DOCUMENT RESUME

ED 224 831 TM 830 025

AUTHOR Choppin, Bruce
TITLE Latent Trait Models for Answer-Until-Correct Tests. Methodology Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 82
GRANT NIE-G-80-0112
NOTE 55p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; *Computer Assisted Testing; Computer Programs; *Educational Testing; *Guessing (Tests); *Latent Trait Theory; Measurement Techniques; *Multiple Choice Tests; Research Methodology; Test Items
IDENTIFIERS *Answer Until Correct; Rasch Model

ABSTRACT

The answer-until-correct procedure has made comparatively little impact on the field of educational testing due to the absence of a sound theoretical base for turning the response data into measures. Three new latent trait models are described. They differ in their complexity, though each is designed to yield a single parameter to measure student achievement. The simplest, a "partial credit" model, has a single difficulty parameter for each item. This model takes no account of the variations in distractor attractiveness from item to item, nor of which distractors were actually selected by the respondent. The second model treats the test as a sequence of distinct steps, each of which has a difficulty parameter. This method does not assume that all items have the same logical structure with regard to difficulty. It takes no account of which distractors are selected. The third model is an extension of the second. In this model, the step difficulty values for an item vary in terms of which distractors were previously selected. A technical manual describing software developed for an effective and efficient program for administering answer-until-correct tests using microcomputer systems is reported as Appendix 1. (Author/PN)

Deliverable - November 1982

METHODOLOGY PROJECT

LATENT TRAIT MODELS FOR ANSWER-UNTIL-CORRECT TESTS

Bruce Choppin
Study Director

Grant Number
NIE-G-80-0112, P3

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
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 NIE position or policy.

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

1. Introduction

Though they are convenient to use and have some desirable psychometric properties, multiple choice tests have been very widely attacked. Three specific criticisms that have been made against conventional multiple choice tests are:

1) They they face the testee with three or four times as many incorrect statements as correct ones and provide no feedback to help the student learn the correct answers.

2) That they encourage random guessing.

3) That they are inefficient and that little information is gained about the student from his response to a single item.

The "answer-until-correct" testing mode (Brown, 1965; Hanna, 1975) is designed to overcome these problems. In this mode the student is presented with instant feedback to a response. If the response is correct, the student is directed to continue to the next question, but if the response is incorrect he or she is asked to attempt the item again. This form of testing has the advantage of extracting significantly more information about a student's ability fro a given number of items, and thus makes it easier to distinguish between different levels of partial knowledge or part mastery. It has also been suggested that this response mode reduces the incidence of

random guessing behavior among students, and has the additional benefit that (most of the time) the final answer chosen by the student to an item is also the correct one. There is, a priori, reason to believe that this response, the one that receives positive reinforcement, is the one most likely to be remembered.

A number of research studies have focused on the characteristics and usefulness of answer-until-correct testing. For example, Merwin (1959), Brown (1965) and Frary (1980) investigated various scoring procedures. None of the more complex alternatives they tried appeared to improve significantly on Brown's simple approach of reducing the total score by one point for every incorrect distractor selected. Hanna (1975), and Kane & Moloney (1978), investigated the implications of AUC responding for reliability and validity. Hanna suggested that the AUC procedure increased reliability but generally appeared to decrease validity as measured by correlation with a substantive external criterion. The implication is that testwiseness may play a more significant role in AUC tests than on conventional tests. This relates back to Merwin's earlier paper in which he concluded that if test constructors were to reap advantages from the AUC procedure, then item distractors would have to be carefully designed so as to relate in a clear way to the criterion variable.

Much of the earlier work displayed considerable vagueness as to the presumed behavior of the student when taking a test.

A careful reading and analysis of the logic presented suggests that the writers were assuming the relevance of one or the other of two contrasting and incompatible models. The first, which may be

called the partial knowledge model, assumes that the student may know enough about the subject matter with which the item is concerned in order to be able to eliminate one or more of the distractors with some certainty. He is then presumed to guess at random among those that remain. Complete master of the problem involves the certain elimination of all but one of the alternative responses so that the student chooses the correct answer without guessing.

The second model assumes that a student arrives at an incorrect response not through some guessing procedure, but through the application of misinformation. Under the answer-until-correct procedure, such a student having applied his misinformation to obtain the wrong answer, is forced to choose again. The feedback that the first piece of misinformation is incorrect may be important incidental learning. The next choice may be a random guess, or another response selected on the basis of misinformation.

Frary showed that the AUC procedure was effective in discriminating between students when they operated on the basis of partial information, but suggested that the scoring procedure could be improved for students operating the misinformation model. Wilcox (1982) further considers the distinction between the partial knowledge and misinformation models and appropriate rules for scoring tests when the latter operates. Unfortunately, it would appear that in practice many individuals use both strategies when taking tests, and it is difficult to tell when looking at the pattern of results on which items they were employing partial knowledge and on which

misinformation. Questioning students following the administration of an AUC test could help to clarify this issue.

The answer-until-correct procedure has made comparatively little impact on the field of educational testing in the seventeen years since Brown's paper for two reasons:

(a) the lack of convenient and appropriate technology for providing instant feedback to the student, since clinical administration of tests is prohibitively expensive; and

(b) the absence of a sound theoretical base for turning the data into measures, for while Brown's system appears to work in practice, there is no model to substantiate it or check its validity.

On the first issue, there have been a number of recent developments. Answer-until-correct tests currently in use (on an experimental or regular basis) use one of three different feedback technologies. The first approach requires an answer sheet preprinted in invisible ink, so that when the student responds (using a special pen) a portion of the preprinted material becomes visible, and the student obtains the appropriate feedback. The second method involves having the student erase a shield printed over the top of the feedback information again on a specially prepared answer sheet. Each of these approaches requires some special equipment for preparing the answer sheets which have to be customized to fit a particular test. However, this equipment is now fairly generally available, and the answer sheets produced from it are not unduly expensive.

The third approach involves testing by the computer. This method is potentially superior to the other methods because it allows the

recording of the sequence in which particular responses are chosen. The first two methods described allow only the inference that the correct response was chosen last, but do not easily allow the earlier incorrect responses to be ordered. Until very recently the computer was far too expensive to be considered seriously as a test administering device, but the rapid development of terminals and in particular of inexpensive micro processers opens up new possibilities.

The computer is able not only to record the sequence in which distractors are selected, but also to accumulate other information (e.g., how long was the delay between each response), and continually update estimates of the student's level of performance and the measurement precision. It is also able to provide more or less detailed feedback under the control of the test constructor, and to provide the feedback in an entirely standard fashion so that no inadvertant clues are presented. During the last year, the CSE team has devoted considerable effort to developing an effective and efficient program for administering answer-until-correct tests using Apple microcomputer systems. We have designed this system so as to be useful to teachers who currently have access to Apple or similar computers. The system has also been valuable in collecting answer-until-correct data for use in our psychometric research, and it records on disk, in a standard format, considerable information about the students' attempts at the test including his or her expressed confidence in each of the initial responses to each item.

The technical manual describing the software we have developed to accomplish this is attached to the present report as Appendix 1. A

somewhat simplified description designed to be used as a teacher's manual is currently in preparation.

The rest of this paper will be devoted to describing the latent trait models which address the second of the problems mentioned earlier, the absence of a sound theoretical base for turning the response data into a measure.

2. Latent Trait Models

Three new latent trait models will be described in the remainder of this paper. They differ from one another in their complexity, though each is designed to yield a single parameter to measure student achievement.

The simplest, a "partial credit" model has a single difficulty parameter for each item. It is the latent trait analogue for Brown's (1965) integer scoring scheme based on the number of attempts needed to reach the correct response. The scoring is from 1.0 for a correct response on the first attempt to 0.0 for failure in (m-1) attempts, where there are m alternatives presented for an item. This model takes no account of the variations in distractor attractiveness from item to item, nor of which distractors were actually selected by the respondent.

The second latent trait model treats the test as a sequence of distinct steps each of which has a difficulty parameter. A single five-way multiple choice item can be regarded as comprising four steps, with each successive step after the first being attempted if, and only if, the preceding one is failed. The scoring is 1/0 for each step, with steps not attempted being coded as incomplete data. This

produces four difficulty parameters for each item, but a single and more precise ability estimate for the individual. The method does not assume that all the items have the same logical structure with regard to difficulty, but it takes no account of exactly which distractors are selected.

The third model is an extension of the second. In this model, the step difficulty values for an item vary in terms of which distractors were previously selected. Thus for a five-way multiple choice item there is one difficulty parameter at the first step, four at the second, six at the third, four at the fourth. This give a total of fifteen difficulty parameters for a single five-way multiple choice item. It should in general give a better fit than the model described above because it treats the distractors individually, but it requires more data for the necessary calibration of the item parameters.

To some extent, the utility of these models is going to depend on the relative preponderance of the two styles of student behavior discussed earlier. Under partial knowledge, distractor elimination and random guessing (style A) the noise introduced by guessing precludes the possibility of very precise measurement, and the first model described may well prove as effective as either of the others. Where item responses based on correct information or misinformation (style B) dominate, we would expect that models two and three would provide more precise and valid measures of student performance.

Each of the models described is based on the simple one-parameter Rasch logistic model. This is for two reasons. Firstly, as argued in

a separate report to NIE, the Rasch model seems the logical choice in a situation which involves the construction of new test instruments, since it focuses attention on meeting the logical rquirements for objective measurement. Secondly, the main alternative, the three-parameter logistic model, has severe practical limitations even when applied to regular test data. Estimating techniques are primitive, and very large samples are required in order to obtain stable parameter estimates. The three-parameter model has been found useful in describing large bodies of existing data derived from tests of varied quality, but such data sets do not exist in the AUC format. Since obtaining sufficient data for adequate item calibration is anticipated to be a problem even for the Rasch model, it appeared sensible to concentrate initial efforts in this direction.

Model (i): Fixed Partial Credit

The model is $E(X_{vi}) = \dfrac{e^{(\alpha_v - \delta_i)}}{1 + e^{(\alpha_v - \delta_i)}}$

where: $E(X_{vi})$ is the expected score of person v on item i

α_v is a parameter describing the abiliuty of person v

δ_i is a parameter describing the difficulty of item i

and the scoring function $X_{vi} = \dfrac{m_i - g_{vi}}{m_i - 1}$

where m_i is the number of alternative choices on item i (of which 1 is correct and (m-1) are incorrect)

and g_{vi} is the number of attempts by person v on item i until the correct alternative is chosen. If the (m_i-1)th attempt fails then $X_{vi}=0$.

The rationale for this scoring scheme is based on a "partial knowledge" distractor elimination model. If a correct response is chosen at the first attempt, then it is assumed that the student was able to eliminate all the distractors, and so he or she gets full credit. If the first attempt fails, but the second attempt is correct, it is assumed that he or she could eliminate all the distractors but one, so that credit of $\frac{m-2}{m-1}$ is awarded. (The number of distractors is (m-1)).

Although this equal-interval scoring function may appear somewhat arbitrary it is analogous to that frequently adopted in elementary scaling techniques (e.g., Likert scales). Moreover, Andersen (1977) has shown that for the model to retain specific objectivity, successive scoring categories must be equidistant. The immediate advantage of this is that the "raw score" by a student who has worked through the set of items is a sufficient statistic for the ability (and frequently may be used instead of it--hence the viability of the scheme proposed by Brown).

Parameter estimation is approached via a modification of the Rasch PAIR estimation algorithm (Choppin, 1982). For two items i and j, the relative difficulty can be estimated by

$$\delta_i - \delta_j \approx \log b_{ji} - \log b_{ij}$$

where, on this occasion, b_{ij} is the sum, over all people in the sample, of $X_i(1-X_j)$ and b_{ji} is similarly defined. (It can be seen that this reduces to the standard PAIR algorithm in the case of 1/0 scoring.)

$X_i(1-X_j)$ represents the product of an estimate of the extent to which item i is mastered multiplied by an estimate of the extent to which item j is not mastered. It may be viewed, for each subject as a measure of the extent to which item i is easier than item j. The ratio:

$$\frac{E\left[X_i(1-X_j)\right]}{E\left[X_j(1-X_i)\right]} = e^{(\delta_j - \delta_i)}$$

a value independent of α

which is why the accumulation of data over persons to estimate these expectations works.

The algebra for maximum likelihood estimation, and for controlling the model via the squared matrix B* exactly duplicates that laid out in Choppin (1982), except that the formulae presented there for the standard errors of the δ-values are no longer appropriate. (Corrected formulae have not yet been developed, so the values reported by PAIR are used as conservative guides.) Once the items are calibrated, the estimation of person ability again follows the PAIR procedure.

Model (ii): Step Calibration

In this model, the probability of person v responding correctly to item i at the gth attempt, given that he or she makes the attempt, is:

$$\text{Prob.}\left[X_{vig} = 1\right] = \frac{e^{(\alpha_v - \delta_{ig})}}{1 + e^{(\alpha_v - \delta_{ig})}}$$

where X_{vig} = 1 if the gth attempt at item i is successful, and

= 0 otherwise

α_v is again a parameter describing the ability of person v and δ_{ig} is a parameter describing the difficulty of the gth step on item i.

