

Improving the quality of computer scored tests in agronomic courses¹

R. R. Weil and W. Kroontje²

ABSTRACT

Student evaluation can be greatly facilitated by using computer scored and analyzed multiple choice tests. Test parameters can aid the instructor in the interpretation of student scores, and item analyses can provide him with valuable feedback for improving both teaching and writing. The caliber of testing influences the caliber of learning, and at college level it is more appropriate to evaluate complex learning outcomes than it is simply to test recall of information. Therefore we are justified in making an effort to evaluate attainment of higher levels of knowledge on multiple choice tests and to de-emphasize rote memorization. This paper describes ways of increasing the quality and usefulness of computer scored tests by examples from an introductory soil science course. The authors use test and item analysis; define the variable to be tested; and make a table of specifications to assure that the test reflects a balance, characteristic of the course, of both subject matter and level of knowledge. Most of their examples show how to write better test items, but they also give examples of ways to reduce the guessing factor and how to increase the value of the distractor. Some techniques for writing test items for high-level knowledge are: 1) using terms in problems rather than defining the terms, 2) presenting novel rather than familiar situations, 3) using problems that require applying knowledge, and 4) asking for interpretations of data requiring both knowledge and reasoning.

Additional index words: Examinations, Test construction, Item analysis, Multiple choice testing, Complex learning outcomes.

AT MANY land-grant universities, enrollments in basic agricultural science courses have increased dramatically over the past few years. Classes have jumped in size from 20 or 30 students to "megalectures" of several hundred students. Teaching methods have had to undergo considerable modification in order to cope with fundament-

al changes in the classroom situation. The challenge is to maintain or improve the quality of instruction and validity of student evaluation in these large classes. Among the diverse approaches taken by agronomist educators, certain practices seem to have gained widespread acceptance. One evaluation tool is the use of computer-scored achievement tests. At best this type of testing can result in fast, fair, and accurate evaluation of student achievement and readily accessible feedback on the progress of the teaching process. At worst it reduces testing and achievement to the level of rote recall of unconnected facts. This paper discusses ways of increasing the quality and usefulness of computer scored tests used in agronomic education. It is based largely on the authors' experiences with the introductory soil science course at Virginia Polytechnic Institute and State University (VPI & SU).

Computerized Test Analysis

In addition to simply compiling student scores, the computer can be used to analyze certain qualities of the test itself. The Learning Resources Center at VPI & SU analyzes the test parameters shown in Table 1 for each test form administered.³ Of these parameters, the mean number right, the standard deviation of the number right, the KR-20 reliability estimate (6), and the standard error of measurement warrant particular attention. Unless there is a penalty for wrong answers, the mean number omitted should be near zero. The effectiveness of the test in ranking examinees according to their

Table 1—A typed copy of the whole-test analysis portion of the output of the Learning Resources Center Test Scoring and Analysis Program. (Fall 1974 Final Exam in Introductory Soil Science (AGRN 3110), consisting of 80 test questions.)

Total score statistics by forms	Test forms	
	A	B
No. of students	103	111
Mean no. right	44.02	42.89
SD of no. right	11.737	9.471
Mean no. omitted	0.08	0.23
SD of no. omitted	0.332	0.536
Reliability estimate (KR-20)	0.894	0.832
SE of measurement	3.827	3.886

¹Contribution of the Agronomy Dep., Virginia Polytechnic Institute and State Univ., Blacksburg, VA 24061.

²Instructor and professor of agronomy, respectively.

³A user's guide to the VPI & SU test and analysis system. Learning Resources Center, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

knowledge is greatest if the mean number right is halfway between the score expected from chance guessing and a perfect score. For example, on a test with an average of five choices per question, chance guessing should result in 20% correct and thus a class average of 60% $(100\% \text{ to } 20\%)/2 + 20\%$ would be most desirable. The accuracy with which the test can rank examinees is enhanced by having a large standard deviation of the number right (i.e., a wide spread of scores) and a small standard error of measurement. The KR-20 is based on internal inter-item correlations between the test items and is an estimate of how similar the scores would be if the same group of students were tested again on the same subjects (6). The KR-20 is only valid if all the items measure similar abilities so that students doing well on one part of the test would probably do equally well on another. Where this is not the case, for instance if half the items involved calculations and half involved definitions, two separate but more homogeneous tests would provide a more valid reliability estimate. Also, longer tests tend to give a more reliable measure of student achievement than shorter ones. An instructor may want to consider the reliability estimate and standard error of measurement on test scores when making distinctions between students.

