. Journal of Educational Technology Systems, 20, 251-260.

Martinez, M. E., & Katz, I. R. (submitted for publication). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture.

Messick, S., Beaton, A., & Lord, F. (1983). A new design for a new era (NAEP Report 83-1). Princeton, NJ: Educational Testing Service.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. Educational Measurement: Issues and Practice. 8(3), 11-15.

Shavelson, R. J., Webb, N. M., & Lehman, P. (1986). The role of symbol systems in problem-solving: A literature review (CSE Report No. 269). Los Angeles: UCLA Center for the Study of Evaluation.

Snow, R. E. (1980). Aptitude processes. In R. E. Snow, P. A. Federico, & W. E. Montague (Eds.), Aptitude, learning, and instruction, Volume 1: Cognitive process analyses of aptitude (pp. 27-63). Hillsdale, NJ: Erlbaum .

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), Educational Measurement (3rd ed.). New York: Macmillan.

Snow, R. E., & Peterson, P. L. (1985). Cognitive analyses of tests: Implications for redesign. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error analysis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition (pp. 453-488). Hillsdale, NJ: Erlbaum.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Appendix

Sample Items

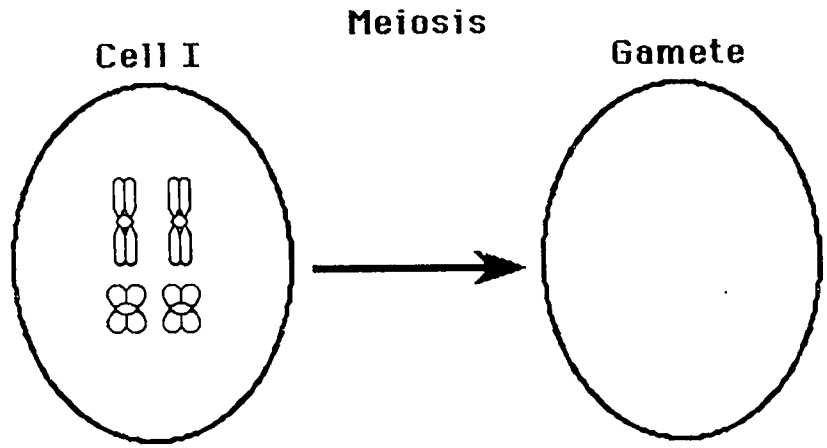
GRE Bio bio frceldv2 #14 of 20 Status: Not Attempted ID: student Time - 0:01:00

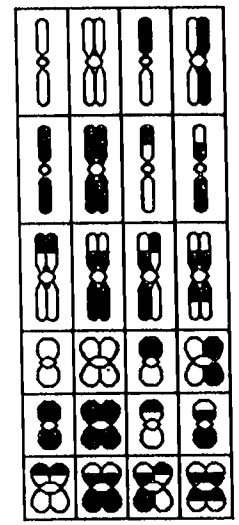
Cell I has the normal diploid chromosome complement of an organism.

Using the chromosomes on the right, show the chromosome complement of one of the expected gametes from this organism. Assume there is no crossover involved.

MOVE OBJECT

ERASE





Each of the chromosomes can be used once, more than once, or not at all.

START OVER

MARK FOR REVIEW

PREV NEXT

NAVIGATE

To move an object, position the crosshairs on the object and click.

GRE Bio bio frtransl #11 of 20 Status: Not Attempted ID: student Time - 0:03:31

The table below depicts the genetic code used during translation, in which sets of three codons in a m-RNA molecule specify the amino acid sequences in the course of protein synthesis.

Complete the translation of the m-RNA segment given the position of alanine.