For a five-way multiple choice item there are five possible sets of observation vectors $\underline{X}$, with asterisks indicating missing data (i.e., attempts that do not occur).

	g =	1	2	3	4
Correct at first attempt:	$\underline{X}$ =	1	*	*	*
Correct at second attempt:	$\underline{X}$ =	0	1	*	*
Correct at third attempt:	$\underline{X}$ =	0	0	1	*
Correct at fourth attempt:	$\underline{X}$ =	0	0	0	1
Failure at fourth attempt:	$\underline{X}$ =	0	0	0	0

If the raw data to be analyzed consists of code numbers for the successful attempt on each item, then it must be transformed into the above format for the calibration analysis. For example, suppose that an individual required (2, 1, 1, 4, 5, 3) attempts to find the correct answers to a six item five-way multiple choice test. The recoding of this vector would yield:

0 1 * *	1 * * *	1 * * *	0 0 0 1	0 0 0 0	0 0 1 *

a vector of 24 elements. A set of such vectors from the different persons attempting the test can be analyzed almost as a standard Rasch model problem--providing the PAIR algorithm (Choppin, 1982) is used to allow for the embedded missing data. The deviation from the standard Rasch procedure is necessitated by the violation of the local independence assumption for AUC data. While it remains important that between items this independence is maintained, it is clear that within an item the different X-values cannot be independent. As shown above, only m possible patterns out of the 3^m theoretically possible on each item ever occur and certain combinations such as (1,0) are impossible. This invalidates the maximum likelihood estimation procedure which assumes that the elements of the B matrix for item pairs are essentially independent.

The full theoretical implications of this are still being explored, but a convenient "fix" in order to calibrate the items is to use instead of ML a least squares procedure based on a modified B* matrix. This B*, instead of being simply the square of matrix B as before, is now screened to remove the contaminating dependence within items.

In the standard PAIR algorithm

$$b^* = \sum^{k} b_{ik}b_{jk}$$

and since $b_{ii} = b_{jj} = 0$, b^*_{ij} is independent of b_{ij}.

In PAIR as modified for AUC tests

$$b^*_{ij} = \sum^{k} v_{ik}b_{ik}v_{kj}b_{kj}$$

where v_{ik} are the elements of a screening matrix such that

$v_{pq} = 0$ if responses p and q relate to the same item

and $v_{pq} = 1$ otherwise.

Least squares estimation procedure applied to the B* matrix yields calibrations for the δ_{ig} values (i = 1,k ; g = 1, m-1).

The estimation of person ability, the usual goal in such exercises, is somewhat different than in the standard Rasch model. Apart from rare failures at the final attempt, each student will score one point on each item and thus will have a raw score of k.

However, this raw score will be based on different numbers of "attempts", and individual step difficulties will be higher on some items than on others. Therefore α_v is estimated by the solution of

$$r_v - \sum \frac{e^{\alpha_v}}{e^{\alpha_v} + e^{\delta_{ig}}} = 0$$

where the summation extends over the item steps actually attempted, and r_v is the observed raw score (usually k). This equation can always be solved to produce a unique LS estimation of α_v , but may be inefficient since its (iterative) solution is required for each observed score pattern. Monte Carlo simulation could compare the variation in α with the scoring function proposed by Brown (1965), to see whether the exact iterative solution is worthwhile.

The standard errors of such estimates depend upon the number of attempts made. Thus someone who usually responds correctly at the first attempt will be measured with less precision than someone who typically requires two or three attempts. Data in which the mean number of attempts per item is 2.0 (a typical value) will yield standard errors of measurement only 0.7 times as large as with a

conventional test with the same number of items. From this it can be seen that major increases in precision can only be achieved by substantially increasing the number of alternatives per question, so that the number of attempts made before success will also increase. A valuable experiment would thus be to try this procedure on a test for which each item had eight or ten alternatives. This has not yet been done.

Model (iii): Distractor Calibration

This model is an extension of (ii) to allow for differences among the distractors. The item step difficulty parameter now describes the difficulty of the item at each step taking account of which distractors have already been eliminated.

Thus δ_{i1} indicates the difficulty of item i at the initial step when all distractors are present

$\delta_{i2.A}$ indicates the difficulty of item i at the second step when distractor A was chosen at the first

$\delta_{i3.BC}$ indicates the difficulty of item i at the third step after distractors B and C have been chosen (in whatever order)

With this notation, the model becomes

$$\text{Prob}\left[X_{vig.\underline{F}} = 1\right] = \frac{e^{(\alpha_v - \delta_{ig.\underline{F}})}}{1 + e^{(\alpha_v - \delta_{ig.\underline{F}})}}$$

The analysis and estimation procedures essentially follow those for model (ii) except that the response data must be coded in

different format. For a five-way item (for which the corrent respone is E, and the distractors are labeled A-D), the structure of the parameters to be estimated is:

δ_{i1}	$\delta_{i2.A}$	$\delta_{i2.B}$	$\delta_{i2.C}$	$\delta_{i2.D}$	$\delta_{i3.AB}$	$\delta_{i3.AC}$	$\delta_{i3.AD}$	$\delta_{i3.BC}$	$\delta_{i3.BD}$	$\delta_{i3.CD}$	$\delta_{i4.ABC}$	$\delta_{i4.ABD}$	$\delta_{i4.ACD}$	$\delta_{i4.BCD}$

Response data for an individual who chose responses A, C , F, in that order, getting the item right at the third attempt, would be coded

0	0 * * *	* 1 * * * *	* * * *

It should be noted that this coding scheme is severely constrained. There is at most one entry in each block, and a "1" entry effectively terminates the vector. Thus the range of possible response patterns is limited, and again the local independence principle is violated.

Estimation procedures can follow the sequence described in model (ii) first to calibrate the item step values, and secondly to estimate the person ability parameters. However, it is apparent that the procedure is somewhat unwieldy. For each item the number of difficulty parameters to be estimated is given by $(2^{m-1} - 1)$ where m is the number of alternative responses in the item format. Inadequate calibration of the parameters due to insufficient data can spoil the overall measurement of person ability (viz: person measurement with

the Lord-Birnbaum three-parameter model and small data sets). A six item five-way multiple choice test such as that described under model (ii) would require the estimation of 90 item difficulty parameters under model (iii) as opposed to 24 under model (ii). For this model, in contrast to model (ii), it would seem wise to restrict item formats to not more than three or four alternatives.

3. Trial Data Analysis

Calibration procedures for models (i) and (ii) have been programmed in FORTRAN using variations of the PAIR algorithm described above. Both programs have demonstrated their ability to recover the parameter values used to generate artificial "fitting" data. Two data sets from AUC tests each comprising several hundred cases have been analyzed using these programs. One test is a junior high school science test under development in England. The second is a college level psychology test used in a private California university. The results are still being studied.

Model (iii) requires the coding of which distractors were selected in which sequence, and this is only practicable with a clinically administered or computer administered test. For this reason we have devoted considerable time to developing a software package that will administer AUC tests in schools, and store the results in a format suitable for aggregation and subsequent analysis. Details of this package are given in the Appendix.

REFERENCES

Andersen, E.B. Sufficient statistics and latent trait models. Psychometrika, 1977, 42, 69-81.

Brown, J. Multiple response evaluation of discrimination. The British Journal of Mathematical and Statistical Psychology, 1965, 18, 125-137.

Choppin, B.H. A fully conditional estimation procedure for Rasch model parameters. Draft report to NIE, 1982.

Frary, R.B. The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. Applied Psychological Measurement, 1980, 4, 1, 79-90.

Hanna, G.S. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. Journal of Ecu tional Measurement, 1975, 12, 3, 175-178.

Kane, M., & Moloney, J. The effect of guessing on item reliability under answer-until-correct scoring. Applied Psychological Measurement, 1978, 2, 1, 41-49.

Merwin, J.G. Rational and mathematical relationships of six scoring procedures applicable to three-choice items. Journal of Educational Psychology, 1959, 50, 4, 153-160.

Wilcox, R.R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74.

APPENDIX

INTERACTIVE COMPUTER PROGRAMS FOR CONFIDENCE-MARKING AND ANSWER-UNTIL-CORRECT TESTING

Raymond Moy and Chih-Ping Chou
Center for the Study of Evaluation, UCLA

Introduction

In traditional scoring of multiple-choice tests, an item score of one is given if the examinee selects the correct answer, and zero if any other alternative is chosen. The problems with such a score assignment procedure are twofold. On the one hand, because of the limited number of distractors available, it is possible for an examinee to obtain a score of one simply through random selection of an alternative and without any knowledge of the correct answer. On the other hand, many students with partial knowledge will receive a score of zero even though they are able to reduce the number of answer alternatives to a smaller subset than those originally presented. Assuming that the correct answer is included in this subset, such students do not deserve one point full credit if they guess the correct answer, nor do they deserve a score of zero if they miss it. A more accurate score, reflecting their state of partial knowledge lies somewhere in between. The net result of the zero-one method of scoring is a reduced efficiency of measurement, because reliability is decreased from having assigned ones to students who do not really know the answers, and zeros to those who have partial knowledge.

It may, therefore, be possible to improve upon traditional zero-one scoring if some method could be devised to obtain more detailed information about the examinees' state of partial knowledge. Although it might be possible to have an examinee give rationales for choosing a particular answer alternative, this is not practical in large scale testing efforts, nor will it be easy to assign objective partial score credit to such open-ended responses. Instead, various objective techniques have been suggested which may yield useful information. Among these techniques are elimination scoring, confidence marking, and answer-until-correct. All of the techniques are based on examinee interactions with the item distractors, or obtaining information about how examinees view the correctness of their answer choices.

In elimination scoring, the examinee is asked to indicate those alternatives which he or she thinks is definitely incorrect. A score of one is assigned if, and only if, all distractors are correctly eliminated and partial scores may be assigned on a weighted basis for correctly eliminating some of the distractors. Various methods for assigning partial credit have been proposed (e.g., Arnold & Arnold, 1970; Coombs, 1953; or Cross & Thayer, 1979), however, all methods are rather arbitrary since none are based on explicit descriptions of the relationship between choice of distractors and the ability of interest. The methods differ, though, in how they deal with the possibility of guessing behavior and misinformation (i.e., eliminating the correct answer as wrong).

Aside from the problems of deciding partial credit scores for various types of elimination patterns, there is also a significant problem in getting examinees to respond properly to the task. There is a tendency among examinees to be much too conservative when faced with expressing their confidence in their answers (e.g., Ebel, 1968; Hritz & Jacobs, 1970). If this is the case, then the ability estimates from this procedure may be negatively biased.

Confidence marking procedures require the examinee to either select a correct answer and provide a confidence judgment in the answer (as exemplified in studies by Shaughnessy, 1979; Sieber, 1979) or to assign probabilities of correctness for each answer alternative (Koriat, Lichtenstein, & Fischhoff, 1980; Rippey & Donato, 1978). These confidence markings can then be used to score examinees' partial knowledge of individual items.

Like elimination scoring, the validity of confidence marking procedures depends on the examinees' responding properly and accurately to the task. Personality characteristics which lead to expressions of over or under confidence would be problematic, as would be variation across examinees in the interpretation of specific confidence ratings.

The answer-until-correct technique avoids requiring examinees to make judgments for individual answer alternatives and instead allows the examinees to select what they consider to be the correct answer and to continue choosing among the distractors until the correct answer is selected. The number of attempts an examinee takes before reaching the correct answer is taken to be indicative of the extent of the examinee's partial knowledge.

In contrast to the confidence marking and elimination procedures, AUC testing requires some method of providing feedback to the examinees that tells them whether their answers are correct or not. This means that either special answer sheets or individualized testing sessions would be required.

Aside from these logistic problems, there is also difficulty in interpreting the relationship between number of attempts and the ability of interest. Although it is commonly agreed that the fewer number of attempts the greater the partial credit that should be awarded, an overall scoring algorithm which will maximize scaling validity has not yet been devised. This is due to the fact that information regarding the relative difficulty of distractors needs to be specified and, as of yet, item writing technology is not refined enough to accomplish the task.

In practical applications of AUC testing, it has been the practice to simply use the number of attempts as the basis for scoring. Whereas Gillman and Ferry (1972) found that split-half reliability for this method of scoring was substantially increased over zero-one scoring, Hanna (1975) and Taylor, West, and Tinney (1975) found little or no improvement. One possible resolution to these conflicting findings is that improvements through the use of AUC scoring are dependent on the properties of the items and their distractors. Kane and Maloney (1978) have shown that when all but two distractors are eliminated as incorrect by all examinees, and when random guessing takes place among the n-1 alternatives, zero-one scoring is more efficient.

In contrast to the approach of assigning partial credit on the basis of the number of attempts, Wilcox (1981, 1982) proposes using AUC information to yield correction for guessing estimates. Under this conceptualization, the ability of interest is the proportion of items an examinee is able to answer correctly, with no credit for partial knowledge. However, partial knowledge will affect the probability of getting an answer correct through guessing, and it is this probability which is estimated from AUC information.