The Item Analysis

The computer's statistical analysis, then, not only helps the instructor evaluate student achievement, but also aids him in evaluating how successful he was in constructing the test. This role of computer scoring is illustrated in the item analyses of Example 21 in this paper (Table 2). Given for each response are the number of students, their percent of the total class, and a discrimination index (R) showing the correlation between each student's response to that item and his total score on the test. The percent of the class choosing the correct answer can indicate to the instructor the relative difficulty of the item. This parameter also provides useful feedback to the instructor on misconceptions (wrong responses) most prevalent in the class. If no students choose a response, no information is gained by including that response in the item.

The discrimination index (point biserial correlation) is a measure of how well the item discriminates between the good and the poor students. A correlation of 0.30 or higher for the correct response is desirable. Negative index numbers are expected for the wrong response if students marking these incorrect choices received low total scores on the exam. A low or negative index for the correct response suggests that the poor students tended to

Table 2—The item analysis for Example 21. (Adapted from output of the Learning Resources Center Test Scoring and Analysis Program.)

Response	(1)	(2)	(3)	(4)†	(5)
No. students	7	1	1	79	15
% of class	7	1	1	77	15
Disc. index (R)	-0.29	-0.07	+0.03	+0.57	-0.47

† Correct response.

get the item right while the better students tended to miss it. In our experience a negative discrimination index for the keyed response, or positive indexes for the other responses, generally indicate 1) that the question was misinterpreted by the good students (usually justifiably so) or 2) that an error was made on the key. Though the item analysis may be useful for identifying effective discriminators the statistics are of limited reliability if the question is too easy or too difficult. The analysis is not trustworthy if more than 85% or less than 30% of the class got the question correct or if the class size is much less than 50 students. In general, items in the midrange of difficulty provide the most information on differences in student achievement.

The computer printout includes a list of students by name and/or identification number with their raw scores, percent scores and T-scores.⁴ This information, and the printout histogram of the frequency distribution of raw and T-scores, are very helpful in assigning letter grades to students and keeping records of test grades.

The data compiled by the computer are, of course, only as meaningful as the tests administered. Writing a good multiple choice test takes considerably more effort than writing short-answer or essay tests. Thus, much of the time saved in scoring the test should be devoted to writing the test. This effort is justified both for evaluation and by the observation that the caliber of the exams determines the quantity and quality of studying invested by the students. Gronlund (4) states that "the anticipation of a test arouses greater learning activity and the nature of the expected test channels and directs the type of learning that takes place." Although the student-teaching relationship in a large class is generally less personal than in a small class, the evaluation of student performance and grading remain important to their motivation and careers. The agronomist-instructor has, therefore, an obligation to strive to be as scientific about evaluating

⁴The T-score (X_T) is calculated by transforming the raw score (X_O) to its equivalent in a normalized distribution with a mean of 50 and a standard deviation of 10. $X_T = (\sigma_t/\sigma_o) X_O - [\sigma_t/\sigma_o] \bar{X}_O - \bar{X}_t$; where σ_t and σ_o are the standard deviations of the raw scores and T-scores distributions, respectively and \bar{X}_O and \bar{X}_t are the means of the raw scores and T-scores respectively (1).

Table 3—A simplified table of specifications for a 30-item test in which higher-level learning outcomes are emphasized

Subject area	Level of knowledge				Total
	Recall	Comprehension	Application	Synthesis	
	No. of items				
Clay structure	1	3	--	1	4
Clay genesis	1	1	--	1	3
Physical properties of clays	--	3	1	1	5
Cation exchange	--	3	--	2	5
Anion exchange	--	3	--	--	3
Reserve acidity	--	1	1	1	3
Cation exchange capacity of soil	1	1	2	1	4
Base sat'n	--	1	2	--	3
Total	3	15	6	7	30

his student's performance as he is in evaluating soil types or crop varieties. Computer scoring and analysis of tests can be a valuable tool to achieve this goal.

Defining the Variable to be Tested

One of the first steps in testing student achievement is to identify the variables to be measured. Generally, this is best done in conjunction with planning the behavioral objectives of the course. Tests may not be the most pleasant aspect of teaching, but they are strong motivators for learning. A student's approach to studying is often molded by what he perceives the tests will be like. Therefore, an appropriate choice of variables to be measured by a test is a significant part of the teaching process.