MOVE DETECT

ERASE

1st Position (5' end)	2nd Position U	2nd Position C	2nd Position A	2nd Position G	3rd Position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Glu	A
	Val	Ala	Glu	Gly	G

Ala Leu Arg Lys Asn Met Asp Phe Cys Pro Gln Ser Glu Thr Gly Trp His Tyr Ile Val

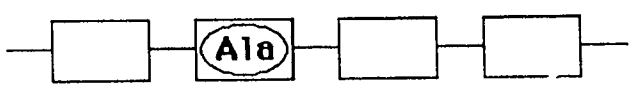
START OVER

MARK FOR REVIEW

PREV NEXT

NAVIGATE

5'-C-U-C-A-G-C-G-U-U-A-C-C-A-U- 3'



To move an object, position the crosshairs on the object and click.

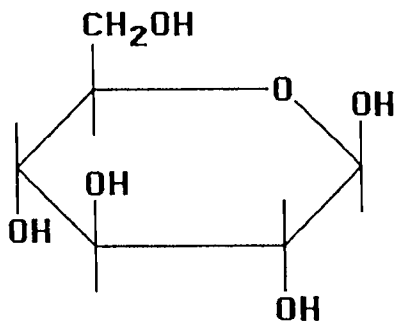
GRE Bio bio frdandll #15 of 20 Status: Not Attempted ID: student Time - 0:04:09

Given the D-glucose below, construct its L-glucose stereoisomer using the template shown.

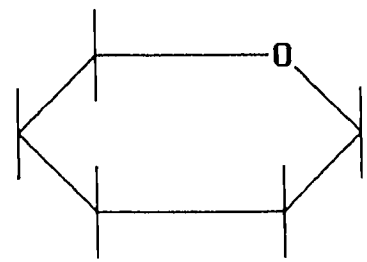
MOVE DETECT

ERASE

D-glucose



L-glucose



CH_2OH

OH

START OVER

MARK FOR REVIEW

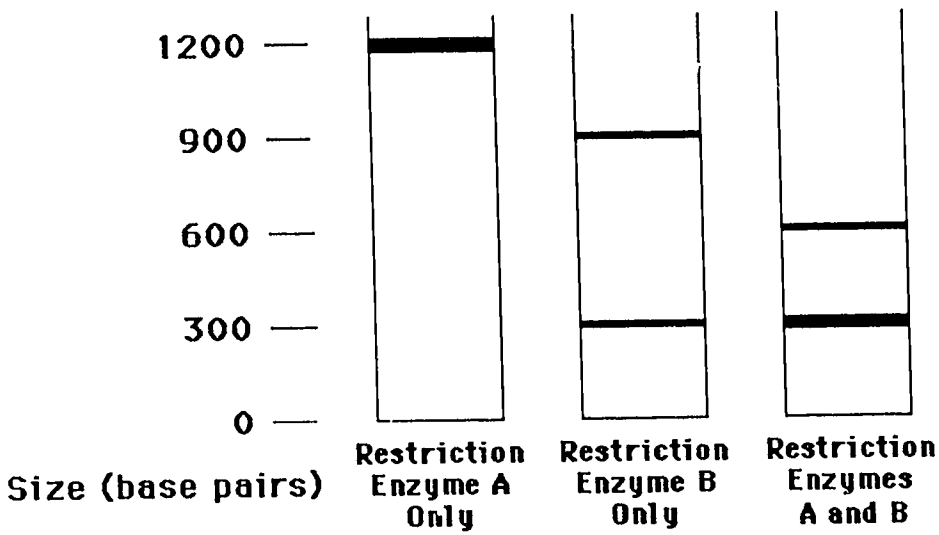
PREV NEXT

NAVIGATE

To move an object, position the crosshairs on the object and click.

The following electrophoretic patterns were obtained following digestion of a plasmid with restriction enzyme A, restriction enzyme B and restriction enzymes A and B together.

Construct a plasmid showing the cutting sites of the two restriction enzymes together.



Plasmid
(1200 base pairs)

Cutting Site of A A

Cutting Site of B B

MOVE DETECT

ERASE

START OVER

MARK FOR REVIEW

PREV NEXT

NAVIGATE

To move an object, position the crosshairs on the object and click.

54020 08913 • Y13M 7 • 209076