Whether one chooses this latter conceptualization of ability or the partial credit conceptualization is a question of the meaning of one's scale and is not a matter of one being correct and the other incorrect. Quite simply, they are two different ways in which AUC information can be utilized to improve on zero-one scoring.

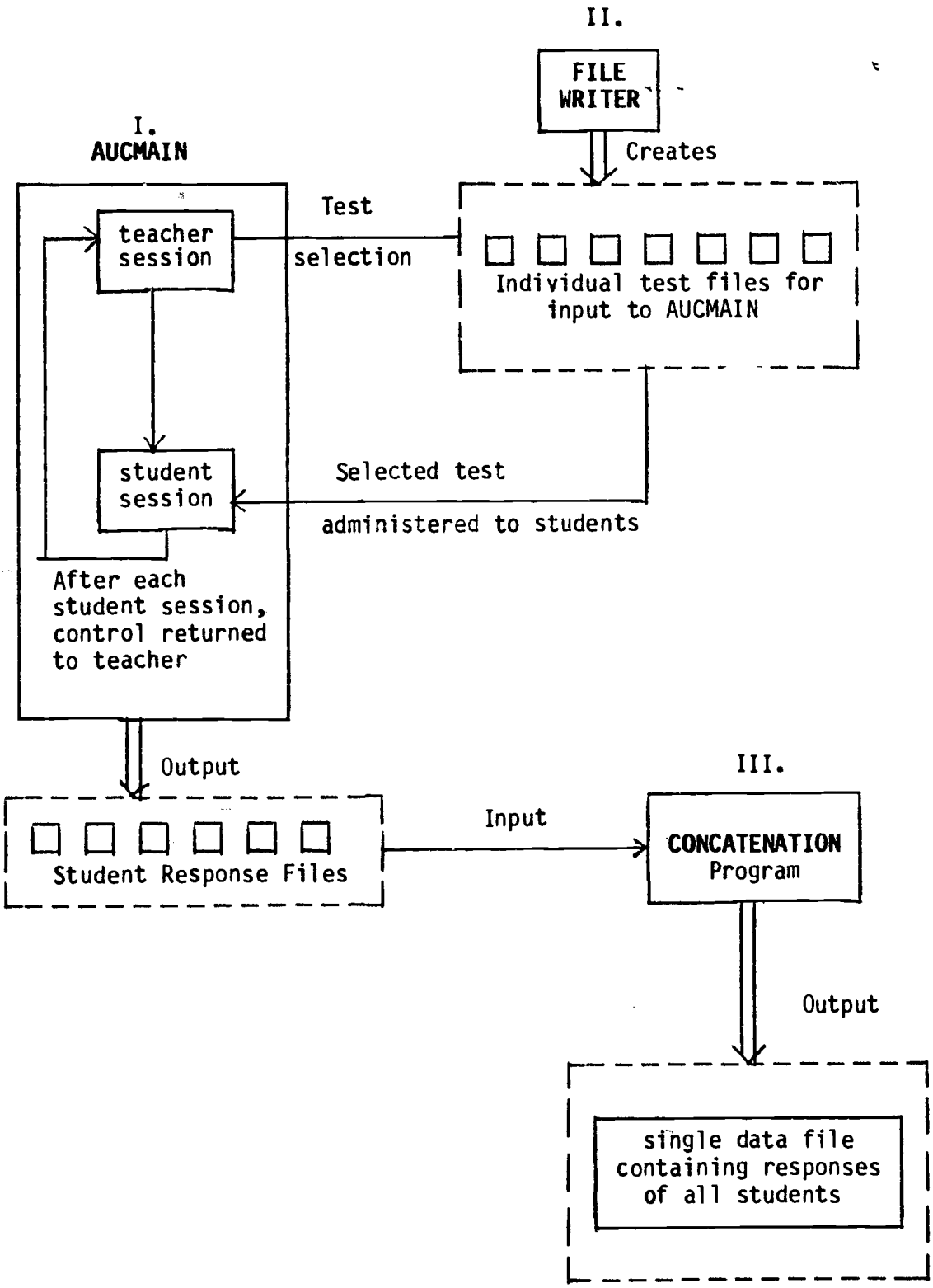
In order to more fully investigate the value of AUC information, substantial amounts of data must be gathered and the logistic problems of providing AUC feedback to the examinees must be solved. Toward this end, an interactive program was developed to follow an AUC format. The program was designed to allow AUC testing on a number of different tests to students of a wide range of ability levels. The rest of this report will describe in greater detail the overall design of the program, the options available in a typical program run, the mechanics of inputting new tests, and the production of output for data analysis.

The AUC Program

Three programs have been developed for gathering A-U-C test data: (1) the AUCMAIN program for administering the tests, (2) a test FILE WRITER program for creating new tests as input to AUCMAIN, and (3) a CONCATENATION program for creating a single data file containing responses from all students being administered a particular test. Figure 1 shows how these three programs are related to each other.

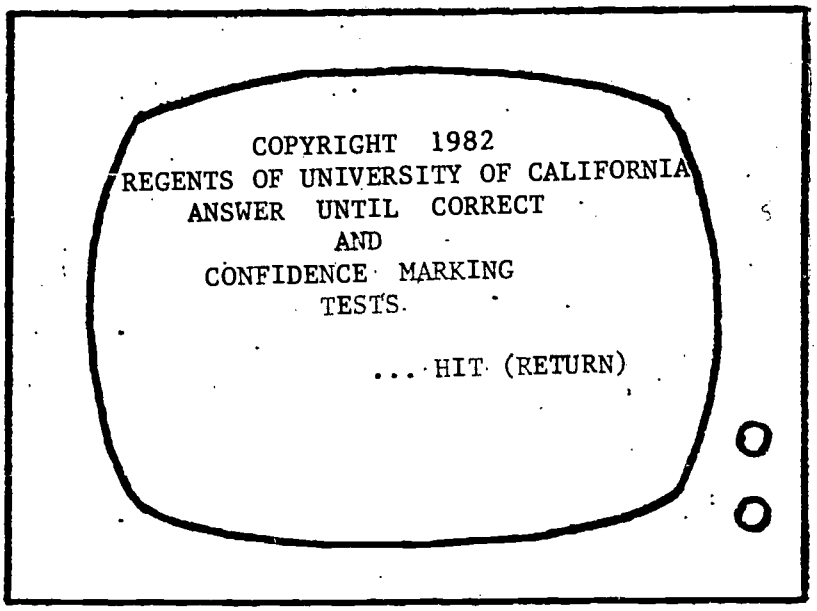
Figure 1

Interrelationship of AUCMAIN (I), FILE WRITER (II), and CONCATENATION (III) programs.

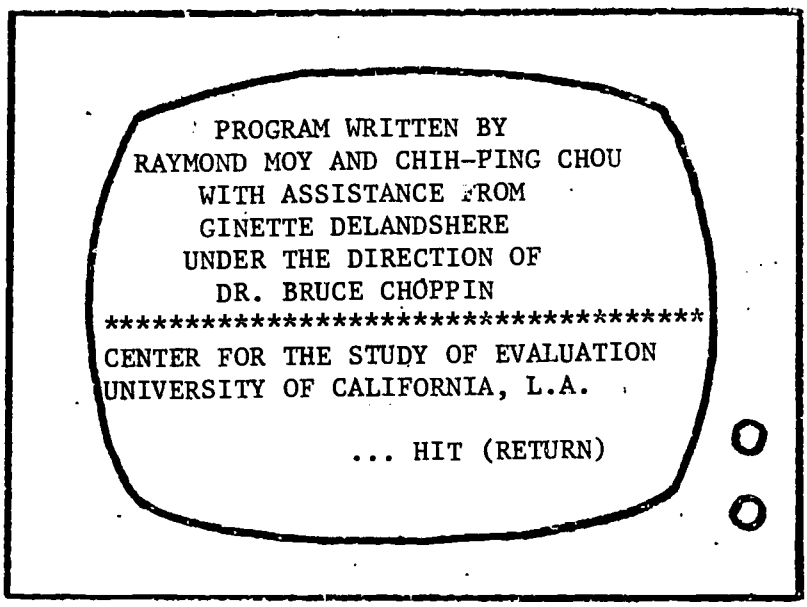


The AUCMAIN program. The AUCMAIN program contains two sections: the first section is designed to interact with the teacher who is given a decription of administration procedures and requires teachers to specify session parameters which will identify and control the administration of tests to students. The second section is the actual test session controlled by the examinee.

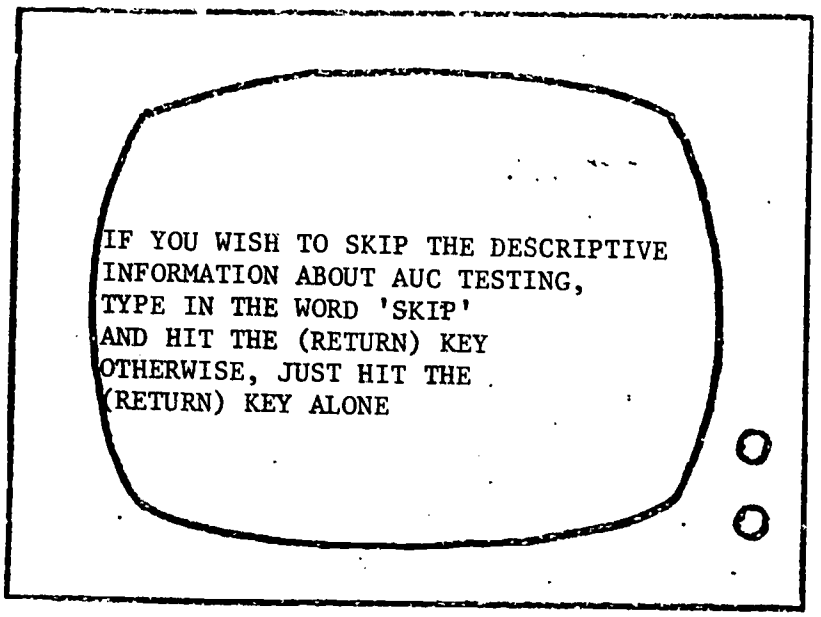
Teacher session. The AUCMAIN program is self-booting once the AUC disk is mounted and the computer turned on. The screen will show:



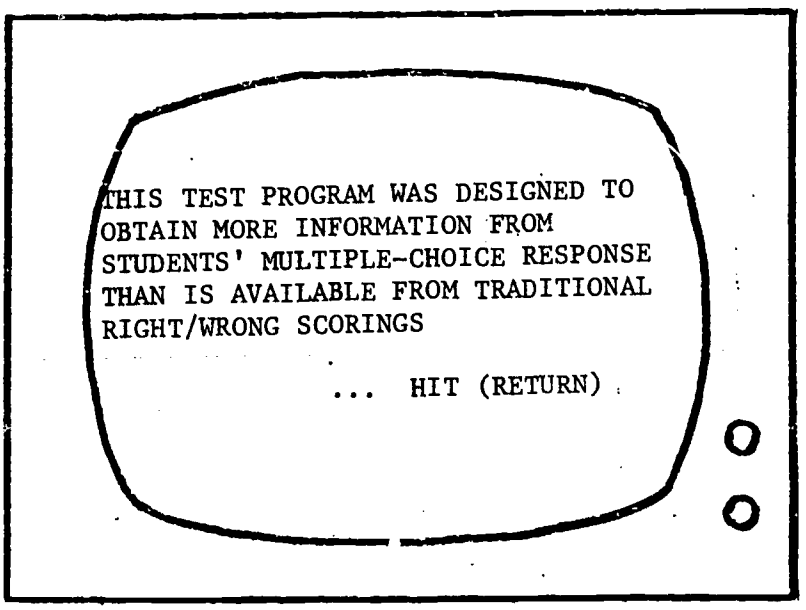
Teachers should then hit the <RETURN> key to view the next screen:

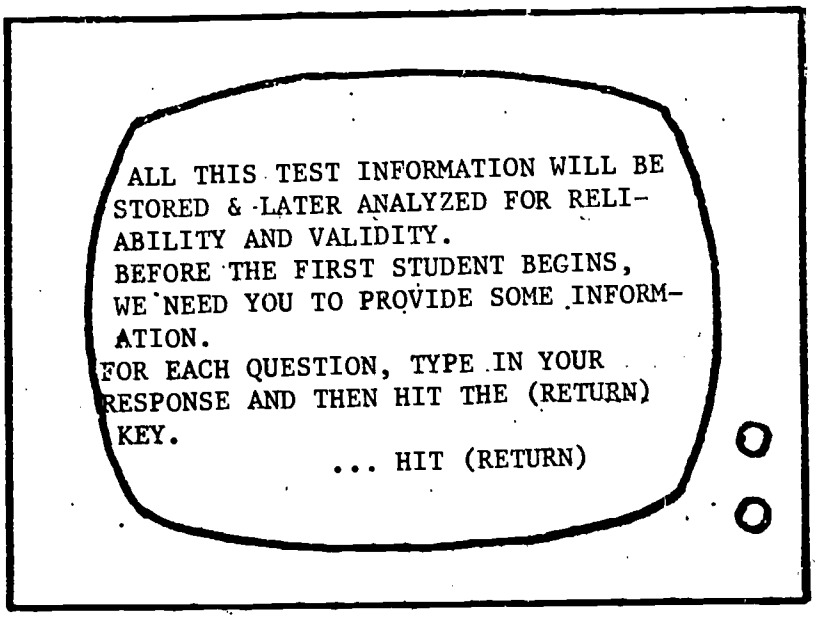


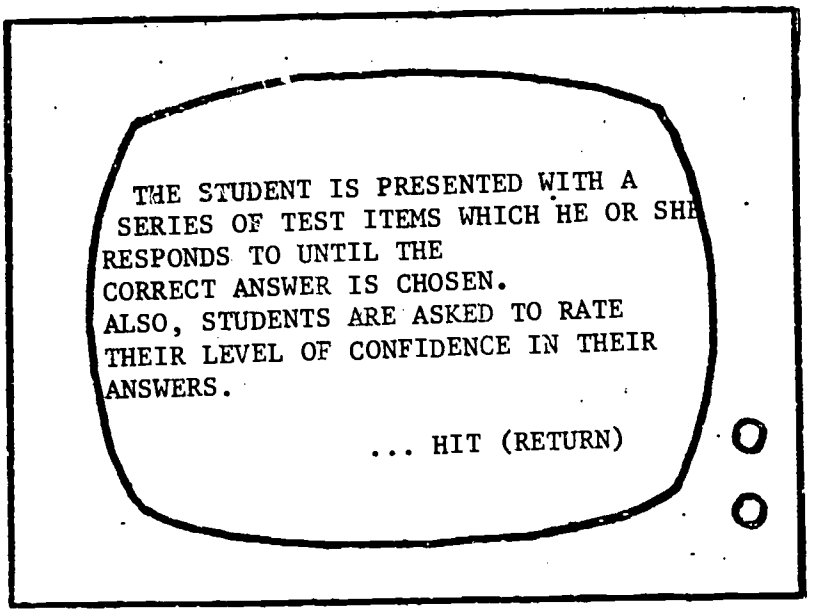
After the <RETURN> key is hit again, the program will ask whether teachers would like to have a description of the AUC testing technique:



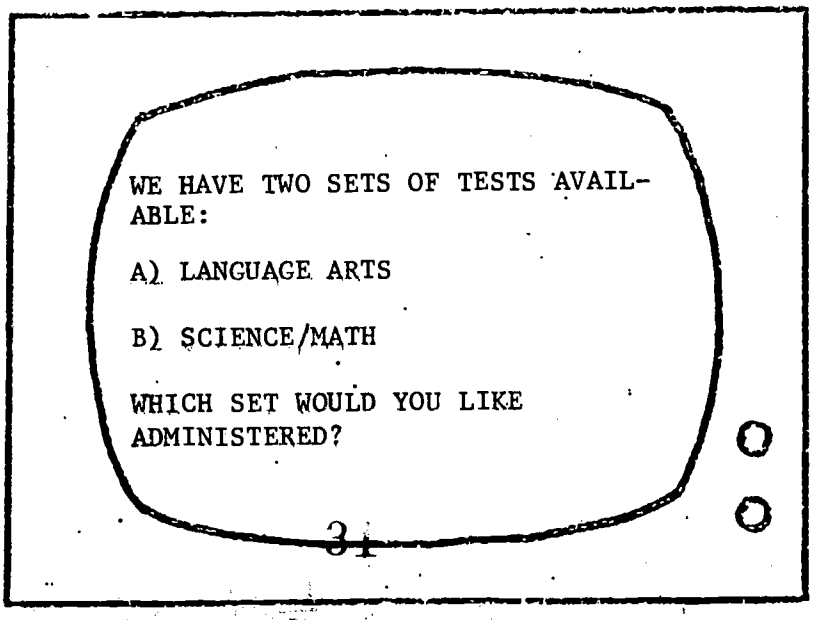
If the word 'skip' is entered the program will proceed to the test selection screen. Otherwise, AUC descriptive information is presented on the following screens. Teachers hit the <RETURN> key to proceed from one screen to the next.



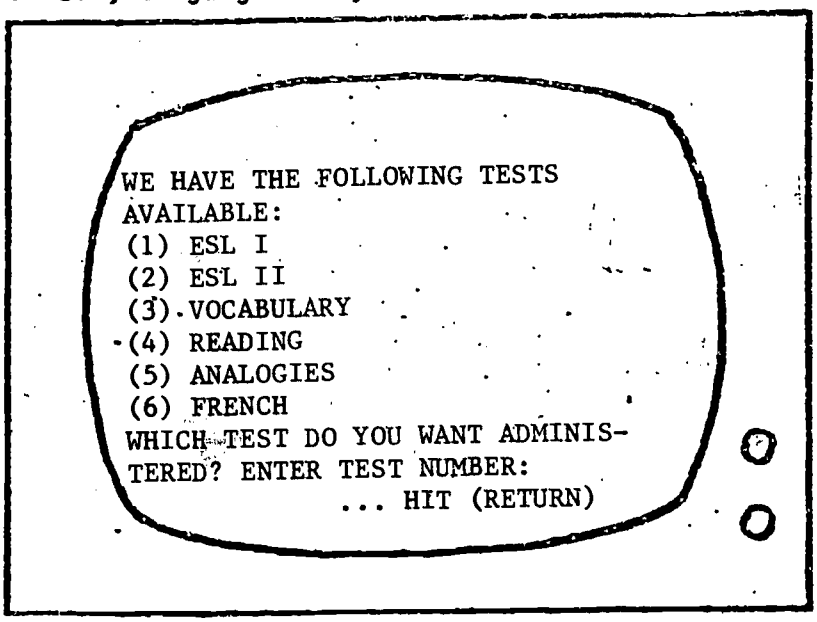




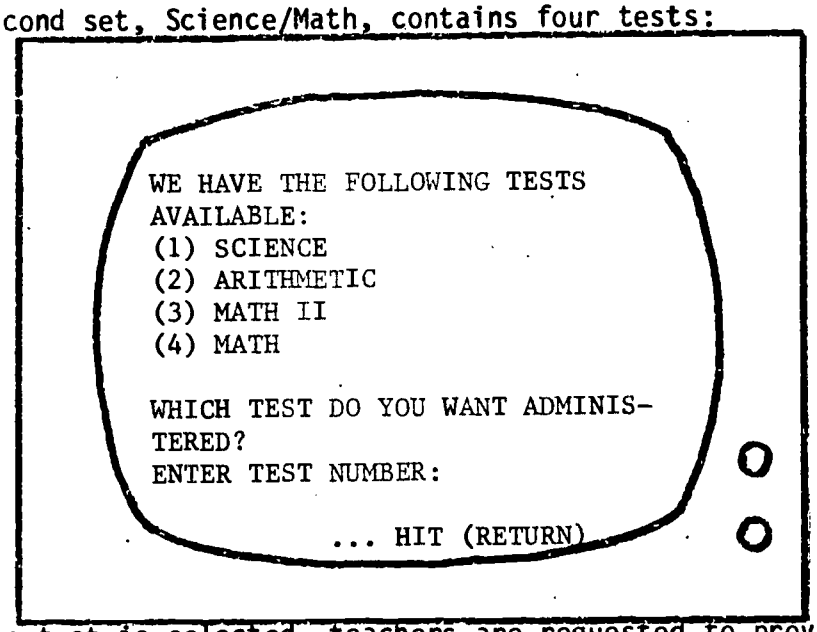
Following the AUC descriptive information, the test selection screen is provided. Teachers are asked to select one of two test sets: Language Arts or Science/Math.



The first set, Language Arts, consists of six tests:

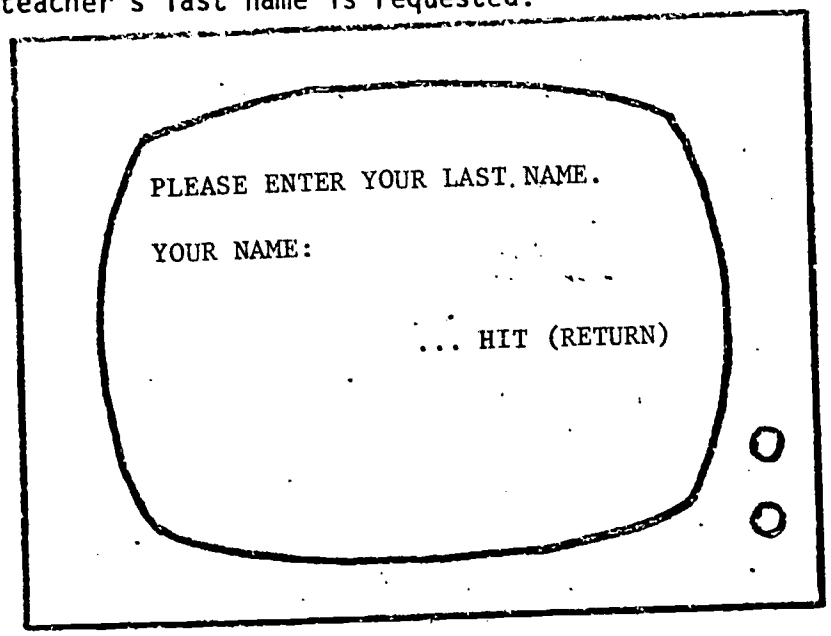


The second set, Science/Math, contains four tests:

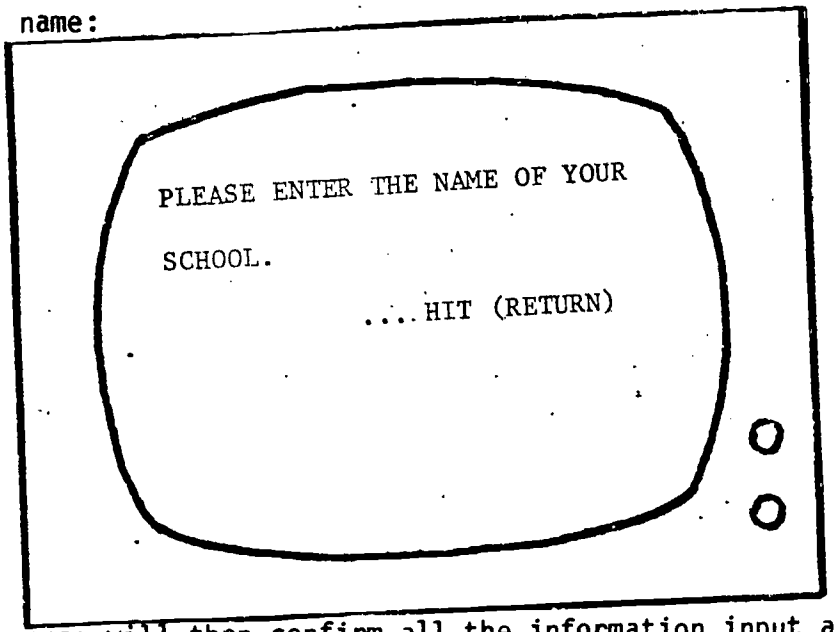


After a test is selected, teachers are requested to provide information which will be used to help identify student response files.

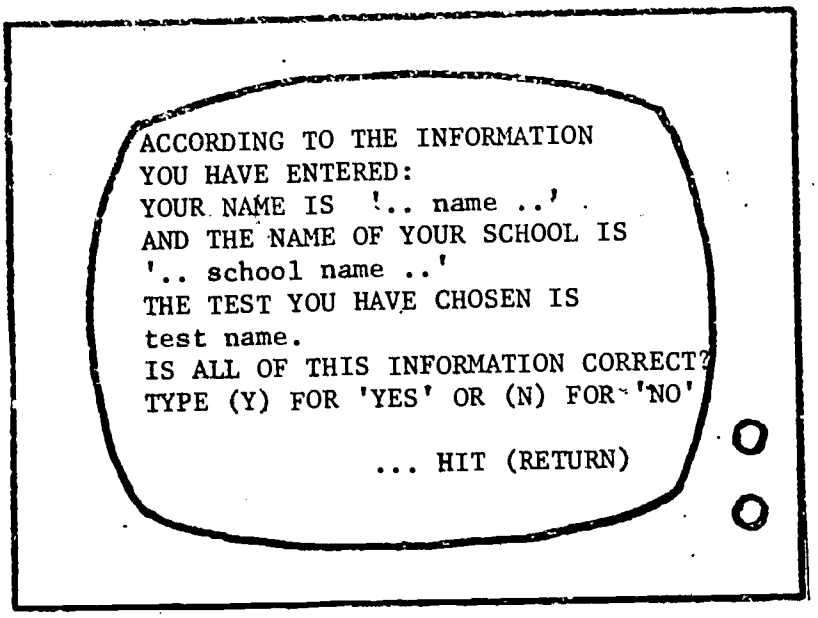
First, a teacher's last name is requested:



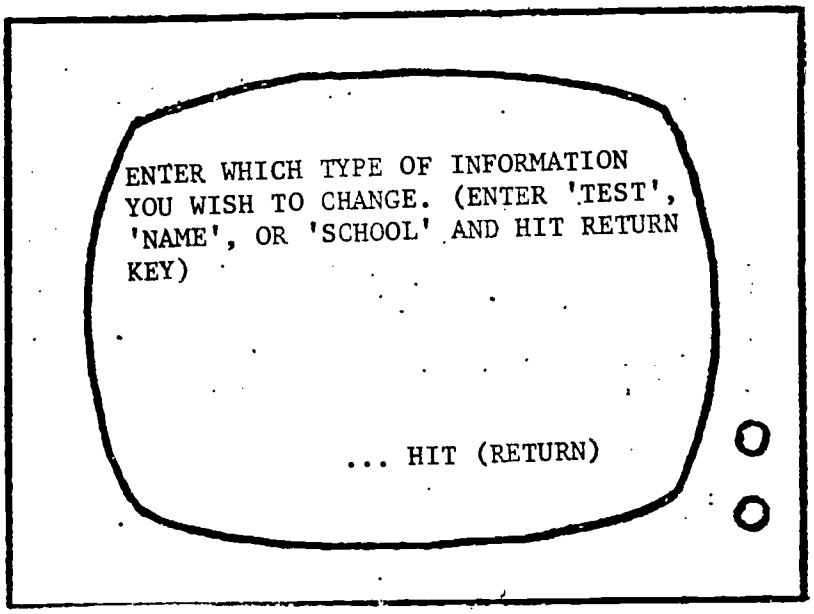
then the school name:



The AUC program will then confirm all the information input as follows:

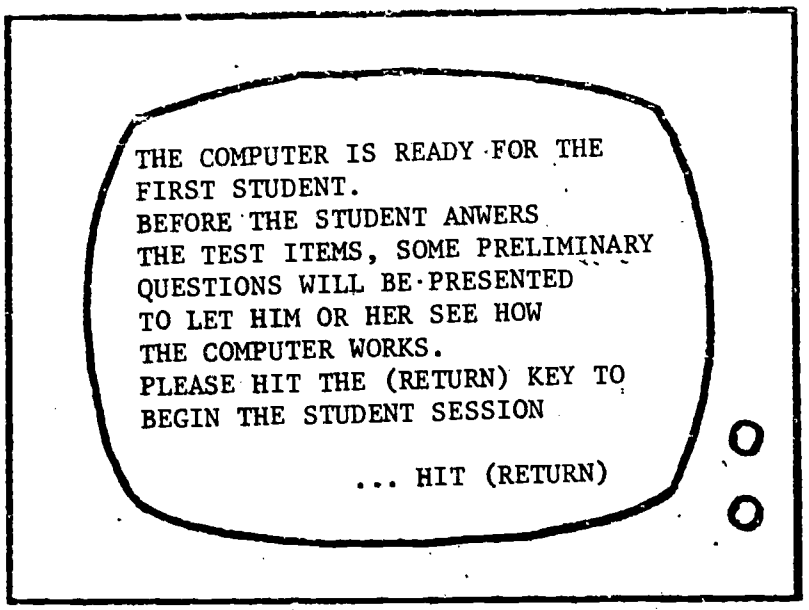


Teachers can type <Y> to confirm the information and proceed with the student session. If corrections are required, teachers should type <N> and hit <RETURN>. The screen will then print out the following question:

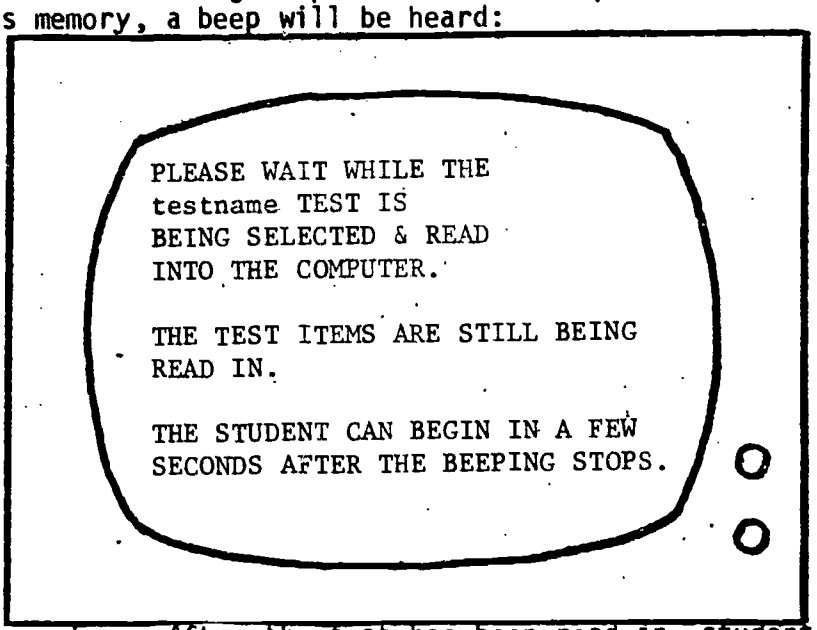


For example, if the teacher wishes to change the test selection, he or she should type 'test' and then hit <RETURN>. The program will go back to the test selection session, and then present the information again for confirmation. The program proceeds to the student session after all the information is entered correctly.

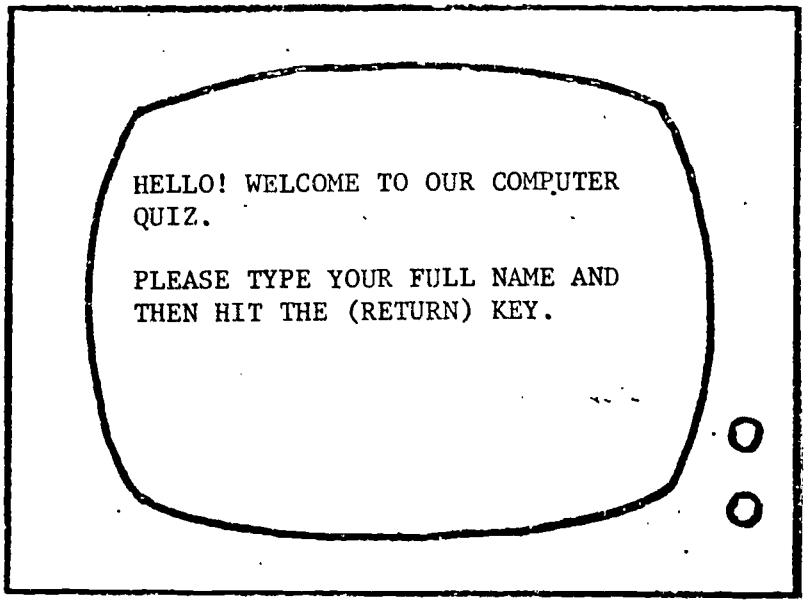
The student test session begins after the following messages:

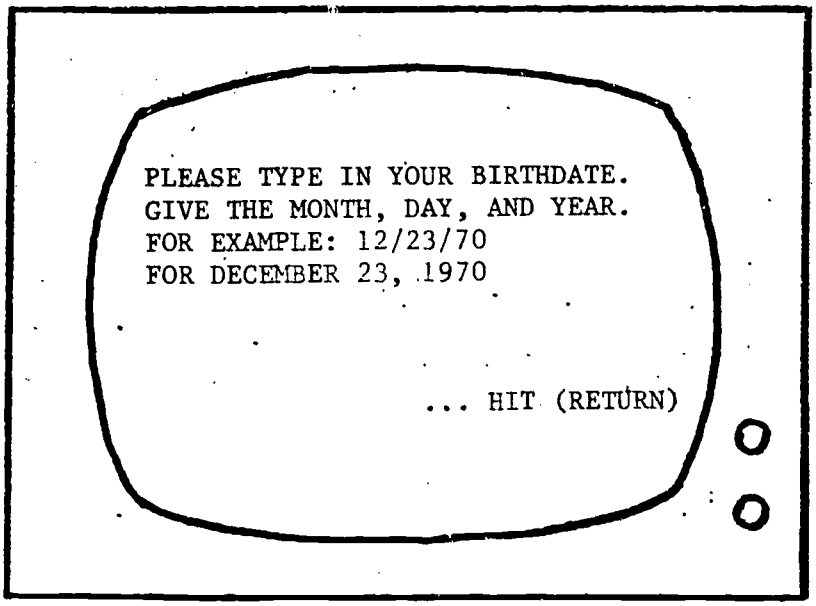


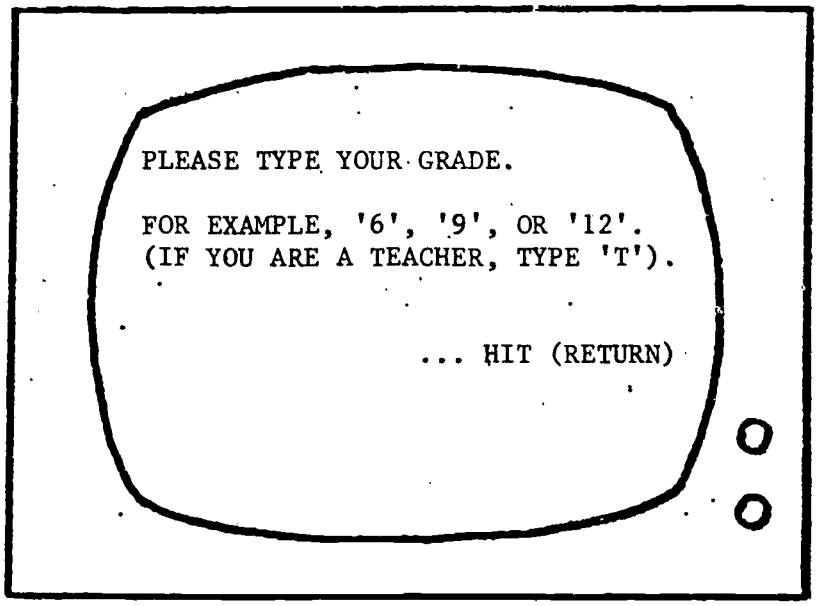
The computer will then load the selected test and asks teacher to stand by while this is being completed. As each question is read into the computer's memory, a beep will be heard:



Student session. After the test has been read in, students are requested to provide information which will be used for identification purposes. Also during this time, students will have an opportunity to get acquainted with the computer and learn how to interact with it. Student are asked for their names, birthdates, and grades:

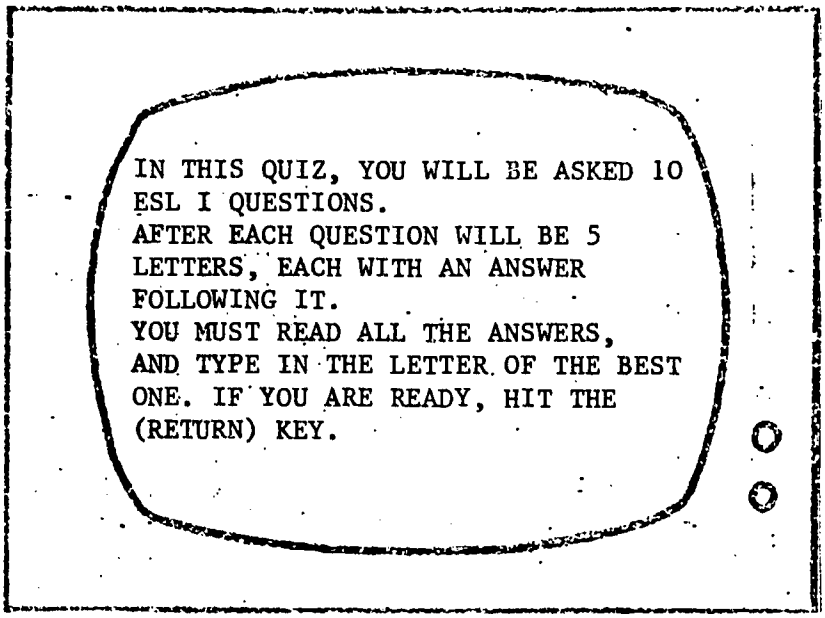

HELLO! WELCOME TO OUR COMPUTER QUIZ.
PLEASE TYPE YOUR FULL NAME AND THEN HIT THE (RETURN) KEY.