While specific test items deal with particular subject matter areas, they generally measure one or more general variables or types of learning outcomes. Bloom et al. (2) refers to these as levels of knowledge and arranges them in a hierarchical order. The least sophisticated of these levels of knowledge is memorization of information. In a multiple choice test this may be further reduced to simple recall. Insofar as the ability to recognize words and phrases often has a very short half life and may not be indicative of a true understanding of the ideas involved, test items which measure only this low level of knowledge are of little value. Yet, the authors' informal observations of teacher-made tests in agronomy and other fields suggests that the majority of multiple choice test items merely measure rote memorization. This tendency is often cited by critics of machine-scored tests who justifiably consider the memorization of facts by itself to be an inappropriate educational objective for a college course.

However, multiple choice tests can be written to measure much more sophisticated educational objectives. They can measure understanding of concepts and comprehension of why things happen.

They can require students to apply principles in problem solving, interpret scientific data, and use factual knowledge in reasoning and making judgments. If these types of behavior comprise the objectives of the course, then test items should be carefully constructed to measure them. Rote memorization and recall should be asked for only where it is a legitimate educational objective.

Making out a table of specifications before writing the test ensures that the test reflects both the balance of subject matter and levels of educational objectives in the course. The procedure also helps the instructor guard against a tendency to write items that test simple recall. Table 3 is the table of specifications for a test in a soil science course. The example in Table 3 is relatively simple and more or fewer distinctions can be made in categorizing both levels of knowledge and subject matter areas according to the preference of the instructor. Although writing a table of specifications may seem to be a formality, the authors have found that the practice has had a useful disciplining effect on test writing. Even with a table of specifications to guide him, the instructor must be on guard against the tendency to test on "picky" detail, which though educationally insignificant, often presents the most obvious possibilities for easily written multiple choice items.

Writing Better Test Items

With the application of a little imagination, questions can be devised to measure almost any type of educational objective. The following discussion presents a number of guidelines and examples. We have found these techniques to be effective in constructing multiple choice test items that measure variables described in our tables of specifications.

Where the variable to be measured is the student's comprehension of factual information, the test item can be written simply. Still, care should be taken not to use the same phrases or examples as those in the textbook or lectures. The student should comprehend the information and not just recognize a familiar sounding sentence. This is particularly true for definitions of terms. Since the term cannot be easily paraphrased, it should appear in the stem of the question with a paraphrase of the textbook definition as one of the several alternative answers. This approach is also consistent with the idea that knowing the nature of a phenomenon is probably more educationally relevant than knowing its name. Thus, testing knowledge of the definition of mycorrhizae is better approached in Example 2 than in Example 1 below:

Example 1.

Symbiotic associations of certain fungi and plant roots are called

- (1) rhizobia
- (2) nodules
- (3) actinomycetes
- (4) mycorrhizae

Example 2.

What are mycorrhizae?

- (1) Root excretions which can make nutrients more available to plants.
- (2) A class of plant-parasitic nematodes.
- (3) Microorganisms found in the soil which produce antibiotics useful in medicine.
- (4) A mutually beneficial partnership between the roots of some higher plants and certain fungi.
- (5) The microscopic feeder roots or root hair responsible for most nutrient uptake.

To go one step further, a test item might ask the student to recognize the importance of a phenomenon, not only to give its definition:

Example 3.

The beneficial effects of mycorrhizae on plant growth are especially evident where

- (1) soil moisture is very low.
- (2) soils are of low fertility.
- (3) leguminous crops have been grown.
- (4) soil compaction has been a problem.
- (5) poor irrigation practices have increased soil alkalinity.

While every scientific discipline has its lexicon of technical terms which must be mastered, the emphasis can be placed upon the ability to use these terms, not merely to recall their definitions. To answer the question in Example 4, one must know about summer fallow, runoff, evaporation, etc., as well as to know how they are related.

Example 4.

Summer fallow is used primarily to reduce the loss of soil moisture by

- (1) run-off
- (2) evaporation
- (3) transpiration
- (4) percolation

Using a somewhat novel approach to a subject favors those students who most thoroughly understand the material. For example, in lecture, nitrification might be defined by the reaction:



A common approach to testing comprehension of this process might be to ask the student to choose the proper reaction equation from a list of alternatives. However, a more effective (and more interesting) test item of a high level of comprehension would not repeat the lecture presentation, but might involve analysis and application of knowledge, in Example 5.

Example 5.

Suppose a new herbicide being tested was found to inhibit the activity of only the particular species of bacteria which forms NO_3^- in the soil. What compound would most likely be first to accumulate to an unusually high concentration in a soil treated with this new chemical?

- (1) ammonium
- (2) nitrate
- (3) amino acid
- (4) nitrite
- (5) nitrosamine

The caveats about rote memorization and testing on insignificant details are especially relevant to matching items. These are really variations of the multiple choice format and can be easily adapted to machine scoring. A series of statements or questions are listed in a left-hand column and labeled with consecutive item numbers. In a right-hand column the list of alternatives appears as would the choices for any other multiple choice item. The matching format is effective only if the alternatives are homogeneous enough so that to the intelligent, but unprepared student, each one could plausibly be matched with any of the items in the left-hand column. To avoid confusion, a brief statement should appear in the test identifying the following items as being of the matching type and indicating that any given choice may be used any number of times. Below are illustrations of a poor series (Example 6) and a good series (Example 7) of matching type items.

Example 6.

- | | |
|---|---------------|
| 1. A variety of soybean. | (1) O, Ca, Mg |
| 2. A poorly drained soil series in Indiana. | (2) Wayne |
| 3. Elements abundant in dolomite | (3) N, P, K, |
| 4. The "fertilizer elements." | (4) Brookston |

Example 7.

- | | |
|---|---------------------|
| 1. highest C.E.C. | (1) Allophane |
| 2. least weathered | (2) Montmorillonite |
| 3. most common in volcanic soils | (3) Feldspar |
| 4. most commonly responsible for "fixing" P | (4) Gibbsite |
| 5. greatest shrink-swell capacity | (5) Kaolinite |

Problem items involving calculations are often useful in determining how well students can apply certain concepts and principles. Again, the problem should be presented from a slightly different angle than that to which the students have been previously exposed. Even with the widespread use of electronic calculators, the mathematics of the problem should be kept as simple as possible unless the purpose of the item is to measure mathematical skills. Information which the student might think necessary to work the problem should be given if memorization of this information is not being

tested. Students who understand the definition of pH, the concept of gram molecular weight, and the meaning of the ionization constant (K_w) should be able to correctly answer the item in Example 8 with almost no calculations.

Example 8.

A soil solution with 1.7×10^{-8} grams of OH^- ions per liter
 a solution with 1×10^{-5} grams of H^+ ions per liter (Note that atw of O = 16; atw of H = 1; Avogadro's number = 6.023×10^{23} ; $K_w = 10^{-14}$).

- (1) is more acid than
- (2) is more basic than
- (3) has a lower pH than
- (4) has the same pH as
- (5) both 1 and 3

The difficulty of such items can be regulated to some degree by providing more or fewer facts, formulas, and hints. Extraneous information may also be given to increase the difficulty. In Example 8 "Avogadro's number" is irrelevant to the solution of the problem and the knowledgeable student should recognize it. It may be best to supply the information needed to work all the problems of the test at the end or beginning of the test, especially if all the items called for calculations can be put on the same page with that information. In this way the student must decide what is relevant to a particular problem and what is not.

The ability to judge the relevancy of information as well as its truth or fallaciousness, is one of the main skills tested by essay questions. In a multiple choice format this ability can be measured by writing items that have true statements for both distractors and correct answers. In Example 9 below, each choice is a true statement of information the student should be familiar with, but only number 2 explains the phenomenon in question.

Example 9.

Why may some plants growing in a loam soil wilt on a hot, sunny day following several days of rainy weather?

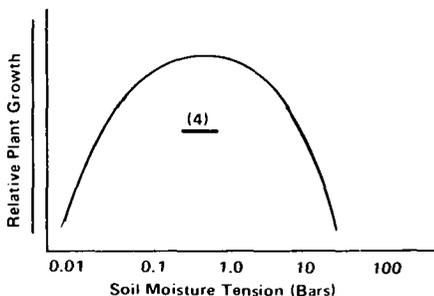
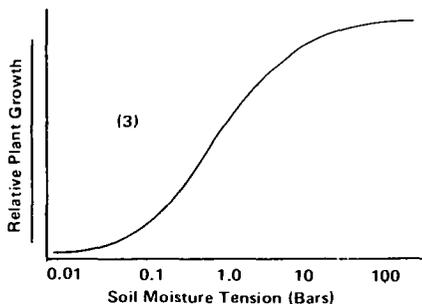
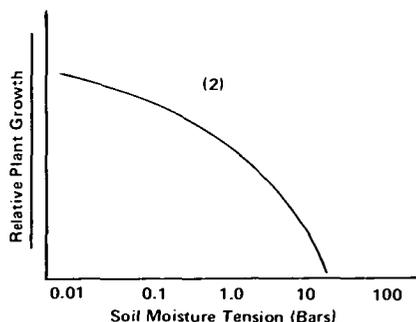
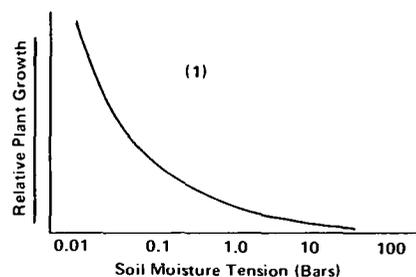
- (1) Many plants wilt at about 15 bars soil moisture tension.
- (2) Atmospheric energy, instead of soil moisture energy, may determine plant moisture status.
- (3) Stomata tend to close in the presence of a steep water potential gradient between the interior and exterior of the leaf.
- (4) In many plants an increase in temperature affects respiration rates more than rates of photosynthesis.
- (5) Field capacity is often reached on loam soils within a day or two of a rainstorm