PLEASE TYPE IN YOUR BIRTHDATE.
GIVE THE MONTH, DAY, AND YEAR.
FOR EXAMPLE: 12/23/70
FOR DECEMBER 23, 1970
... HIT (RETURN)

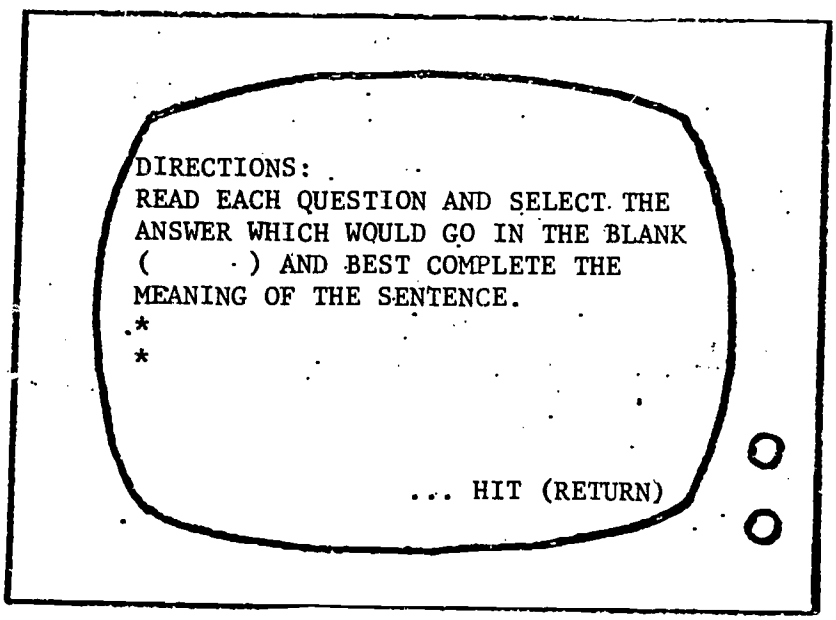

PLEASE TYPE YOUR GRADE.
FOR EXAMPLE, '6', '9', OR '12'.
(IF YOU ARE A TEACHER, TYPE 'T').
... HIT (RETURN)

If <T> is typed, no response file is created at the end of the session.

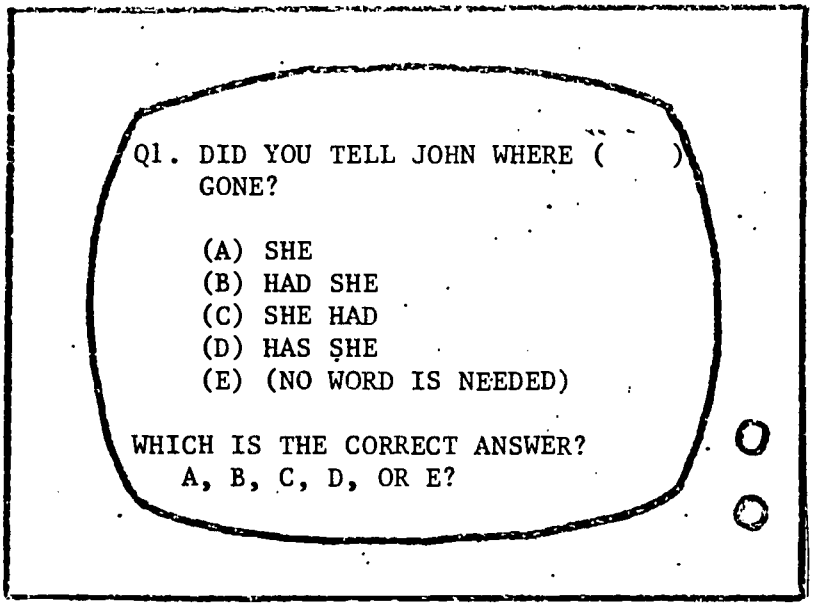
Students will then get a short description of the test that is going to be administered. Using the ESL I test as an example, the student will see the following screen:



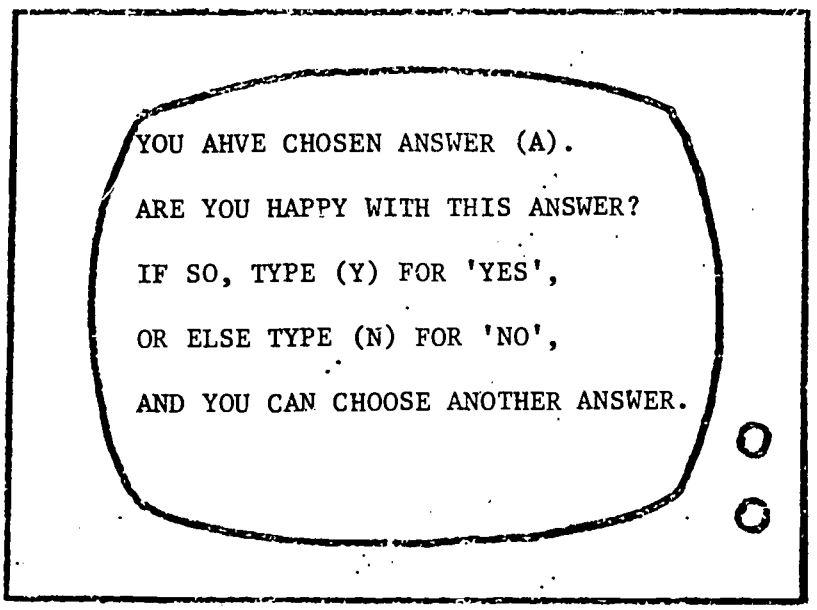
When the student is ready and hits the <RETURN> key, the directions for the ESL I test are presented:



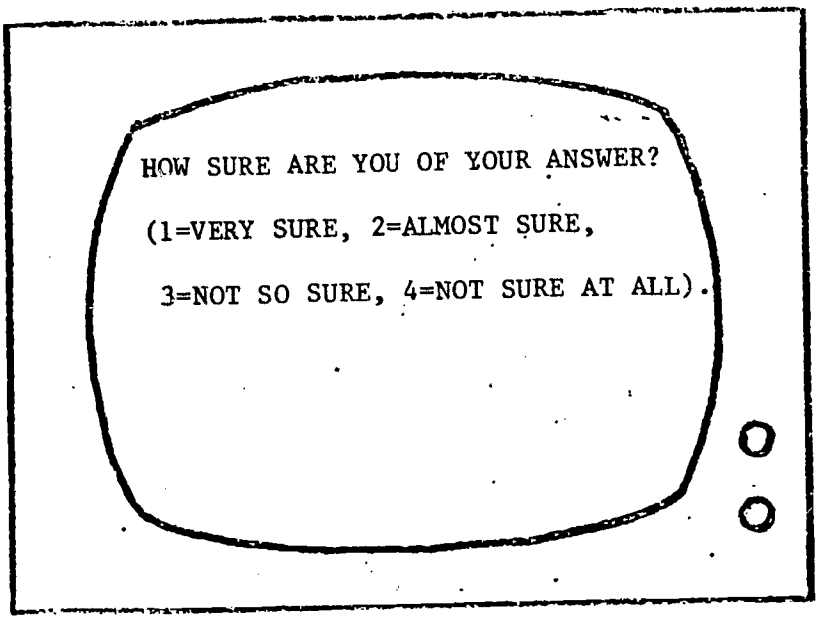
The first item of the test will then be presented:



In this program, students have as many chances as they need to answer an item correctly. Each time an answer is provided, the screen will present the answer just made, and allow students to make changes if desired. For instance, if answer <A> is chosen, the program will print out the following statements on the screen:

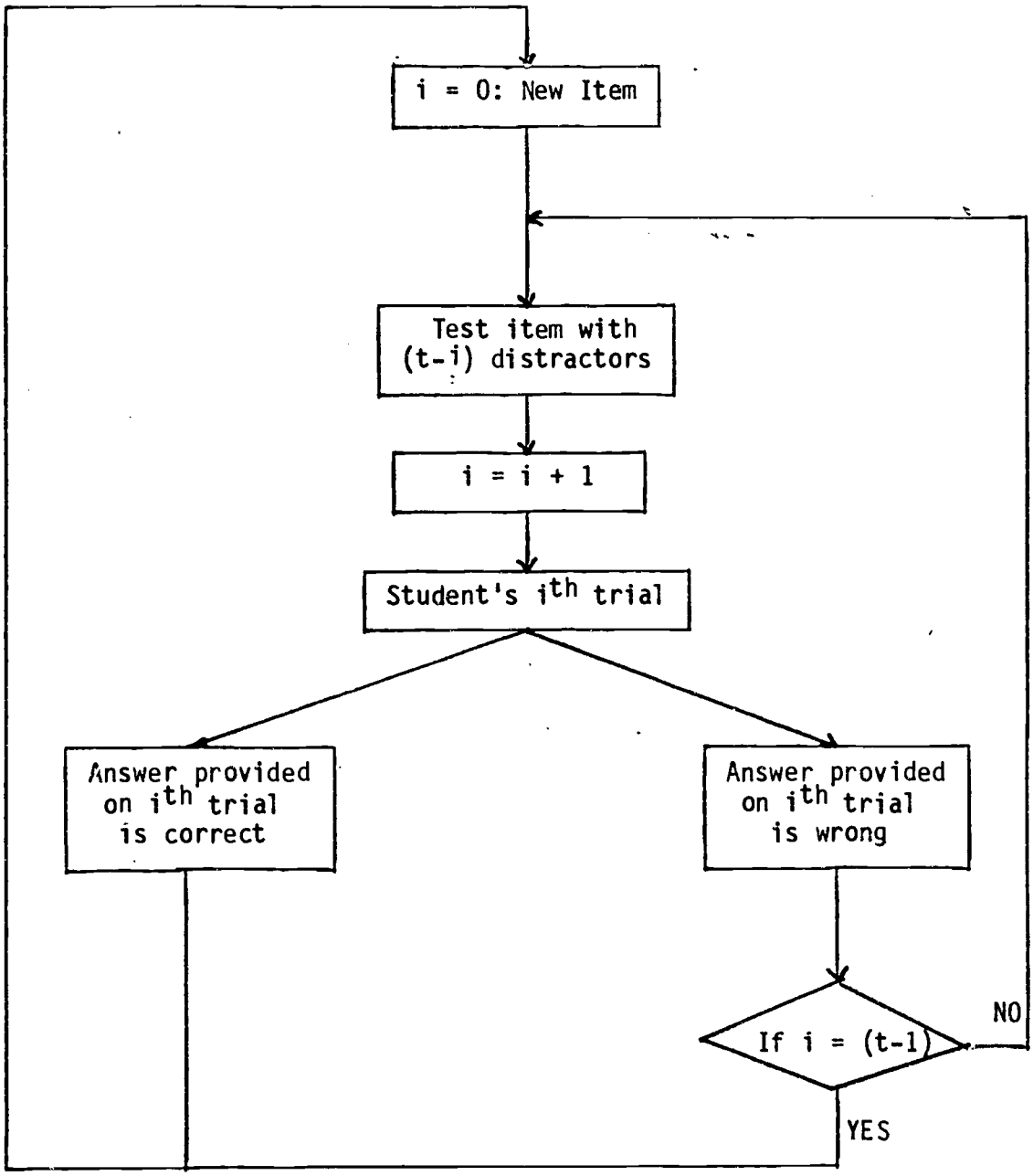


If students type <N> at this point, the question is presented again along with the available choices. For the first attempt of each item, students are asked about the level of confidence in their answer:



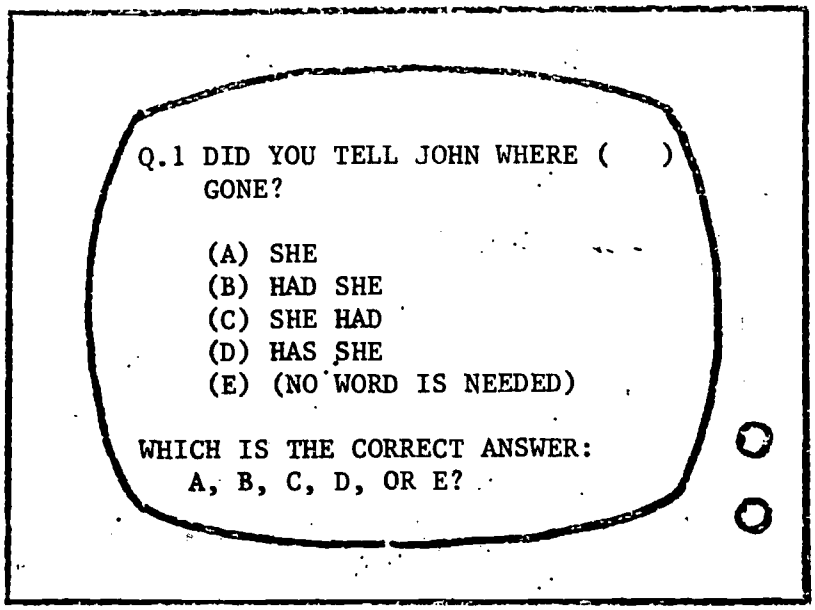
For subsequent attempts, the confidence-marking part is skipped.