An important aspect of any scientific study lies in the understanding of the relationships between variables. Although the theoretical relationships may be explained verbally in lecture, they can often also be described graphically. Perception of important relationships may, therefore, be measured

by items having various graphs as possible answers. They may be very simple choices between "increasing" or "decreasing" or they may involve certain significant nuances as in the following example. Here it is important that the student realize that plant growth is generally retarded by both very dry and very wet soil conditions. While only the last choice illustrates this relationship, all four curves have shapes commonly encountered in biological work, thus minimizing guessing on the basis of recollection of a familiar shape.

Example 10.

The relationship between the growth of most plants and soil moisture tension is best illustrated by graph number _____.



Another method to determine understanding of relationships is to ask the student to predict the effect of changing certain variables based upon the principles taught in the course. If a student has been taught that cultivation of row crops tends to destroy granular structure and decrease pore space, Example 11 would measure his comprehension of these concepts and of the relationship between bulk density and porespace.

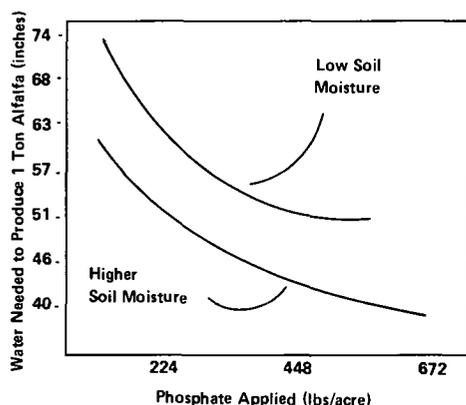
Example 11.

The bulk density and % pore space of a silt loam on which corn has been cropped in the conventional manner for the last 12 years are 1.52 g/cc and 43%, respectively. If you measured the variables on an adjacent plot which had been left undisturbed, which of the following would you expect to find?

	Bulk Density	% Porespace
(1)	1.52	43%
(2)	1.57	48%
(3)	1.46	45%
(4)	1.56	39%
(5)	1.47	40%

While graphs, tables, and diagrams can be effectively used as alternative answers to an item, they can also be used in the question part of an item to determine the student's ability to interpret scientific data. This may be done with data familiar to the student, thus emphasizing that understanding the data, not memorizing it, is most important. The figure in Example 12 was adapted from the course textbook (3), but the interpretations asked for in the test item are not those discussed in the textbook.

Example 12.



Examine the above figure and decide which of the following statements are supported by the data?

- (1) Phosphorous fertilizer was utilized more efficiently at the lower soil moisture level.
- (2) Phosphorous, but not water, was a limiting factor at *both* levels of moisture.
- (3) The response to added water would be greater at lower levels of P.
- (4) Compared to increasing from 224 to 336 lbs. of P, switching to the higher level of moisture

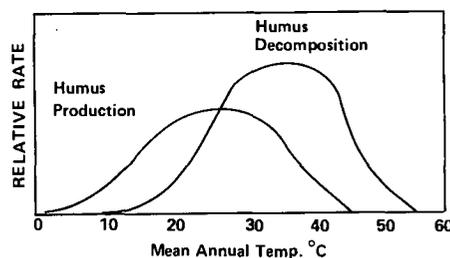
would have a *smaller* effect on the amount of water needed to produce 1 ton of alfalfa.

- (5) Water use efficiency increased as phosphorous fertilizer was added under *both* soil moisture regimes.

An item like Example 12 obviously cannot be answered after a quick glance, but must be read thoroughly and carefully analyzed with respect to the accompanying figure. The time needed to answer items that test a considerable depth of understanding should be taken into consideration when deciding on the length of the test and the time allotted. Our experience has been that while 40 or 50 simple multiple choice items are appropriate for a 50 min exam period, only about 30 items of a more complex nature can be answered by most students in that time. Having too many items on a test rushes the students and increases the incidence of mindless guessing. This especially reduces the effectiveness of difficult or complex items since the test-wise student who answers all the simple questions first will be too rushed to think carefully when he gets around to the items which require the most thought.

In Example 12 the student must consider the data from a number of viewpoints and decide which of a diverse set of statements is supported by the data. A simpler approach to writing an item involving data interpretation is illustrated by Example 13. Here only one characteristic of the data need be analyzed and the answers can be read much more rapidly. However, these data did not come from the course textbook and the student had no previous exposure to the graph. An important objective of our soils course is to train young professionals to be able to interpret data and communicate within the discipline. To answer Example 13, the student must have a general conception of the meanings of humus production and decomposition and their relationship in determining the amount of

Example 13.



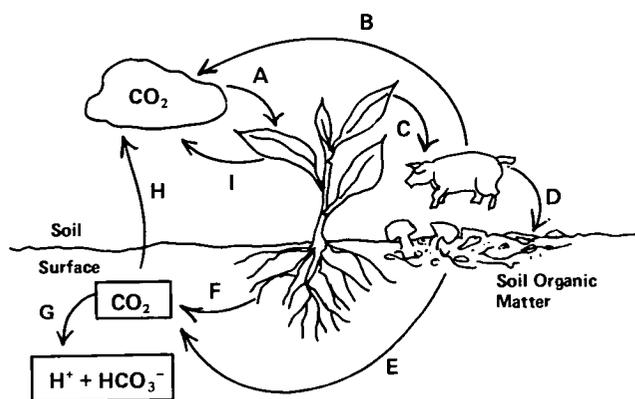
The data used to draw the above figure pertain to aerobic soils. At what mean annual temperature is the humus content of such soils the greatest?

- (1) 10°C
- (2) 15°C
- (3) 25°C
- (4) 30°C
- (5) 40°C

humus accumulation in the soil (i.e., the difference between production and decomposition). In addition, the item calls for some skill in reading graphs.

Often diagrams are more efficient than words for describing relationships in nature. Also, for concepts which were taught verbally, a visual presentation may be useful in establishing that degree of novelty necessary to de-emphasize the role of rote recall. In place of written definitions, a student might be asked to match number features on a diagram (e.g., of a soil profile, stem cross section, etc.) with one or more technical terms. Example 14, below, uses a simple diagram to test the student's knowledge of the role of respiration in the carbon cycle and the primary carbonaceous product of respiration, CO_2 .

Example 14.

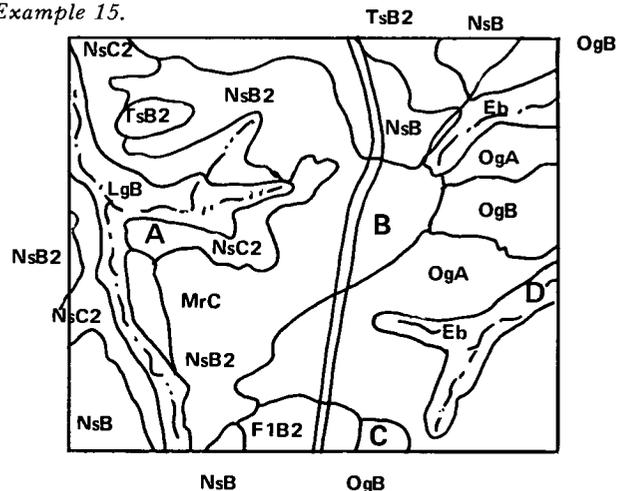


Which of the above processes represents respiration?

- (1) A, C, I, and F
- (2) B, F, and D
- (3) A, B, G, and H
- (4) A, B, C, and H
- (5) B, E, F, and I

Where the concepts illustrated by a diagram (or data set) are important enough to warrant more than one test item, a series of questions may be asked in reference to a diagram in what Gronlund (5) calls an interpretive exercise. For example, if the ability to utilize soil survey reports is a major course objective, the following type of item series might be useful. Note that a copy of the appropriate Soil Survey Report would be available to the student. Thus, rather than being asked to memorize any information, he is asked to use the published information to make judgements.

Example 15.



A tracing of a portion of sheet 8 of the Orange County, Virginia, Soil Survey Report.