After each response, students will receive feedback on whether they are correct or not. If the answer is correct, the next item will be presented. On the other hand, if the answer is wrong, students stay on the item. For each additional attempt, the answers previously selected are eliminated from the distractors remaining for that item. The answer-until-correct procedure can be illustrated by the following flow chart:

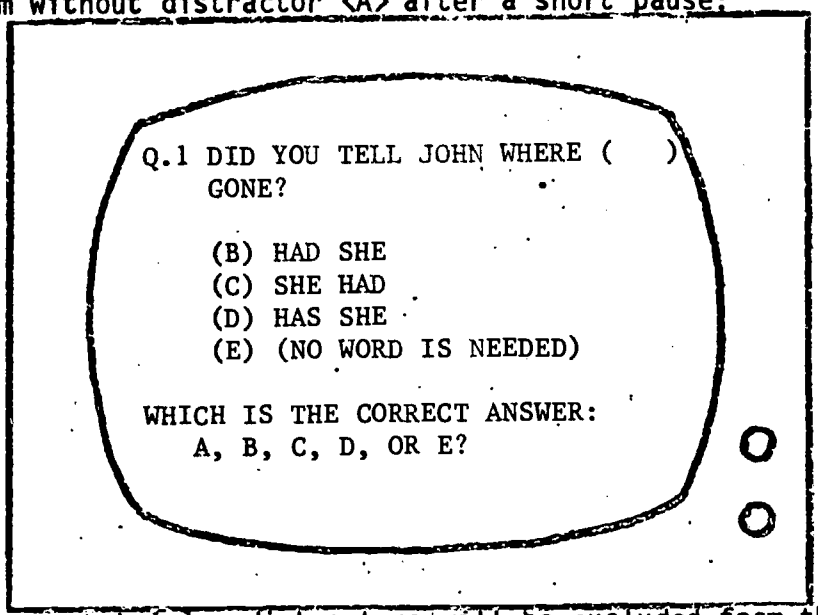


i: number of trials attempted by student
t: number of total distractors in an item

Using our example test item, this would proceed as follows. First a new item is presented:



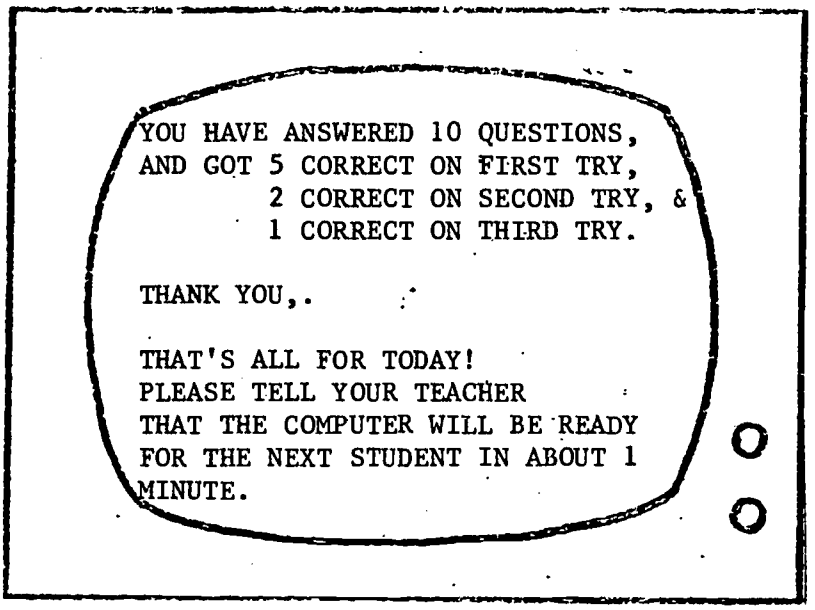
If answer <A>, which is incorrect, is selected, the student will see the same item without distractor <A> after a short pause:



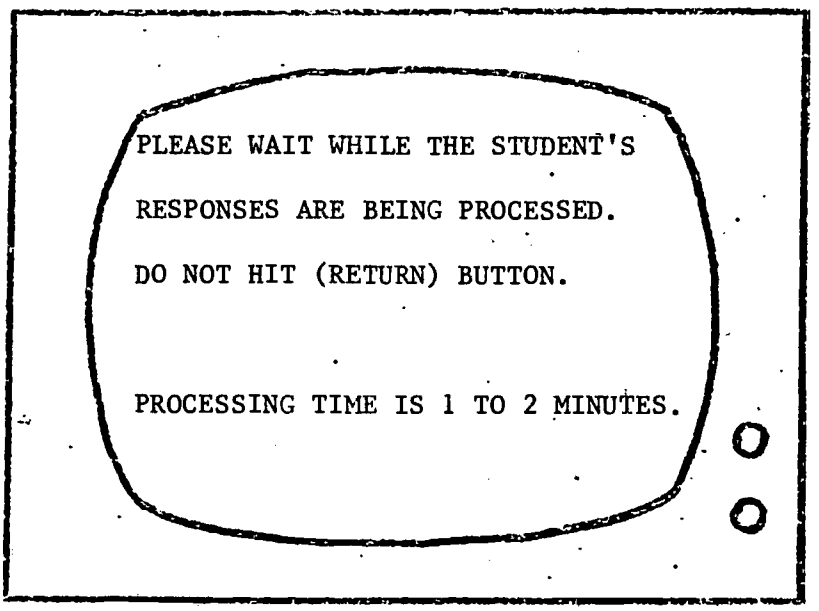
For the subsequent trials, distractors will be excluded from the available choices after they are selected.

Students are allowed to proceed to the next item under the following conditions: (1) the present item is answered correctly, (2) all the incorrect answers have been chosen, or (3) the response time is longer than the time limit allowed, which is 120 seconds.

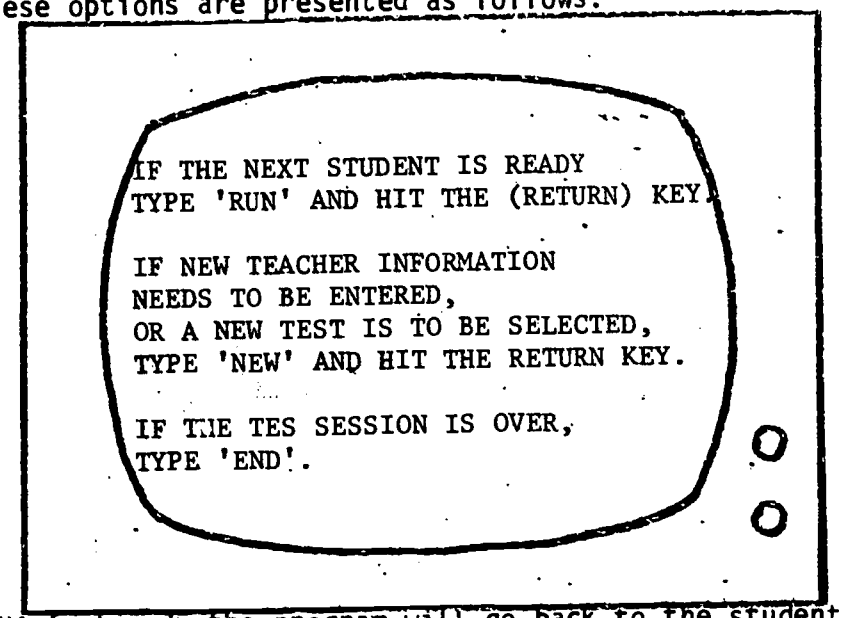
At the end of the test session, the screen will present a summary of the test results. For example, one student's results might be presented as follows:



These results will remain on the screen for about 45 seconds. After the elapsed time, the screen will then show the following message while the computer clears out old variable values from memory and stores student's responses on disk:

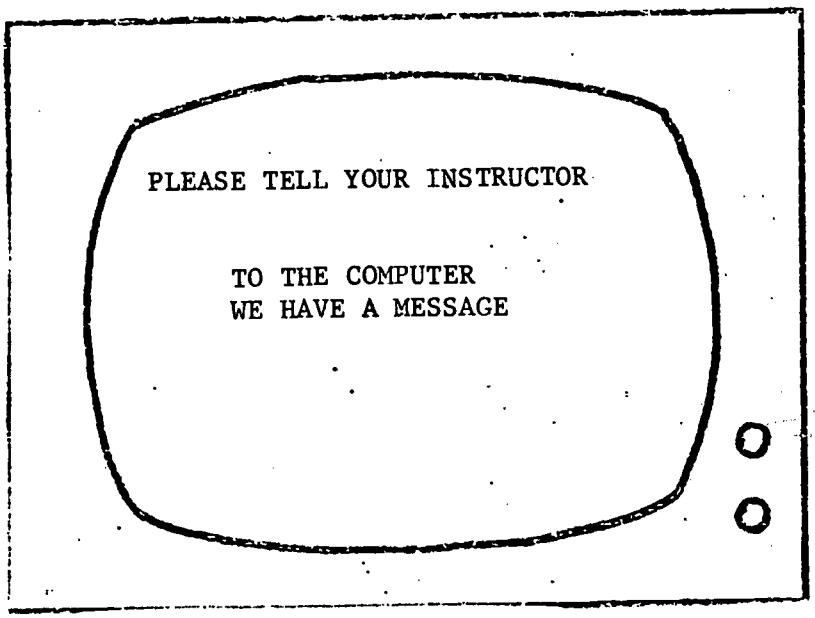


The <RESET> key should never be touched during this stage, otherwise the data of the student who just finished the test will be ruined. As soon as the data is saved, control of the program returns to the teacher. The teacher then has 3 options: (1) to run another student on the same test, (2) to select another test, or (3) to end the program. These options are presented as follows:

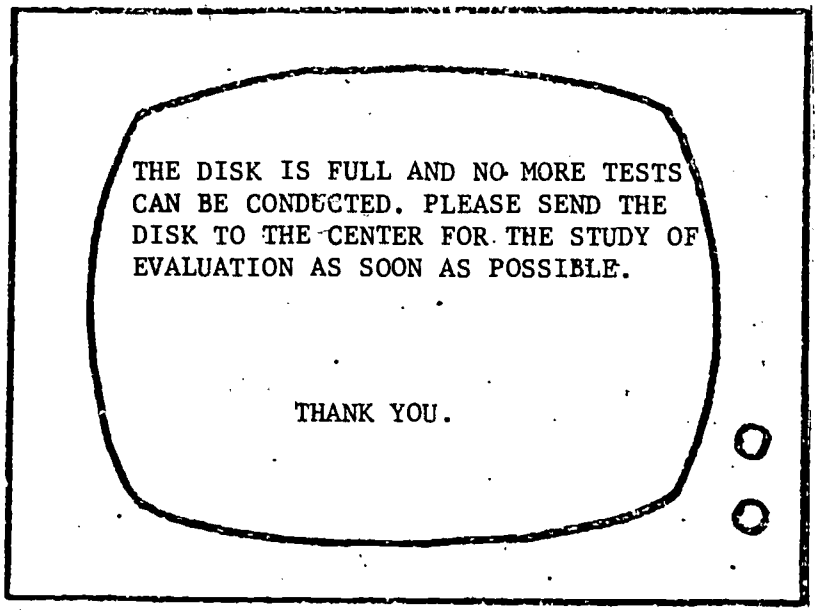


If 'RUN' is typed, the program will go back to the student session. If 'NEW' is typed, the program will go to the very beginning of the program when the teacher is asked to supply new parameters for a program run. The AUC program can be stopped by typing 'END'.

Another feature of the AUC program is the detection of whether there is enough space to store student data. If the disk is full, the following messages will be presented to the student:

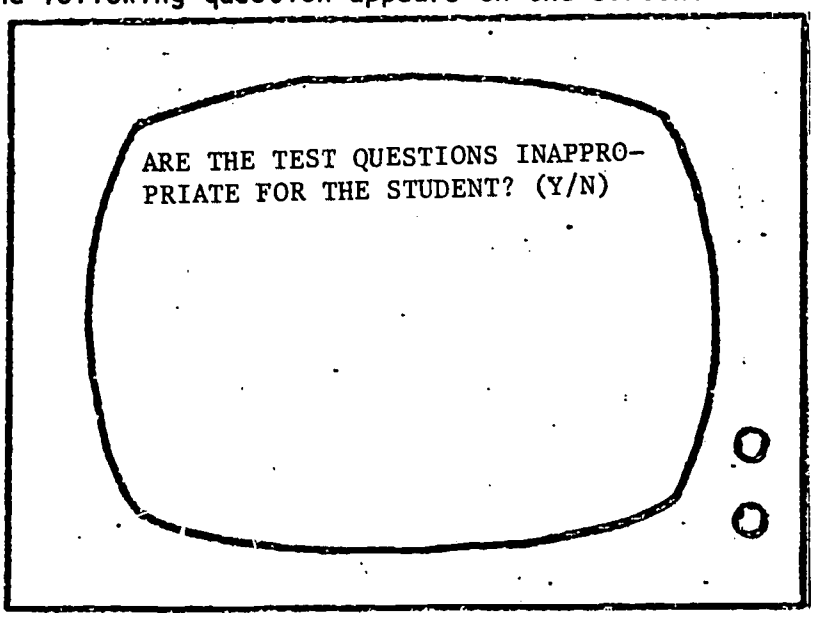


This message remains on the screen for 60 seconds and then the teacher will receive the following message:



In addition to the AUC response patterns and confidence level responses, the AUC program also keeps track of the time it takes a student to respond to each distractor. The maximum time recorded for each response is ninety-nine seconds. This should be adequate for most examinees since the average response time on the first trial is less than fifty seconds.

One last feature installed in the AUCMAIN program is that it allows teachers to interrupt a test when they feel the test being administered is inappropriate. If the teacher holds down the <CNTRL> button, and hits the <F> key at the same time right after an item is presented, the following question appears on the screen:



If the teacher responds 'YES', the program skips down to the last step of the program where teacher is given three options of running a new student, selecting a new test, or ending the program. If the teacher responds 'NO', then the program proceeds with the last question presented.