1. The area outlined above represents about _____ acres.
(1) 25, (2) 250, (3) 575, (4) 900, (5) 1,400
2. Point _____ would be a good pond site.
(1) A, (2) B, (3) C, (4) D
3. The A horizon is probably the shallowest at point
(1) A, (2) B, (3) C, (4) D
4. This area is in the _____ region.
(1) Ridge and valley, (2) Coastal Plain,
(3) Piedmont, (4) Highland plateau
5. The land owner wants to sell several adjacent homesites with a minimum of loss in his potential corn production. The most practical and best suited area for the homesites would be near point
(1) A, (2) B, (3) C, (4) D

Of course, the visual approach to item writing need not be limited to simple diagrams. Reproductions of photographs or crop distribution maps can be used in test items. If the classroom environment permits, several items on a test might refer to color slides to be shown at the beginning or end of the test period. This is particularly useful for the identification of plant diseases or crop species, etc. Actual specimens can also be used in conjunction with multiple choice tests, if the class size and logistics permit suitable access by each student. For instance, five different numbered soil monoliths could be set up in the classroom. Several test items could be written asking students to choose the soil best suited for septic systems; the soil most probably from a floodplain; the soil with a platy structure; etc. Even where the use of such visual presentations is impractical, test items can be made more interesting and educationally relevant by posing practical problems likely to be encountered on the job. These "story problems," if they are not long and cumbersome, can provide a means of asking a question indirectly enough so as to avoid sug-

gesting the answer to the less knowledgeable student. This is more similar to the circumstances in which information is supplied in real life than is the very simple, direct question. For instance, Example 16 asks a question about soil consistence similar to Example 17, but more narrowly focuses the student's attention.

Example 16.

The terms hard, friable, and plastic are used to describe the consistence of a soil at different _____ respectively.

- (1) organic matter contents
- (2) moisture contents
- (3) pH values
- (4) degrees of compaction
- (5) temperatures

Example 17.

A highway engineer wanted to know the consistence of the subsoil at a certain location on a proposed highway route. He went to the spot in question, dug down to the B horizon, picked out 6 clods of soil and mailed two in a cardboard box to each of 3 soils consultants. Within a week he got his answers. One classified the clods as "hard", another said they were "friable" and the third reported the consistence as "plastic". The discrepancy in the replies was probably due to

- (1) differences in subjective judgement.
- (2) temperature differences.
- (3) moisture differences.
- (4) the use of different systems of classifying consistence by engineers, agronomists and geologists.
- (5) the rapid oxidation of organic matter.

Reducing the Guessing Factor

The foregoing discussion centered on ways of making multiple choice test items educationally meaningful. But even a question well designed to test a high level of knowledge of a relevant concept will be of little value if a large proportion of the responses are the result of chance or test-taking savvy. Whether the examinee responses are based on the educational variables meant to be tested or on irrelevant clues and blind chance is largely determined by the nature of the choices provided. The number of choices obviously will influence the probability of guessing the correct one and the ability of the item to discriminate between good and poor students. One of the disadvantages of True-False items is that there is a 50% chance of guessing the correct answer. This probability can be reduced (and often the level of learning outcome tested can be increased) by asking the student to choose whether the statement is true or false *and why*. In the example below the probability of guessing the correct answer has been reduced to 25% and there has been an increase in the amount of potential feedback on why some students answered the item incorrectly.

Example 18.

Where all other soil forming factors are the same, mineral soils formed under deciduous forests generally have *more* humus in their profiles than soils with native grassland vegetation.

- (1) True, because woody tree roots do not oxidize as rapidly as fine grass roots.
- (2) True, because deciduous trees return their leaves to the forest floor each year.
- (3) False, because organic matter decomposes less rapidly in grassland climatic regions than in forest climatic regions.
- (4) False, because compared to a grassland, a forest has a greater proportion of its organic matter present in the aboveground standing vegetation.

Other questions which require only bivariate answers (e.g., yes or no, increases or decreases, etc.) can be ameliorated by combining several questions into one to increase the item's discriminatory power and reduce the probability of guessing. Another advantage of this technique is that it often requires the student to synthesize information learned at different times during the course. Example 19 involves a synthesis of four concepts: (1) the increased solubility of manganese with lowered pH, (2) the increased solubility of the reduced form of manganese over the oxidized, (3) the effect of drainage on soil aeration, and (4) the role of anaerobic bacteria in reducing manganese.

Example 19.

Examine the table below. In which flower pot would the potential for Mn toxicity be greatest?

Pot no.	(1)	(2)	(3)	(4)	(5)	(6)
Drain hole?	No	Yes	No	Yes	No	Yes
Soil sterilized?	Yes	Yes	Yes	No	No	No
Soil pH	5.5	6.5	6.5	6.5	5.5	5.5

Note that nine combinations were possible but only six were presented. The number of choices for any item is limited by the number of spaces on the machine-scored answer form and by the point at which the advantages with respect to guessing and discrimination are counterbalanced by the disadvantages of time consuming and confusing complexity. There is no reason to attempt to make all items on a test have the same number of choices.

Even with four to five alternatives to choose from, a test-wise but unprepared student may have a good chance of guessing the correct answer if the distractors contain some common clues. Writing effective distractors is perhaps the most difficult part of constructing multiple choice tests and it is all too easy to unwittingly "give away" the correct answer. Care should also be taken to make each item independent of the others so that clues to one item will not be provided in some other item. Today's college student is experienced at taking mul-

multiple choice tests. He knows that a True-False question with words like "always" or "never" is usually false, that given choices 1, 2, and 3, the middle one is the best bet; that correct answers are usually longer, more carefully worded, and qualified, etc. To be effective, the test writer must be aware of these and other foibles such as illustrated by the following item about art history.

Example 20.

Romanesque architecture was

- (1) characterized by light, pointed arches.
- (2) the forerunner fo the Classic style.
- (3) characterized by round, heavy arches.
- (4) reminiscent of the massive buildings of ancient Rome.
- (5) 3 and 4.

While most agronomists would have little familiarity with architecture, probably they would have little trouble in identifying choice 5 as the correct response. The item contains a specific determiner (the word Romans is similar to Romanesque) and an opposite determiner (the arches are probably either pointed or rounded) to aid the unprepared student. It also contains the choice "3 and 4" which, like "all of the above" is the correct choice more often than not.⁵

Not only can such clues be avoided, but we have found that planting false clues helps to separate the well prepared student from the unprepared, but experienced test-taker. From the item analysis for Example 21 (Table 2) it is obvious that the response, "1 and 2" was very attractive to those trying to guess on the item. Most of the students who missed this item choose response 5 and the highly negative discrimination index for that response indicates that as a group they scored poorly on the whole exam.

Example 21.

The _____ is usually positively correlated with loss of soil by erosion.

- (1) infiltration rate
- (2) aggregate stability
- (3) organic matter content
- (4) length of slope
- (5) 1 and 2

Increasing Feedback Value of Distractors

Finally, distractors can be written that do more than distract. With data from an item analysis distractors can reveal what errors or misconceptions are prevalent among the students and thus can serve as valuable feedback to the instructor. It is worthwhile, therefore, to write distractors that are mean-

ingful in the sense that they represent particular learning difficulties. Problems, for example, may be calculated in several *incorrect* ways to arrive at good distractors. In other cases, it may be possible to ask a question on an open-ended written test and use some of the common, but incorrect responses as distractors for a future multiple choice test item.

Example 22.

What is the % base saturation of a soil described in the table below:

Element	Grams exchangeable in the sample	Atw.
H ⁺	13	1
Mg ²⁺	36	24
K ⁺	59	39
Ca ²⁺	70	40
Al ³⁺	54	27

(1) 3.37, (2) 24.1, (3) 29.6, (4) 34.2, (5) 51.9

If a large proportion of the class chooses the first response, the problem most likely lies with inadequate training in calculating percentages, not in knowledge of base saturation, for 3.37 is 100 times the reciprocal of the correct answer. The choice of response 2 indicates that the student failed to convert atomic weight into meq for each element. Students who mistakenly considered Al³⁺ to be included as a base would arrive at a base saturation of 34.2% (response 4). Combining the errors in reasoning which lead to response 2 and 4 would result in the choice of response 5.

LITERATURE CITED

1. Angoff, W. H. 1971. Scales, norms, and equivalent scores. In R. L. Thorndike (ed.) Educational measurements. 2nd Ed. American Council on Education, Washington, D. C.
2. Bloom, B. S., M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. 1956. Taxonomy of educational objectives: Cognitive domain. David McKay Co., Inc., N. Y.
3. Brady, N. C. 1974. The nature and properties of soils, 8th Ed. Macmillan Co., Inc., N. Y. p. 211.
4. Gronlund, N. E. 1968. Constructing achievement tests. Prentice Hall, Inc., Englewood Cliffs, N. J.
5. ———. 1971. Measurement and evaluation in teaching, 2nd ed. Macmillan Co., Inc., N. Y.
6. Guilford, J. P., and B. Fruchter. 1973. Fundamental statistics in psychology and education, 5th ed. McGraw-Hill Book Co., N. Y.
7. Wesman, A. G. 1971. Writing the test item. In R. L. Thorndike (ed.) Educational measurement. American Council on Education, Washington, D. C.
8. Wood, D. A. 1961. Test construction, development and interpretation of achievement tests. Charles E. Merrill Books, Inc., Columbus, Ohio.

⁵For a more complete list of irrelevant clues, see Wood (8), Gronlund (5), and Wesman (7).