Output Files

After a test is administered to a student, an output file is created and named with the following format:

TESTNAME STUDENTNAME BIRTHDATE

For example:

MATH JOHN DOE 5/16/65

MATH II SALLY BUCK 4/21/67

In the first line of each output file are the student's name, the teacher's name, the school, student's birthdate, and grade level. A period is used as a separator character inserted between each variable (see Figure 2 for an example output file of the 10 item Math II test). Following the first line, are the student responses, one line per question. Up to k responses, where k is the number of question alternatives, are stored on a line in the same order as the student selected them. The last response in each line is always the correct answer. An exclamation mark ends each line. In the event that the student does not respond at all to a question, then only an exclamation mark will appear on the data line.

Following the response choices for the n questions are the response times in seconds, that it takes the examinee to select a particular alternative. Again, there is one line allocated per question. There is a one-to-one correspondence between each line of responses and each line of response times. Within a line, response

Figure 2

Example Examinee MATH II Output Produced
by a Single Run of AUCMAIN Program.

(Output File is Saved on Disk as MATH II YING LU 05/17/67.)

```
YING LU .CHU.UCLA.Ø5/17/67.9
 B!
 A!
 D!
 B!
 BC!
 E!
 D!
 !
 C!
 C!
 21.!
 7.!
 23.!
 23.!
 22.12.!
 22.!
 2Ø.!
 !
 23.!
 42.!
 1!1!1!1!1!1!1!!3!1!
```

times are separated by periods. Finally, in the last line of the output file are the confidence ratings for the first response to each question. Confidence ratings are only obtained for the student's first choice for each question, so there is only one rating per question. Ratings are separated by exclamation marks. It should be noted that in the example output file in Figure 2, the examinee did not respond to Question 8.

At the same time the output file is saved, the file name (including the test name, student name, and birthdate), is appended to a master file which includes the names of all examinees taking the same test on the same disk (an example file is shown in Figure 3). There is a master file for each available test named

AUC(testname)

For example:

AUCMATH

AUCMATH II

The master files are subsequently used to concatenate all responses for all examinees on all disks into a single data file for the purposes of overall analysis of test responses. The program which has been developed to do this is called AUCFILE and is described below.

Figure 3

Contents of AUCMATH II Master File of All Examinees
Taking MATH II Test on a Single Disk.

```
MATH II DELWIN CHIN APRIL 18
MATH II SEAN MOORE 7/20/66
MATH II FRANK DAMIANI 8/8/66
MATH II AARON SEELER 11/25/66
MATH II PEDRAM MADDAHIAN 2 2 79
MATH II ANNE HOLMES 9/2/66
MATH II YING LU  05/17/67
MATH II SHARON SMASON 6/9/67
```

Concatenation of Files with the AUCFILE Program

A program entitled, AUCFILE, has been created to concatenate the student files on a test into a single data file. After AUCFILE has been loaded into the computer, the disk or disks containing the student files are inserted into either Disk Drive I or II. When the AUCFILE program is run, the user is queried about which tests need to be concatenated. The program then uses the master files (created by the main program and updated with each test run) on the disks to control the reading and concatenating of student responses. The concatenated file is saved as (testname)DATA. For example:

MATHDATA

MATH IIDATA

These files are always written to the disk in Drive I.

The format of the file is such that responses, confidence ratings, and response times follow a fixed format. Each student has three records. The first record contains information about the student (name, teacher, school, birthdate, and grade). The second record contains the item responses and the confidence ratings. A column is allocated for each item alternative. Once a correct answer is selected, blanks are inserted for the remaining distractors. Following an item's responses is the confidence rating for the item. The third record contains the response times for each alternative. Two columns are allocated for each alternative, so the maximum possible time is 99 seconds. As with Record Card 2, blanks are inserted for the remaining alternative choices after the correct answer is selected. An example concatenated file appears in Figure 4.

Figure 4

Contents of MATH IIDATA: Concatenated and Formatted Responses of Examinees Taking MATH II Test from Several Different Disks

```
ERIK KNUTZEN.CHU.UCLA.8/13/65.12
B    2A    1D    1B    1C    1E    1D    1C    2C    1C    1
32        29        32        15        12        25        24        48
25        51
DELWIN CHIN.GINETTE.SUMMER.APRIL 18.7
B    2A    1D    2B    1C    2E    1CD   3DC   2DC   3C    2
Ø8        13        2Ø        16        Ø9        Ø7        1ØØ6      1712
3105      5Ø
SEAN MOORE.GINETTE.UCLA.7/2Ø/66.11
B    2A    2D    2B    2C    2E    2CBD  2BADEC2BDEC 3C    3
14        11        29        2Ø        16        43        371612    42Ø5Ø41ØØ
148120203 62
FRANK DAMIANI.CHU.UCLA.8/8/66.11
B    2A    1D    1B    1C    1E    1D    1C    1C    1C    1
19        13        26        Ø9        Ø7        13        19        45
22        92
AARON SEELER.CHU.UCLA.11/25/66.1Ø
B    1A    1D    1B    1C    1E    1CD   1BC   3BC   2BC   1
Ø8        Ø6        14        22        24        Ø6        1615      43Ø9
34Ø8      2226
PEDRAM MADDAHIAN.GINETTE.UCLA.2 2 79.7
B    1A    1D    1B    1C    1E    1CBAD 1DC   1BC   2BC   1
16        Ø5        35        21        14        15        26453131  7318
3Ø14      4532
ANNE HOLMES.GINETTE.UCLA.9/2/66.11
B    1A    1D    1B    1C    1E    1D    1C    2D    1     !
45        21        54        13        13        25        31        85
67
YING LU .CHU.UCLA.Ø5/17/67.9
B    1A    1D    1B    1BC   1E    1D    1     !C    3C    1
21        Ø7        23        23        2212      22        2Ø
23        42
SHARON SMASON.CHU.UCLA.6/9/67.9
B    1A    1D    1B    1BC   1DE   1D    1BAC  1BDC  1C    1
21        18        36        18        3513      2717      23        3ØØ71Ø
241706    43
MING TSENG.CHU .UCLA.Ø1/24/68.'8'
B    1DBCA 1D    1B    1C    1CDE  1CD   1DEBC 1EDABC1BC   1
19        12441212  46        17        14        18Ø9Ø6    2225      31270906
4814Ø6Ø6Ø13Ø41
SHEREE CHAN.CHU.UCLA.1Ø/31/66.1Ø
B    1DA   1D    1B    1C    1E    1D    2BDC  1BC   1C    1
Ø5        Ø716      35        16        Ø7        12        21        Ø12ØØ6
2817      3Ø
```

Creating New Files for Use as Input Tests to AUCMAIN Program

A program named FILEWRITER has been created to create input files for the AUCMAIN program. If any new tests are to be input into the program, the following format must be followed:

Line(s)	Contents
1	Title of test
2	Number of items
3	Number of choices per item
4 - 9	Directions for taking the test - up to six lines long. Dummy characters must be typed in lines not occupied by directions.
10 -	- Start in line 10 the stem of question 1: Q1 (item stem) - continue on next line as needed. Each line should not exceed 34 spaces in length.
	- Response alternatievs must begin with an open parenthesis, (,: (A) (distractor) - continue on next lines as needed. Each line should not exceed 34 spaces.
	- The correct answer must follow the last distractor of each question. it must be starred: *B
	- After the correct answer, start the next question on the next line (Q. 2).
	- Repeat until all questions are typed in.
	- End the entire file with a '!'.

The total possible lines for each question is 23 lines; within this limit, a stem or distractor can be up to 10 lines long. An example test following this format is presented in Figure 5. Unfortunately, one limitation of the FILEWRITER program is that commas may not be used anywhere in the file.

Figure 5

Contents of MATH II: Test File Input for AUCMAIN Program

```
MATH II
 10
 5
 CHOOSE THE BEST POSSIBLE ANSWER
 FOR THE FOLLOWING MATH QUESTIONS.
 YOU DO NOT NEED ANY MATERIAL OR
 CALCULATOR TO FIND THE CORRECT
 ANSWER.
 *
 Q.1  ONE SET OF FACTORS FOR 56 IS
 (A) 2*3*7
 (B) 8*7
 (C) 2*26
 (D) 4*13
 (E) 9*6
 *B
 Q.2  WHICH NUMBER IS THE MISSING
 FACTOR?
 2*2*    *8 = 64
 (A) 2
 (B) 3
 (C) 5
 (D) 8
 (E) 12
 *A
 Q.3  WHICH ONE OF THESE EQUATIONS
 IS TRUE?
 (A) (8*5) = (8+5)
 (B) (8+2)/4 = (4+2)/8
 (C) (6-2)*5 = (2*5)-6
 (D) (2+6)*5 = (5*8)
 (E) (5*6)+2 = (5*6)-2
 *D
 Q.4  WHAT IS THE MISSING NUMBER
 IN THE SEQUENCE?
 35; 31; 27;    ;19;
 (A) 24
 (B) 23
 (C) 15
 (D) 14
 (E) 11
 *B
 Q.5  WHAT IS THE NEXT NUMBER IN
 THE SEQUENCE?
 3;3;4;5;5;6;7;7;8;9;   ;
 (A) 11
 (B) 10
 (C)  9
 (D)  8
 (E)  7
 *C
```

```
Q.6  ANOTHER WAY TO REPRESENT
647 IS...
(A) 6 + 4 + 7
(B) (6+4+7)*100
(C) (6*10)+(4*10)+(6*10)
(D) (6*10)+47
(E) (6*100)+(4*10)+(7*1)
*E
Q.7  WHICH OF THE FOLLOWING PERIOD
OF TIME IS CLOSEST TO AN HOUR?
(A) 23 MINUTES 50 SEC.
(B) 36 MINUTES 58 SEC.
(C) 43 MINUTES 10 SEC.
(D) 71 MINUTES 12 SEC.
(E) 99 MINUTES  2 SEC.
*D
Q.8  MR. JONES LEAVES HIS HOUSE
EVERY MORNING AT 6.30 A.M. TO
GO TO WORK. HE HAS TO DRIVE
72 MILES AND HIS CAR AVERAGES
48 MILES AN HOUR. AT WHAT TIME
DOES HE ARRIVE AT WORK?
(A) 7.00 A.M.
(B) 7.30 A.M.
(C) 8.00 A.M.
(D) 8.30 A.M.
(E) 9.00 A.M.
*C
Q.9  YOU HAVE TO BUY LEMONADE FOR A
PARTY. EACH BOTTLE COSTS 75 CENTS.
HOW MANY BOTTLES WILL YOU BE ABLE
TO BUY IF YOU HAVE 10 DOLLARS TO
SPEND?
(A) 10
(B) 12
(C) 13
(D) 14
(E) 15
*C
Q.10 LAST MONTH JIM WORKED 3 HOURS
A DAY FOR 20 DAYS. HE WAS PAID 4
DOLLARS AN HOUR. HE ALSO BOUGHT
2 RECORDS FOR 8 DOLLARS EACH. HOW
MUCH MONEY DOES HE HAVE LEFT?
(A) 240 DOLLARS
(B) 232 DOLLARS
(C) 224 DOLLARS
(D)  80 DOLLARS
(E)  64 DOLLARS
*C
!
```

REFERENCES

Arnold, J.C. & Arnold, P.L. On scoring multiple-choice exams allowing for partial knowledge. The Journal of Experimental Education, 1970, 39, 8-13.

Coombs, C.H. On the use of objective examinations. Educational and Psychological Measurement, 1953, 13, 308-310.

Cross, L.H. & Thayer, N.F. A new method for administering and scoring multiple-choice tests: Theoretical and empirical considerations. Unpublished manuscript, Virginia Polytechnic Institute and State University, 1979.

Ebel, R.L. Blind guessing on objective achievement tests. Journal of Educational Measurement, 1968, 5, 321-325.

Gilman, D. & Ferry, P. Increasing test reliability through self-scoring procedures. Journal of Educational Measurement, 1972, 9, 205-207.

Hanna, G. Incremental reliability and validity of multiple-choice tests with an answer until correct procedure. Journal of Educational Measurement, 1975, 12, 175-178.

Hritz, R.J. & Jacobs, S.S. Risk-taking and the assessment of partial knowledge. Paper presented at the Annual Meeting of the American Psychological Association, Miami Beach, Florida, September 1970.

Kane, M. & Moloney, J. The effect of guessing on item reliability under answer until correct scoring. Applied Psychological Measurement, 1978, 2, 41-49.

Koriat, A., Lichtenstein, S., & Fischhoff, B. Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 1980, 6, 107-118.

Rippey, R. & Donato, J. Interactive confidence test scoring and interpretation. Educational and Psychological Measurement, 1978, 38, 153-157.

Shaughnessy, J. Confidence-judgment accuracy as a predictor of test performance. Journal of Research in Personality, 1979, 13, 504-514.

Sieber, J. Confidence estimates on the correctness of constructed and multiple-choice responses. Contemporary Educational Psychology, 1979, 4, 272-287.

Taylor, J., West, D., & Tinning, F. An examination of decision-making based on a partial credit scoring system. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Wilcox, R.R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414.

Wilcox, R.R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74.